



Electronic lexicography in the 21st century: Smart lexicography

Proceedings of the eLex 2019 conference

edited by

Iztok Kosem
Tanara Zingano Kuhn
Margarita Correia
José Pedro Ferreira
Maarten Jansen
Isabel Pereira
Jelena Kallas
Miloš Jakubíček
Simon Krek
Carole Tiberius

Sintra, Portugal, 1–3 October 2019

elex.link/elex2019



Electronic lexicography in the 21st century.
Proceedings of the eLex 2019 conference.

edited by

Iztok Kosem
Tanara Zingano Kuhn
Margarita Correia
José Pedro Ferreira
Maarten Jansen

Isabel Pereira
Jelena Kallas
Miloš Jakubíček
Simon Krek
Carole Tiberius

published by

Lexical Computing CZ s.r.o., Brno, Czech Republic

proofreading by

Paul Steed

licence

Creative Commons Attribution ShareAlike 4.0
International License

Sintra, October 2019

ISSN 2533-5626



ORGANIZERS

Univerza v Ljubljani



ACKNOWLEDGEMENT

We would like to thank our sponsors and supporting institutions for supporting the conference.

SPONSORS



A. S. Hornby Educational Trust



elex.link/elex2019

CONFERENCE COMMITTEES

Organising Committee

Tanara Zingano Kuhn
Margarita Correia
José Pedro Ferreria
Maarten Janssen
Isabel Pereira
Jelena Kallas

Miloš Jakubíček
Iztok Kosem
Simon Krek
Carole Tiberius
Ondřej Matuška
Teja Goli

Scientific Committee

Andrea Abel
Špela Arhar Holdt
Vit Baisa
Gerhard Budin
Nicoletta Calzolari
Lut Colman
Paul Cook
Margarita Correia
Gilles-Maurice de Schryver
María José Dominguez Vazquez
Patrick Drouin
Jose Pedro Ferreira
Edward Finegan
Thierry Fontenelle
Polona Gantar
Yongwei Gao
Radovan Garabik
Alexander Geyken
Kris Heylen
Ales Horak
Miloš Jakubíček
Maarten Janssen
Jelena Kallas
Ilan Kernerman
Maria Khokhlova

Annette Klosa-Kückelhaus
Svetla Koeva
Iztok Kosem
Vojtěch Kovář
Simon Krek
Michal Kren
Tanara Zingano Kuhn
Margit Langemets
Lothar Lemnitzer
Robert Lew
Pilar León Araúz
Nikola Ljubešić
Henrik Lorentzen
Tinatin Margalitzadze
Stella Markantonatou
John P. McCrae
Amalia Mendes
Michal Boleslav Měchura
Julie Miller
Victor Mojela
Monica Monachini
Orion Montoya
Sara Može
Christine Möhrs
Chris Mulhall

Carolin Müller-Spitzer
Lionel Nicolas
Sussi Olsen
Vincent Ooi
Isabel Pereira
Jordi Porta
Adam Rambousek
Laurent Romary
Klaas Ruppel
Roser Sauri
Tanneke Schoonheim
Hindrik Sijens
Emma Sköldbberg
Nicolai Hartvig Sørensen
Egon Stemle
Kristina Štrkalj Despot
Arvi Tavast
Carole Tiberius
Yukio Tono
Lars Trap Jensen
Agnes Tutin
Tamas Varadi
Carlos Valcárcel Riveiro
Serge Verlinde
Piotr Zmigrodzki

TABLE OF CONTENTS

Practice of Smart LSP Lexicography: The Case of a New Botanical Dictionary with Latvian as a Basic Language <i>Silga SVIĶE, Karina ŠĶIRMANTE</i>	1
Challenges in the Semi-automatic Reversion of a Latvian-English Dictionary <i>Daiga DEKSNE, Andrejs VEISBERGS</i>	18
Zapotec Language Activism and Talking Dictionaries <i>K. David HARRISON, Brook Danielle LILLEHAUGEN, Jeremy FAHRINGER, Felipe H. LOPEZ</i>	31
Resource Interoperability: Exploiting Lexicographic Data to Automatically Generate Dictionary Examples <i>María José DOMÍNGUEZ VÁZQUEZ, Miguel Anxo SOLLA PORTELA, Carlos VALCÁRCEL RIVEIRO</i>	51
Croatian Web Dictionary – Mrežnik – Linking with Other Language Resources <i>Lana HUDEČEK, Milica MIHALJEVIĆ</i>	72
Representation and Classification of Polyfunctional Synsemantic Words in Monolingual Dictionaries and Language Corpora: The Case of the Croatian Lexeme <i>Dakle</i> <i>Virna KARLIĆ, Petra BAGO</i>	99
Reengineering an Online Historical Dictionary for Readers of Specific Texts <i>Tarrin WILLS, Ellert Þór JÓHANNSSON</i>	116
Assessing EcoLexiCAT: Terminology Enhancement and Post-editing <i>Pilar LEÓN-ARAÚZ, Arianne REIMERINK, Pamela FABER</i>	130
Lexical Tools for Low-Resource Languages: A Livonian Case-Study <i>Valts ERNŠTREITS</i>	161
Ontological Knowledge Enhancement in EcoLexicon <i>Juan Carlos GIL-BERROZPE, Pilar LEÓN-ARAÚZ, Pamela FABER</i>	177

Smart Lexicography for Low-Resource Languages: Lessons Learned from Buddhist Sanskrit and Classical Tibetan	
<i>Ligeia LUGLI</i>	198
The Russian Academic Neography Information Retrieval Resource	
<i>Marina N. PRIEMYSHEVA, Yulia S. RIDETSKAYA, Kira I. KOVALENKO</i>	213
A Thesaurus of Old English as Linguistic Linked Data: Using OntoLex, SKOS and lemon-tree to Bring Topical Thesauri to the Semantic Web	
<i>Sander STOLK</i>	223
SASA Dictionary as the Gold Standard for Good Dictionary Examples for Serbian	
<i>Ranka STANKOVIĆ, Branislava ŠANDRIH, Rada STIJOVIĆ, Cvetana KRSTEV, Duško VITAS, Aleksandra MARKOVIĆ</i>	248
eDictionary: the Good, the Bad and the Ugly	
<i>Marijana JANJIĆ, Dario POLJAK, Kristina KOCIJAN</i>	270
DiCoEnviro, a Multilingual Terminological Resource on the Environment: The Brazilian Portuguese Experience	
<i>Flávia Cristina CRUZ LAMBERTI ARRAES</i>	291
Modelling Specialized Knowledge With Conceptual Frames: The TermFrame Approach to a Structured Visual Domain Representation	
<i>Špela VINTAR, Amanda SAKSIDA, Katarina VRTOVEC, Uroš STEPIŠNIK</i> ...	305
The Lexicographer's Voice: Word Classes in the Digital Era	
<i>Geda PAULSEN, Ene VAINIK, Maria TUULIK and Ahti LOHK</i>	319
Repel the Syntruders! A Crowdsourcing Cleanup of the Thesaurus of Modern Slovene	
<i>Jaka ČIBEJ, Špela ARHAR HOLDT</i>	338
Communities of Related Terms in a Karst Terminology Co-occurrence Network	
<i>Dragana MILJKOVIC, Jan KRALJ, Uroš STEPIŠNIK, Senja POLLAK</i>	357
How Can App Design Improve Lexicographic Outcomes? Examples from an Italian Idiom Dictionary	
<i>Valeria CARUSO, Barbara BALBI, Johanna MONTI, Roberta PRESTA</i>	374

TEI Encoding of a Classical Mixtec Dictionary Using GROBID-Dictionaries	
<i>Jack BOWERS, Mohamed KHEMAKHEM, Laurent ROMARY</i>	<i>397</i>
TEI Lex-0 In Action: Improving the Encoding of the Dictionary of the Academia das Ciências de Lisboa	
<i>Ana SALGADO, Rute COSTA, Toma TASOVAC, Alberto SIMÕES</i>	<i>417</i>
Aggregating Dictionaries into the Language Portal Sõnaveeb: Issues With and Without Solutions	
<i>Kristina KOPPEL, Arvi TAVAST, Margit LANGEMETS, Jelena KALLAS.....</i>	<i>434</i>
LeXmart: A Smart Tool for Lexicographers	
<i>Alberto SIMÕES, Ana SALGADO, Rute COSTA, José João ALMEIDA</i>	<i>453</i>
Make My (Czechoslovak Word of the) Day	
<i>Michal ŠKRABAL, Vladimír BENKO</i>	<i>467</i>
Collecting Collocations for the Albanian Language	
<i>Besim KABASHI</i>	<i>478</i>
Investigating Semi-Automatic Procedures in Pattern-Based Lexicography	
<i>Laura GIACOMINI, Paolo DIMUCCIO-FAILLA.....</i>	<i>490</i>
ELEXIFINDER: A Tool for Searching Lexicographic Scientific Output	
<i>Iztok KOSEM, Simon KREK.....</i>	<i>506</i>
Lexicographic Practices in Europe: Results of the ELEXIS Survey on User Needs	
<i>Jelena KALLAS, Svetla KOEVA, Margit LANGEMETS, Carole TIBERIUS, Iztok KOSEM..</i>	<i>519</i>
Language Varieties Meet One-Click Dictionary	
<i>Egon W. STEMLE, Andrea ABEL, Verena LYDING</i>	<i>537</i>
Identification of Languages in Linked Data: A Diachronic-Diatopic Case Study of French	
<i>Sabine TITTEL, Frances GILLIS-WEBBER</i>	<i>547</i>
Challenges for the Representation of Morphology in Ontology Lexicons	
<i>Bettina KLIMEK, John P. MCCRAE, Julia BOSQUE-GIL, Maxim IONOV, James K. TAUBER, Christian CHIARCOS</i>	<i>570</i>

Proto-Indo-European Lexicon and the Next Generation of Smart Etymological Dictionaries: The Technical Issues of the Preparation	
<i>Jouna PYYSALO, Fedu KOTIRANTA, Aleksi SAHALA, Mans HULDEN</i>	<i>592</i>
Converting and Structuring a Digital Historical Dictionary of Italian: A Case Study	
<i>Eva SASSOLINI, Anas Fahad KHAN, Marco BIFFI, Monica MONACHINI, Simonetta MONTEMAGNI</i>	<i>603</i>
Challenges and Difficulties in the Development of Dicionário Olímpico (2016)	
<i>Rove CHISHMAN, Aline Nardes dos SANTOS, Bruna da SILVA, Larissa BRANGEL</i>	<i>622</i>
The ELEXIS Interface for Interoperable Lexical Resources	
<i>John P. MCCRAE, Carole TIBERIUS, Anas Fahad KHAN, Ilan KERNERMAN, Thierry DECLERCK, Simon KREK, Monica MONACHINI, Sina AHMADI.....</i>	<i>642</i>
Improving Dictionaries by Measuring Atypical Relative Word-form Frequencies	
<i>Kristian BLENSENIUS, Monica von MARTENS.....</i>	<i>660</i>
Planning a Domain-specific Electronic Dictionary for the Mathematical Field of Graph Theory: Definitional Patterns and Term Variation	
<i>Theresa KRUSE, Laura GIACOMINI</i>	<i>676</i>
Text Visualization for the Support of Lexicography-Based Scholarly Work	
<i>Shane SHEEHAN, Saturnino LUZ</i>	<i>694</i>
Validating the OntoLex-lemon Lexicography Module with K Dictionaries' Multilingual Data	
<i>Julia BOSQUE-GIL, Dorielle LONKE, Jorge GRACIA, Ilan KERNERMAN</i>	<i>726</i>
Towards the Automatic Construction of a Multilingual Dictionary of Collocations using Distributional Semantics	
<i>Marcos GARCIA, Marcos GARCÍA-SALIDO, Margarita ALONSO-RAMOS</i>	<i>747</i>

SkELL Corpora as a Part of the Language Portal Sõnaveeb:

Problems and Perspectives

*Kristina KOPPEL, Jelena KALLAS, Maria KHOKHLOVA, Vít SUCHOMEL,
Vít BAISA, Jan MICHELFEIT* 763

A Corpus-Based Lexical Resource of Spoken German in Interaction

*Meike MELISS, Christine MÖHRS, Maria Ribeiro SILVEIRA,
Thomas SCHMIDT* 783

Automating Dictionary Production: a Tagalog-English-Korean Dictionary from Scratch

*Vít BAISA, Marek BLAHUŠ, Michal CUKR, Ondřej HERMAN,
Miloš JAKUBÍČEK, Vojtěch KOVÁŘ, Marek MEDVEĎ,
Michal MĚCHURA, Pavel RYCHLÝ, Vít SUCHOMEL* 805

An Open Online Dictionary for Endangered Uralic Languages

Mika HÄMÄLÄINEN, Jack RUETER 819

The Semantic Network of the Spanish Dictionary During the Last Century: Structural Stability and Resilience

Camilo GARRIDO, Claudio GUTIERREZ, Guillermo SOTO 831

Towards a Graded Dictionary of Spanish Collocations

Marcos GARCÍA SALIDO, Marcos GARCIA, Margarita ALONSO-RAMOS 849

Designing an Electronic Reverse Dictionary Based on Two Word Association Norms of English Language

*Jorge REYES-MAGAÑA, Gemma BEL-ENGUIX, Gerardo SIERRA,
Helena GÓMEZ-ADORNO* 865

Towards Electronic Lexicography for the Kurdish Language

Sina AHMADI, Hossein HASSANI, John P. MCCRAE 881

Introducing Kosh, a Framework for Creating and Maintaining APIs for Lexical Data

Francisco MONDACA, Philip SCHILDKAMP, Felix RAU 907

Enriching an Explanatory Dictionary with FrameNet and PropBank

Corpus Examples

*Pēteris PAIKENS, Normunds GRŪZĪTIS, Laura RITUMA,
Gunta NEŠPORE, Viktors LIPSKIS, Lauma PRETKALNIŅA,
Andrejs SPEKTORS 922*

Karst Exploration: Extracting Terms and Definitions

from Karst Domain Corpus

Senja POLLAK, Andraž REPAR, Matej MARTINC, Vid PODPEČAN 934

LexiCorp: Corpus Approach to Presentation of Lexicographic Data

Vladimír BENKO 957

Porting a Crowd-Sourced German Lexical Semantics Resource

to Ontolex-Lemon

Thierry DECLERCK, Melanie SIEGEL..... 970

Practice of Smart LSP Lexicography: The Case of a New Botanical Dictionary with Latvian as a Basic Language

Silga Sviķe, Karina Šķirmante

Ventspils University of Applied Sciences, Inženieru Street 101, Ventspils, LV-3601, Latvia
E-mail: silga.svike@gmail.com, karina.krinkele@gmail.com

Abstract

The article provides an insight into the project “A New Botanical Dictionary: Terms in Latvian, Latin, English, Russian, and German” implemented in the second half of 2017 and in 2018 within the Ventspils University of Applied Sciences (VUAS) internal call for proposals “Development of Scientific Activity at the VUAS”. The VUAS Faculty of Translation Studies in collaboration with the Faculty of Information Technologies in their scientific and research work along with other Latvian universities aim to occupy a niche in the branch of applied linguistics, therefore the research is related to this discipline and offers solutions in practical lexicography.

The study describes a new botanical dictionary (NBD) – a mobile application prototype – with Latvian as a basic language. An insight into the macrostructure of the dictionary and the structure of entries is given. The research deals with questions concerning IT solutions in general (simple) and semantic search in particular. It also introduces a general search – a morphological approach developed by the authors of the research specifically for the Latvian language; this approach is used to search for Latvian botanical terms in both singular and plural forms. The extracted and linked data methodology developed by the authors is described in detail, as well as the NBD technical solutions and architecture, technologies used, database model, and additional features.

Keywords: LSP lexicography; botanical dictionary; mobile application

1. The Need for a New Botanical Dictionary

One of the indicators of a well-structured and successful process of developing and coordinating field-specific terms is using qualitative, topical and useful terminology resources related to a particular field (TTC, 2007: 38). Although approximately 30% of Latvian lexicography consists of dictionaries of a terminological nature (Helviga & Peina, 2016: 127), the translators’ need for them is still not satisfied (Balode, 2012: 40; Sviķe, 2018: 228-241); besides, the importance of specialized dictionaries for society in general should be noted. (Fuertes-Olivera & Tarp, 2014: 2) The need for compiling a new botanical dictionary is proved by the fact that more than half a century has passed since in 1950 the first issue of Galenieks’s *Botanical Dictionary* (Latvian: *Botaniskā vārdnīca*) was published, thus it is necessary to compile a new dictionary of botanical terms with Latvian as a basic language. Within this study, the term NBD means a terminological work in the form of a multilingual translation dictionary (mobile

application) that can be used when translating from and into different languages. As the Latvian part of the dictionary has more specific implementations and offers a wider range of solutions (e.g. search options: see Section 4.2, definitions retrieved from *www.tezaurs.lv*), the Latvian language is defined as the basic language of the dictionary, while the other languages (English, Russian, German) as contrasted languages.

Plant names are an important part of botanical terms. However, some of the currently available electronic dictionaries and databases (e. g. the database of terms compiled and approved by the Terminology Commission of Latvian Academy of Sciences – *www.termini.lza.lv*) do not include the names of several important genera and species, like translations of the Latvian *ārstniecības izops* (*hyssop* in English) and *zilā vizbulīte* (*liverleaf* in English) into German and English (see *termini.lza.lv*). Translations of the names of many crops and economically important plants into English, Russian, and German are also not found in the electronic encyclopaedia *Latvian Nature* (see *Latvijasdaba.lv*), which mostly includes the names of Latvian species of flora. A conceptually new botanical term dictionary is needed not only for professional translators, but also for media professionals, science students, and natural science teachers or students.

Before compiling the dictionary, a survey and a statistical processing of survey data were conducted to identify potential users of a future product. The conclusions drawn from the analysis of survey data (see Sviķe, 2018) were taken into account when developing the prototype of a mobile application. One of the respondents' preferences was an electronic botanical dictionary with an offline option, so a mobile application with the dataset included in a local application database was considered to be the right solution. Initially, the intention was to develop an Android version of the dictionary, as, for instance, in the period from June 2018 – July 2019 in the Latvian market around 65–75% of smartphones were Android devices, and only 24–32% were iOS ones (see *Statista.com*). The situation could be similar elsewhere in the world. However, during the upcoming stages of improving the mobile app, the production of an iOS version will also be considered by using the Cross-Platform Mobile Development App “Ionic” or other possibilities. New approaches to the structure, as well as the functionality of the NBD, are described in the following sections of the study. The aim of this article is to show practical lexicographic solutions for the development of a new botanical dictionary (mobile application), specifying the problems encountered when using the Latvian language as the main one, and offering innovative solutions in developing search functions.

Compilation of the dictionary was conducted within two stages and financed by the VUAS. The first stage was implemented during the project “New Botanical Dictionary: Lexicographic Concept and Working Model” (project duration – five months), when the term search functionality, plant and flower structure visualization and linkage with terms, representation of pictures and literature lists were developed. The second stage was implemented during the project “New Botanical Dictionary: Supplementation of

Lexicographic Material and Modernization of the Mobile Application Prototype” (project duration – six months), by introducing possibilities to change the interface language from Latvian into English and vice versa; adding images of seed and root structure, of simple and compound leaves; creating interactivity between visual and search parts; making a list of publications; supplementing entries with photos; and introducing semantic search.

2. The Macrostructure of the NBD

This section provides only an insight into the macrostructure of a dictionary to explain the overall structure of the app.¹ A brief overview of the macro- and microstructure of the dictionary in this article is also required to describe programming solutions in the following paragraphs.

The macrostructure of the dictionary (see Figure 1) includes the main body – the lexicographic database level of the mobile application (at the presentation level the user sees the term search view when starting the mobile app) – and several sections (mobile application menus):

1. About the NBD – a view providing the description of the project, the authors’ names and useful information about the mobile application in Latvian, English, Russian, and German.
2. Entry structure – describes all the components of the entries and their functions, as well as features used for increasing functionality: simple and semantic search.
3. The Designations Used – a view showing the table of designations and abbreviations used in the app, as well as explanations and translations into all contrasted languages.
4. Pictures – containing the following parts: Plant, Flower, Root, Seed and Leaf Structure – a view showing a picture of a plant or its part, where one can interactively translate the term of the chosen part of a plant into any of the contrasted languages.
5. Sources Used – a view showing all the sources used when developing the botanical dictionary prototype.
6. Publications – a view containing links to articles on botanical terminology that are potentially useful for users of the dictionary.

¹ Reported at the international scientific conference “Meaning in Translation Illusion of Precision” (*Semantiskais aspekts tulkošanā: precizitātes ilūzija*) organized by Riga Technical University in Riga, May 16–19, 2018.

7. Semantic search – searching only according to the scientific (Latin) name of the plant included in the dictionary (a detailed description is given in Section 4.2.2).
8. Selection of app language – changes the language of the user interface (English or Latvian).

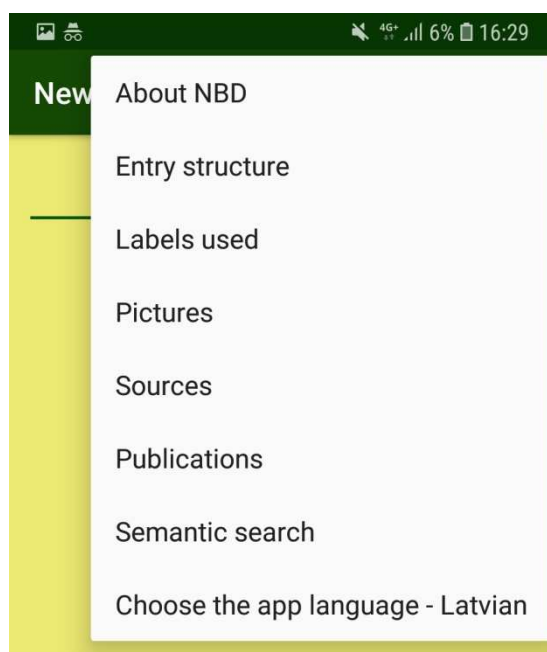


Figure 1: Menu of the mobile application in English

Initially, the macrostructure of the dictionary did not include the section of publications and the selection of app language; those were added during the second stage of compiling the dictionary, when it was supplemented and upgraded. However, this does not exclude the possibility of adding other useful sections to the macrostructure (such as external plant image databases or plant and plant structures schemes) during subsequent stages of upgrading the dictionary prototype and supplementing the language material.

3. The Microstructure of the NBD

Sylviane Granger (2012: 2) lists the six most significant innovations offered by the electronic medium: (1) corpus integration; (2) larger and better data; (3) efficiency of access; (4) customization; (5) hybridization; and (6) user input. The NBD compilers have attempted to include at least five of these, as follows: (1) linking a consolidated corpus of dictionaries with the corpus integration; (2) additional data from other free-access sources according to users' preferences: *www.tezaurs.lv* and *www.wikipedia.org*; (3) internal hyperlinks (from the main view of the mobile app to included pictures) and external hyperlinks (to external sources); (4) lexicographic surveys as a form of

customization; (5) some aspects of hybridization, such as linking the encyclopaedic and linguistic approaches (in the further processing of automatically extracted definitions). These aspects are more clearly evident in the structure of entries elaborated for the dictionary.

Terms included in the dictionary form a so-called *block* structure on the home view of the mobile app. On the home view, the dictionary shows the equivalents of a word searched in all the contrasted languages, thus creating a block. The entries consist of the following structural elements² (see Figure 2):

1. A word or words searched in the input field (box).
2. The functional search button is on the right of the input field.
3. Below the functional search button there is a photo icon which, when being touched, shows on the smartphone screen a photo of the plant that was saved in the resource directory created during the development of the mobile app (only for entries with an image saved in the app database).
4. Below the input field, there is a block of contrasted languages (arranged under each other) and term equivalents. In the NBD, after terms in Latin (put in italics), translations into other languages are arranged in alphabetical order. For German, Latvian, and Russian equivalents grammatical references are also given: gender (female – *f* (*femininum*), male – *m* (*maskulinum*), neuter – *n* (*neutrum*)), singular – *sg* (*singularis*) and plural – *pl* (*pluralis*).

The explanatory part of an entry appears below the language block.

5. A hyperlink to the entry description in the free-access multilingual encyclopaedia *Wikipedia*. This function was mentioned by potential users of the dictionary in the lexicographic survey (see Sviķe, 2018). However, not all plant names included in the dictionary could be provided with a hyperlink to the plant photos, so it was decided to include photos in the dictionary itself; the user can view a photograph of the plant by using a pictogram.
6. A hyperlink to *www.tezaurs.lv* and definition of an entry retrieved from *www.tezaurs.lv* – the website consolidating different Latvian dictionaries – with the help of specially developed software. The abbreviations and markings used are shown below the explanatory section.
7. Glossary of Latin abbreviations used in the dictionary.

² The microstructure of NBD was discussed at the conference “The Word: Aspects of Research” (*Vārds un tā pētīšanas aspekti*) organized by Liepāja University in Liepāja, November 29–30, 2018.

8. Indications and markings of taxonomic levels: the word or words searched in the entry are coloured in the related colours.
9. Explanation of “T*” marking.
10. The INFO section (not shown in Figure 2) – a commentaries part made by the compilers of the dictionary for a relevant entry. In the future, it will be possible to keep in the INFO section not only the corrected or updated definitions automatically retrieved from *www.tezaurs.lv*, but also other comments about the entry. To implement this idea, during the subsequent phases of the dictionary compilation project it is necessary to analyse all the definitions automatically retrieved and to develop new definitions for those cases when an automatic retrieval is not accurate or is incorrect (a detailed description of this is given in Section 4.2.4).

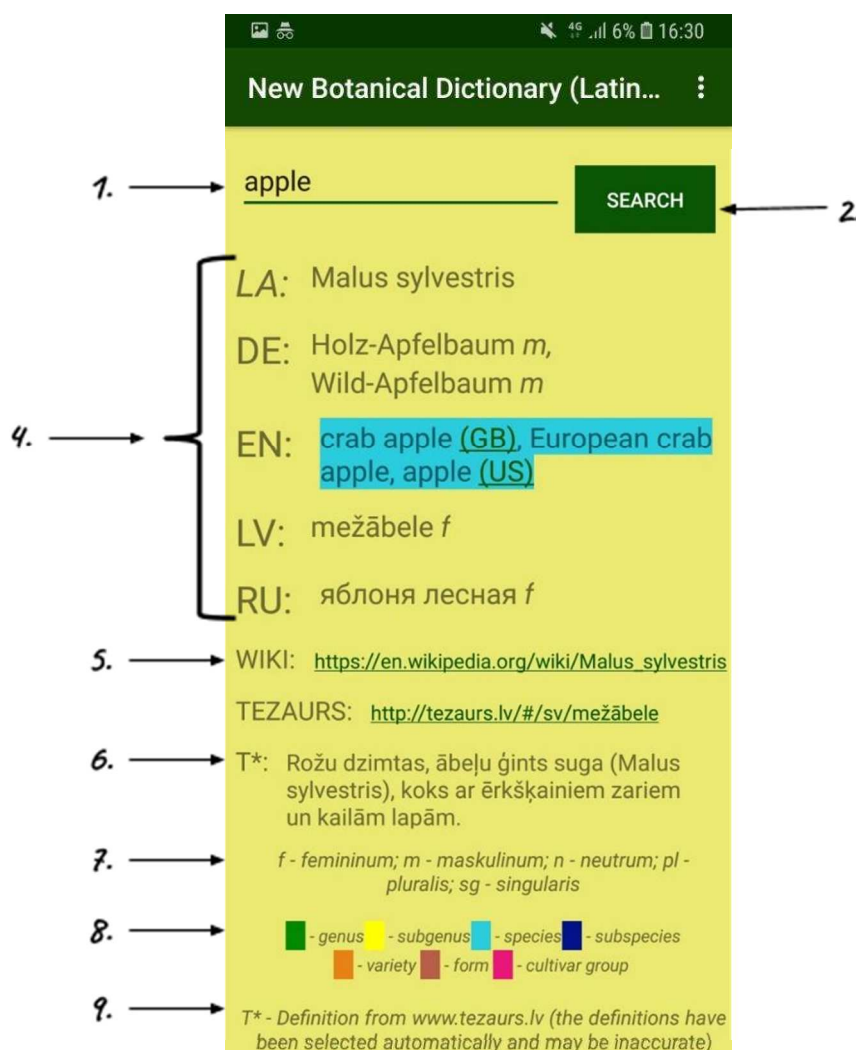


Figure 2. Term searching view

Figure 2 also shows accompanying explanatory notes (especially useful for translators) for English plant names – GB and US, which indicate the use of a plant name in Great Britain or the USA. When expanding the variety of entries, it has been found that in the same language one plant is named and called differently in various countries, e.g. *shadbush* in Germany is more often referred to as *Felsenbirne*, but in Austria as *Edelweißstrauch*. That is why the following markings were introduced: German equivalents used in Germany have a country code (DE), in Austria – (AT), in Switzerland – (CH). For English equivalents the codes US (United States) and GB (Great Britain) are used. The two-letter country codes were selected according to ISO 3166-1. Of course, such country codes are just one of the solutions offered in the dictionary (app prototype) that would make it easier for translators to choose the best equivalent. However, during the next stages of improving the dictionary it should be decided how to distinguish national varieties.

It is intended to supplement the NBD with the sixth innovation mentioned by S. Granger – user input – by adding extension and reduction signs “+/-”, for instance, to the definition and INFO sections. Thus, the vocabulary user will be able to open or hide the information section of an entry. To create the “show and hide” functionality of large or small texts, the expandable TextView component of Android may be used. Possible further solutions – entering data by users, and thus personalizing the app, e.g. by adding comments in one of the sections if the user needs it. However, further improvements of the mobile app prototype require additional funding, deeper research, and extra programming work to develop a new system module. New solutions will be described in future research done by the authors.

4. The Functionality of the NBD Technical Solution

4.1 Technologies Used and Database Model

After researching the most popular OS (operating systems) of mobile phones, the Android mobile platform was selected with the start level Android API 19. The open source Android Studio was used for the development of the application. The sqllite (small local store) database was used to store data, because one of the requirements for a mobile application is the ability to operate with entries without using the internet (as desired and emphasized by potential dictionary users in the previously mentioned lexicographic survey), and without any need to keep data on the distributed server. For application testing the ASUS ZenFone 2 Laser mobile (with Android 6.0.1 version and API 23 level), Oneplus 5 A5000 mobile (with Android 8.1.0 version and API 27 level) and Samsung Galaxy Tab 9,6 E (with Android 4.4.4 version and API 19 level) devices were used.

The database model is based on the dictionary document structure with the following information: (1) a term, its designation and priority in Latvian; (2) a term in Latin; (3) a term in English; (4) a term, its designation and priority in German; (5) a term, its

designation and priority in Russian; (6) a wiki link (automatically generated); (7) notes to relate the term to its visualization in interactive structure images; (8) a definition from *www.tezaurs.lv* (automatically retrieved); (9) a link to the term in *www.tezaurs.lv* (automatically generated); and (10) an info field. The dictionary document content is automatically imported into the application's *insert.sql* resource file, where all INSERT SQL queries with each term's parameters are stored by the script developed based on the Java programming language and Apache POI (Java API for Microsoft Documents). During the process of compiling the application the *insert.sql* resource file is read and executed to create a local database, tables and records. When the dictionary document is updated with some new entries, the developed script automatically updates the database and creates a new android archive **.apk* file.

4.2 The Functionality of the NBD

Special attention was paid to improving the functionality of the NBD in the second phase of dictionary supplementation and mobile application modernization. The related improvements mainly concerned a maximally user-friendly and simplified search in the main section of the dictionary, which was done by developing a special morphological approach for Latvian terms and elaborating a semantic search function – when the user sees the visualization of a taxonomic link between the plant name searched for and the language material included in the dictionary database. A semantic search function offers searching for a taxonomic category represented by a (Latin) name of the plant, i.e. higher and lower taxonomic units (genus – subgenus – species, etc.) included in the dictionary (see Figure 3 below).

4.2.1 A Simple Search Option and its Elaboration

An improvement and special development of the search function is related to the specifics of Latvian as a basic language of the dictionary. Traditionally the Latvian names of plants species are used in the singular, but genus names in the plural. Paragraph 1 of the “Botanic Term-Building Principles”, approved by the Botanical Terminology Subcommittee of the Terminology Commission of the Latvian Academy of Sciences, states that in Latvian the genus names of organisms should be put in the plural and species names in the singular (LZA TK TJ No. 10, 2004: 22), so the Latvian names of plants genus included in the dictionary are given only in the plural forms. However, in spoken language the singular form of a genus is often used (although incorrect), so the dictionary has a search function for both cases. As word endings in Latvian in the singular and plural forms are different, the programmer had to find a solution for cases when the user enters the word in the search box in either singular or plural forms. Thus the possibility of listing both forms in the database and getting the needed form through a simple lookup is not used in the app, although that might seem a simpler solution.

As Latvian is a flexive language, there are very different ways that word endings can change (see Table 1). The most difficult are the cases when one word has two or more different endings in the grammatical category of number – singular or plural. This section of the article offers an overview on how to make it possible to find words with both endings in the application database, and what was the programmer’s approach and solution to this issue. In the lexicographic survey (Sviķe, 2018), one of the users’ preferences was a simple search function as well as the ability to search the database, even if the word was entered in the search box slightly differently. In order to improve the functionality of the app, it was necessary to introduce an additional function – a search option regardless of the singular or plural form of the Latvian word is entered. When implementing the dictionary development project, a method was developed that performs an change in ending recognition algorithm and finds the combination of corresponding changes in word endings, as in Table 1.

In most cases the ending *-as* in plural changes to *-a* in singular (e.g. *aronijas* (pl.) and *aronija* (sg.)), but there are also some more difficult cases: for example, if in plural the ending is *-ņi*, then in singular that could be *-nis* (*alkšņi* (pl.) and *alksnis* (sg.)) or *-ņš* (*amoliņi* (pl.) and *amoliņš* (sg.)). To implement the solution for the singular and plural substitutions, all ending variants were stored in the application resource file. First of all, the word is searched for in the database with no substitution of an ending. If the query returns a positive result from the database, this means the word was found, and the process of translating into others languages and searching for a definition starts. If the word is not found, ending substitution starts: (1) the last symbols of the word are compared with the endings (shown in Table 1), and the algorithm starts searching for a suitable ending pattern; (2) if there is more than one corresponding pattern of an ending, a list with all of them is created, if only one pattern is suitable, then it is stored in the list as the only element; (3) regarding the list of suitable ending patterns, the ending of a word is substituted for an ending from the list, and the algorithm checks whether the changed word is available in the database. If so, the word substitution is successful and the process of translating and searching for a definition can begin. If not, then the next pattern in the list is checked. For small databases (such as the NBD, with 2,000 entry words in Latvian and their equivalents in contrasted languages) the algorithm is quick, but for larger databases the algorithm update may be needed.

Plural	Singular	Example
-s	-a	aronijas → aronija
-s	-e	purenes → purene, lapegles → lapegle
-či	-cis	lakači → lakacis
-dži	-dzis	dadži → dadzis
-i	-s	artišoki → artišoks, bērzi → bērzs
-i	-š	ceriņi → ceriņš, augstiņi → augstiņš
-ji	-is	ķirbji → ķirbis
-li	-lis	āmuļi → āmulis, paegli → paeglis, fizāli → fizālis
-lli	-llis	amariļli → amaryllis
-ši	-sis	bukši → buksis, oši → osis
-ši	-tis	sunīši → sunītis, jānīši → jānītis, žibuliši → žibulītis
-šļi	-slis	grīšļi → grīslis
-šņi	-nis	alkšņi → alksnis
-ži	-dis	skābarži → skābardis
-ņi	-nis	doņi → donis, apiņi → apinis
-i	-us	zeltieti → zeltietus

Table 1: Change of word endings

In order to improve the search function, different cases of endings changing from plural into singular were analysed (plural → singular). Considering these changes, as well as the fact that some words also have changes of consonant in the root, e.g. *alkšņi*, *ķirbji*, the search methodology was adapted to the tradition of using Latvian plant names – plant genus. All consonant substitutions are included in Table 1, and the methodology of changing endings is the same as described above.

The algorithm developed during the study also performs its function in reverse, from singular to plural. The word searched for is displayed in the app's input or search window, but after the recognition of a change in ending this word appears in the results section. The related algorithm is developed for the material compiled in Latvian, i.e. for the Latvian part of the application, but in further stages of improving the app it could also be developed for other languages used in the dictionary.

4.2.2 Semantic Search

One of the most characteristic features of hybrid and printed dictionaries is an innovative search function (Tono, 2009: 65). Such solutions are also found in the NBD, for which a semantic search system was elaborated. During the implementation of the project, work was performed on the representation of taxonomic categories, e.g. a link between genus and species, or a display of the semantic search function (referring to plant names). A new section, semantic search, was created, which performs a semantic search only according to the scientific (Latin) name of the plant included in the dictionary. The algorithm that was developed can successfully process simplified cases (see Figure 3).



Figure 3: Semantic search (Rosa view after a semantic search)

Figure 3 shows that the dictionary includes two species of roses – *Rosa rugosa* and *Rosa canina*. The user sees a visualized link in the taxonomic categories between the genus and species.

When implementing the project, the main task related to the semantic search was to verify whether it is possible to use this function in the application. The results of the study confirm this possibility: the algorithm is able to perform semantic search, but only for the Latin plant names included in the database. The algorithm currently being developed performs data selection from a database taking into account a Latin equivalent of the term searched, in this case – the scientific name of a plant. For example, searching for *Rosa* (at the highest taxonomic level – genus), the app searches for terms in the database at lower taxonomic levels, where the first part of terms includes the keyword *Rosa*. For the scientific names of plants in Latin, the names of lower taxonomic levels will always include the first name of the highest level (for example, genus *Rosa* and species *Rosa rugosa*). When searching from a lower taxonomic level to a higher

taxonomic level, the database searches for a word at a higher taxonomic level; this identifies the semantic tree root of the word being searched for. Moreover, in order to build a full semantic tree of a the word being searched for, both the taxonomic level and the lower levels of the word are searched for. In forming the algorithm, a “tree” data structure is used to store the selected data at specific taxonomic levels and make it easier to display semantic search results in the semantic search view of the mobile app. Usually such challenging tasks are performed by groups of computer linguists and lexicographers within long-term projects (implemented over several years). The two phases of the NBD prototype project lasted for less than a year, so the development of the semantic search function could be implemented during future upgrades of the app. This task should be carried out within possible future projects along with broadening the research task and implementing it in a more detailed way (and also for other botany terms, not just for plant names) and offering specific solutions to the related problems.

4.2.3 An Extracted and Linked Data Methodology

This subsection provides an insight into the automated data selection methodology from free-access resources, and shows how linkage with other sources of information was performed.

As mentioned earlier, one of the application’s features is retrieving the definitions from the Latvian Definition Information System *www.tezaurs.lv* (referred to as Tezaurs in this subsection). For the purposes of the project a script was developed to retrieve the definitions of all dictionary entries in Latvian from Tezaurs by using the *tezaurs.lv* API (Application Programming Interface). The developed script automates the definition retrieval from Tezaurs and stores the results in the dictionary document – in MS Word or Google Sheets format (Microsoft Word was used for storing entry units within the first stage of the project, but Google Drive Sheets was used during the second stage). The script was written using Java programming language and the external library was developed using Apache POI (Java API for Microsoft Documents), with this needed to retrieve words from a Sheets or Word document and to store entry definitions in the same document. The Tezaurs API returns HTML code with tags, and filtering of results is necessary. The script algorithm includes three data filtering and processing methods:

1. The result stream from the Tezaurs API is filtered using the external library JSoap (Java HTML parser) and by using an eliminator for the division of HTML tags “div”, “sv_Sense”, “span”, “sv_NO”. It is important to note that also multiple definitions might be retrieved, and it is necessary to automate choosing the right one. This is done by comparing the scientific (Latin) names, because most definitions in Tezaurs include the scientific names. For example, when searching for a definition of the Latvian *ābols* (*apple* in English), Tezaurs retrieves three definitions in Latvian from which only one is related to *ābols*. 1. *Sulīgs daudzsēklu auglis, raksturīgs ābelēm, bumbierēm, cidonijām, pīlādžiem u.*

- c. (in English: juicy multi-seeded fruit, common for apple trees, pear trees, quince trees, rowan-trees, etc.); 2. *Āboliņš* (it is a Latvian plant name (*clover* – in English), not an apple as required by the NBD); 3. *Parastais ķirbis* (it is a *field pumpkin* regionally called *ābols*, not an apple as required by the NBD).
2. If the word is not found in the Tezaurs database then additional filtering is carried out to search for synonyms in the NDB database and look for definitions of a specific synonym. For example, when the word searched is *ziemasteres* (Latvian plant name of a genus *Symphytotrichum*), but the Tezaurs API retrieves only the link to an entry *miķelītes* (*aster* in English), it is necessary to retrieve the definitions by using the link, because the definitions of both words given in Tezaurs are the same.
3. The Tezaurs API does not retrieve the definition of a word if this word was not entered in the correct form (plural or singular). In such cases the algorithm developed for word substitution from plural to singular or opposite is used. For example, when looking for the plant name *akanti* (*bear's-breech* in English), the Tezaurs API retrieves no results, but when substituting the term *akanti* to its singular form, *akants*, the Tezaurs API retrieves the definition. This implementation includes the algorithm described in Section 2, above.

After filtering and processing the data (when the Tezaurs API is used), the retrieved definitions are stored in a database table column “def_tez”. The developed program is intended to be used only for obtaining definitions automatically from the Tezaurs database, and is not responsible for the correctness of the definitions and relevance to the term searched.

4.2.4 Analysis and Correction of Automatically Retrieved Definitions

The definitions included in the entries and automatically retrieved from *www.tezaurs.lv* are important additional information, as the NBD provides both a translation and explanation of the entry words it includes. It should also be noted that the study revealed that the definitions which are retrieved automatically are only a temporary solution in providing an explanatory function of the dictionary. The desire to link the newly compiled electronic dictionary with other existing lexicographic sources was mentioned by respondents in a lexicographic survey conducted before the dictionary was developed (Sviķe, 2018: 228-241). An insight into the problem of automatically retrieved definitions was given at the international scientific conference “The Word: Aspects of Research” (*Vārds un tā pētīšanas aspekti*) organised by Liepāja University on November 29-30, 2018 in Liepāja. The study concluded that the automatically retrieved definitions have many inaccuracies and even errors, so they need to be corrected and aligned with their information layout. As an example, the definition of *aronijas* (*chokeberry* in English) retrieved automatically from *tezaurs.lv* is translated into English and described below:

*Rose family (genus "Aronia") deciduary shrubs with glossy, elliptical leaves, white flowers, black berries, 3 species (native to eastern part of North America, from Ontario to Florida), all introduced in Latvia.*³

The derived definitions in the original – Latvian – language are given in footnotes (for comparison). First of all, it should be noted that there is a mistake at the beginning of the definition – *Rose family* ("*Aronia*" genus), because the scientific (Latin) name in Latvian should not be put in quotation marks. Similarly, the wording of the definition needs to be corrected. It should also be noted that the fruits of chokeberries are pomes. The correct definition translated into English would be:

*The rose family genus of a deciduary plant. Shrubs have glossy, whole and elliptical leaves. Flowers are white. Fruits are black pomes. The genus has 3 species.*⁴

By analysing the definitions automatically retrieved from *tezaurs.lv*, a methodology has been elaborated for developing a basic variant of the definition, where the definition has been applied and adapted to the taxonomic level of a plant name in the NBD, as shown with the following example of *lotus*.

When searching for the word *lotosi* (*lotus* in English) in *tezaurs.lv*, the following definitions were found, which describe the order, the family of this order and the genus (*lotosi* in Latvian):

1. Divdīgļlapju klases gundegu apakšklases rinda ("Nelumbonales"), kurā ir tikai viena dzimta; 2. Šīs rindas dzimta ("Nelumbonaceae") ar 1 ģinti; 3. Šīs dzimtas ģints ("Nelumbo"), kurā ir 2 sugas, ūdensaugi ar lielām lapām un krāšņiem ziediem, kas sakņojas zemē, bet zieds atveras virs ūdens.

The translation of the definitions into English is: 1. An order of dicotyledon class, crowfoot sub-class ("Nelumbonales") with only one family; 2. Family of this order ("Nelumbonaceae") with 1 genus; 3. The genus of this family ("Nelumbo"), consisting of 2 species, aquatic plants with large leaves and bright flowers rooted in the ground and the flowers opening above the water.

Since the NBD requires a definition that characterizes genus, the third definition is appropriate, but there is still a need for corrections. The above definitions could be combined into one by correcting them as follows:

Lotosu dzimtas (Nelumbonaceae) ģints. Ģintī ir 2 sugas. Ūdensaugi ar lielām lapām un krāšņiem ziediem, kas sakņojas zemē, bet zieds atveras virs ūdens. (In English - *The*

³ In Latvian: Rožu dzimtas ģints ("Aronia"), vasarzaļi krūmi ar spīdīgām, eliptiskām lapām, baltiem ziediem, melnām ogām, 3 sugas (Ziemeļamerikas austrumu daļā no Ontārio līdz Floridai), visas introducētas Latvijā.

⁴ In Latvian: Rožu dzimtas ģints vasarzaļi augi. Krūmi, ar spīdīgām, veselām un eliptiskām lapām. Ziedi balti. Augļi melni āboli. Ģintī 3 sugas.

genus of a lotus-lily family (Nelumbonaceae). The genus has 2 species. An aquatic plant with large leaves and bright flowers rooted in the ground and the flowers opening above the water).

Pursuant to the taxonomy category that specifies the NBD entry – genus, the higher taxonomic name of a genus is added in Latvian, in genitive case – *lotosu* – (which replaces the pronoun *šīs*) and the word *family* with its scientific (Latin) name in brackets without quotation marks (but italicized) according to the Latvian punctuation traditions. The scientific (Latin) name of the genus is not needed in the definition, as the scientific (Latin) name is included in the translating section of the dictionary after the label LA, so in the definition it was deleted. The word *kurā* used as the link (*..kurā ir 2 sugas.*) is replaced by the taxonomic category *ģintī* (in English – genus). In order to maintain a structure similar to the corrected Aronia definition, an auxiliary clause was not used, but a new sentence has been started in which the word *ģints* is written with a capital letter. The description of the plant with the word *ūdensaug*s is also given in a new sentence.

As mentioned before, the Tezaurs API was used for to retrieve a definition from the *www.tezaurs.lv* database. The result is the HTML output stream of the Tezaurs API filtering using the JSoup parser and specific HTML tags. In this case, the verification by scientific (Latin) name in all retrieved definitions is carried out. If the definition consists of the words “Šis” or “Šī” (“This” in English in plural and singular forms), then it is a wrong definition, so processing is necessary. The concatenation of both definitions is done by cutting out the repeating parts of the concatenated definition. The algorithm works with multiple definitions as well.

The examples described may be one of the possible solutions for further reviews and corrections of new definitions done by the dictionary compilers. It is certainly important to verify the correctness of all definitions. The explanation should include the most important features only – this is a lexicographic axiom (Baldunčiks, 2012: 118). It should also be noted that for plant names, which make up the majority of the NBD entries, there is no strict difference between the encyclopaedic and philological definition formulation approach described by Melita Stengrevica (Stengrevica, 1998: 115-120). Without describing the appearance, lifestyle or function of the plant concerned, the meaning of the name of the plant cannot be formulated. The definitions added in offline mode provide the dictionary user with a concise, precise definition of the essential features of the denoted realia. However, in online mode it is possible to quickly access more information by using hyperlinks. Due to the limited length of this article, these aspects have not been addressed, but further research by the authors is certainly required in order to retrieve, combine, correct, and write definitions.

5. Short Summary, Conclusions and Future Plans

This research paper describes a prototype of the mobile app – a dictionary structure that includes a basic part and a visual part (images with terms). The paper specifically analyses problematic cases that required some special solutions, i.e. the development of search function in Latvian both in singular and plural, as well as the semantic search for displaying the taxonomic categories of plant names.

The authors of the study have researched 18 different types of changes in ending in the language material collected in the database (e.g. the plural ending *-ni* in singular might be *-nis* (*alkšņi* (pl.) and *alksnis* (sg.)), or *-š* (*amoliņi* (pl.) and *amoliņš* (sg.)), therefore, a new methodology for processing language material was developed.

The study concludes that automatically retrieved data (definitions) should still be reviewed by an experienced lexicographer in collaboration with an industry expert to develop an optimal language material (definition) solution. It is still necessary to test the already developed NBD functionality and evaluate users' feedback, as well as implement possible corrections and improvements.

During the next stages of improving the application it will also be necessary to include a feature that could hide an automatically retrieved definition, e.g. by adding an information extension and reduction function (+/-). The authors hope that in the near future users will receive the NBD app described in this article, which is intended to be supplemented with 2,500 to 3,000 entries, and the dictionary will be useful not only for translators, but also for students of science, educators, and all others interested in the world of flora.

6. Acknowledgements

The research is supported by funding received within the VUAS internal call for proposals “Development of Scientific Activity at the Ventspils University of Applied Sciences” for years 2017 and 2018.

7. References

- Baldunčiks, J. (2012). Pārskats par nozīmīgākajām vienvalodas vārdnīcām: skaidrojošās vārdnīcas, svešvārdu vārdnīcas, etimoloģijas vārdnīca, slenga vārdnīca. In: A. Lauzis (ed.). *Vārdnīcu izstrāde Latvijā 1991–2010*. Pētījums J. Baldunčika vadībā. Rīga: Latviešu valodas aģentūra, pp. 108–190.
- Balode, I. (2012). Vācu-latviešu un latviešu-vācu leksikogrāfija (1991–2010). In: A. Lauzis (ed.). *Vārdnīcu izstrāde Latvijā 1991–2010*. Pētījums J. Baldunčika vadībā. Rīga: Latviešu valodas aģentūra, pp. 16–61.
- Fuertes-Olivera, P.A. & Tarp, S. (2014). *Theory and Practice of Specialised Online Dictionaries. Lexicography versus Terminography*. Berlin/Boston: Walter de

- Gruyter.
- Granger, S. (2012). Introduction: Electronic lexicography – from challenge to opportunity. In: S. Granger, M. Paquot (eds.) *Electronic Lexicography*. New York: Oxford University Press, pp. 1–11.
- Helviga, A. & Peina, E. (2016). Mūsdienu latviešu terminogrāfijas raksturojuma daži teorētiskie un praktiskie aspekti. In J. Baldunčiks, I. Jansone, A. Veisbergs (eds.) *Latviešu valodas vārdnīca*. Valsts valodas komisijas raksti. (8). Rīga: Zinātne, pp. 127–158.
- Latvijasdaba.lv. *Latvijas daba: sugu enciklopēdija*. Accessed at: www.latvijasdaba.lv (31 May 2019)
- Statista.com. *Mobile Operating System Market Share Latvia June 2018 – July 2019*. Accessed at: www.statista.com (22 July 2019)
- LZA TK TJ No. 10 (2004). *LZA TK Terminoloģijas Jaunumi*. Nr. 10. V. Skujiņa (resp. ed.). Rīga: LZA Terminoloģijas komisija.
- Stengrevica, M. (1998). Daži vārdi par Latviešu literārās valodas vārdnīcu (sakarā ar pēdējā sējuma iznākšanu – pēcvārda vietā). In: *Linguistica Lettica*. 2. Rīga: Latviešu valodas institūts, pp. 115–120.
- Sviķe, S. (2018). A New Dictionary of Botanical Terms: Data Analysis of a Lexicographic Survey. *Economics World*. May-June 2018, 6 (3), pp. 228–241.
- termini.lza.lv. *Latvijas Zinātņu Akadēmijas Akadēmiskā terminu datubāze*. Accessed at: www.termini.lza.lv (31 May 2019)
- tezaurs.lv. *Electronic dictionary developed by Artificial Intelligence Laboratory (AILAB)*. Accessed at: <https://tezaurs.lv/> (25 July 2019)
- Tono, Y. (2009). Pocket Electronic Dictionaries in Japan: User Perspectives. In: H. Bergenholtz, S. Nielsen & S. Tarp (eds.) *Lexicography at a Crossroads. Dictionaries and Encyclopedias Today, Lexicographical Tools Tomorrow*. Bern: Peter Lang, pp. 33–67.
- TTC (2007). Valsts aģentūra “Tulkošanas un terminoloģijas centrs” (State Agency “Translation and terminology Centre”). *Latviešu valodas terminoloģijas resursu kvalitātes un pieejamības apzināšana dažādās zinātnes un praktiskās darbības nozarēs*. Available at: http://www.vvc.gov.lv/export/sites/default/LV/publikacijas/vardnicu_petijums_TTC_16102007.pdf

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



Challenges in the Semi-automatic Reversion of a Latvian-English Dictionary

Daiga Deksne¹, Andrejs Veisbergs²

¹ Tilde, Vienības gatve 75a, Rīga, Latvia, LV-1004

² University of Latvia, Visvalža iela 4a, Rīga, Latvia, LV-1050

E-mail: daiga.deksne@tilde.lv, andrejs.veisbergs@lu.lv

Abstract

The electronic version of the Latvian-English dictionary has been significantly supplemented over the last year with new linguistic material from corpora, databases and other sources. In contrast, the English-Latvian dictionary can be considered outdated as its electronic version was updated 10 years ago. This motivated us to create a semi-automatic process for reversion of the Latvian-English dictionary in order to supplement the English-Latvian dictionary with missing entries. Some of the major challenges for automatic reversion were as follows: grouping translations by part of speech, deciding to which entry the example should be attached, and ordering translations with similar meaning. By using automatic scripts it was possible to create reversed entries of quite good quality within a short time. Three groups of entries were prepared for manual post-editing: new entries with a single translation, new entries with a more complex structure, and existing entries with additional new content. The tasks for post-editing are: to check the suitability of the chosen headword, part of speech and translation order, to group the translations having the same meaning, and to move examples after appropriate translations.

Keywords: electronic dictionaries; bilingual dictionary reversing; phraseology

1. Introduction

The electronic version of the Latvian-English dictionary has undergone a series of significant revisions and has been supplemented with a significant number of entries that users previously lacked. There are words that have recently entered the language (both Latvian and English neologisms), words that have spelling variants, and words that are frequently found in the corpus but not found in dictionaries. It is common practice not to include regular derivatives in a dictionary. But as not every user of electronic dictionaries is a grammar expert and able to derive the needed word on his or her own, it is helpful to have some regular derivative forms included as well, such as deverbalized nouns, participles, and feminine forms of nouns. At present the Latvian-English dictionary comprises 54,465 entries, 139,796 translations and 23,617 usage examples. It can be considered to be the most up-to-date Latvian bilingual dictionary.

The English-Latvian dictionary was published in 1995, its electronic version was slightly updated in 2009. It comprises 52,202 entries, 118,723 translations and 32,510 usage examples. This dictionary can be considered outdated, and this motivated us to create an automatic process for reversion of the Latvian-English dictionary in order to supplement the English-Latvian dictionary with missing entries.

2. Studies of Reversion

Numerous reports describe attempts at compiling dictionaries by semi-automatic reversion of the opposite direction dictionaries. The language pairs of target dictionaries involve languages from the same language group, for example Estonian-Finnish (Langemets et al., 2017), as well as languages of different groups, such as English-Albanian (Newmark, 1999), Estonian-Dutch (Tamm, 2002), Latvian-English (Veisbergs, 2004), and Slovenian-English (Krek et al., 2008). The main motivation is to save time and the very valuable human lexicographers' resources, and to get the maximum benefit from abundance of examples and translation equivalents in the source dictionary. However, it is also noted that the process does not always go smoothly, and some, often unexpected, manual post-editing is required (Veldi, 2010).

3. Dictionary Structure

Both dictionaries are monodirectional, aimed at the Latvian user, but their microstructures differ. Usually every entry of both dictionaries starts with a single headword. There can be several headwords as well if they are absolute synonyms or phonetic variations. Generally the headword is in its canonical form. For the English-Latvian dictionary, the headword is followed by the pronunciation written using phonetic alphabet. Such information is not included in the Latvian-English dictionary. Within a given part of speech, translation equivalents are grouped into senses in the English-Latvian dictionary. Translation equivalents are grouped into senses in the Latvian-English dictionary as well, but part of speech information is absent as normally the Latvian ending clearly signals the part of speech (except in some minor cases). The most frequent senses are placed first. The sense may contain information about the usage domain (e.g., 'biol.'), register (e.g., 'slang') or some comment about usage context. Examples and their translation equivalents are included at the end of the particular sense. Idiomatic expressions and their equivalents are given at the end of an entry.

4. The Process of Reversion

4.1 Retrieval of data from the Latvian-English dictionary

The first step of reversion consists in the retrieval of words, translation equivalents and examples from the Latvian-English dictionary. The dictionary is internally stored in an XML format (Deksne et al., 2013). The XML tag names describe all pieces of information found in the microstructure of the dictionary entry. The following example (see Figure 1) shows an entry with three senses, each having a single translation, and two examples for the first sense as well as comments clarifying the second and third senses and the usage information for the second.

```
<entry title="lēģenda">
  <title>lēģenda</title>
  <mean digits="1" symbol="." />
  <transl>legend</transl>
  <from_sample>l. vēsta</from_sample><to_sample>legend goes</to_sample>
  <from_sample>lēģendu krājums</from_sample><to_sample>legendary</to_sample>
  <mean digits="2" symbol="." /><comment>(spiega)</comment><usage>pārn.</usage>
  <transl>cover story</transl>
  <mean digits="3" symbol="." /><comment>(skaidrojums)</comment>
  <transl>caption</transl>
</entry>
```

Figure 1: Sample xml entry for the entry *lēģenda* ‘legend’.

To ease the reversion process, the data is transformed into a tabular format. Separate files are created for translations and for translation examples. Each line of the first file contains the title of an entry and the translation. The same entry title is on several lines if the entry contains several translations (see Figure 2). It is not important to preserve the division in senses, as translations of the Latvian-English dictionary will be the headwords of different entries in the English-Latvian dictionary.

lēģenda	legend
lēģenda	cover story
lēģenda	caption

Figure 2: Translations from the entry *lēģenda* ‘legend’ in a tabular format.

Each line of the second file contains the title of an entry, the example and the translation of the example with some optional comment (see Figure 3). There are several lines with the same entry title and the same example if the particular entry contains several examples or the example contains several translations. In a dictionary a word in an example may be abbreviated to the first letter of an entry title. In the further process, it will be expanded to a full word.

lēģenda	l. vēsta	legend goes
lēģenda	lēģendu krājums	legendary

Figure 3: Samples from the entry *lēģenda* ‘legend’ in a tabular format.

The entries of the English-Latvian dictionary are prepared in a similar way. An automatic process will be used to ignore translations and examples that are already in the dictionary.

4.2 Determining part of speech

Latvian words from the Latvian-English dictionary are morphologically analysed using the morphological analyser developed by Tilde (Deksne, 2013), by which their part of speech is determined. This is where the problems start: often a word is not in its basic form and it is attributed to various parts of speech or a part of speech which the

corresponding English word will not have, e.g. *izglītības* in Latvian is a noun in genitive. The English counterpart/equivalent ‘educational’ is (and should be labelled) as an adjective. Several parts of speech are attributed to the Latvian word *ātri*, but one has to choose adverb, since the English equivalent ‘quickly’ is an adverb. For many Latvian words the part of speech is undetermined, as they have not been included in the morphological analyser’s dictionary. Among them are non-traditional compounds, foreign words, abbreviations, non-literary vocabulary, and so on. For 3,631 words out of 55,920 the morphological analyser does not return part of speech information.

The algorithm for choosing the most appropriate part of speech is the following:

- if the part of speech is unknown but a word ends with *-ot*, *-ēt* or *-ties*, it is a verb;
- if the part of speech is unknown but a word ends with *-ošs*, *-īgs* or *-isks*, it is an adjective;
- if the part of speech is unknown but a word ends with *-ējs*, *-tājs*, *-isms*, *-ists*, *-ums*, *-īnš* or some other common noun ‘suffix + ending’ pattern (35 in total), it is a noun;
- if a word is the past active participle in masculine singular nominative form with a definite ending it receives the part of speech ‘noun’, as such words have completed the process of nominalization; for example, the participle *pieaugušais* (‘the grownup’) in the dictionary is included as a noun;
- if a word is a noun in genitive it receives the part of speech ‘adjective’; for example, *vietniekvārda* (‘pronominal’);
- if a word is the adjective in masculine plural nominative form with an indefinite ending it receives the part of speech ‘adverb’; for example, *viesmīlīgi* (‘hospitably’);
- if a word is the adjective in masculine nominative form with a definite ending it receives the part of speech ‘noun’; for example, *ļaunais* (‘the evil one’);
- other words keep the part of speech assigned by the morphological analyser if the current word form coincides with the basic form.

4.3 Adding examples and translations

Supplementing a dictionary based on the principle of nesting is complicated. For digital purposes a bilingual dictionary based on the alphabetic principle may be more convenient, although that would mean changing the whole pattern, which is not feasible in the short term.

Automatic joining of examples to the entry is the hardest task in the process. Should common phrases and multiword units (MWU) (Fellbaum, 2016) be included as separate entries, or as examples in the existing entry or as examples of contextual use? Which component of the MWU should we choose as the headword for joining? Which is the dominant word, e.g. in the collocation ‘the language of the proceedings’? The problem is similar to that of deciding on keywords in the treatment of idioms in lexicography (Yong & Peng, 2007; Mulhall, 2010), and it is well known that users are not sure where to find idioms (Atkins & Varantola, 1998: 30).

English pattern	Latvian pattern	% of all examples	English example	Latvian equivalent
adj. + noun	noun	12.48%	‘folic acid’	<i>folijskābe</i>
noun + noun	noun	7.93%	‘savings account’	<i>krājkonts</i>
adj. + noun	adj. + noun	6.37%	‘jolly crowd’	<i>jautra kompānija</i>
noun + noun	noun + noun	5.94%	‘sports hall’	<i>sporta halle</i>
adj. + noun	noun + noun	5.31%	‘normative act’	<i>tiesību akts</i>
‘to’ + verb + adv./particle	verb	2.58%	‘to pay off’	<i>atpirkties</i>
adj. + noun	participle + noun	2.06%	‘decisive battle’	<i>izšķiroša kauja</i>
‘to’ + verb + adj.	verb	1.94%	‘to get fat’	<i>aptaukoties</i>
‘to’ + verb + det. + noun	verb + noun	1.68%	‘to call the police’	<i>izsaukt policiju</i>
‘to’ + verb + adv.	verb	1.63%	‘to beat back’	<i>atsist</i>
‘to’ + verb + prep.	verb	1.44%	‘to blend in’	<i>iederēties</i>
noun + prep. + noun	noun + noun	1.26%	‘field of action’	<i>darbalauks</i>
‘to’ + verb + ‘a/the’ + noun	verb	1.24%	‘to get a fright’	<i>izbēties</i>
‘to’ + verb + noun	verb	1.03%	‘to shed light’	<i>izgaismot</i>

Table 1: The most popular structural correspondences of English examples and their Latvian equivalents.

There are 28,155 examples in the Latvian-English dictionary. The English and Latvian examples often present different syntactical patterns (see Table 1 for the most popular structural correspondences). The most popular correspondences are as follows: 12.48% MWUs that have the construction ‘adjective + noun’ and 7.93% MWUs with the structure ‘noun + noun’ are translated as ‘noun’ in Latvian; 2.58% of examples with the structure ‘verb + adverb/particle’ are translated as ‘verb’. English phrasal verbs are translated into Latvian predominantly as prefix-verbs (Veisbergs, 2013: 110-112). We generally see Latvian compounds as corresponding to English MWUs.

The issue of compounds is complicated both theoretically (Burger, 2007; Scalise, 2010) and practically, and increased due to the possibility of hyphenation (The Chicago, 2010; Vrbinc & Vrbinc, 2011: 256). First, while Latvian compounds by definition are written together (which ensures their separate entry status), this is not the case in English. Second, in both languages compound spelling often fluctuates both diachronically and synchronically, with a general tendency for two-component phrases to merge into a compound. In both languages normativizing tendencies (Levin-Steinmann, 2007: 37) exist but are hard to follow. This uncertainty and asymmetry in contrastive aspect has been noted by Čermak (2007: 20).

Thus we have to decide which compounds can be considered full entry words. In the existing English-Latvian dictionary many compounds frequently appear only as contextual examples, while the first component does not have a Latvian translation, for example, ‘citric acid’ (in Latvian *citronskābe*) is included within the entry with a headword ‘citric’. It seems worth avoiding the “categorical bias” (Granger et al., 2012) and leaving some decisions as to where to place the word, compound or MWU for post-editorial work.

The automatic process starts with putting the content of the tab separated files of both dictionaries into hash tables. The data from the Latvian-English dictionary is treated in a reversed way, i.e. the key of a hash table is an English word/phrase and the value is a concatenation of the corresponding Latvian words/phrases. Only word/phrase pairs not existing in the English-Latvian dictionary are considered.

The phrases containing all content words with an initial capital letter are considered to be headwords. Phrases with the capital letters usually are some named entities like ‘Little Red Riding Hood’, ‘the Atlantic Ocean’, and ‘the Book of Psalms’. The single words are considered to be headwords as well. We accept Latvian phrases consisting of one or two words as translations.

The most complex part of the process is to sort out examples. We ignore phrases containing more than five words. It is too risky to decide automatically which word of a phrase should be taken as a headword of an entry. We avoid full sentence-like examples. They frequently have almost word-for-word translation and do not provide any new information. For example, we do not process the example ‘this accusation is unfounded’ (in Latvian, *šīs apvainojums nav dibināts*). We avoid such phrases by

looking for words ‘is’, ‘am’, ‘are’, ‘were’, ‘was’, ‘has’, ‘have’, and ‘had’ in the middle of a phrase or by checking if a phrase starts with ‘I’, ‘you’, ‘he’, ‘she’, ‘we’, ‘they’, ‘are’, and ‘is’. Of course, in such a manner we could filter out some valuable examples as well, but with our abundance of examples the potential loss is far smaller than the benefit of quality assurance.

For some popular entry headwords the automatic process assigns up to 80 examples, and these examples are not found in the existing English-Latvian dictionary. Of course, this is too many for a single entry. We thus set a maximum limit and print out only the first ten examples per entry. The most example-rich headwords are ‘time’, ‘work’, ‘right’, ‘way’, ‘call’, ‘cut’, ‘stand’, ‘covered’, ‘place’, ‘cover’, ‘pay’, ‘side’, ‘plant’, ‘hand’, ‘look’, ‘hold’, ‘throw’, and ‘day’. In our first experiments the process assigned numerous examples to both the stop words and common verbs. To avoid this, we compiled the lists of the stop words and common verbs which we do not choose as entry headwords for particular examples. The stop word list contains pronouns, prepositions, numerals, and some adverbs. The common verb list contains such verbs as ‘make’, ‘be’, ‘give’, ‘get’, ‘put’, ‘push’, ‘pull’, ‘take’, ‘become’, ‘come’, ‘grow’, ‘turn’, ‘set’, ‘run’, ‘keep’, ‘bring’, ‘fall’, ‘let’, ‘make’, ‘break’, ‘play’, ‘draw’, and ‘use’.

In our final version, the algorithm for processing examples is the following:

- in a two-word phrase starting with a capital letter the first word is considered as a headword (e.g., for the examples ‘Devonian era’ and ‘Devonian period’ the headword ‘Devonian’ is chosen);
- we delete stop words from the beginning and end of the example and common verbs from the beginning if followed by a stop word, then we look for an appropriate headword in the remaining text string:
 - if a single word is left we consider it as a headword for the entry in which the current example is included (e.g., for the example ‘to accustom oneself to’ the headword ‘accustom’ is chosen);
 - if a text string starts with a common verb we take the word after the verb as a headword (e.g., for the example ‘to make suffer’ the headword ‘suffer’ is chosen);
 - if in the middle of a text string one finds the words ‘into’, ‘to’, ‘of’, ‘on’, ‘with’, ‘by’, ‘from’, ‘in’, ‘for’, or ‘a’ and there are two words before one of them having an attributive ending, the word without an attributive ending is considered as a headword (e.g., for the example ‘additional edition of copies’ the headword ‘edition’ is chosen), otherwise all words before are taken as a head phrase (e.g., for the example ‘a hard nut to crack’ the head phrase ‘hard nut’ is chosen);

- if the phrase starts with the word ‘to’ we take the next word as a headword (e.g., for the example ‘to adjust the fire’ the headword ‘adjust’ is chosen);
- if the first word of a two-word phrase is a hyphenated compound and it is a headword in the existing dictionary we take it as a headword for the current example otherwise the second word is chosen (e.g., for the example ‘colour-blind person’ the headword ‘colour-blind’ is chosen, but for the example ‘computer-composed music’ the headword ‘music’ is chosen);
- for the other two-word examples we take the last word as a headword unless the number of examples for that headword has exceeded ten; then we take the first word as a headword (e.g., we choose the headword ‘limit’ for the example ‘credit limit’, but the headword ‘credit’ for the example ‘credit line’ as the word ‘line’ has too many examples);
- for the remaining examples, we take the last word of the example as a headword.

As headwords are chosen from examples, they are frequently not in their base form, like most of the headwords in the English-Latvian dictionary are. In order not to create too many separate entries unnecessarily, small adjustments are performed to the chosen headwords. If the headword is a verb in the simple past or present participle form and the corresponding root form is a headword in the English-Latvian dictionary, we take the root form for a headword (‘praised’ → ‘praise’, ‘praying’ → ‘pray’). If the headword has the plural ending and the corresponding singular form is a headword in the English-Latvian dictionary, we take the singular form for a headword (‘activities’ → ‘activity’). If the headword has a comparative or superlative ending and the corresponding base form is a headword in the English-Latvian dictionary, we take the base form for a headword (‘smallest’ → ‘small’).

Entries in the XML format are generated from the processed data. Translations with the same part of speech are grouped together in an entry. Groups are sorted alphabetically by part of speech abbreviation, e.g., the first are adjective translations and the last are verb translations. All examples are at the end of an entry, as it is impossible to determine after which translation a particular example should be. There is a single exception with the non-verb phrases. If an example does not start with the particle ‘to’ it is moved to the previous part of speech translation group.

4.4 Merging the new entries with the entries from the English-Latvian dictionary

We store the dictionary data in the Microsoft SQL Server database on a permanent basis. For editing purposes, the data is exported to a plain text file. The new content and the existing dictionary are in the same XML format. The unique identifier of the

entry is its headword. The entry is left unchanged if the existing dictionary does not contain the entry with a specific headword or an example that equals the new content's headword. Otherwise we try to merge the new content with the existing entry, although there are some restrictions. We merge the existing dictionary entry and the entry with new content if both entries have translations with a single part of speech grouped in a single meaning only, and if the existing entry does not have examples. It would require too much manual work to merge entries with a more complex structure. For the tags containing a new content the colour attribute is added. This allows users to keep track of the part that is automatically included in an entry. When the XML format is transformed to the HTML format tags with a colour attribute provide a good visual indicator of the new content (see Figure 4). These entries still require post-editing, possibly with regard to changing the order of translations or grouping some translations in the separate senses.

agnomen [æɡ'nəʊmən] *n* palama, iesauka; pavārds
agriculturist [ˈæɡrɪ'kʌltʃərɪst] *n* agronomš; agronome, lauksaimniece, lauksaimnieks
agrimony [ˈæɡrɪməni] *n* dadzis; ancītis; hemp agrimony – krastkaņepe
agronomist [ə'ɡrɒnəmɪst] *n* agronomš; agronome

Figure 4: Existing entries automatically updated with new content.

Depending on the outcome of the merging process regarding the structure of an entry, we define three different post-editing tasks of various complexity:

- 1) for the new entries with a single translation and an optional example, the suitability of the chosen headword and part of speech should be checked (11,500 such entries);
- 2) for the new entries with several translations and/or examples, the suitability of the chosen headword, part of speech and translation order should be checked; translations should be grouped in meanings; every example should be moved after the appropriate translation (2,992 such entries);
- 3) for the existing entries with some additional content, new translations and examples should be moved to the appropriate position (6,368 such entries).

4.5 Separating senses of translations

Automatic separation of translations into senses is impossible. Thus, a convenient editing format is defined for manual processing. A special script is developed for transforming entries of the first and second tasks to a tabular format with six columns. The first column is reserved for the headword. If an entry has several headwords they are separated by a vertical line '|'. The second column contains the part of speech abbreviation. The third column contains the sense number for grouping of translations. Numeration of senses is organized within the framework of the part of speech. By

default, this column contains sense number ‘1’. The fourth column contains the translation. The fifth column contains one or several English examples separated by ‘|’. The sixth column contains one or several Latvian translations of examples separated by ‘|’. Not all columns are filled. Each line contains either the fourth column with the translation or the fifth and sixth columns with an example(s) and its translation(s) (see Figure 5).

agitated	a	1	uzbudināts		
agitated	a	1	uztraukts		
agitated	a	1		to be agitated get agitated	uztraukties
agonizing	a	1	mokošs		
agonizing	a	1	mokpilns		
aiding	n	1	palīdzēšana		
aiding	n	1	veicināšana		
Aids	abbr	1	AIDS		
Aids	abbr	1	Aids		
Aids	abbr	1		contract Aids get Aids	saslimt ar Aidu

Figure 5: New entries in tabular format prepared for post-editing.

The editor is asked 1) to correct the headword if it is not appropriate; 2) to check the part of speech; 3) to correct the sense number if an entry has translations with a different sense; 4) to move lines with examples directly after the appropriate translation. If some translations seem very distant from the headword or are used in a very narrow context the line should be deleted. Any spreadsheet application can be used for editing, and we use Microsoft Excel for this task.

5. Results and Discussion

The manual post-editing has not been completed yet. But the automatic part of the dictionary reversion process has prepared the rough material of quite good quality in a short time. The creation of the scripts for the reversion process took less than a month. As a result, 11,500 new entries have been created containing one translation equivalent or usage example and 2,992 new entries with more complex structure. These entries will help fill the gaps in the English-Latvian dictionary. The addition of new translation equivalents or examples to the existing 6,368 entries is not vital but enriches the dictionary, making its content more up to date, allowing the user to choose from a wider range of translation equivalents or to better understand the meaning of some unknown English word by exploring the newly added usage examples. Usage examples, multi-word terms or idiomatic expression meanings that have translations with different structures (frequently a single word) are especially valuable, for example: ‘to appear publicly for the first time’ (*debitēt* in Latvian), ‘employee buy-out’ (*uzņēmuma pārdošana darbiniekiem* in Latvian), ‘to lie like a trooper’ (*šausmīgi melot* in Latvian). Though some researchers have spoken in favour of omitting idioms when encoding

dictionaries (Hausmann, 2004), it seems they can contribute to a better overall reflection of the linguistic system of the language as well as improve users' choice and production capability.

The first results of the post-editing process reflect the quality of automatically generated entries. Of the first 610 post-edited entries containing one translation equivalent or usage example 64% did not require any editing, while 2% contained the wrong part of speech; for 16% of entries the part of speech tag was added as it was unknown before; 8% of entries were deleted as inappropriate; headwords of 4% of entries were corrected; the translations of 2% of entries were corrected; and the examples of 2% of entries were corrected.

The existing version of the English-Latvian dictionary is available online at <https://www.letonika.lv/groups/default.aspx?g=2&r=10331062&f=1>. After the post-editing process is completed the new version will be available at the same address.

6. Acknowledgements

The research has been supported by the European Regional Development Fund within the project "Neural Network Modelling for Inflected Natural Languages" No. 1.1.1.1/16/A/215.

7. References

- Atkins, B.T.S. & Varantola, K. (1998). Language Learners Using Dictionaries: The Final Report on the EURALEX/AILA Research Project on Dictionary Use. In B.T.S. Atkins (ed.) *Using Dictionaries: Studies of Dictionary Use by Language Learners and Translators*. Lexicographica Series Maior, Number 88, Tübingen: Max Niemeyer Verlag, pp. 21–81.
- Burger, H. (2007). Semantic aspects of phrasemes. In H. Burger., D. Dobrovol'skij, P. Kuehn, N.R. Norrick (eds.) *Phraseologie*. Vol.1. Berlin, New York: Walter de Gruyter, pp. 90–109.
- Čermak, F. (2007). Idioms and morphology. In H. Burger., D. Dobrovol'skij, P. Kuehn, N.R. Norrick (eds.) *Phraseologie*. Vol.1. Berlin, New York: Walter de Gruyter, pp. 20–26.
- Deksne, D. (2013). Finite State Morphology Tool for Latvian. In Mark-Jan Nederhof (ed.) *Proceedings of the 11th International Conference on Finite State Methods and Natural Language Processing*. St. Andrews: Scotland, pp. 49–53.
- Deksne, D., Skadina, I., & Vasiljevs, A. (2013). The modern electronic dictionary that always provides an answer. In I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets, M. Tuulik (eds.) *Electronic lexicography in the 21st century: thinking outside the paper: proceedings of the eLex 2013 conference*. Ljubljana: Trojina, Institute for Applied Slovene Studies; Tallinn: Eesti Keele Instituut, pp. 421–434.
- Fellbaum, C. (2016). Treatment of Multi-Word Units. In P. Durkin (ed.) *The Oxford*

- Handbook of Lexicography*. Oxford: Oxford University Press, pp. 411-424.
- Granger, S., & Lefer, M. A. (2012). Towards more and better phrasal entries in bilingual dictionaries. In R. Vatvedt Fjeld & J. M. Torjusen (eds.) *Proceedings of the 15th EURALEX International Congress*. Oslo: University of Oslo, pp. 682-692.
- Hausmann, F. J. (2004). Was sind eigentlich Kollokationen? In K. Steyer (ed.) *Wortverbindungen – mehr oder weniger fest*. Institut für deutsche Sprache. Jahrbuch 2003. Berlin, New York: de Gruyter, pp. 309-334.
- Krek, S., Šorli, M., & Kocjančič, P. (2008). The Funny Mirror of Language: The Process of Reversing the English-Slovenian Dictionary to Build the Framework for Compiling the New Slovenian-English Dictionary. In E. Bernal & J. DeCesaris (eds.) *Proceedings of the XIII EURALEX International Congress*. Barcelona: Universitat Pompeu Fabra, pp. 535-542.
- Langemets, M., Hein, I., Heinonen, T., Koppel, K., & Viks, Ü. (2017). From Monolingual to Bilingual Dictionary: The Case of Semi-automated Lexicography on the Example of Estonian-Finnish Dictionary. In I. Kosem et al. (eds.) *Electronic Lexicography in the 21st Century: proceedings of eLex 2017 Conference, 19-21 September 2017, Leiden, Netherlands*, Brno: Lexical Computing, pp. 155-171.
- Levin-Steinmann, A. (2007). Orthographie und Phraseologie. In H. Burger., D. Dobrovol'skij, P. Kuehn, N.R. Norrick (eds.) *Phraseologie*. Vol.1. Berlin, New York: Walter de Gruyter, pp. 36-41.
- Mulhall, C. (2010). A Semantic and Lexical-Based Approach to the Lemmatisation of Idioms in Bilingual Italian-English Dictionaries. In A. Dykstra & T. Schoonheim (eds.) *Proceedings of the XIV EURALEX International Congress*, Ljouwert: Fryske Akademy, pp. 1355-1369.
- Newmark, L. (1999). Reversing a One-Way Bilingual Dictionary. *Dictionaries. Journal of The Dictionary Society of North America*, 20(1), pp. 37-48.
- Scalise, S. & Vogel, I. (eds.) (2010). *Cross-Disciplinary Issues in Compounding*, Amsterdam, Benjamins.
- Tamm, A. (2002). Reversing the Dutch-Estonian Dictionary to Estonian-Dutch. In A. Braasch & C. Povlsen (eds.) *Proceedings of the Tenth EURALEX International Congress*. Vol. 1, Copenhagen: Center for Sprogteknologi, pp. 389-399.
- The Chicago Manual of Style. (2010). 16th ed. University of Chicago Press.
- Veldi, E. (2010). Reversing a Bilingual Dictionary: a mixed blessing? In A. Dykstra & T. Schoonheim (eds.) *Proceedings of the XIV EURALEX International Congress*, Ljouwert: Fryske Akademy, pp. 861-865.
- Vrbinc, A. & Vrbinc, M. (2011). Treatment of multi-word lexical items in the dictionary: the current situation and the potential problems facing dictionary users. *Eesti Rakenduslingvistika Ühingu aastaraamat 7*, pp. 249-263.
- Veisbergs, A. (2004). Reversal as Means of Building a New Dictionary. In G. Williams & S. Vessier (eds.) *Proceedings of the Eleventh EURALEX International Congress*. Vol. 1, Lorient: UBS, pp. 327-332.
- Veisbergs, A. (2013). *English and Latvian Word Formation Compared*. Rīga: Latvijas

Universitātes Akadēmiskais apgāds.

Yong, H. & Peng, J. (2007). *Bilingual Lexicography from a Communicative Perspective*.
John Benjamins.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



Zapotec Language Activism and Talking Dictionaries

K. David Harrison¹, Brook Danielle Lillehaugen²,

Jeremy Fahringer³, Felipe H. Lopez⁴

¹ Swarthmore College, 500 College Ave., Swarthmore PA 19081, and New York Botanical Garden, 2900 Southern Blvd, The Bronx, NY 10458

² Haverford College, 370 Lancaster Ave., Haverford PA 19041

³ Swarthmore College, 500 College Ave., Swarthmore PA 19081

⁴ University of California, San Diego, 9500 Gilman Dr., Literature Dept 0410, La Jolla CA 92093

E-mail: harrison@swarthmore.edu, blilleha@haverford.edu, jfahrin1@swarthmore.edu, lieb@ucla.edu

Abstract

Online dictionaries have become a key tool for some indigenous communities to promote and preserve their languages, often in collaboration with linguists. They can provide a pathway for crossing the digital divide and for establishing a first-ever presence on the internet. Many questions around digital lexicography have been explored, although primarily in relation to large and well-resourced languages. Lexical projects on small and under-resourced languages can provide an opportunity to examine these questions from a different perspective and to raise new questions (Mosel, 2011). In this paper, linguists, technical experts, and Zapotec language activists, who have worked together in Mexico and the United States to create a multimedia platform to showcase and preserve lexical, cultural, and environmental knowledge, share their experience and insight in creating trilingual online Talking Dictionaries in several Zapotec languages. These dictionaries sit opposite from big data mining and illustrate the value of dictionary projects based on small corpora, including having the flexibility to make design decisions to maximize community impact and elevate the status of marginalized languages.

Keywords: lexicography; collaboration; endangered languages; Zapotec

1. Introduction

Dictionaries, whether print or digital, are much more than just an organized collection of words with definitions or translations. At their best, they are living repositories of the collective knowledge base compiled by and for the community that owns it. When shared with outsiders, they can also provide a window into a culture, its traditions, beliefs, and values. Dictionaries can serve as accurate records of the historical and contemporary state of a language, to the extent that they are inclusive of variation and not highly standardized. Finally, dictionaries can shape the future development of a language. They can influence the vitality of a language, expanding its domains of use. And, intentionally or not, they contribute to processes of standardization.

The Zapotec lexicography project draws inspiration from Native American Language online dictionaries such as the Passamaquoddy-Maliseet Language Portal (<https://pmportal.org>) and the Lenape Talking Dictionary (<http://www.talk-lenape.org>). Both of these projects represent an intense effort by these respective communities to reclaim, record, share, and generationally transmit their severely endangered languages. They serve as proof that indigenous languages can thrive thanks to the digital activism of community members and linguists. Methodology—not just final product—is of central importance in the creation of these dictionaries: the collaborative practices as well as the resulting resources can be interventions in contexts where discrimination and detrimental linguistic ideologies conspire to silence languages, such as those described by Sicoli (2011) based on ethnographic research in Southern Sierra Zapotec communities.

In this paper, we describe the ideation, design, building, and sharing of a suite of five Zapotec Talking Dictionaries. We also discuss unintended uses and effects, and what we hope are positive impacts on community practices and ideas about language maintenance.

2. Talking Dictionary project history

In 2003, one of the current study’s co-authors, Harrison, published a print Tuvan Dictionary (Anderson & Harrison, 2003), and upon distributing it to the community received questions and feedback suggesting that some words were missing. Such experiences are ubiquitous—the very same sentiment is noted on page 1 of *Making Dictionaries* (Frawley et al., 2002: 1) in a long list of challenges when making dictionaries. This led to the realization that print was not the best medium for dictionaries, and so we began to put some of our many lexical field recordings into a searchable online format, which became the “Tuvan Talking Dictionary” launched in 2006 (Harrison & Anderson, 2006). We have considerably expanded the platform since then in terms of technological capabilities, design, and community participation. As of 2019, we have 120 Talking Dictionaries in varying stages of development. The total number of lexical entries is 150,000+, with the largest collections being the Gutob Talking Dictionary (Anderson & Harrison, 2016) with 13,338 entries, and the Siletz Dee-ni Talking Dictionary (Anderson & Harrison, 2007) with 10,552 entries. The lexical entries come from a variety of sources: (a) field recordings, (b) recordings made during digital lexicography workshops we have hosted, and (c) recordings made on an ongoing basis by online co-authors working in places such as India, Mexico, and Vanuatu.

The Valley Zapotec Talking Dictionaries began in 2012 with the creation of the Tlacolula Valley Zapotec Talking Dictionary. Another of the current work’s co-authors, Lillehaugen, created this using already existing audio recordings as a mock-up to show members of the San Lucas Quiaviní and Tlacolula de Matamoros communities as a way to gauge interest in developing the dictionary further. During a field trip to Oaxaca during summer 2013, Lillehaugen met with the authorities and community members in

San Lucas and Tlacolula. The feedback was clear: members of both communities were interested in developing the dictionaries further, and both wanted dictionaries that represented only the language variety as spoken in their community. Thus, while the Zapotec of San Lucas Quiaviní and the Zapotec of Tlacolula de Matamoros may be considered dialects of the same language on linguistic grounds—both are classified with the ISO 639-3 code [zab] (Eberhard et al., 2019)—that was not the relevant criterion for interested community members. In response to this, the mock-up dictionary was split into two dictionaries in 2013—the first two Valley Zapotec Talking Dictionaries: Tlacolula de Matamoros (Lillehaugen et al., 2013) and San Lucas Quiaviní Zapotec (Lillehaugen et al., 2109a), the latter co-authored and locally directed by a native Zapotec speaker and co-author of this paper, Felipe H. Lopez. Soon thereafter, two additional communities joined in: San Jerónimo Tlacoahuaya (Lillehaugen & García Guzmán et al., 2019) and Teotitlán del Valle (Lillehaugen and Chávez Santiago et al., 2019). Most recently, the Talking Dictionary for San Bartolomé Quialana was started in summer 2019 (Lillehaugen et al., 2019b).

3. Design and features

The intention in designing the Talking Dictionaries was to create a multimedia resource (audio, video, photo, text, maps) for small languages that went beyond traditional dictionary design and content. The user experience would be paramount, while the back-end design would be secondary. It would be a living, constantly expanding resource that was community-authored, community-owned, and fully attributed by name to all contributors. The interface would be easy to access (online, on smartphones, or even as a paper printout), and would use simple iconography (for example, an ear icon for sound files, as seen in Figure 1). It would be rich in content from the very first encounter: the user would never be confronted with only a blank “search” box on the front page or a null search result. Regardless of what a user searched for, some content would always appear. The back-end design would not be based on specialist or proprietary software, but would use widely available and well-supported database software: we chose MySQL, an open-source relational database management system. (Note that the examples illustrated throughout this text show the English language interface of the Talking Dictionary, since this article is in English. The Zapotec Talking Dictionaries also have a Spanish-language interface.)



Figure 1: Ear icon for sound files

We learned from our experience with other digital lexicography efforts, such as LEXUS (Kemps-Snijders & Wittenburg, 2006), a tool designed at MPI-Nijmegen and launched in 2001. LEXUS prioritized information architecture over user experience. As a result, it proved to be of limited appeal beyond those individual researchers who were uploading their lexical data, and even such specialized users needed to engage in significant learning and adaptation to the unique environment (Wojtylak, 2012). As one reviewer observed: “LEXUS is obviously a tool for the creation and maintenance of a lexicon rather than for the visually appealing presentation of lexical data” (Kochetova, 2009: 244). It remained in circulation until 2017, but is no longer supported. Likewise, Bergenholtz and Bothma (2011: 55) warn of “information death” resulting from an overwhelming presentation of content, especially in a digital context. We hoped to avoid these scenarios of limited user appeal and technical support leading to obsolescence.

Design decisions for any particular Talking Dictionary are made together with local co-authors. For example, the dictionaries were originally designed with a computer web-based interface in mind, but it quickly became clear that most users were accessing these Talking Dictionaries from their smartphones. Future development was thus optimized with smartphone use in mind. As all of the dictionaries are supported from the same back-end structure, solutions for needs that originate in one community can end up providing design enhancements for other Talking Dictionaries.

3.1 Acknowledgement and crediting of expertise

A key feature of the dictionaries is that expertise is identified and credited with every entry (see Figure 2). This overt recognition of knowledge holding is a form of decolonizing lexicography, and a response to anthropological and linguistic practices which erased the names of local experts, as well as the larger issue of the exclusion of indigenous authority, as argued in Anderson and Christen, 2019. In these dictionaries, you hear and see the names of the experts behind each word. Copyright is explicitly mentioned in each dictionary as belonging to the community, and authorship of the dictionaries includes the Zapotec co-authors.

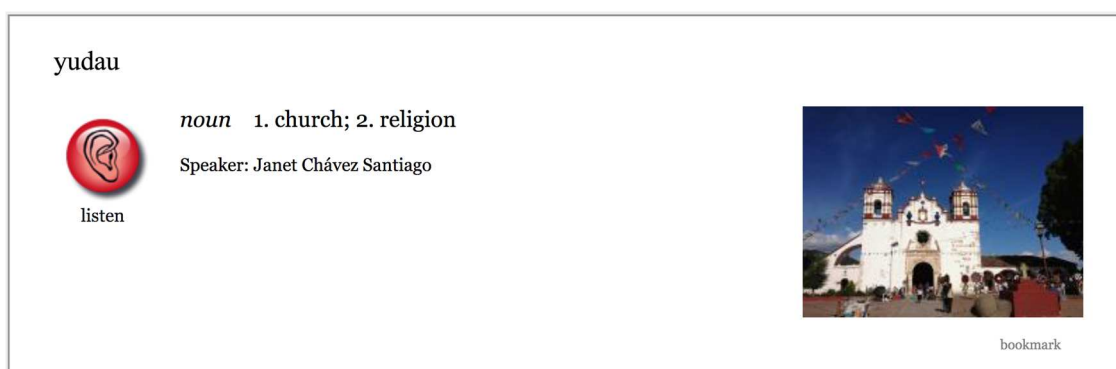


Figure 2: Speaker credit at the entry level
(Lillehaugen & Chávez Santiago et al., 2019: entry 248)

3.2 Semantic domains

Dictionaries can be organized and browsed based on a set of semantic categories (e.g., kinship terms, food, botany) that are dynamic and customizable at the dictionary level. Any new semantic domain can be created and immediately available. This freedom allows for flexibility and creative experimentation with ways of interacting with the words in a dictionary, and for highlighting domains of (specialized) knowledge that are important at the local level. For example, Teotitlán del Valle is a weaving town, so the category “weaving” is crucial and contains scores of entries, as seen in the list on the left in Figure 3. “Weaving” does not (currently) occur as a semantic category in the dictionary for San Jerónimo Tlacoahuaya, an agricultural town, illustrated in the list of semantic domains on the left in Figure 3.

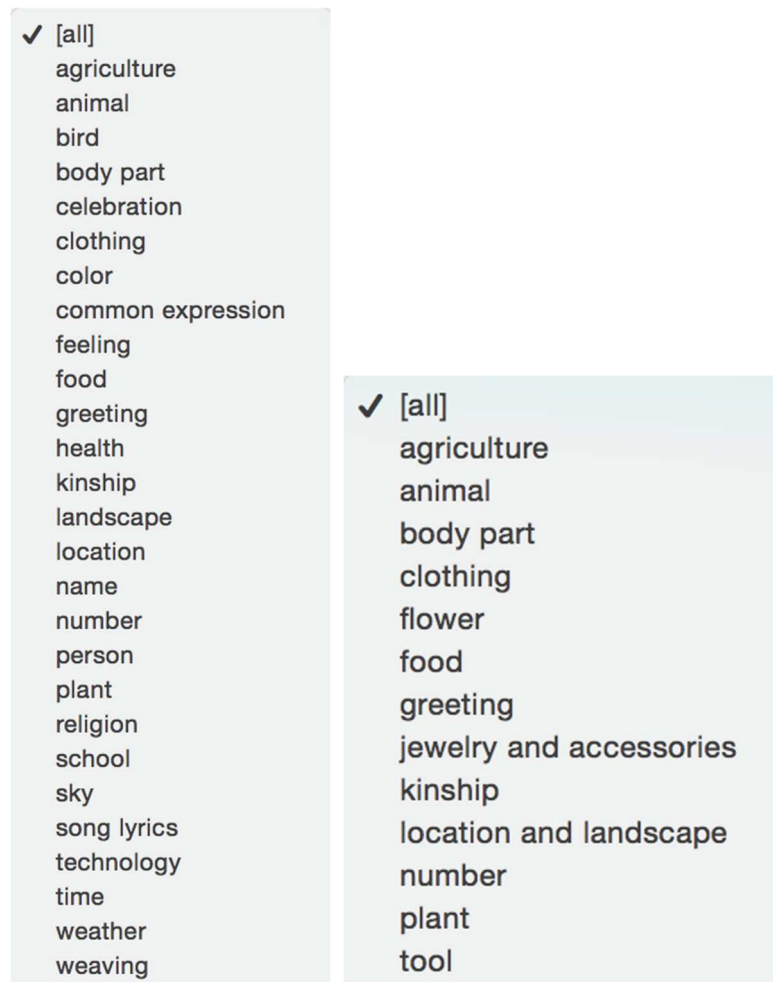


Figure 3: Semantic domains in two Valley Zapotec Talking Dictionaries (left Teotitlán; right Tlacoahuaya)

3.3 Dictionaries without standardized orthographies

Our design anticipates the possibility of making a dictionary for a language with an emerging orthography or no orthography. Some communities that want to develop dictionaries lack an orthography and may not have plans to develop one, or may still be working towards a consensus about the best orthography. We can accommodate multiple orthographies (showing various alternate spellings under a single entry), or none at all. In the case where no orthography exists, we do not attempt to devise one, as we believe this work is best done by the community itself, not by outsiders. Instead we can use IPA transcription (which we fully acknowledge is not appealing to indigenous speech communities), while leaving blank for future use the field where an orthographic spelling would go. Another option is to use the spelling preferences of the speaker, whether or not those preferences are part of a systematic set of decisions on how to spell the sounds of the language.

The San Lucas Quiaviní Zapotec dictionary uses the orthography defined in Cali Chiu (Munro et al., 2008). All other Valley Zapotec dictionaries exist in the absence of a standardized orthography. For these dictionaries, we use the spelling preferences of the speaker. This means that words and sounds may be spelled inconsistently throughout the dictionary, and single words may be spelled more than one way. While this may make some lexicographers and linguists uncomfortable, it reflects the current practice and linguistic reality of these speech communities, where individuals are used to deciphering personal spelling decisions, like those used, for example on store signage or social media. Trying to read Zapotec on Facebook and Twitter is a frequent experience for co-author Felipe H. Lopez, and two examples of Twitter exchanges across orthographic differences in Valley Zapotec can be seen in Figures 4 and 5 in Lillehaugen (2019). Moreover, “[c]ommunity discussions about standardized orthographies can sometimes become unproductive, and these debates can even impede other advances in increasing the use of the language. In some cases, these disagreements can turn into ‘orthography wars’ (Hinton, 2014), draining the precious time and energy of the activists involved” (Lillehaugen, 2016: 367).

The Valley Zapotec practice around spelling represents the full continuum between idiosyncratic, ad hoc spelling choices on one end, and fully developed, orthographic systems based on a phonological analysis of the language on the other. Our Talking Dictionaries support writing systems at any point in this continuum, and are flexible to change as decisions about writing choices in a community emerge. The design of the dictionary allows us to include multiple spellings in a single entry and multiple entries for a single word, each potentially with a different spelling. For example, ‘flower’ in the San Jerónimo Tlacoahuaya dictionary is currently spelled both *gie* and *gie’*, the latter marking the word-final glottal stop. In the San Bartolomé Quialana dictionary we find even more diversity in the spelling of ‘flower’, which currently includes: *gi*, *gui*, *glli*, and *gyi’i*.

This feature has had the added benefit of representing dialect variation within a pueblo, without being forced to choose one of the forms as the headword. Multiple parallel entries and pronunciations, where no one is primary or authoritative, can exist side by side. While we currently have between one and four pronunciations for each word, as Garrett points out, hearing multiple voices is a benefit to those who might be using a dictionary in language learning efforts.

In the online Yurok dictionary, we have tried to include audio examples of as many words and short phrases as possible, spoken by as many fluent speakers as possible. Users report greatly appreciating a chance to hear the range of variation that would have been present in the speech community when Yurok was still used as a first language in many households... Users can hear recordings as spoken by six fluent speakers recorded in the 2000s. (Garrett, 2019: 201)

3.4 Integration of multimedia

The heart of each entry is an audio recording of the word by a native speaker, and this is one reason why spelling and orthographic choices need not be critical constraints. In addition to translations in multiple languages, other types of multimedia can be included in any particular entry. Entries may include images or photos. These might be the photographs or artwork of participants in the dictionaries, such as in Figure 4 or Figure 5, or other images available for use under creative commons licenses, such as those from iNaturalist.org in Figure 6, and the line drawings from the ILV / SIL Artwork for Literacy in Mexico (ILV, 2004) in Figure 7. The experience of browsing the Talking Dictionaries is enriched by these visual complements, which is also akin to real-life experiences that speakers have with images when forming cognitive representations of lexical items, especially in the domain of specialized knowledge (Faber, 2012: 225-226).

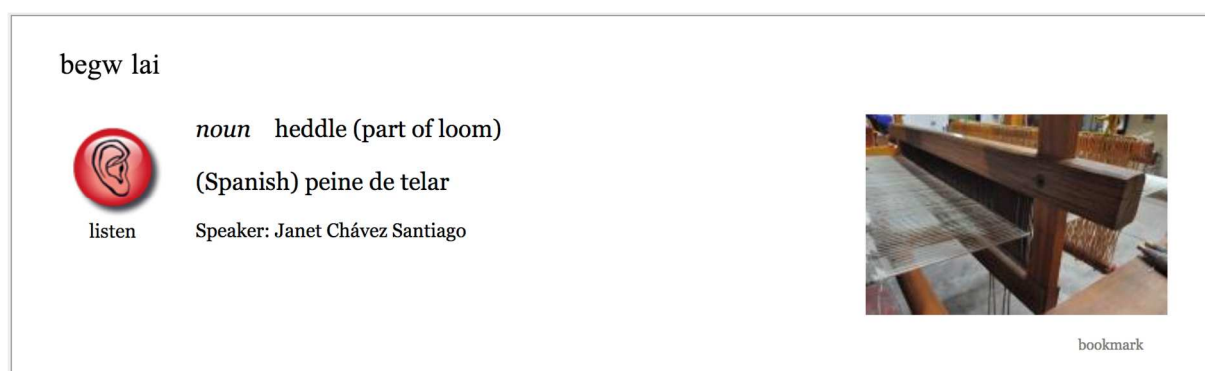




Figure 4: Original photography (Lillehaugen & Chávez Santiago et al., 2019, entry 921)

¡Zac rsily!

Good morning!

Speaker: María Mercedes Méndez Morales

 listen




bookmark


Figure 5: Original artwork by María Mercedes Mendez Morales (Lillehaugen & García Guzmán et al., 2019, entry 2858)

byub [byùù'b]

leaf-cutter ant

Scientific name: *Atta cephalotes*, Speaker: Felipe H. Lopez

 listen




Example: Photo by momoto-erick / iNaturalist, License: CC-BY-NC via [inaturalist.org](https://www.inaturalist.org)


bookmark

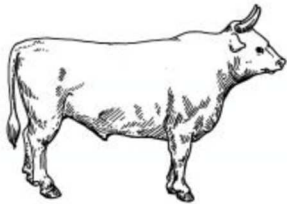
Figure 6: Creative commons photos (iNaturalist.org) in lexical entry (Lillehaugen et al. 2019a, entry 3764)

guan [gùu'ann]

bull; ox

 listen





Example: Guan zugua lo neziu

Speaker: Felipe H. Lopez

bookmark

Figure 7: Creative commons artwork (ILV, 2004) in lexical entries (Lillehaugen et al., 2019a, entry 458)

The entry in Figure 7 illustrates a further feature of the platform, which allows a tweet to be linked from, and embedded in, an entry. These languages have very little in terms of a native speaker written corpus, and most of that corpus is “born digital” on Twitter. In fact, many of the same Zapotec co-authors of the dictionaries are also individuals who write on Twitter in their language (see Lillehaugen, 2016 & 2019). Twitter is not the only domain of digital language activism, as Zapotec language can also be found on YouTube in a variety of contexts including language lessons created by native speaker teachers, such as those created by Talking Dictionary co-author Moisés García Guzmán (<https://www.youtube.com/user/BnZunni>), Zapotec language materials created for the purpose of language documentation, including expressly for the purpose of illustrating a lexical entry, and other Zapotec language materials that aren’t expressly for teaching or documenting Zapotec language. The video embedded in the entry in Figure 8 is a 3-minute long episode of a documentary web series (Dizhsa Nabani; García Guzmán et al., 2018). This episode illustrates the preparation of beans with narration in Zapotec, and adds significant cultural context to the entry.

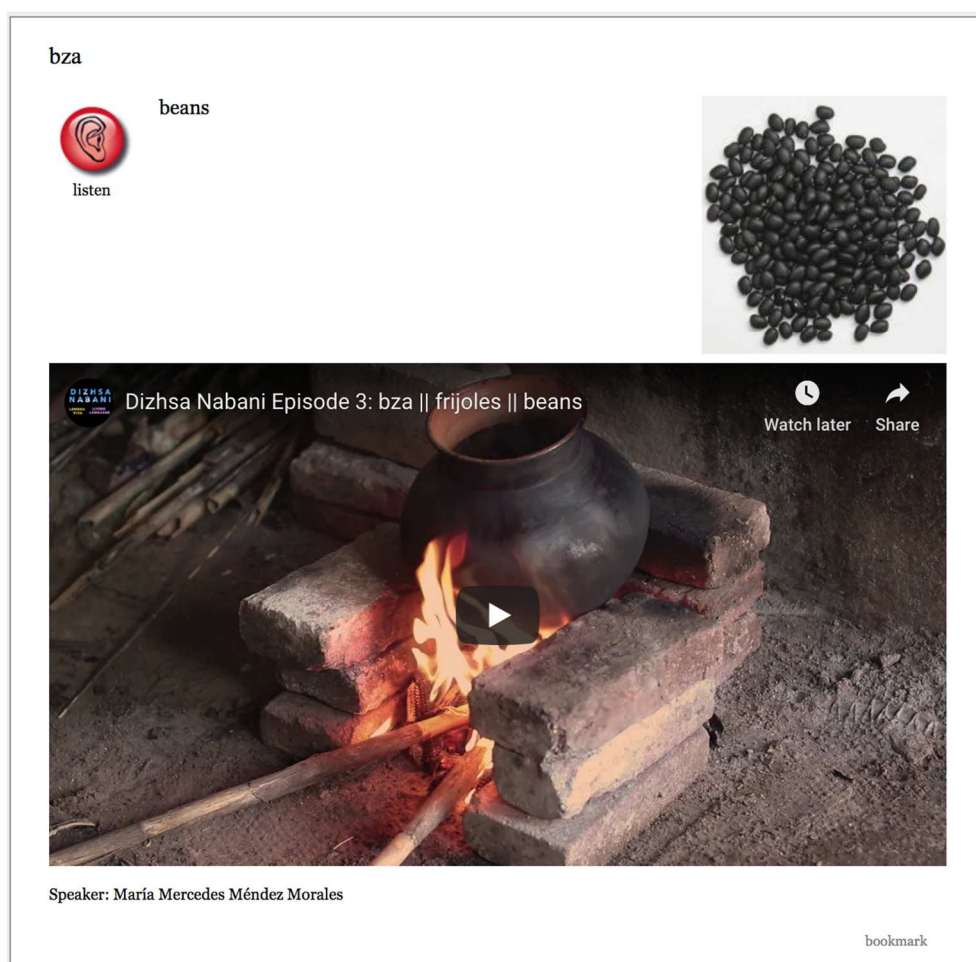


Figure 8: Embedding and linking of YouTube videos in lexical entries (Lillehaugen & García Guzmán et al., 2019, entry 3471)

Consistent with Biesaga’s (2017: 232) observation regarding the use of illustrations in her sample study, multimedia is used extensively for entries in the semantic domain of plants in the Zapotec Talking Dictionaries. When videos are associated with entries for names of plants, they often include a (monolingual) Zapotec scientific explanation, illustrating the more encyclopaedic nature of some of the multimedia components (cf. Biesaga, 2017: 1). In a Zapotec Talking Dictionary being developed by another team for a Northern Sierra Zapotec language, Macuiltianguis (Foreman & Martinez et al., 2019), the embedded videos are used to show verb paradigm information in an accessible format, revealing some of the range of possibilities teams are exploring in the utilization of embedded multimedia.

4. Methodology

The Zapotec Talking Dictionaries have grown since 2013, through the work of linguists, undergraduate students, and Zapotec speakers. Each Zapotec dictionary has local Zapotec co-authors, and the work is highly collaborative, with many community members participating in the creation of the dictionaries. Intense periods of work, usually in the summer, also serve as training and research experience for undergraduate students and Zapotec co-authors. While the original intention was to set up work during the summer that the Zapotec partners would continue during the year, we have come to accept that the natural rhythm of the project varies over the course of the year, with lots of additions of words over the summer, and slower work during the academic year that may focus on technical advances, corrections, and linguistic analysis.

Three of the co-authors of the San Lucas Quiaviní Zapotec Talking Dictionary (Munro, Lopez, and Lillehaugen) were involved in the creation of two print dictionaries on Tlacolula Valley Zapotec: the tri-lingual San Lucas Quiaviní Zapotec dictionary (Munro & Lopez et al., 1999) and the fourth volume of the textbook *Cali Chiu?* (Munro et al., 2008), which uses a revised, simplified orthography based on Munro and Lopez et al. (1999). These print dictionaries were our starting points for the San Lucas Quiaviní Talking Dictionary.

For both this and the other Zapotec Talking Dictionaries we use a variety of elicitation techniques based on: (i) legacy published sources that are out of copyright or, if copyrighted, up to 10% can be used in fair use; (ii) prepared word lists, such as SIL’s Rapid Word Collection (<http://www.rapidwords.net/>), or photos from iNaturalist.org; (iii) community generated lists (e.g. photo elicitation); (iv) born-digital corpora (like Zapotec Twitter); and (v) pedagogical materials that are published or informally circulated. Finally, (vi) the more established Zapotec Talking Dictionaries can now serve as starting wordlists for the newer ones.

New lexical data can be gathered by photo elicitation techniques or thematic conversations (around kinship, foodways, etc.). Part of our workflow involves having

large public workshops in the pueblos, where interested community members are invited to record words or otherwise contribute to the project. These workshops facilitate a type of crowdsourcing, which, as Čibej et al. (2015: 71) put forward, “could have lasting consequences on the nature of lexicographic work... as well as the perception, use, and life-cycle of the lexicographic product”. Such community workshops may be held outside in public spaces, as seen in Figure 9, or in collaborators’ homes, as shown in Figure 10, both from Teotitlán de Valle. On more than one occasion, local experts have arrived with their own, often extensive, word lists and dictionaries that they have compiled out of a love for the language, such as the one that can be seen in Figure 10, created by Froilán Carreño Gutiérrez, which contained an impressive number of names of local avifauna. The hand drawn images in Figure 5 came from another personal dictionary created by the teacher María Mercedes Mendez Morales.



Figure 9: Talking Dictionary workshop in Teotitlán del Valle. Photo credit Brook Lillehaugen.

As Zapotec speakers across generations are involved in the creation of the Talking Dictionaries, the methods employed have also facilitated intergenerational language learning. For example, while documenting the semantic domain of weaving in Teotitlán del Valle, Janet Chávez Santiago, the local director of the Talking Dictionary, learned specialized terminology previously unknown to her. Likewise, Felipe H. Lopez learned

many names for medicinal plants as part of dictionary work in San Lucas Quiaviní. In the creation of the dictionaries, even fluent speakers can learn and share knowledge. We have also observed dictionary work fostering meta-linguistic conversations between speakers. While language revitalization work is complex and there is no simple panacea, we value every positive step forward, each small shift in ideology, and all these small moments of language learning.

5. User experience and creative uses

We view the Zapotec Talking Dictionaries as living projects: as both resources and sites for collaboration. As such, user experience with the dictionaries and novel, unexpected uses are particularly inspiring. We cannot always know ahead of time exactly how individuals will want to interact with the dictionary. As attested in the following quote from Teotitlán del Valle Zapotec Talking Dictionary co-author Janet Chávez Santiago, sometimes this may change for a user over time:

When I started teaching Zapotec as a second language I did not have any kind of pedagogical material printed or online. I basically was creating my own material as my classes were developed. In 2012, when I met Professor Lillehaugen for the first time, I told her about the lack of material in my Zapotec variant. She mentioned to me about the Talking Dictionary platform and, without hesitating, offered me her support to create one for my Zapotec variant. At the very beginning I saw the Talking Dictionary as a resource to help my students to practice and learn the language outside the classroom, but soon I understood that it is more than what you can listen to and see on your screen as the final “product”. For me, the Talking Dictionary is a tool that helps to document the language and, most importantly, thanks to the technical and academic support, it is a tool that reunites and involves a community of speakers that reflect on the language before recording any word, and on how the speakers want to show and teach the language to the outside through the dictionary.

(Janet Chávez Santiago, personal communication June 6, 2019)



Figure 10: Janet Chávez Santiago recording words with Froilán Carreño Gutiérrez, seen holding his personal dictionary of Teotitlán del Valle Zapotec. Photo credit Brook Lillehaugen.

5.1 Comparing languages

Just as Chávez Santiago noted that her view of the dictionary changed over time, so has our perception of particular features of the Talking Dictionaries. One such example is the “compare languages” feature, illustrated in Figure 11 for the word ‘guava’. This feature was originally designed with the linguist user in mind, as a way to give insight into the complex dialect continuum of the Tlacolula Valley. When speakers could not remember a particular word, we showed this view and asked if they would like to hear it in other varieties. After listening to other ways of saying the word in nearby pueblos, speakers were often reminded of the word in their language, saying things like, *ah, that sounds like what my grandmother used to say*. Moreover, participants were interested in hearing the words in other varieties even when they knew the word in their own language. After the fact, we realized this should not have surprised us, given the intense level of both active multilingualism / multi-dialectalism and passive understanding of other varieties that exists in the Tlacolula Valley. The dictionary became another space where speakers could hear multiple varieties of Zapotec, echoing events like the weekly

market in Tlacolula, where many Zapotec languages and varieties can be heard side by side.

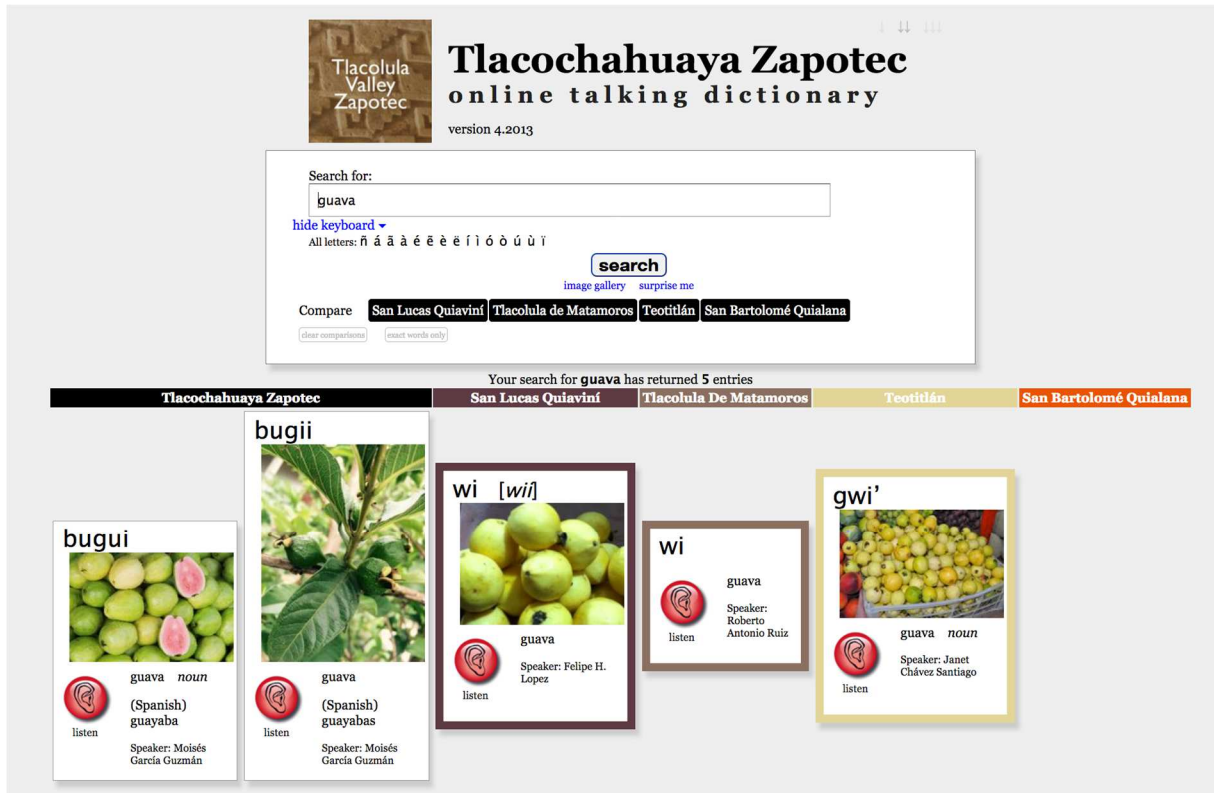


Figure 11: The compare dictionaries feature for “guava”.

5.2 Comparing languages

Another area of unexpected development was growing synergies with other digital platforms. As described above, tweets can be embedded in entries. Likewise, each entry has a unique URL, and these entries can be shared through social media. This type of two-way relationship can also be seen in relation to the documentation of the natural world and biodiversity. We utilize photographs from crowd-sourced naturalist sites like iNaturalist.org and Fishbase.org. This, too, creates a two-way relationship with a global network of scientists, conservationists, and amateur or expert naturalists. Fishbase, in particular, has declared an interest in including “vernacular” names for species, and is open to contributions in Zapotec or any other language.

Two of the co-authors of this paper are also involved in an online text explorer for a corpus of Zapotec language texts written in the Mexican Colonial period: Ticha (<http://ticha.haverford.edu>; Lillehaugen et al., 2016). As this digital scholarship project focuses on a historical corpus of Zapotec texts, the team wanted to be careful to “not reinforce [the] harmful false ideology that Zapotec language and people are only of the past, frozen in time” (Broadwell et al., to appear). The team describes how the Zapotec Talking Dictionaries were used as one intervention:

One way we addressed this was by bringing Zapotec voices to the site. Figure 3 [Figure 12 in this text] shows one of the resources available on Ticha: a vocabulary of the most common words found in the corpus... Wherever possible, we connect these lexical entries for historical forms of words with their modern counterparts, by linking entries in Ticha's Vocabulary with entries in online Talking Dictionaries for several Valley Zapotec language varieties... The design came out of one of the in-person workshops in Oaxaca. As the room full of Zapotec speakers from different communities in the Valley of Oaxaca worked through understanding one of the Colonial era texts together, a pattern of practice emerged. For each word, speakers would go around the table, saying the modern cognate in their variety of Zapotec. The text was read, performed—even echoed—in a multitude of modern Zapotec languages. Ticha's Vocabulary is our attempt to realize this in a digital format. (Broadwell et al., to appear)

A

Search

A	B	C	E	G	H	I	L	M	N	O	P	Q	R	S	T	U	V	X	Y	Reverse Index
----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------------------

=a sp. var. of =ya 'I; me'

aca no, not (negative particle). (Also attested as **acua**, **haca**.)

acua sp. var. of **aca** 'no, not'

alari item (used to mark items in a list). See also **alarini**.

alarini item (used to mark items in a list). See also **alari**. (Also attested as **alatini**, **latini**.)

alatini sp. var. of **alarini** 'item'

ana now. (Also attested as **na**, **yana**.)

anachi 'today'

Hear it in San Lucas Quiaviní Zapotec: **na**

anachi today. (Composed of **ana** 'now', **chi** 'day'.) (Also attested as **anachihi**, **anachi**, **yanachi**.)

Hear it in San Lucas Quiaviní Zapotec: **nazhi**

Hear it in Tlacoachahuaya Zapotec: **an chi**

Hear it in Teotitlán del Valle Zapotec: **nadxi**

Figure 12: Connections to historical corpus on Ticha

5.3 Connecting communities in Oaxaca and the diaspora

As is the case for many indigenous communities in southern Mexico, there is a large diaspora community of Zapotecs in the United States. In the case of Valley Zapotec communities, California (or “Oaxacalifornia”), and especially the greater Los Angeles

area, has become home to hundreds of thousands of Zapotec people (Lopez & Runsten, 2004). The Zapotec Talking Dictionaries, thus, serve—and are created by—transnational communities. Felipe H. Lopez notes that he appreciates being able to point members of his community to the Talking Dictionary when they ask him for resources on how to write their language—and he starts by pointing out that this community exists “on both sides of the border”:

The Talking Dictionary is a very useful and important resource for my community both in San Lucas Quiaviní, Oaxaca, and for many of those who live in California in order to preserve the language. Some members of the pueblo of Quiaviní, on both sides of the border, have requested help to either learn how to write the language or for written material to teach Zapotec to their children or to learn it themselves. Most people who speak the language don't write. For instance, a young woman who leads a local group of young Zapotecs in San Lucas asked me to help them write poems and local stories in Zapotec. In another case, an artisan couple wanted to incorporate Zapotec writing in their promotions for their local textiles and promote their work in various public social spaces in the City of Oaxaca. They reached out to me to help them with written materials in Zapotec. More recently, a Quiaviní woman asked me to help her to translate a Spanish song into Zapotec. Also, within the diaspora community there has been an interest in learning more about the Zapotec language. For example, a Zapotec college student reached out to help her get a Zapotec dictionary or other written materials that would help her to learn her parent's language. Additionally, a couple of parents with young children living in Los Angeles have requested material that they can give their children to learn Zapotec.

As the co-author of the print dictionary of his language (Munro & Lopez et al., 1999), Lopez often receives such requests. He further noted, “before the online dictionary, there was little I could do [in response to these requests] because the San Lucas Quiaviní dictionary is very hard to obtain since it is no longer in print and is expensive. Now with this Talking Dictionary, I can refer people to it, not only to see written Zapotec but also to listen to it. Most of all, it is free.”

5.4 Reaffirming kinship relations and Zapotec identity

Given the diasporic nature of Valley Zapotec communities, it may not be surprising that Zapotec users pay close attention to the identity of the individual speaking for each entry. Many users want to make sure they understand who this person is and how they are related to them, displaying knowledge of kinship relations across the diaspora. This tracking and re-affirming of kinship ties is also a reaffirmation of ethnic identity and mirrors other modes of affirmation of belonging to a pueblo.

Even within the pueblo, the identity of speakers is of central importance in the user experience with the Talking Dictionaries, as these words and this knowledge are never separate from those who came before, as expressed clearly by Moisés García Guzmán, Secretary of Culture of the pueblo of San Jerónimo Tlacochoauaya, and co-author of the Talking Dictionary for his language (Lillehaugen & García Guzman et al., 2019):

The Talking Dictionary personally means that we are able to document all this knowledge in a digital platform and share it with others, and that ultimately leads to better preservation efforts for the language and the community. It's been great to share words that at the same time remind me of stories that my grandmother taught me and be able to link them.

Moisés García Guzmán, personal communication, June 6, 2019

6. Looking forward

We plan to continue developing and expanding the Talking Dictionary platform as a living cultural repository that is community-owned, inclusively co-authored, and fully attributed. The history of lexicography may have largely belonged to the empowered gatekeepers of knowledge enforcing static linguistic norms (Mugglestone, 2011). But the future of lexicography—as envisioned in our work, and in the other papers in this volume—looks very different. It is a future of words technologized, yet remaining under collective ownership and individual authority. Digital lexicography has the potential for constant expansion and can reflect dynamic language change and variation. Finally, by supporting local agency over linguistic resources, the Talking Dictionaries can play a positive role in community-based language revitalization and maintenance. As Moisés García Guzmán said to us: “This is the best tool that we could ever have to save our languages.”

7. Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 1461056 *REU Site: Building Digital Tools to Support Endangered Languages and Preserve Environmental Knowledge in Mexico, Micronesia, and Navajo Nation*. Additional support has been provided by Haverford College, the Center for Peace and Global Citizenship at Haverford College, Living Tongues Institute for Endangered Languages, Swarthmore College, National Geographic Society, Endangered Language Fund, Fundación Alfredo Harp Helú, Biblioteca de Investigación Juan de Córdova, and the Tri-College Department of Linguistics at Haverford, Bryn Mawr, and Swarthmore Colleges.

Thanks to Kate Riestenberg and two anonymous reviewers for their comments on a draft of this paper. Special thanks to Jaime Metzger for her editorial support and especially to Janet Chávez Santiago, Moisés García Guzmán, Aurora Sánchez Gómez,

and Floriana Hernández Martínez for allowing us to work with them and learn alongside them. *Xtyozën yuad!*

8. References

- Anderson, J. & Christen, K. (2019). Decolonizing Attribution: Traditions of Exclusion. *Journal of Radical Librarianship* 5, pp. 113–52.
- Anderson, G. D.S. & Harrison, K. D. (2003). *Tuvan Dictionary*. München: Lincom Europa.
- Anderson, G. D.S. & Harrison, K. D. (2007). *Siletz Talking Dictionary*. Living Tongues Institute for Endangered Languages. <http://siletz.talkingdictionary.org>
- Anderson, G. D.S. & Harrison, K. D. (2016). *Gutob Talking Dictionary*. Living Tongues Institute for Endangered Languages. <http://www.talkingdictionary.org/gutob>.
- Bergenholtz, H., & Bothma, T. J. D. (2011). Needs-adapted Data Presentation in e-Information Tools. *Lexikos / AFRILEX-reeks* 21: 53-77.
- Biesaga, M. (2017). Pictorial Illustrations in Encyclopaedias and in Dictionaries – a Comparison. In I. Kosem, C. Tiberius, M. Jakubíček, J. Kallas, S. Krek, & V. Baisa (eds.), *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference*, pp. 221-236.
- Broadwell, G. A., García Guzmán, M., Lillehaugen, B. D., Lopez, F. H., Plumb, M. H. & Zarafonetis, M. (to appear). Ticha: Collaboration with indigenous communities to build digital resources on Zapotec language and history. *Digital Humanities Quarterly*.
- Eberhard, D. M., G. F. Simons, & Fennig, C. D. (eds.). (2019). *Ethnologue: Languages of the World*. Twenty-second edition. Dallas, Texas: SIL International. Online: <http://www.ethnologue.com>.
- Frawley, W., Hill, K. C. & Munro, P. (2002). *Making Dictionaries: Preserving Languages of the Americas*. Berkeley: University of California Press.
- García Guzmán, M., Brashear, H., Deutch, L., Evans, S. K., Funari, V., Goldberg, K., Jauregui-Volpe, M., Lillehaugen, B. D., Ogborn, E., Palmarini, L. & Rodgers, C. (producers). (2018). *Dizhsa Nabani – Lengua Viva – Living Language*. United States: Haverford College. Online: <https://doculabs.haverford.edu/dizhsanabani/>.
- Garrett, A. (2019). Online Dictionaries for Language Revitalization. In L. Hinton, L. Huss, & G. Roche (eds.) *The Routledge handbook of language revitalization*. Abington-on-Thames: Routledge, pp. 197-206.
- Harrison, K. D. & Anderson, G. D. S. (2006) *Tuvan Talking Dictionary*. <http://tuvan.talkingdictionary.org>
- Hinton, L. 2014. Orthography wars. In M. Cahill & K. Rice (eds.), *Developing orthographies for unwritten languages*. Dallas: SIL International Publications, , pp. 139–168.

- Instituto Lingüístico de Verano, A. C. (2004). *Arte para la Alfabetización en México (Art for Literacy in Mexico)*. Online: <http://www.mexico.sil.org/es/publicaciones/arte4lit>.
- Kemps-Snijders, M. & Wittenburg, P. (2006). LEXUS- a web-based tool for manipulating lexical resources. International conference on Language Resources and Evaluation, Genoa, Italy. Accessed at: https://pure.mpg.de/rest/items/item_61151_3/component/file_61152/content. (15 July 2019)
- Kochetova, K. (2009). LEXUS: from Max Planck Institute for Psycholinguistics Nijmegen. *Language Documentation and Conservation*. Vol. 3, No. 2 (December 2009), pp. 241-246.
- Čibej, J., Fišer, D. & Kosem, I. (2015). The role of crowdsourcing in lexicography. In I. Kosem, M. Jakubiček, J. Kallas, & S. Krek (eds.), *Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference*. Ljubljana/Brighton: Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd., pp. 70-83.
- Faber, P. (2012). *A Cognitive Linguistics View of Terminology and Specialized Language*. Berlin/Boston: De Gruyter Mouton.
- Foreman, J., Martinez, M. with D. Arellano, R. Cabrera, L. Closner, K. Grimaldo, & F. Pérez Ruiz. 2019. *Macuiltianguis Zapotec Talking Dicitonary*, version 0.1. Living Tongues Institute for Endangered Languages. <https://talkingdictionary.swarthmore.edu/macuiltianguis/>
- Lillehaugen, B. D. (2016). Why write in a language that (almost) no one can read? Twitter and the development of written literature. *Language Documentation and Conservation* 10: 356-392. Online: <http://hdl.handle.net/10125/24702>.
- Lillehaugen, B. D. (2019). Tweeting in Zapotec: social media as a tool for language activists. In J. C. Gómez Menjívar & G. E. Chacón (eds.) *Indigenous Interfaces: Spaces, Technology, and Social Networks in Mexico and Central America*, 202—226. Tucson: University of Arizona Press.
- Lillehaugen, B. D., Antonio Ruiz, R. & Antonio Ruiz, J. with C. Batten, H. M. Felker, A. Mannix, K. D. McCormick, & R. E. Weissler. (2013). *Tlacolula de Matamoros Zapotec Talking Dictionary*, pilot version. Living Tongues Institute for Endangered Languages. Online: <http://www.talkingdictionary.org/tlacolula>.
- Lillehaugen, B. D., Broadwell, G.A., Oudijk, M.R., Allen, L., Plumb, M.H., & Zarafonetis, M. (2016). Ticha: a digital text explorer for Colonial Zapotec, first edition. Online: <http://ticha.haverford.edu/>.
- Lillehaugen, B. D. & Chávez Santiago, J. with A. Freemond, N. Kelso, J. Metzger, K. Riestenberg, & K. D. Harrison. (2019). *Teotitlán del Valle Zapotec Talking Dictionary*, version 2.0. Living Tongues Institute for Endangered Languages. Online: <http://www.talkingdictionary.org/teotitlan>.
- Lillehaugen, B. D. & García Guzmán, M. with K. Goldberg, M. M. Méndez Morales, B. Paul, M. H. Plumb, C. Reyes, C. G. Williamson, & K. D. Harrison. (2019). *Tlacoachahuaya Zapotec Talking Dictionary*, version 2.1. Living Tongues

- Institute for Endangered Languages. Online: <http://www.talkingdictionary.org/tlachahuaya>.
- Lillehaugen, B. D., Lopez, F. H. & Munro, P. with S. M. Deo, G. Mauro, & S. Ontiveros. (2019a). *San Lucas Quiavini Zapotec Talking Dictionary*, version 2.0. Living Tongues Institute for Endangered Languages. Online: <http://www.talkingdictionary.org/sanlucasquiavini>.
- Lillehaugen, B. D., Sánchez Gómez, A. & Hernández Martínez, F. with S. M. Deo, G. Mauro, S. Ontiveros, & K. D. Harrison. (2019b). *San Bartolomé Quialana Zapotec Talking Dictionary*, version 1.0. Living Tongues Institute for Endangered Languages. Online: <http://www.talkingdictionary.org/quialana>.
- Mosel, U. (2011). Lexicography in endangered language communities. In P. Austin & J. Sallabank (eds.) *The Cambridge Handbook of Endangered Languages (Cambridge Handbooks in Language and Linguistics)*. Cambridge: Cambridge University Press, pp. 337–353.
- Mugglestone, L. (2011). *Dictionaries: A Very Short Introduction*. Oxford: Oxford Univ. Press.
- Munro, P., Lillehaugen, B. D. & Lopez, F. H. (2008). *Cali chiu? A course in Valley Zapotec*, vol. 1–4. Lulu publishing, www.lulu.com.
- Munro, P. & Lopez, F. H. with O. V. Mendez, R. Garcia, & M. R. Galant. (1999). *Di'csyonaary x:tee'n Dii'zh Sah Sann Lu'uc (San Lucas Quiavini Zapotec dictionary / Diccionario zapoteco de San Lucas Quiavini)*. Los Angeles: UCLA Chicano Studies Research Center Publications.
- Sicoli, M. (2011). Agency and ideology in language shift and language maintenance. In T. Granadillo and H. A. Orcutt-Gachiri (eds.) *Ethnographic contributions to the study of endangered languages*. Tucson: University of Arizona Press, pp. 161-176.
- Wojtylak, K. (2012). LEXUS for creating lexica Version 3.0. Accessed at <https://www.mpi.nl/corpus/manuals/manual-lexus.pdf> (15 July 2019)

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



Resource Interoperability: Exploiting Lexicographic Data to Automatically Generate Dictionary Examples

María José Domínguez Vázquez¹, Miguel Anxo Solla Portela²,

Carlos Valcárcel Riveiro³

¹ Department of English and German Philology and Galician Language Institute, University of Santiago de Compostela

² Department of English and German Philology, University of Santiago de Compostela

³ Department of English, French and German Philology, University of Vigo

Email: majo.dominguez@usc.es, miguel.solla@usc.es, carlos.valcarcel.riveiro@uvigo.es

Abstract

This paper describes the different design and development stages of the MultiGenera and MultiComb prototypes for the multilingual automatic generation of dictionary examples that contain nominal argument patterns at the phrasal and sentence levels. The main objective of MultiGenera is the development of a simulator for the automatic generation of phrases in Spanish, German and French, which is based on the argument patterns of ten valency nouns. The second one, MultiComb, aims to automatically generate the phrasal and sentence contexts of the previously selected nouns in MultiGenera. In the present study we focus on the description of resource interoperability and a set of tools developed to support the methodology of both projects.

Keywords: Valency Dictionary; Argument Patterns; Natural Language Generation; WordNet; Semantics and Ontologies

1. Introduction

The advances in the automatic generation of the natural language have allowed the development of many applications following different methodologies, and thus it has been possible to generate many varied texts, from meteorological forecasts to song lyrics. However, in many cases the texts generated lack meaning or coherence. The *MultiGenera* and *MultiComb* projects were launched to help tackle these problems by exploring the potential of the information contained in valency dictionaries and take advantage of the opportunities offered by WordNet for lexical data extraction. This article presents the different steps taken in developing the tools and prototypes within these projects, focused on the automatic generation of noun phrases and their sentence contexts in Spanish, German and French.

The next section explains in more detail the core principles of the MultiGenera and MultiComb projects. Section 3 focuses on the main features of the PORTLEX dictionary and on how the workflow for this project led to the idea of developing

MultiGenera and MultiComb (for more information see Domínguez Vázquez, Lindemann & Valcárcel Riveiro, 2018). In section 4, the combined theoretical and methodological approaches for the automatic generation of linguistic data are explained. This section describes how prototypical lexical units are obtained for filling in argument slots. Furthermore, it presents the process of lexical expansion, a phase prior to automatic generation, and the role of WordNet ontologies for this purpose. The functionalities and uses of the developed tools (APIs, LEMATIZA, COMBINA and XERA) are also presented in this section. Finally, a brief summary of the main ideas discussed will serve as the conclusion of this work.

2. General framework

The main goal of the MultiGenera project is to develop a tool for automatically generation of nominal phrases in Spanish, German and French. Some pre-project tests (Valcárcel Riveiro & Domínguez Vázquez, 2016) led us to the idea that the semantic acceptability of automatically generated noun phrases may be improved by providing enriched phrasal and sentence contexts. This assumption is actually at the basis of the MultiComb project, which aims to offer a simulator for creating acceptable sentence contexts for noun phrases in the three languages involved: Spanish, German and French. It is therefore a question of progressing from a valency noun with its different arguments to a sentence that contains it.

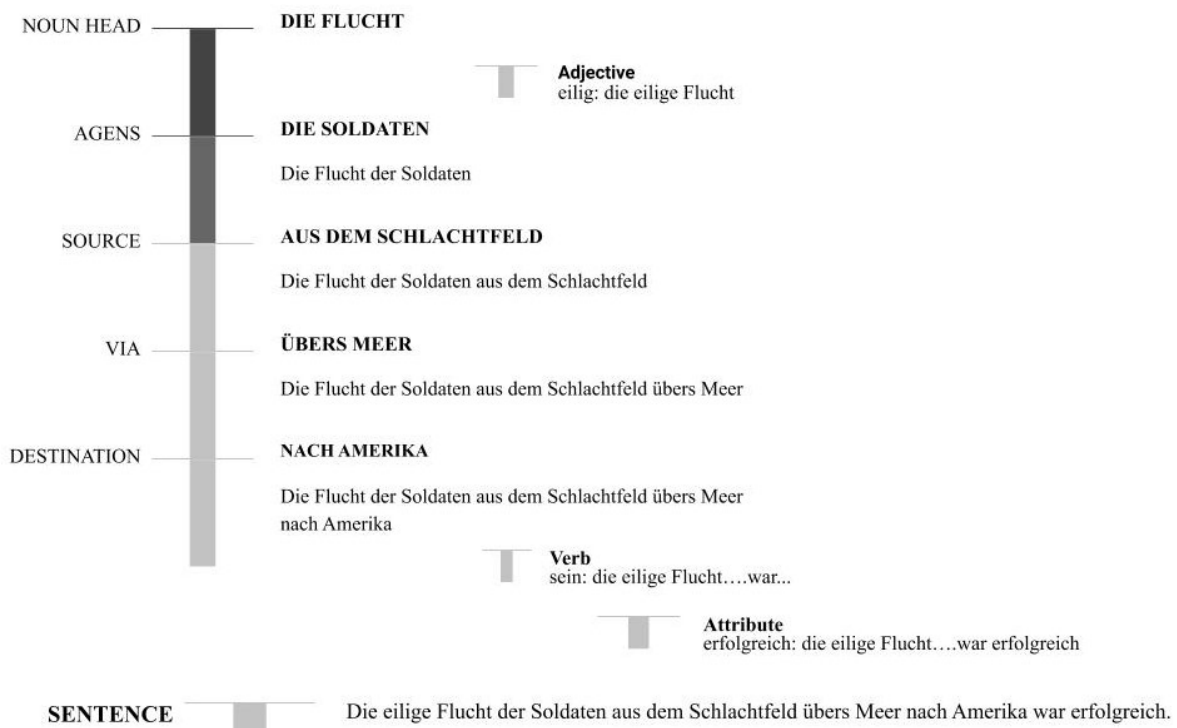


Figure 1: Progression in building examples¹

¹ Literal translation of the example in Table 1: ‘The hasty escape of the soldiers from the battlefield by sea to the Amerika was successful’.

The development of both projects is fed by different theoretical and methodological approaches from different linguistic theories, such as Valency Grammar, Prototypes Theory, Meaning-to-Text Theory and Natural Language Processing (NLP). Furthermore, our combined method utilizes i) the automatic extraction of data from NLP resources, ii) the analysis of corpora, co-occurrence databases and wordnets, iii) as well as the outcoming evaluation produced by both generators.

This paper presents a way of exploiting existing lexicographic information (see section 3) to generate new lexicographic data based on custom-made tools (MultiTools²) and on resource interoperability. Specifically, the following tools have been developed in the current phase:

- 1) Three query APIs, one for each language³, were designed with the aim of extracting lexical data from queries pointing to the semantic relations of WordNet and to the ontologies linked to the synsets in the EuroWordNet model (see 4.3.1). They provide the results in a standard data exchange format (JSON).
- 2) LEMATIZA⁴ analyses exported documents from corpora and provides the lemma of the inflected form of each argument. Each lemma is linked to all the possible queries to the API for the corresponding language. This tool significantly reduces time spent in formulating queries with a semi-automatic query selection (see 4.3.2).
- 3) Another application, COMBINA⁵ makes it possible to combine or crosscheck the results of several API queries. Most of the time, the typology of classes available with simple queries does not conform to an ‘ontology’ of classes based on linguistic semantics. However, a combination of queries offers an enormous variety of possibilities and manages to fine-tune the results with great precision. In addition, these new classes are easily reusable (and even perhaps implementable as a new ontology linked to wordnets) (see 4.3.3).
- 4) A prototype of a generator of noun phrases, XERA⁶, is also being developed for the three languages (see 4.4).

In relation to the foregoing it should be noted that exploring data bootstrapping from NLP resources is interesting for MultiGenera and MultiComb, and therefore for the resources on which they are based. Resource interoperability is understood here in two directions:

² <http://portlex.usc.gal/develop/>

³ The API functionalities are described in the following links, from which queries can also be launched. Spanish API: <http://portlex.usc.gal/develop/es/api/>; French API: <http://portlex.usc.gal/develop/fr/api/>; German API: <http://portlex.usc.gal/develop/de/api/>.

⁴ <http://portlex.usc.gal/develop/lematiza/>

⁵ <http://portlex.usc.gal/develop/combina.php>

⁶ <http://portlex.usc.gal/develop/xera.php>

- 1) The use of data from, for example, WordNet ontological features, PORTLEX's argument patterns (see Section 3) and the dictionaries from the FreeLing tagger (Padró, 2011) for the development of our generators. so that the inflector, although it is also custom-made, reuses FreeLing's dictionaries.
- 2) The use of our generators and tools to improve other resources or design new ones. Thus, for example, resources on lexical selections are offered in JSON format so that they can be used directly by other applications. A further illustration of the intended interoperability is the possible exploitation of our APIs and tools, such as COMBINA and LEMATIZA.

3. The PORTLEX dictionary as a starting point for developing MultiGenera and MultiComb

PORTLEX⁷ is an online valency dictionary of noun phrases with application in language production. It compiles multilingual data in German, Galician, Spanish, Italian and French. The main features of this resource are:

- (1) **valency based** (Engel, 2009): PORTLEX provides detailed information on the nominal phrase from the point of view of valency grammar. This dictionary primarily concerns deverbal (EVALUACIÓN 'evaluation', INVESTIGACIÓN 'research', etc.) and deadjectival nouns (SINCERIDAD 'sincerity', TRANQUILIDAD 'tranquillity', etc.), but also non-derivative nouns that present valency patterns such as PROBLEMA 'problem', GANA 'desire, craving', among others. The specific arguments and semantic roles constitute first-order elements in the entries microstructure. On the one hand, a series of roles are defined to identify the semantic function of the nouns' arguments (e.g. 'that which performs an action', 'that which is affected', etc.) as well as their syntactic function (*subiectivus*, *obiectivus*, etc.). On the other hand, the semantic description also resorts to a list of semantic features ('animate', 'institution', 'object', 'situation', etc.) associated with the valency arguments and present in the different formal realizations of each argument.
- (2) **online** (Klosa, 2013; Müller-Spitzer, 2014) and **semi-collaborative** (Abel & Meyer, 2013; Melchior, 2014): Regarding its medial features, this dictionary was developed as an online and continuously updated resource based on hypertextualization, user interaction and combined access. It is not a finished work, but is constantly updated thanks to its semi-collaborative nature.
- (3) **modular, multilingual** and **cross-lingual** (Domínguez Vázquez & Valcárcel Riveiro, 2019; Gouws, 2014): Domínguez Vázquez & Valcárcel Riveiro (2019: 140)

⁷ <http://portlex.usc.gal/portlex/>

describe these features as follows: “The PORTLEX dictionary covers five languages contrasted with each other. Indeed, its database is designed to include more languages. It contains a specific module for each language in which data relating to each one of them is stored. These modules are linked to each other through a mother dictionary (Gouws, 2014) where Spanish is the pivot language. This allows the alignment of the data of each language and enables their contrastive display according to the user’s needs. In this way, PORTLEX can be defined not only as a multilingual dictionary, but above all also as a cross-lingual dictionary [...]”.

A valency dictionary should provide syntactic and semantic information that helps its users to improve their linguistic production in a foreign language. Therefore, any valency dictionary must describe the different argument realizations of a lexeme, their combining rules and the syntactic-semantic restrictions attached to them, since its aim is to provide users with a complete and detailed description of argument patterns (Domínguez Vázquez, 2018). In order to get a broad dataset PORTLEX relied on corpora for the different languages described in the dictionary and thoroughly analysed them. The examination of the compiled corpus-data allowed the observation that many extracted examples or surface realizations did not meet the requirements of a valency dictionary and, in this sense, we encountered difficulties related to the following issues:

- i. The time-consuming corpus-based compilation of all the noun surface realizations. In this case, the search for certain realizations functioning as noun complements, such as adjectives and compounds, is very time consuming, since they are either scarcely represented in the large corpora used or are not found in them even though they do exist.
- ii. The tedious description of the noun argument patterns, i.e. the compilation of all possible combinations and syntactic-semantic restrictions for each argument along with their different surface realizations in the five languages of the PORTLEX dictionary. The combination patterns of the German noun FLUCHT ‘flight’/‘escape’ well exemplifies such cases, since it presents four arguments: A1: argument with the role ‘that which performs the action’, A2: Argument with the role ‘origin’, A3: Argument with the role ‘transit’ and A4: Argument with the role ‘destination’.

A1	A2	A3	A4
1. Genitive	1. von + dative	1. durch + accusative	1. in + accusative
2. von + dative	2. von ... aus	2. über + accusative	2. auf + accusative
3. Adjective	3. aus + dative	3. via	3. nach + dative
4. Compound	4. Compound		4. zu + dative
			5. bis + preposition + dative

Figure 2: Arguments and surface realizations of the German noun FLUCHT.

In its current state the dictionary describes 61 patterns for the noun FLUCHT, such as the following:

16 monoargumental patterns
$A1_1$ = Die Flucht der Tiere $A1_2$ = Die Flucht von 231 Migranten $A1_3$ = Die väterliche Flucht $A1_4$ = Die Einwohnerflucht $A2_3$ = Die Flucht aus Spanien $A2_4$ = Die Stadtflucht
31 biargumental patterns
$A1_1 + A2_1$ = Die Flucht der Familie aus Spanien $A1_4 + A2_1$ = Die Tierflucht aus dem Zoo $A1_1 + A3_1$ = Die Flucht der Gefangenen durch den Wald $A1_2 + A4_3$ = Die Flucht von DDR-Bürgern nach West-Berlin $A2_3 + A3_2$ = Die Flucht aus Prag über Salzburg $A3_1 + A2_3$ = Die Flucht durch einen Tunnel aus dem Gerichtssaal $A3_3 + A4_3$ = Die Flucht via Jugoslawien nach Österreich $A4_3 + A1_2$ = Die Flucht nach Amerika von Carl Schurz
13 triargumental patterns
$A1_4 + A2_1 + A4_4$ = Die Lehrerflucht von öffentlichen zu privaten Schulen $A2_3 + A3_1 + A4_1$ = Die Flucht aus der Erdgeschosswohnung durch das Fenster in den Innenhof. $A1_2 + A3_2 + A4_3$ = Die Flucht von EU-Bürgern über Thailand nach Japan
1 Tetrargumental pattern
$A1_5 + A2_3 + A3_1 + A4_5$ = Die Flucht von Räufern aus China durch Europa bis in die Schweiz

Figure 3: Argument patterns of the German noun FLUCHT.

As Figure 3 shows, the main difficulties arise in describing the combinatorial arguments, i.e. the interaction of each involved argument in all their realizations and distribution possibilities.

- iii. Corpus-extracted data often do not suit the requirements of a valency dictionary. This is mainly due to the fact that most corpora are not semantically tagged. This is a real concern, as the head of an argument, which represents a certain semantic role (Engel, 1996), must present specific semantic features accordingly, regardless of its formal realization (prepositional phrase, adjective phrase,

apposition, compound name, etc.). As shown in Figure 2, for example, the German lexeme FLUCHT has four different surface realizations for its agent complement (A1). The use of a compound noun ‘agent’+FLUCHT (*Die Einwohnerflucht*) is one of these possible realizations. However, a query on the German web 2013 (deTenTen 13) using Sketch Engine⁸ retrieves all kinds of compounds (*Die Weiterflucht* or *die Berufsflucht*), since these can't be semantically filtered. In fact, most extracted compound nouns do not contain any agent in their first element. A syntactic-semantic analysis of the 100 most frequent lemmas in the mentioned search (Figure 4) shows that a semantic analysis leads us to reject many of them, and this is because the agent of FLUCHT has to feature the semantics characteristics ‘human’, ‘animal’ or ‘vehicle’.

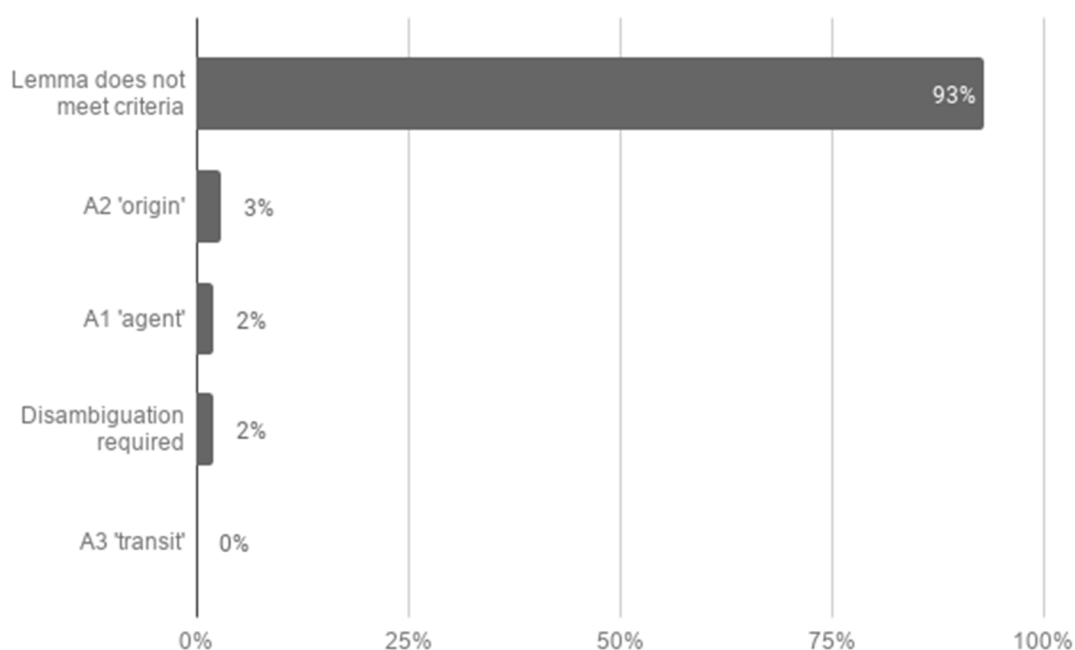


Figure 4: Semantic analysis of the compound nouns retrieved for FLUCHT (deTenTen13)

These cases, in which two or more noun arguments present the same formal realization, are quite frequent. Since we obtain argumental realizations from corpora thanks to their grammatical annotation, in many cases the results show occurrences that are formally similar to the argumental realization that we are searching for, but that actually correspond to another, different semantic role. Thus, very often observing the semantic features of a corpus realization is the only way to determine to which semantic argument it belongs. This means that a human review of the entire list of a query results is necessary to find the examples which can represent a specific semantic role.

⁸ <https://www.sketchengine.eu/>

And it is precisely here where MultiGenera’s strength lies, because this project tackles not only the semantic roles of arguments, but also the distinctive semantic features shared within the lexical paradigms involved in their slot-filling. For this reason, it is not enough to pick up the lexical units retrieved by queries in large corpora (it is not even always representative due to metaphorical uses of the nominal head or their arguments, context dependence for interpretation). The project aims to solve this problem by first identifying the semantic prototypes involved in the roles of the arguments. Ultimately, the purpose is thus the creation of semantically coherent paradigms for the generation of natural language that are independent of context⁹.

4. MultiGenera and MultiComb: theoretical and methodological approaches

4.1 Starting Point

We start from a combined approach for the collection and analysis of data on noun phrases for Spanish, German and French (see section 1). This procedure allows combining valency grammar, the lexical prototype theory, semantic classes and natural language processing (information retrieval and extraction, as well as natural language generation). The automatic generation of the nominal phrase and its arguments relies specifically on a combined method, which is based on the following methodological phases shown in Figure 5:

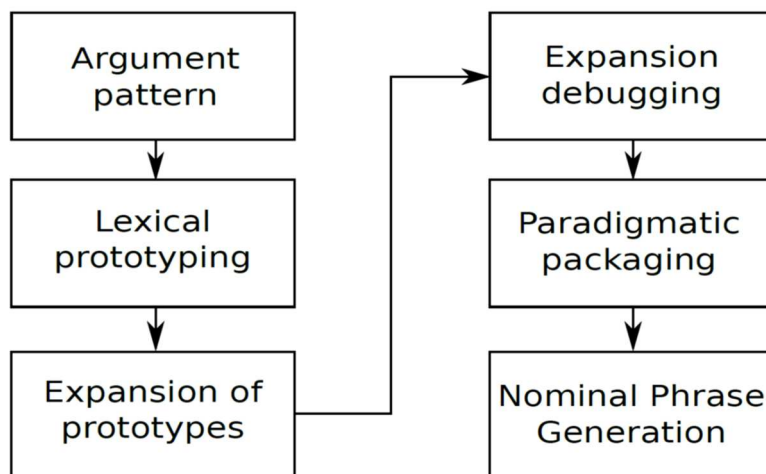


Figure 5: Combined method phases

⁹ MultiComb project deals with the context generation.

In the following sections we will focus on the argument pattern and the lexical prototyping phases (4.2), as well as on the procedure for the prototypes expansion (4.3) and the generation of nominal phrases (4.4).

4.2 Argument pattern and lexical prototyping

The PORTLEX dictionary is used to obtain syntactic and semantic patterns of noun arguments in Spanish, German and French:

Argument s	A1	A2	A3	A4
Semantic role	‘that which performs an action’	Location: origin	Location: transit	Location: destination
Semantic feature	[animate]	[place], [locality], [territory]	place], [locality], [territory]	[place], [locality], [territory]

Table 1: Argument structure of A1 und A2 and semantic features for the German noun FLUCHT.

Argument patterns in PORTLEX provide the parameters for the route queries in Sketch Engine’s corpora. There queries are designed to identify lexical units that could fill the argument slots of the nouns selected. To illustrate it we will provide the following example with FLUCHT: we search precisely for the slot-filling nouns for A2 (semantic role ‘origin’; see Table 1) in coappearance with the preposition *aus* (Table 2). A detailed semantic examination of the examples obtained from CQL¹⁰ queries is carried out following a frequency criterion. Lexical units such as *DDR* ‘GDR’, *Ghetto* ‘ghetto’, *Troja* ‘Troja’, *Haus* ‘home’, *Frankreich* ‘France’, *Ost-Berlin* ‘East Berlin’, *Ostgebieten* ‘eastern territories’ and *Kriegsgefangenenlager* ‘POW camp’ appear frequently in the Sketch Engine corpus as examples for A2-Nouns and thus are, according to our methodological approach, prototypical slot-candidates. The identification of these lexical prototypes makes it possible to define the main semantic classes involved in the slot-filling of each noun argument. This proceeding enables, from these lexical prototypes, to propose the main semantic classes from among the categories of a custom-made linguistic ontology with semantic classes (Table 2):

¹⁰ Corpus Query Language (see <https://www.sketchengine.eu/documentation/corpus-querying/>)

		Lexical prototypes	1st Level	2nd Level	3rd Level	4th Level
FLUCHT aus+dative +		Warschauer Ghetto	situation	location	territory	
		Haus	situation	location	building	
		Kriegsgefangenenlager	situation	location	building	
		Wohnung	situation	location	building	
		Troja	situation	location	locality	proper name
		Ost-Berlin	situation	location	locality	proper name
		Venedig	situation	location	locality	proper name
		Frankreich	situation	location	territory	proper name
		Deutschland	situation	location	territory	proper name
		Italien	situation	location	territory	proper name

Table 2: Example of semantic annotation of lexical prototypes for the argument pattern A23
FLUCHT + aus.

By prototyping we get to establish not only the most representative semantic classes of the different argument patterns, but also the constraints involved in the lexical selection of the focal pattern, such as in the following example for the semantic role ‘source’ of the argument pattern A2₃ (FLUCHT aus + dative):

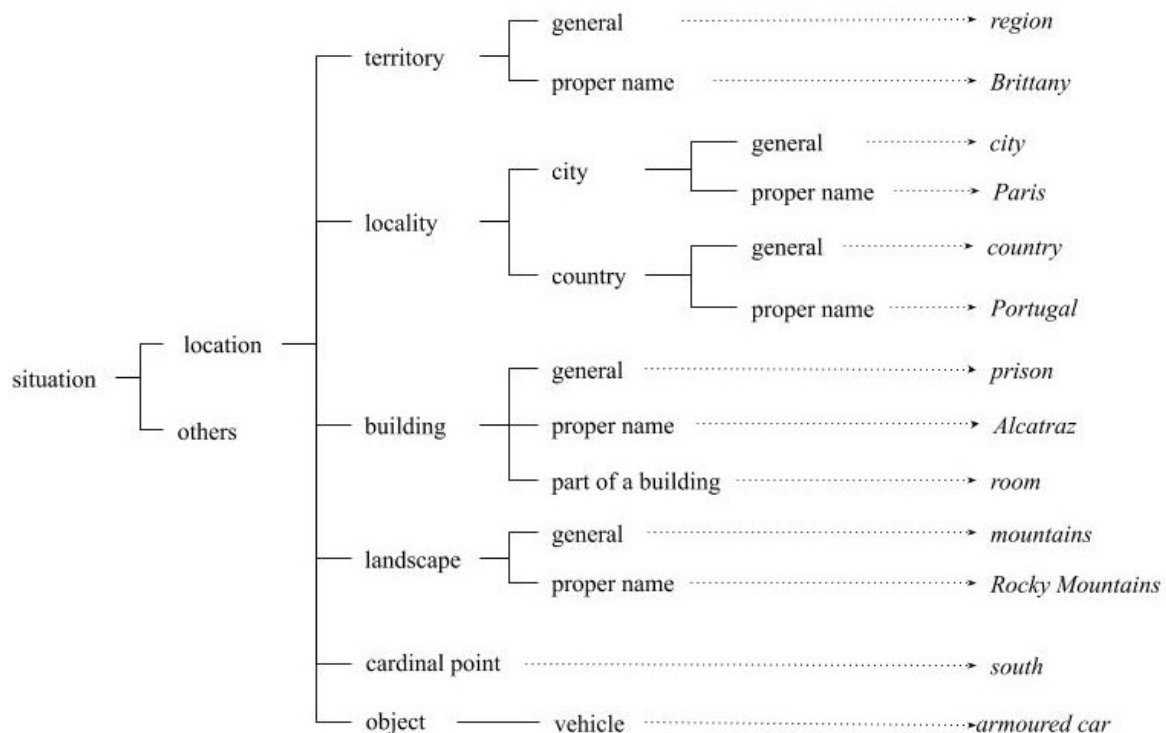


Figure 6: Prototypical semantic classes of FLUCHT + aus

4.3 Expansion of prototypes

4.3.1 Resorting to WordNet

The semi-automatic extraction of lexical candidates for the paradigmatic axis of each argument relies on the fact that the synsets of the wordnets following the EuroWordNet model of the Multilingual Central Repository (MCR)¹¹ (González Agirre & Rigau, 2013) are associated with semantic or cognitive features categorized in different ontologies. In particular, we are dealing with Suggested Upper Merged Ontology¹² (SUMO) (Niles & Pease, 2001), Top Concept Ontology¹³ (Top) (Álvez et al., 2008), WordNet Domains¹⁴ (Bentivogli et al., 2004), Basic Level Concept (Izquierdo et al., 2007) and Epinonyms (Gómez Guinovart & Solla Portela, 2018). Therefore, it is necessary to identify the categories that resort to a concrete wordnet and enable us to fill in the valency slots according to the required semantic feature. For this, besides the ontologies already mentioned, we also use the semantic primes (Miller et al., 1990), i.e., the semantic primitives that organize the lexicographic files of nouns in WordNet, and even the semantic relations among synsets.

Nevertheless, the difficulty in establishing these connections arises from the fact that the cognitive organization of the ontological classifications in the wordnets of the MCR and Galnet¹⁵ (Galician WordNet development interface) do not exactly follow a fully adequate organization for the linguistic description required for MultiGenera. In spite of this, many of the semantic classes defined for our project also constitute categories or general classes in ontologies that are already present in the MCR, such as Top, SUMO, WordNet Domains or Epinonyms. The difficulty consists, therefore, in establishing the appropriate channels for obtaining lexical repertoires with finer semantic granularity to fill in the argument slots of each surface realization. But, in addition, the decision to resort to WordNet has entailed a series of initial tasks, since at the beginning of MultiGenera and MultiComb only Spanish had a wordnet linked to the aforementioned ontologies, as part of the MCR. Thus, the first step undertaken was the creation of databases for French and German. This was done by extracting the alignment between lexical variants and identifying offsets of the meaning from the WordNet Libre du Français¹⁶ (WOLF) (Sagot & Fišer, 2008) and with data from the Extended Open

¹¹ <http://adimen.si.ehu.es/web/MCR>

¹² <http://www.adampease.org/OP/>

¹³ http://globalwordnet.org/gwa/ewn_to_bc/ewnTopOntology.htm

¹⁴ <http://wndomains.fbk.eu/>

¹⁵ <http://sli.uvigo.gal/galnet/index.php?lg=en>. We link to the multilingual web interface of the Galician wordnet to explore the synsets.

¹⁶ <https://gforge.inria.fr/projects/wolf/>

Multilingual WordNet¹⁷ (Bond & Foster, 2013). Both have been made available on the Galnet interface after being converted to the EuroWordNet format of the MCR. In this way, the links with the categories of the ontologies discussed above are available to operate in the three languages of the project. Since syntactic arguments perform semantic roles with their respective ontological-semantic features, we can turn to a lexicon, in this case wordnets, to fill in the argument slots of the selected nouns with lexical units. Expansions of the lexical prototypes described earlier can be made by connecting their semantic classes with the categories of ontologies linked to WordNet in combination with other selection criteria based on the internal structure of this lexical-semantic network. In such a way, through queries in the wordnets, we obtain series of synsets with a meaning that meets the semantic requirements of the lexical paradigms of a noun argument. From these synsets we extract the variants of each language to integrate them into the lexical paradigm of the argument concerned. These connections between semantic classes and WordNet ontological categories can be made using two custom-made designed tools: LEMATIZA and COMBINA.

Figure 7 illustrates that the Semantic Prototype Class (SPC) [situation, location, locality, proper name] is connected with three categories from three different ontologies linked to the wordnets by intersecting the synsets that share these categories.

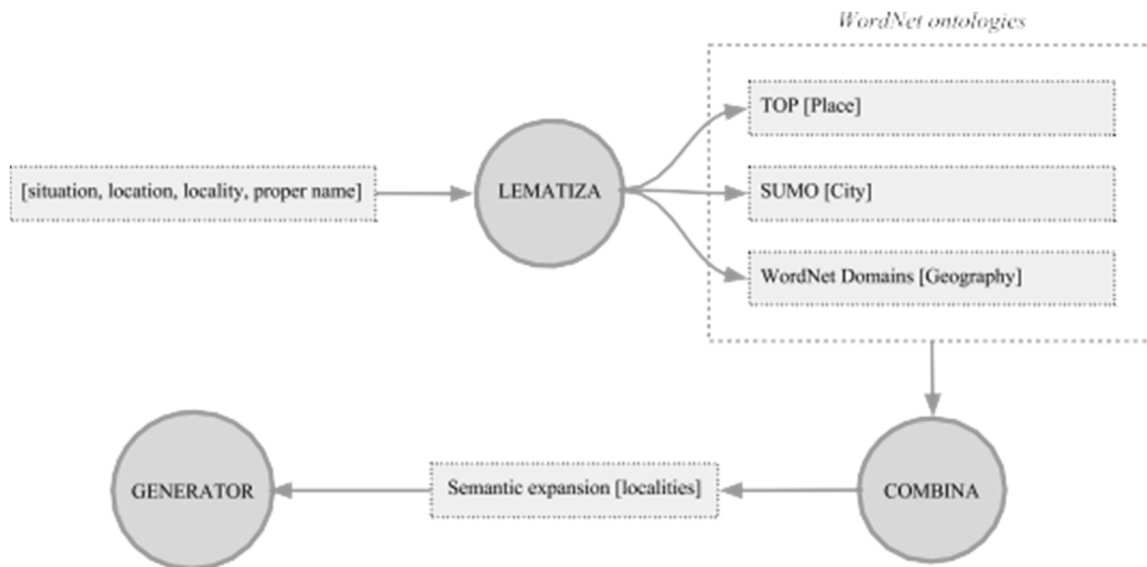


Figure 7: Tools for semantic analyses and expansion by using the wordnets

This procedure allows us to obtain a lexical selection or paradigm with the same semantic characteristics of the initial lexical-semantic prototype. The debugging of the lexical expansion establishes the paradigmatic axis that supports the lexical selection in the automated generation of phrasal contexts. Below the functionalities of the LEMATIZA (4.3.2), COMBINA (4.3.3) and XERA (4.4) will be explained in more detail.

¹⁷ <http://compling.hss.ntu.edu.sg/omw/summx.html>

4.3.2 LEMATIZA

LEMATIZA aims to ease more appropriate queries in the APIs (see section 2). This robust tool allows introducing both concordances and frequency lists, as exported from Sketch Engine, in any of the three languages involved. LEMATIZA returns lemmas from the inflected forms of argument realizations retrieved from CQL queries in Sketch Engine. Each resulting lemma is searched, in turn, in the WordNet of the corresponding language and the output shows each of the synsets in which it is present. In addition, and importantly, this tool provides links to API queries pointing to the ontological categories of each synset, as well as to internal queries to its direct hypernym and hyponyms (see Figure 8) and all its hyponymic descendants. Since LEMATIZA offers links for all the synsets of a lemma, a process of manual disambiguation needs to be carried out to identify the meaning according to that specific usage in the corpus. Disambiguated query links are combined to get the lexical selection for each argument. Moreover, this also allows us to validate the semantic categories of the ontology that we build in order to semantically organize, structure and, when possible, reuse all the lexical selections of our projects.

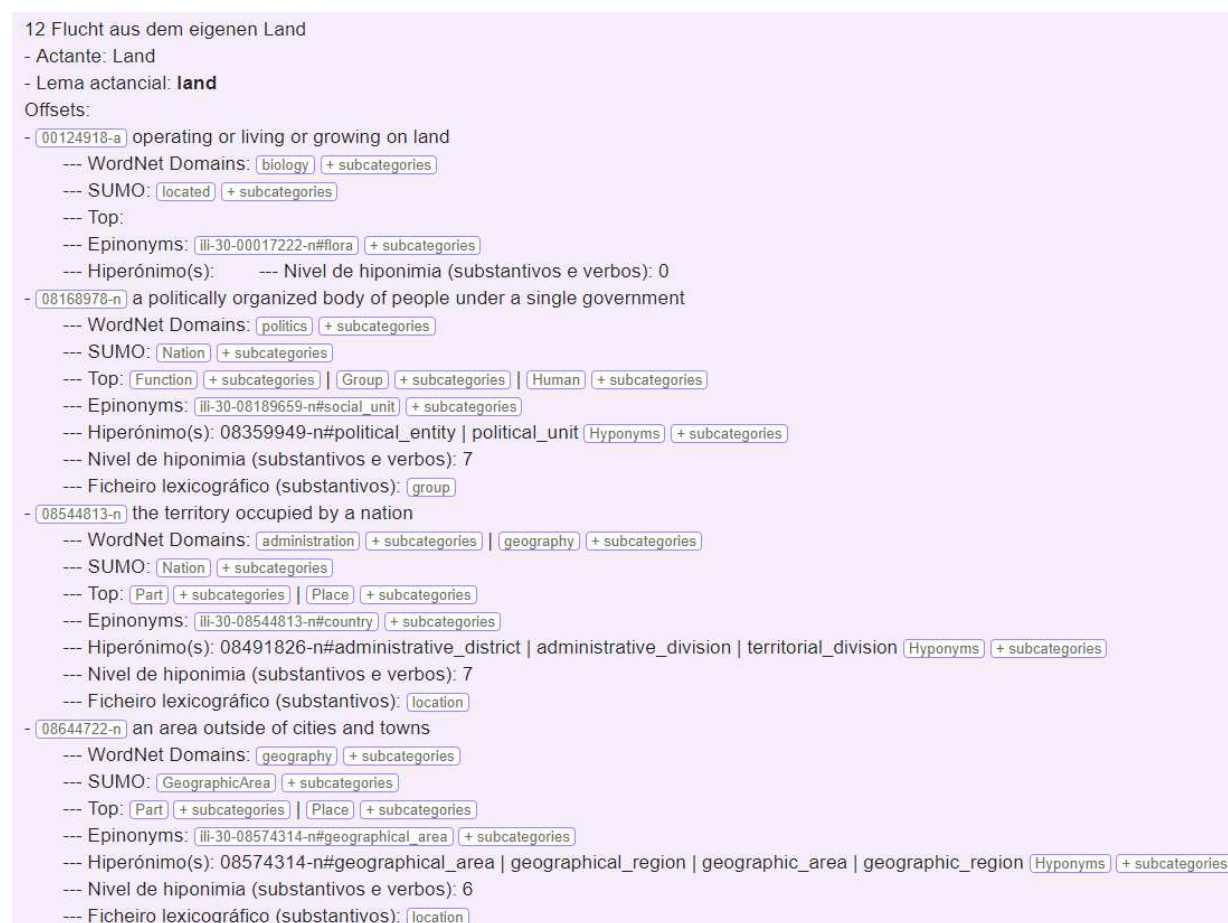


Figure 8: Screenshot (incomplete) of the data retrieved from LEMATIZA

4.3.3 COMBINA

For its part, the COMBINA tool has been developed with the purpose of integrating the API results more accurately. It combines the data from different API queries in the same language, either to add the data from one query to those of another or others (through the combined lemmas option) or to obtain the intersection of the results from different queries (shared lemmas). Figure 9 shows a COMBINA search for German lexemes that belong to the class ‘Buildings’. An example of the results is shown in Table 3.

Figure 9: Screenshot of COMBINA

74 02977936-n Kasino	81 03007130-n Kirche	88 03078506-n Kommunikationszentrum
75 02984203-n Kathedrale	82 02820798-n Klasmühle	89 03089753-n Konferenzzentrum
76 02984061-n Kathedrale	83 03043274-n Klinik	90 03092314-n Konservatorium
77 03032252-n Kino	84 02667576-n Kloster	91 03093427-n Konsulat
78 03028079-n Kirche	85 03054311-n Klubhaus	92 03093427-n Konsulatgebäude
79 02984061-n Kirche	86 04018399-n Kneipe	93 03540595-n Krankenhaus
80 02984203-n Kirche	87 03056288-n Kohlenkeller	94 03043274-n Krankenhaus

Table 3: Results retrieved from COMBINA by crossing API queries.

The results are provided in text format, but also in JSON so that they can be used directly by other applications (such as the prototype generator of MultiGenera, XERA). The debugging of the results constitutes the expanded lexical paradigms used for the automated generation of noun phrases.

4.4 Generation of the nominal phrase: phrasal and sentence context

All these previous steps lead to the design of the generator prototype for noun patterns, XERA¹⁸ (see Figure 10). This tool generates nominal phrases using packaged lexical files built from the results of COMBINA searches. In query mode, it currently uses direct queries to an API or results from COMBINA in JSON format as input for lexical selections. The entire process is performed in real-time. Specific inflectors have been developed for each language, which provide the appropriate form for each context; that is, the inflection of case (only in German), gender and number for determinants, nouns (and the compounds argument + nucleus in the case of German) and adjectives (in German with formal variation depending on the determination, case and gender of the noun they accompany). The code that produces the inflected forms reuses the dictionaries¹⁹ of the well-known tagger FreeLing. The presence of each lemma is verified and inflected forms are obtained by checking the morphosyntactic tags from the corresponding dictionary. In addition, in the case of German, at the moment we also run FreeLing so that it can, sometimes, offer the division into primary lemmas when compound forms are provided from a German wordnet. When the elements are inflected, the concordances and possible restrictions on the usage of all the words in the phrase are verified. The specific contractions of each language are carried out by means of functions that were specifically developed for this purpose.

¹⁸ A more user-friendly interface will be designed in a later phase.

¹⁹ See <https://github.com/TALP-UPC/FreeLing/blob/master/COPYING>

Seleccione a lingua de traballo

Deutsch ☒

español ☐

français ☐

Substantivo nuclear

Geruch ☐

Geschmack ☐

Schmerz ☐

Anwesenheit ☐

Diskussion ☐

Frage ☐

Text ☐

Tod ☐

Zunahme ☐

Flucht ☒

Estrutura argumental

[Determinante] + Flucht + [Determinante] + [Actante-N1G] ☐

[Determinante] + Flucht (sg.) + von + [Determinante] + [Actante-N1D] ☐

[Determinante] + [Actante-A1N] + Flucht ☐

[Determinante] + [Kompositum = Actante-1 + Flucht] ☐

[Determinante] + Flucht (sg.) + von + [Determinante] + [Actante-N2D] ☐

[Determinante] + Flucht (sg.) + von + [Actante-N2D] ☐

[Determinante] + Flucht (sg.) + aus + [Determinante] + [Actante-N2D] ☐

[Determinante] + Flucht (sg.) + aus + [Actante-N2D] ☐

[Determinante] + [Kompositum = Actante-2 + Flucht] (sg.) ☐

[Determinante] + Flucht (sg.) + über + [Determinante] + [Actante-N3A] ☐

[Determinante] + Flucht (sg.) + über + [Actante-N3A] ☐

[Determinante] + Flucht (sg.) + durch + [Determinante] + [Actante-N3A] ☐

[Determinante] + Flucht (sg.) + durch + [Actante-N3A] ☐

[Determinante] + [Kompositum = Actante-3 + Flucht] (sg.) ☐

[Determinante] + Flucht (sg.) + zu + [Determinante] + [Actante-N4D] ☐

[Determinante] + Flucht (sg.) + zu + [Actante-N4D] ☐

[Determinante] + Flucht (sg.) + in + [Determinante] + [Actante-N4A] ☐

[Determinante] + Flucht (sg.) + in + [Actante-N4A] ☐

[Determinante] + Flucht (sg.) + nach + [Actante-N4D] ☐

[Determinante] + [Kompositum = Actante-4 + Flucht] ☐

Vai →

Figure 10: Example of argument patterns on the generator interface

The following screenshot shows the automatic generation for a search of the type “buildings you can flee from”, expressed in German with the preposition *aus*.

163	02927161-n	die Flucht aus diesem Fleischmarkt die Flucht aus dem Fleischmarkte keine Flucht aus den Fleischmärkten
164	08571898-n	die Flucht aus dem Flohmarkt keine Flucht aus dem Flohmarkte keine Flucht aus den Flohmärkten
165	02945813-n	diese Flucht aus dem Flüchtlingslager die Flucht aus den Flüchtlingslagern
166	02687821-n	keine Flucht aus der Flugzeughalle die Flucht aus diesen Flugzeughallen
167	03061505-n	keine Flucht aus der Flugzeugkancel diese Flucht aus den Flugzeugkanceln
168	02715513-n	jene Flucht aus dem Foyer eine Flucht aus den Foyers

Figure 11: Screenshot of XERA: automatically generated noun phrases

After this phase we will have to integrate the adjectives candidates to the Lexical Functions (LF) (Alonso Ramos, Tutin & Lapalme, 1995; Mel'čuk, 1996; Barrios Rodríguez, 2010) in the nominal phrase and generate the sentence context. For this purpose, the selection of LF is based on frequency criteria according to corpora data from Sketch Engine. Returning to the example of FLUCHT, we observe that this noun frequently appears combined with adjectives such as *überstürzt* 'hastily', *dramatisch* 'dramatic', *heimlich* 'secret', *feige* 'cowardly', *missglückt* 'unsuccessful', *schleunig* 'rapid', etc. From this initial frequency selection, the adjectival lexical items are allocated to classes according to the LF, for example as Magn-speed (*überstürzt*, *schleunig*) and Antibon (*feige*, *dramatisch*, *missglückt*), and then we debug and package for each LF²⁰. In this way, we get more natural examples of the nominal phrase:

Magn-speed: *Eine/Die/Jene/Jede [überstürzte schleunige,] Flucht*

Antibon: *Eine/Die/Jene/Jede [feige, dramatische, missglückte,] Flucht*

In the next step we focus on the selection of verbs for each of the central structures (see Table 4). We follow the same procedure as before. In this way we generate sentence contexts with the examples which represent the most frequent valency patterns.

²⁰ Evidently, these paradigmatic sets associated with LF will depend not only on each noun, but also on the specific lexical restrictions of each of the three languages.

subject (NP: Flucht) + verb
<i>gelingen, führen, beginnen, scheitern, enden, verlaufen, geschehen</i>
subject + verb + direct object (NP: Flucht)
<i>ergreifen, antreten, schlagen, planen, verhindern, ermöglichen</i>
subject (NP: Flucht) + copula + attribute
<i>sein</i>
subject + copula + attribute (PP: Flucht)
<i>auf der Flucht sein, sich auf der Flucht befinden</i>
subject + verb (reflexiv) + prepositional complement (PP: Flucht)
<i>sich auf die Flucht begeben, sich auf die Flucht machen</i>
subject + verb + direct object + prepositional complement (prep. + NP: Flucht)
— direct object (accusative) + preposition + accusative
<i>jmdn. in die Flucht schlagen, jmdn. in die Flucht treiben, jmdn. in die Flucht zwingen</i>
— direct object (accusative) + preposition + dative
<i>jmdn. zur Flucht gezwungen, jmdn. zur Flucht verhelfen, jmdn. an der Flucht hindern</i>
— indirect object (dative) + preposition + dative
<i>jmdm. bei der Flucht helfen</i>

Table 4: Sentence frame for the German noun FLUCHT²¹.

Along with the debugging of the phrasal context generation and sentence context there is a combined testing and control phase. This is required because the occurrence of some type of LF might show restrictions concerning the presence of a semantic class of verbs or with some of their arguments or modifiers. For example, with a result from MultiGenera we can obtain the completely acceptable nominal phrase such as a). However, its use in a sentence frame such as b) would be unacceptable from a semantic and communicative point of view:

- | | |
|---|--|
| <p>a) <i>die gelungene Flucht der Deserteure</i>
 ‘the successful escape of the
 deserters’</p> | <p>b) <i>Die gelungene Flucht der Deserteure war
 eine Katastrophe.</i>
 ‘The successful escape of the deserters was
 a catastrophe’</p> |
|---|--|

5. Conclusions

This paper deals with the different design and development stages of prototypes for the automatic generation of linguistic data, which can be directly applied to obtain examples that provide noun argument patterns at phrasal and sentence levels. We focus in particular on the description of the combined method for three languages (Spanish, German and French). The tools presented here make it easier to explore ontologies linked to wordnets and to automate lexical selection procedures in the slot-filling of nominal arguments in the three languages. The final prototype for generating noun phrases is provided with both packaged lexical files and API queries in WordNet

²¹ NP: *Flucht* appears in a nominal phrase. PP: *Flucht* appears in a prepositional phrase.

following the semantic characteristics of the nominal arguments concerned. Certainly, deploying all these developments for the three languages has also been an added challenge due to its contrastive approach. The developments implemented for MultiGenera and MultiComb would not have been possible without the use of a series of tools that were not initially conceived for the generation of natural language. However, the outputs of both projects can also be used freely to improve these or other tools. In this way, the custom-made tools, the packaged lexical files and all the data concerning the combinatorial relations of nominal arguments and its restrictions could be especially useful for new developments.

6. Acknowledgements

This work has been developed within the framework of the projects “Generación multilingüe de estructuras argumentales del sustantivo y automatización de extracción de datos sintáctico- semánticos” - MultiGenera (BBVA Foundation grants for Scientific Research Teams, 2017) and “Generador multilingüe de estructuras argumentales del sustantivo con aplicación en la producción en lenguas extranjeras” - MultiComb (funded by FEDER / Ministry of Economy, Industry and Competitiveness - State Research Agency / Project FFI2017-82454-P).

7. References

- Abel, A. & Meyer, C. (2013). The Dynamics Outside the Paper: user contributions to online dictionaries. In I. Kosem et al. (eds.) *Electronic lexicography in the 21st century: thinking outside the paper: proceedings of the eLex 2013 conference*. Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut, pp. 179-94.
- Alonso Ramos, M., Tutin, A. & Lapalme, G. (1995). Lexical functions of explanatory combinatorial dictionary for lexicalization in text generation. In P. St. Dizier & E. Viegas (eds.) *Computational Lexical Semantics, Studies in Natural Language Processing*. Cambridge: University Press. Révision de la présentation au Second Seminar on Computational Lexical Semantics, Toulouse, 1992, pp. 351-366.
- Álvez, J., Atserias, J., Carrera, J., Climent, S., Laparra, E., Oliver, A. & Rigau, G. (2008). Complete and consistent annotation of wordnet using the top concept ontology. In N. Calzolari, K. Choukri, B. Maegaard, J. Ariani, J. Odiijk, S. Piperidis & D. Tapias (eds.) *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco: European Language Resources Association (ELRA).
- Barrios Rodríguez, M. A. (2010). El dominio de las funciones léxicas en el marco de la teoría sentido-texto. *Estudios de Lingüística del español*, 30, pp. 1-477.
- Bentivogli, L., Forner, P., Magnini, B., Pianta, E. (2004). Revising WordNet domains hierarchy: Semantics, coverage, and balancing. In *Proceedings of COL-ING Workshop on Multilingual Linguistic Resources*, MLR '04. Stroudsburg, PA, USA:

- Association for Computational Linguistics, pp. 101-108.
- Bond, F. & Foster, R. (2013). Linking and extending an open multilingual wordnet. In *51st Annual Meeting of the Association for Computational Linguistics: ACL- 2013*, Sofia.
- Domínguez Vázquez, M. J. & Valcárcel Riveiro, C. (2019). PORTLEX as a multilingual and cross-lingual online dictionary. In M. J. Domínguez Vázquez, M. Mirazo Balsa & C. Valcárcel Riveiro (eds.) *Studies on multilingual lexicography*. De Gruyter: Berlin, pp. 135-158.
- Domínguez Vázquez, M. J., Valcárcel Riveiro, C. & Lindemann, D. (2018). Multilingual generation of noun valency patterns for extracting syntactic-semantic knowledge from corpora (MultiGenera). In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (eds.) *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*. Ljubljana: Ljubljana University Press, pp. 847-854.
- Domínguez Vázquez, M. J. (2018). Was sind Valenzwörterbücher? *Sprachwissenschaft*, 43(3), pp. 309-342.
- Engel, U. (1996). Semantische Relatoren. Ein Entwurf für künftige Valenzwörterbücher. In N. Weber (ed.) *Semantik, Lexikographie und Computeranwendung*. Tübingen: Max Niemeyer, pp. 223-236.
- Engel, U. (2009). *Deutsche Grammatik – Neubearbeitung*. München: Iudicium.
- Gouws, R. (2014). Towards bilingual dictionaries with Afrikaans and German as language pair. In M. J. Domínguez Vázquez, F. Mollica & M. Nied Curcio (eds.) *Zweisprachige Lexikographie zwischen Translation und Didaktik*. Berlin: De Gruyter, pp. 249-262.
- Gómez Guinovart, X. & Solla Portela, M. A. (2018). Building the galician wordnet: methods and applications. *Language Resources and Evaluation*, 52(1), pp. 317-339.
- González Agirre, A. & Rigau, G. (2013). Construcción de una base de conocimiento léxico multilingüe de amplia cobertura: Multilingual central repository. *Linguamática*, 5(1), pp. 13-28.
- Izquierdo Beviá, R., Suárez Cueto, A. & Rigau Claramunt, G. (2007). Exploring the automatic selection of basic level concepts. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*. Shoumen, pp. 298-302.
- Klosa, A. (2013). The lexicographical process (with special focus on online dictionaries). In R. Gouws, U. Heid, W. Schweickard & H. E. Wiegand (eds.) *Dictionaries. An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography*. Berlin, Boston: de Gruyter, pp. 517-524.
- Mel'čuk, I. (1996). Lexical functions: a tool for the description of lexical relations in a lexicon. In L. Wanner (ed.) *Lexical functions in lexicography and natural language processing*. Amsterdam: John Benjamins, pp. 37-102.
- Mel'čuk, I. (2013). *Semantics. From meaning to text, 2*. Amsterdam/Philadelphia: John Benjamins.

- Melchior, L. (2014). Ansätze zu einer halbkollaborativen Lexikographie. *Online publizierte Arbeiten zur Linguistik*, 4, pp. 27-48.
- Müller-Spitzer, C. (ed.) (2014). Using Online Dictionaries. Berlin/Boston: de Gruyter.
- Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D. & Miller, K. J. (1990). Introduction to WordNet: an on-line lexical database. *International Journal of Lexicography*, 3(4), pp. 235-244.
- Niles, I. & Pease, A. (2001). Towards a standard upper ontology. In *FOIS '01. Proceedings of the International Conference on Formal Ontology in Information Systems*. New York: ACM, pp. 2-9.
- Padró, L. (2011). Analizadores multilingües en freeling. *Linguamatica*, 3(2), pp. 13–20.
- Sagot, B. & Fišer, D. (2008). Building a free French wordnet from multilingual resources. In N. Calzolari, K. Choukri, B. Maegaard, J. Ariani, J. Odiijk, S. Piperidis & D. Tapias (eds.) *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco: European Language Resources Association (ELRA).
- Valcárcel Riveiro, C. & Domínguez Vázquez, M. J. (2016). Teste “muerte”: falantes a avaliar a aceitabilidade de frases nominais geradas artificialmente. [<https://carlosvalcarcel.net/2016/11/30/teste-muerte-falantes-a-avaliar-a-aceitabilidade-de-frases-nominais-geradas-artificialmente/>]. Accessed 2 June 2019

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



Croatian Web Dictionary – Mrežnik – Linking with Other Language Resources

Lana Hudeček, Milica Mihaljević

Institute of Croatian Language and Linguistics
Republike Austrije 16, 10000 Zagreb
E-mail: lhudecek@ihjj.hr, mmihalj@ihjj.hr

Abstract

The *Croatian Web Dictionary – Mrežnik* will be a free, monolingual, hypertext online dictionary consisting of three modules (general module for adult native speakers and older schoolchildren, the module for schoolchildren aged 6 to 10, and the module for non-native speakers of Croatian). *Mrežnik* is a corpus-based dictionary, not a corpus-driven dictionary, i.e. the corpus and all data extracted from it serve only as guidelines. The project started on the 1st of March 2017 and the duration of the project is four years. *Mrežnik* is based on these two Croatian corpora: *Croatian Web Repository* (<http://riznica.ihjj.hr/index.hr.html>) and hrWaC – the *Croatian Web Corpus* (<http://nlp.ffzg.hr/resources/corpora/hrwac/>). The paper will focus on the possibilities of linking the *Croatian Web Dictionary – Mrežnik* with other language resources of the Institute of Croatian Language and Linguistics and examples of entries connected to these resources will be shown.

Keywords: Mrežnik; e-lexicography; dictionary links; hyperlinks; Croatian

1. Introduction

Digital or electronic lexicography has gained in importance in the last few years. This can be seen in the increasing number of online dictionaries and publications focusing on the field (Möhre & Töpel, 2011: 199). In the Institute of Croatian Language and Linguistics, the *Croatian Web Dictionary – Mrežnik* is being compiled within the research project IP-2016-06-2141 financed by the Croatian Science Foundation. It is a four-year project and the work started on 1st March 2017. The project will end on 28th February 2021 and the result of the project will be the first Croatian monolingual web-born dictionary.¹ However, the end of the project will not be the end of the compilation of *Mrežnik*. *Mrežnik*² is compiled in TLex, in which the fields have been adapted to the needs of the project. Data extraction from the corpora is performed with the Sketch Engine web tool. The compilation of the dictionary is based on word sketches and a grammar sketch specially developed for the project which allows the display of the lexeme context through WordSketches. The most common collocations are sorted into syntactic categories which enable the discovery of good examples of word usage and collocations. After lexicographic processing is completed, the data will be exported

¹ On the state-of-the-art on Croatian monolingual lexicography see Despot Štrkalj et al., 2019 and on challenges of Croatian e-lexicography see Hudeček, 2018.

² For more about *Mrežnik* see in Hudeček & Mihaljević (2017a, 2017c, 2017d) and Hudeček & Mihaljević (2018a, 2018b).

from TLex to both the web application and the CLARIN European science infrastructure repository (clarin.si repository and the github.com, a software development version control system). *Mrežnik* consists of three modules: the general module (the module for adult and older schoolchildren native speakers of Croatian), the module for younger schoolchildren³, and the module for non-native speakers of Croatian.⁴ The dictionary is corpus-based and the compilers of the dictionary work with these corpora: *Croatian Web Corpus hrWaC* and *Croatian Language Repository*. The lexicographers select data from the corpora as well as from other Croatian dictionaries, websites, and resources. The main aim of the project is to compile a free, monolingual, hypertext, searchable, online dictionary of Standard Croatian with ten thousand dictionary entries in the general module, three thousand words in the module for children, and a thousand words in the module for non-native speakers of Croatian. From the beginning of the project, the plan was that dictionary entries contain links to repositories which will be created as a part of this project, and compiled in parallel with the dictionary as well as with repositories which have already been compiled within other projects conducted at the Institute of Croatian Language and Linguistics.

2. Goals of the *Mrežnik* project

The *Mrežnik* project defines five goals and objectives (Hudeček & Mihaljević, 2018b) two of which are connected to the topic of this paper:

1. Connecting the dictionary with the databases created in parallel with dictionary processing: linguistic advice database (300 pieces of advice for schoolchildren), conjunction database with description of conjunction groups and their modifiers (for all conjunctions in the dictionary), a database of idioms (50 idioms), a database of ethnics and ktetics (300 ethnics and ktetics)⁵, etc.
2. Connecting the basic dictionary with other web sources currently being created at the Institute of Croatian Language and Linguistics – *Valence Database* (<http://ihjj.hr/projekt/baza-hrvatskih-glagolskih-valencija>), *the Database of Collocations* (<http://ihjj.hr/kolokacije/>), *Repository of Metaphors* (<http://ihjj.hr/metafore/>), *Terminology Database Struna* (<http://struna.ihjj.hr/>), *Language Advice*⁶ (<http://jezicni-savjetnik.hr>), and *Better in Croatian* (<http://bolje.hr/>).

³ For more about the module for schoolchildren and non-native speakers see in Mihaljević (2018) and more about the module for non-native speakers of Croatian in Hudeček et al. (2017).

⁴ For more about the structure of the three modules see in Hudeček & Mihaljević (2017a, 2017c, 2017d).

⁵ In Croatian onomastic terminology the term *ethnic* denotes the name of the inhabitants (male and female) of cities, villages, provinces, and countries while a *ktetic* is an adjective derived from the names of cities, villages, provinces and countries.

⁶ The data on this site is, with some additions, from the book by Blagus Bartolec et al. (2015).

Mrežnik is now in its third year. However, after two years of the project these goals were expanded to include some other sources and somewhat modified, as will be shown in the text below.

3. Links in the structure of *Mrežnik*

From the beginning of the compilation of *Mrežnik* links have been considered an important component. In deciding what to link and in which way (data incorporated into the entry, internal or external links) we have taken into consideration the following:

1. "Language is a common good and a common property. Access to information about language should be fast, easy, and intuitive. The electronic dictionary should therefore be a knowledge base with language as its access point, and with simple, yet rich access to (combinations of) linguistic and non-linguistic facts." (...) "At the same time, lexicography must be able to prove itself trustworthy by offering access to sources both for usage and for normative decisions." (Gronvik & Smith Ore, 2013: 243).
2. The possibility of linking data is one of the differences between printed dictionaries and e-dictionaries. Lew (2013: 20), quoting Nesi (1999) and explaining each of the skills that Nesi analyses, says about skill 16 (Understanding the cross-referencing system in print dictionaries, and hyperlinking in electronic dictionaries): "Dictionary users' ability to take advantage of hypertext features of dictionaries is likely to improve with the growing role of the Web in today's life and work. The skill implies awareness of which elements are linked, and what the hyperlinks point to. Principles of user-centered design should ensure that hyperlinks are made evident to the users, but the actual decision of whether to follow a hyperlink needs to be grounded in an awareness of dictionary content and structure."
3. The quality of the dictionary cannot be judged by the number of links. While analysing which elements can contribute to the positive opinion of the users on the dictionary among the elements causing the dictionary to be positively judged Carolina Flinz (2011: 84)⁷ states: "limited use of links, as too many can cause readers to feel lost."

Internal links are links which link one *Mrežnik* entry to another. These links can be divided into two groups:

1. Links linking the lemma (entry word, headword) to another lemma.
2. Links linking the sense of one entry to the sense of another entry.

The difference between these two groups of internal links depends on whether the link refers to a particular meaning or to the whole entry. The links that are placed under sense are links to synonyms, antonyms, hyponyms, co-hyponyms, meronyms, and

⁷ She also stresses careful use of LSP words, which would not be understandable to the average reader. That is the reason why most of the terminology from the *Struna* database is only on hyperlinks and not incorporated into the structure of the *Mrežnik* entry.

masculine/feminine pairs⁸, as is shown on the example of the entry *veslač* (rower, a person who rows) in Table 1. The sense of the entry *veslač* is linked to the sense of the entry *veslačica* (meaning a woman who rows, a female rower). These two entries have similar definitions which differ only in the element *person regardless of gender or male vs female*, as shown in Table 2.

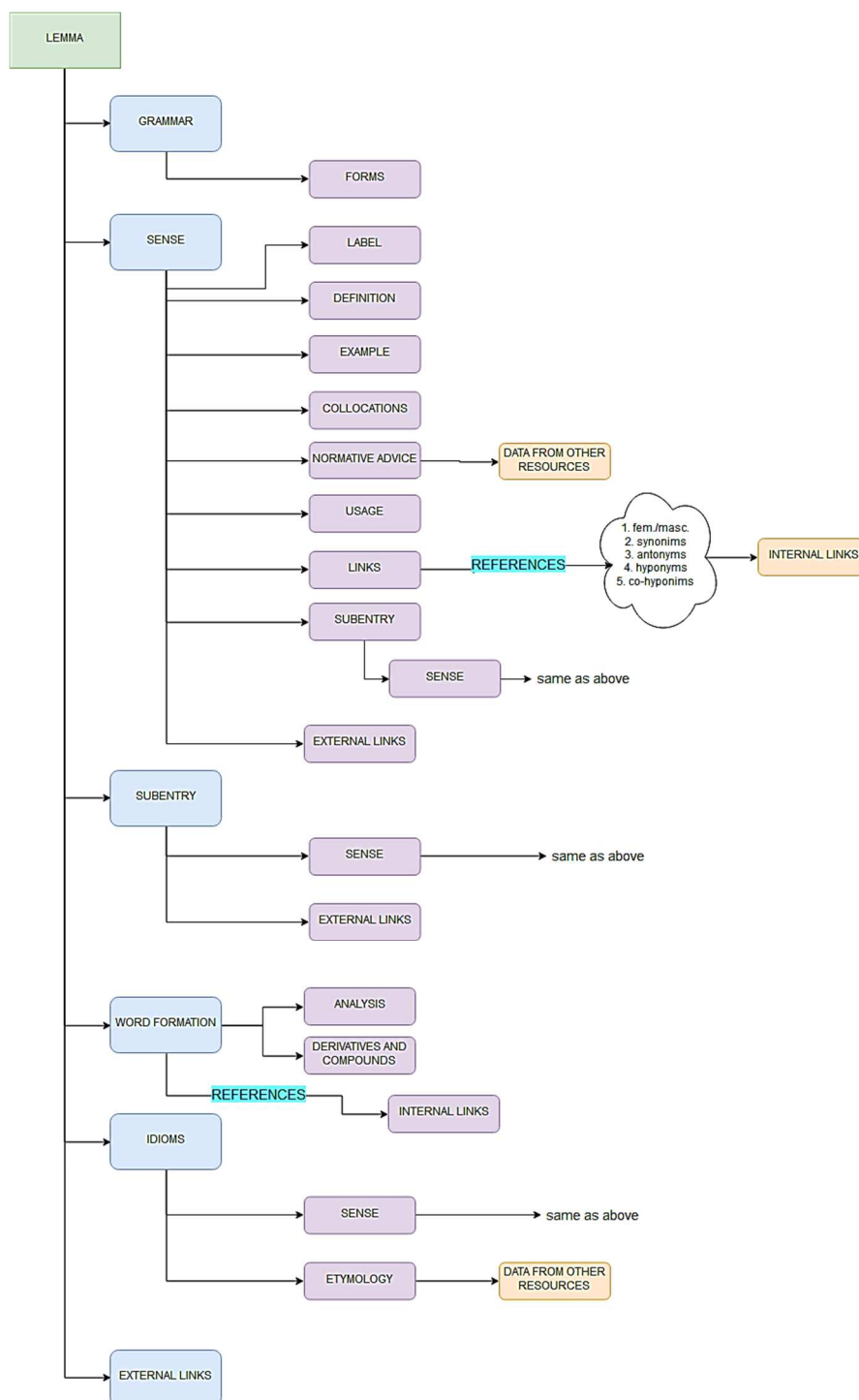


Figure 1. The structure of *Mrežnik* with the position of internal and external links.

⁸ These pairs can (as is the case with *veslač/veslačica*), but do not have to be (as is the case of *medicinska sestra / medicinski tehničar*), linked by word formation.

The structure of *Mrežnik* is shown in Figure 1. in which the position of internal and external links and data integrated from other resources is marked.

Structure elements are illustrated by the entry *veslač* (rower) in Table 1.

	Entry structure
vèslāč veslač im. m. (GA vesláča, DL vesláču, V vèslāču, I vesláčem; mn. NV vesláči, G vesláčā, DLI vesláčima, A vesláče)	grammatical block
Veslač je osoba bez obzira na spol ili muška osoba koja vesla ili se bavi veslanjem.	definition
– Najviše su uspjeha imali veslači mladosti, koji su pobijedili u obje utrke. – Prijelaz Britanaca u četverac mogao bi olakšati posao našim najboljim veslačima, braći Skelin, u borbi za olimpijsko odličje.	examples
Kakav je veslač? brončani, iskusan, juniorski, kvalitetan, mlad, odličan, ponajbolji, regatni, seniorski, uspješan, srebrni Čiji je veslač? britanski, gusarov, hrvatski Što veslač može? imati (uspjeha), nastupati, osvajati (medalju), trenirati, veslati Što se s veslačem može? čestitati mu, omogućiti mu (da treniraju, uvjete), uputiti mu (čestitke) Koordinacija: jedriličari i veslači, kajakaši i veslači, kormilar i veslači, mornari i veslači, olimpijac i veslač, posada i veslači, trener i veslači, vaterpolisti i veslači, odnosi se samo na muškarce: veslačice i veslači	collocations
žensko: veslačica :1	feminine form
tvorenice: veslačev, veslačica, veslački tvorba: vesl-ač	word formation

Table 1: The entry *veslač* (rower) in *Mrežnik*⁹.

The links under the lemma are the links to derivatives and compounds. The word *veslač* is formed by derivation adding the suffix *-ač* to the stem *vesl-*. From the noun *veslač* the derivatives *veslačev* (rower's, belonging to the rower), *veslačica* (female rower), and

⁹ Sense includes definition, examples, collocations, normative notes, pragmatic notes and some links.

veslački (relating to rowers) are derived. Thus, the entry *veslač* is linked to entries *veslačev*, *veslačica*, and *veslački*. This information is not connected to a particular sense of the word. The internal link under sense is the link to the feminine form *veslačica*. Internal links support the systematic nature of definitions in *Mrežnik*, as is shown in Table 2 by comparing the definitions of the entry words *veslač* and *veslačica*.

Definition in <i>Mrežnik</i>	Translation and explanation
Veslač je osoba bez obzira na spol ili muška osoba koja vesla ili se bavi veslanjem.	The rower is a person regardless of the gender or a male person (man or boy) who rows (at this moment) or practices rowing (not necessarily at this moment but is active in rowing).
Veslačica je žena koja vesla ili se bavi veslanjem.	A rower (female) is a female person (woman or girl) who rows (at this moment) or practices rowing (not necessarily at this moment but is active in rowing).

Table 2: Definitions of *veslač* (rower) and *veslačica* (female rower).

4. Integrating external sources into the dictionary entry

From the beginning of the project, the idea was to connect the dictionary entries to repositories and databases which have already been compiled within other projects conducted at the Institute of Croatian Language and Linguistics. These projects are the *Valence Database* (<http://ihjj.hr/projekt/baza-hrvatskih-glagolskih-valencija>), *Repository of Metaphors* (<http://ihjj.hr/metafore/>), *Terminology Database Struna* (<http://struna.ihjj.hr/>) *Language Advice* (<http://jezicni-savjetnik.hr>), and *Better in Croatian* (<http://bolje.hr/>).

The plan was also to create special databases and repositories as a part of the *Mrežnik* project. They are compiled in parallel with the *Mrežnik* dictionary. These repositories are *Language Advice for Schoolchildren* (<http://hrvatski.hr/savjeti>), *Conjunction Repository*, *Repository of Idioms* (<http://hrvatski.hr/frazemi/>), and *Repository of Ethnicities and Ktetics* (<http://hrvatski.hr/etnici-i-ktetici/>).

However, during dictionary compilation, the original plan was somewhat modified. During the compilation of the dictionary, it was decided that the content of the websites *Language Advice*, *Language Advice for Schoolchildren*, *Repository of Idioms*, and *Better in Croatian* should be incorporated into the basic structure of the entry. It was decided that due to the normative nature of *Mrežnik*¹⁰ the language advice component is very

¹⁰ The normative nature of *Mrežnik* is apparent in the selection of entry words, accentuation of entry words, selection of forms in the grammatical block, selection of examples, and language advice in all three modules.

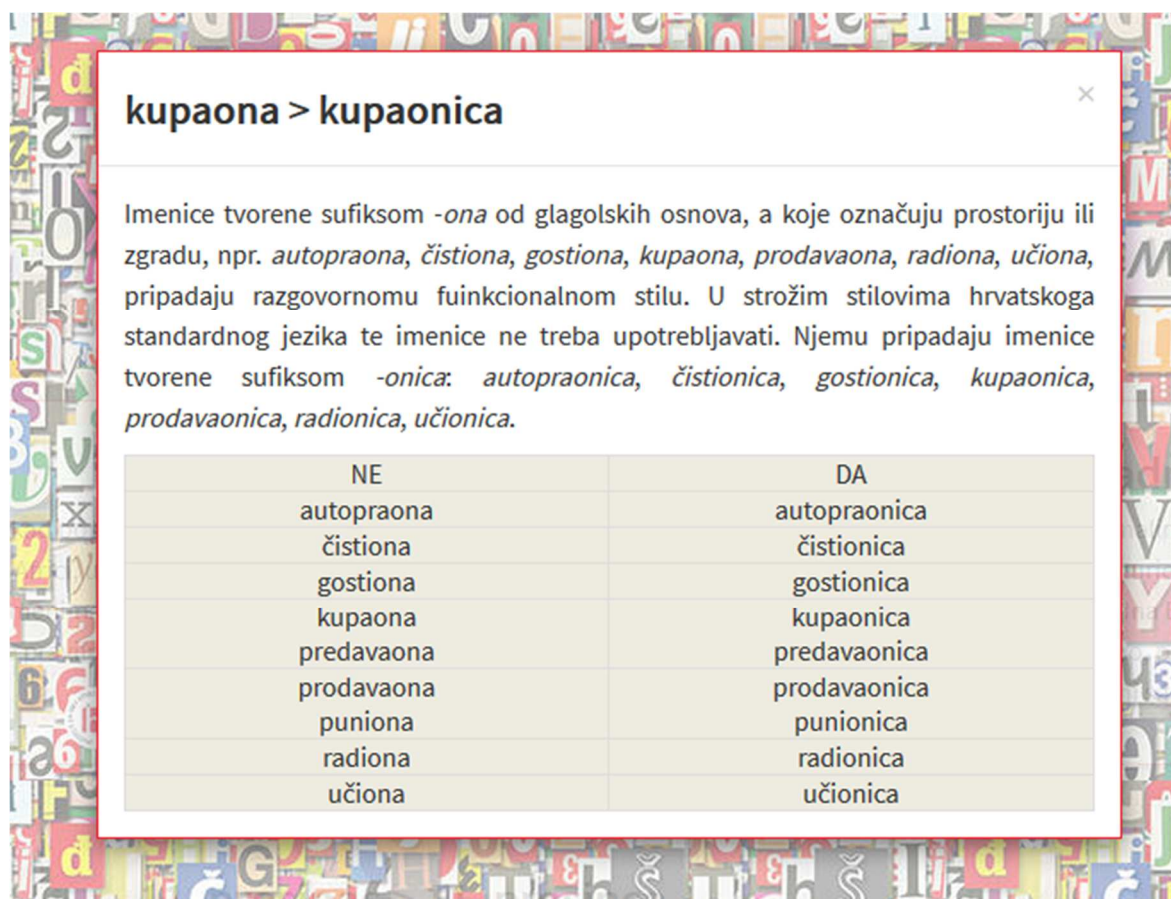
important and should not be presented only as a link. An example of normative advice is shown in Table 3.

	Entry structure
kupaònica kupaonica im. ž. (G kupaònicē, DL kupaònici, A kupaònicu, I kupaònicōm; mn. NA kupaònice, G kupaònicā, DLI kupaònicama)	grammatical block
Kupaonica je prostorija namijenjena održavanju osobne higijene.	definition
Imenice tvorene sufiksom <i>-ona</i> od glagolskih osnova, a koje označuju prostoriju ili zgradu, npr. <i>autopraona</i> , <i>blagovaona</i> , <i>čistiona</i> , <i>gostiona</i> , <i>kupaona</i> , <i>predavaona</i> , <i>prodavaona</i> , <i>radiona</i> , <i>učiona</i> , pripadaju razgovornomu funkcionalnom stilu. U strožim stilovima hrvatskoga standardnog jezika te imenice ne treba upotrebljavati. Njemu pripadaju imenice tvorene sufiksom <i>-onica</i> : <i>autopraonica</i> , <i>blagovaonica</i> , <i>čistionica</i> , <i>gostionica</i> , <i>kupaonica</i> , <i>predavaonica</i> , <i>prodavaonica</i> , <i>radionica</i> , <i>učionica</i> .	normative advice listing words in which <i>-onica</i> should be used instead of <i>-ona</i> .

Table 3: Some elements from the entry *kupaonica* (bathroom) in *Mrežnik*.

The entry *kupaonica* (bathroom) consists of the definition (this entry has only one sense), examples, collocations (What is the bathroom like?, What can we do with the bathroom?, *Koordinacija*:, In connection with the bathroom we mention:). After that, an extensive normative note follows differentiating between the usage of the noun *kupaonica* used in the standard language and the noun *kupaona* used in colloquial speech. As *Mrežnik* is a normative dictionary, dialectal forms *kupatilo*, *banja*, *badecimer* are not mentioned in the advice. The word *kupaonica* is put into a wider context of the language systems as other words with the suffix *-ona* and *-onica* are mentioned.

If we compare the normative advice from *Mrežnik* with the advice on the *Language Advice* site we see that they are the same except for the table given on the *Language Advice* site. In some other cases, there are some differences even in the basic text as it is adapted to every lemma. The advice for the word *kupaonica* from the *Language Advice* webpage is shown in Figure 2.



kupaona > kupaonica

Imenice tvorene sufiksom *-ona* od glagolskih osnova, a koje označuju prostoriju ili zgradu, npr. *autopraona*, *čistiona*, *gostiona*, *kupaona*, *prodavaona*, *radiona*, *učiona*, pripadaju razgovornomu fuinkcionalnom stilu. U strožim stilovima hrvatskoga standardnog jezika te imenice ne treba upotrebljavati. Njemu pripadaju imenice tvorene sufiksom *-onica*: *autopraonica*, *čistionica*, *gostionica*, *kupaonica*, *prodavaonica*, *radionica*, *učionica*.

NE	DA
autopraona	autopraonica
čistiona	čistionica
gostiona	gostionica
kupaona	kupaonica
predavaona	predavaonica
prodavaona	prodavaonica
puniona	punionica
radiona	radionica
učiona	učionica

Figure 2: The advice for the word *kupaona* on the *Language Advice* site.

This piece of language advice is placed in each entry to which it applies and it is somewhat modified if needed so that it mentions the entry word (followed by some other prototype examples to which the rule applies). In *Mrežnik* this piece of advice will appear in the entries of words listed in Table 4.

Colloquial	Standard	English
autopraona	autopraonica	car wash
blagovaona	blagovaonica	dining room
čekaona	čekaonica	waiting room
češljaona	češljaonica	hairdresser's
čistiona	čistionica	drycleaner's
fotokopiraona	fotokopiraonica	photocopying shop
gostiona	gostionica	bar, inn, pub, tavern

Colloquial	Standard	English
kazniona	kaznionica	jail
kladiona	kladionica	betting house
krstiona	krstionica	baptistery
ljevaona	ljevaonica	foundry
praona	praonica	wash-house, laundry-room
predavaona	predavaonica	classroom, lecture-room
prediona	predionica	cotton-mill, spinning mill
prodavaona	prodavaonica	store
propovjedaona	propovjedaonica	pulpit
rađaona	rađaonica	delivery room
radiona	radionica	workshop
skakaona	skakaonica	ski jump hill, diving board
slušaona	slušaonica	listening room
spaliona	spalionica	incineration plant
spavaona	spavaonica	dormitory
štaviona	štavionica	tannery
štediona	štedionica	savings bank
taliona	talionica	smelting plant
učiona	učionica	classroom

Table 4: Words ending in *-ona* and *-onica*.

In each entry, the piece of advice will be modified so the entry word appears in the first place in the first and last sentences of the normative advice, as is shown in Table 5.

Croatian	English
Imenice tvorene sufiksom <i>-ona</i> od glagolskih osnova, a koje označuju prostoriju ili zgradu, npr. x ¹¹ , <i>autopraona</i> , <i>čistiona</i> , <i>gostiona</i> , <i>kupaona</i> , <i>prodavaona</i> , <i>radiona</i> , <i>učiona</i> itd. pripadaju razgovornomu stilu.	Nouns formed by the suffix <i>-ona</i> from verbal stems which denote a room or a building, e.g. x, <i>autopraona</i> , <i>čistiona</i> , <i>gostiona</i> , <i>kupaona</i> , <i>prodavaona</i> , <i>radiona</i> , <i>učiona</i> (car wash, drycleaner's, bar, bathroom, store, workshop, classroom), etc. belong to the colloquial style.
Njemu pripadaju imenice tvorene sufiksom <i>-onica</i> : x, <i>autopraonica</i> , <i>čistionica</i> , <i>gostionica</i> , <i>kupaonica</i> , <i>prodavaonica</i> , <i>radionica</i> , <i>učionica</i> .	To this style ¹² belong those words formed with the suffix <i>-onica</i> belong: x, <i>autopraonica</i> , <i>čistionica</i> , <i>gostionica</i> , <i>kupaonica</i> , <i>prodavaonica</i> , <i>radionica</i> , <i>učionica</i> (x, car wash, drycleaner's, bar, bathroom, store, workshop, and classroom).

Table 5: Normative advice for words ending in *-ona*.

In order to ensure that compilation is conducted in a systematic way we have compiled lists of words belonging to a grammatical or semantic class to which a particular piece of advice applies. Here are some entry words for which such systematic but adapted pieces of advice are given:

- **-ist or -ista**: words ending in *-ist* and not those ending in *-ista* should be used: *aktivist* (activist), *alpinist* (alpinist), *biciklist* (cyclist), *daltonist* (colour blind person), *egoist* (egoist), *harfist* (harp player), *idealist* (idealist), *iluzionist* (ilusionist), *kroatist* (Croatian language specialist), *okulist* (ophthalmologist, eye doctor), *optimist* (optimist), *perfekcionista* (perfectionist), *pesimist* (pessimist), *pijanist* (pianist), *šahist* (chess player), *gitarist* (guitar player), *fagotist* (fagot player), *flautist* (flute player), *vaterpolist* (water polo player),
- **-čičin or -čicin** adjectives derived from female nouns ending in *-čica* should be those ending in *-čičin* and not in *-cičin*: *bacačičin* (belonging to a female thrower), *beračičin* (belonging to a female picker), *boksačičin* (belonging to a female boxer), *crtadžičin* (belonging to a female draughtsman), *dizačičin* (belonging to a female lifter), *djevojčičin* (belonging to a girl), *dostavljačičin* (belonging to a female deliverer), *glasačičin* (belonging to a female voter), *glasoviračičin* (belonging to a female pianist), *gudačičin* (belonging to a female

¹¹ X denotes the entry word.

¹² Meaning the formal style of the Croatian language.

string player), *igračičin* (belonging to a female player), etc.

– **-arov or -arev**: adjectives ending in *-arov* as well as in *-arev* belong to standard Croatian, i.e. their normative status is the same: *alkoholičarev/alkoholičarov* (belonging to an alcoholic), *bankarev/bankarov* (belonging to a banker), *bibliotekarev/bibliotekarov* (belonging to a librarian), *bolničarev/bolničarov* (belonging to a male nurse), etc.

– **-ica or -inja**: advice on when to use female nouns ending in *-ica* and when to use those ending in *-inja*, *antropologica/antopologinja* (female anthropologist), *kandidatica/kandidatkinja* (female candidate), *pedagogica/pedagoginja* (female pedagogue), *psihologica/psihologinja* (female psychologist), etc. is given.

– etc.

In a similar way, data from the database *Better in Croatian* is incorporated into the normative advice. This is illustrated by the entry *poveznica* (link). In the normative advice, the relation between the Croatian term *poveznica* and the English term *link*, which is also used in Croatian, especially in the colloquial style, is explained as shown in Table 6.

	Entry structure
pòveznica poveznica im. ž. (G pòveznicē, DL pòveznici, A pòveznicu, I pòveznicōm; mn. NA pòveznice, G pòveznīcā, DLI pòveznicama)	grammatical block
<i>inform.</i> Poveznica je sličica, riječ ili izraz u dokumentu na internetu koji taj dokument povezuju s kojim drugim dokumentom na internetu.	field label and definition
U engleskome se jeziku nazivom <i>link</i> označuje sličica ili riječi u dokumentu na internetu koje taj dokument povezuju s kojim drugim dokumentom. Umjesto engleske riječi <i>link</i> , koja u hrvatskome pripada samo računalnome žargonu, u standardnome jeziku treba upotrebljavati hrvatski naziv <i>poveznica</i> .	normative advice – stating the relation between Croatian <i>poveznica</i> and English <i>link</i>

Table 6: Some elements from the entry *poveznica* (link).

The piece of advice from the database *Better in Croatian* adapted to the dictionary entry appears in the normative note of these entries: *pisač* (printer), *slagalica* (puzzle), *poslužilac* (server), etc.

It has also been decided that pragmatic data (usage notes), produced as a part of the *Mrežnik* project¹³, should be incorporated into the entry due to the necessary

¹³ At this moment a separate pragmatic database doesn't exist but in the future it would be useful to create it and pragmatic notes from *Mrežnik* could be a starting point.

adjustment to each entry. Pragmatic data will be illustrated by the entry *bog* (an informal greeting coexisting with the greeting *bok*), as shown in Table 7.

	Entry structure
bôg bog <i>usk.</i>	grammatical block
<i>razg.</i> Bog je neformalni pozdrav koji se upotrebljava pri susretu i pri rastanku.	field label and definition
U hrvatskome jeziku u neformalnoj komunikaciji upotrebljavaju se pozdravi <i>bog</i> i <i>bok</i> . Pozdrav bog nastao je skraćivanjem pozdrava <i>pomoz' Bog, Bog s tobom</i> ili <i>daj Bog</i> . Upotrebljava se u primorskim i istočnim dijelovima Hrvatske. Podrijetlo je pozdrava <i>bok</i> nejasno jer se smatra da je ili nastao obezvučivanjem krajnjega suglasnika pozdrava <i>bog</i> ili ga se, što je manje vjerojatno, povezuje s arhaičnim austrijskim pozdravom <i>Bücken, mein Bücken</i> 'naklon, moj naklon'. Upotrebljava se u Zagrebu i sjevernoj Hrvatskoj.	pragmatic (usage) note – stating in which areas of Croatia the greeting <i>bog</i> and <i>bok</i> are used.

Table 7: Pragmatic note (usage) in the entry *bog* (greeting).

In the pragmatic note, after analysing the etymology of the greeting it is stated that the greeting *bok* is used in Zagreb and northern parts of Croatia, while the greeting *bog* is used in the coastal and eastern parts of Croatia. The same pragmatic note appears in the entry *bok*.

Data from the *Repository of Idioms* is also included into the idiom subentry as is shown in the subentry *Ahilova peta* (Achilles' heel) of the entry *peta* (heel) in Table 8.

	Entry structure
péta peta im. ž. (G pétē, D pêti, A pêtu, L pétu, I pétōm; mn. NA pête, G pétā, DLI pétama)	grammatical block
Ahilova peta	subentry
Frazem je nastao prema grčkome mitu o neustrašivome borcu Ahileju, kojega je njegova majka Tetida, u želji da ga učini neranjivim, nakon rođenja umočila u rijeku Stiks. Pritom ga je držala za petu, koja je ostala suha. Tako je peta postala jedino ranjivo mjesto na njegovu tijelu.	etymology of the idiom

Table 8: Explanation of the meaning of the idiom *Ahilova peta* (Achilles' heel) in the entry *peta*.

In the module for schoolchildren advice from the website *Croatian at School* created in parallel with the compilation of *Mrežnik* is used and adapted to each entry, e.g. in the entry *bicikl* (bicycle) the following piece of language advice is incorporated, as shown in Table 9.

Croatian	English
Riječ <i>bicikl</i> pripada hrvatskomu standardnom jeziku, a riječ <i>bicikla</i> pripada nekim hrvatskim dijalektima. Zato su u standardnome jeziku točne rečenice: <i>Vozim novi bicikl.</i> , <i>Kupio sam novi bicikl.</i> , <i>To je moj novi bicikl.</i> , a nisu točne rečenice <i>Vozim novu biciklu.</i> , <i>Kupio sam novu biciklu.</i> , <i>To je moja nova bicikla.</i>	The word <i>bicikl</i> (bicycle) belongs to standard Croatian while the word <i>bicikla</i> belongs to some Croatian dialects. Thus sentences <i>Vozim novi bicikl.</i> , <i>Kupio sam novi bicikl.</i> , <i>To je moj novi bicikl</i> belong to standard Croatian and <i>Vozim novu biciklu.</i> , <i>Kupio sam novu biciklu.</i> , <i>To je moja nova bicikla.</i> do not.

Table 9: Advice for schoolchildren incorporated into the entry *bicikl*.

Although dialectal information does not usually appear in *Mrežnik*, this piece of advice is given due to the frequency of the mistake. For example, if we search for the lemma *bicikl* in hrWaC we find 16,638 occurrences, and if we search for the lemma *bicikla* we find 18,450 occurrences. However, if we look at the concordance of *bicikla*, we see that in many sentences the word *bicikl* was actually used but it has been wrongly lemmatized¹⁴, e.g. under *bicikla* we find these randomly selected sentences which show that the form of the word *bicikl* and not *bicikla* is used:

Voditelj projekta Krešimir Herceg ispred udruge je donirao dva dječja bicikla Dječjem vrtiću Vjeversica u Starom Petrovom Selu, koji je u sklopu Dječjeg vrtića Nova Gradiška.

Za sastavljanje lanca brdskog bicikla potreban je poseban alat.

5. External links

External links are links to other language resources created within the *Mrežnik* project or created separately at the Institute of Croatian Language and Linguistics. The databases compiled in parallel with the *Mrežnik* project are linguistic advice database (300 pieces of advice for schoolchildren), a conjunction database, with descriptions of conjunction groups and their modifiers (for all conjunctions in the dictionary), a database of idioms (50 idioms), a database of ethnics and ktetics (300 ethnics and ktetics). Advice for children and the etymology of idioms is incorporated within the

¹⁴ On a sample of 100 occurrences of the lemma *bicikla* 83 were the occurrences of *bicikl* but were wrongly lemmatized under *bicikl*.

dictionary entry. The database of ethnics and ktetics is linked to the dictionary, entry as is shown in the entry for *bečki* (Viennese) in Table 10 and Figure 3.

	Entry structure
běčkī bečki prid. G bēčkōg(a); ž. bēčkā, s. bēčkō	grammatical block
Bečki je koji se odnosi na Beč i Bečane.	definition
Hrvatski u školi - Etnici i ktetici: http://hrvatski.hr/etnici-i-ktetici/	external link to the database of ethnics and ktetics

Table 10: Entry *bečki* (Viennese) connected to the database of ethnics and ktetics.

<p>Beč, Bečanin, Bečanka, bečki</p> <p>Běč, G Běča, D Běču, A Běč, L u Běču, I Běčom</p> <p>Béčanin, GA Béčanina, DL Béčaninu, V Béčanine, I Béčaninom; mn. NV Béčani, G Béčānā, DLI Béčanima, A Béčane</p> <p>Běčanka, G Běčānkē, DL Běčānki, A Běčānku, V Běčānko, I Běčānkōm; mn. NAV Běčānke, G Běčānkā/Běčānki, DLI Běčānkama</p> <p>běčkī</p> <p>Napomena: U nekim starijim tekstovima i pravopisima s početka XX. stoljeća zapisan je i muški etnik Bečlija. Na kajkavskome području Beč se nekoć nazivao i Dunaj.</p>
--

Figure 3: Entry *Beč* (Vienna) from the database of ethnics and ktetics.

From the planned databases all are connected to some entries in *Mrežnik*. During dictionary compilation it turned out that the database that had been most often linked from *Mrežnik* was the terminological database *Struna*.¹⁵ The entry *broj* (number) in *Mrežnik* is connected to *broj* in *Struna* and the subentries of the entry *broj* in *Mrežnik* (*cijeli broj* (whole number), *glavni broj* (cardinal number), *negativni broj* (negative number), *neparni broj* (odd number), etc. are connected to respective entries in *Struna* as is shown in Figure 4.

¹⁵ *Struna* is a database of Croatian Special Field Terminology. It was officially inaugurated on the web in February 2012. Its aim is to gradually make available to the public the standardized Croatian terminology for all professional domains.

<div> <div>broj</div> <div>im. m. jd.</div> </div>	
definicija	osnovni matematički pojam koji nastaje apstrahiranjem predodžbe o prebrojavanju konačnoga skupa
istovrijednice	engleski: number
podređeni nazivi	cijeli broj, kompleksni broj, prirodni broj, realni broj, slučajni broj, suprotni broj
napomena	Broj također nastaje apstrahiranjem predodžbe o duljini ili ploštini određenih objekata.
razredba	polje: matematika grana: algebra projekt: Izgradnja hrvatskoga nazivlja u matematici – temeljni pojmovi

<div> <div>prirodni broj</div> </div>	
skraćeni oblik naziva	broj
definicija	broj koji pripada skupu $\{1, 2, 3, \dots\}$
istovrijednice	engleski: natural number
podređeni nazivi	glavni broj, prosti broj, redni broj, savršeni broj, složeni broj
simbol	\mathbb{N}
napomena	Ovdje ne navodimo i nulu kao prirodni broj iako je neki matematičari smatraju prirodnim brojem. Skup prirodnih brojeva s nulom uobičajeno se označuje s $\mathbb{N}_0 = \{0, 1, 2, 3, \dots\}$.
razredba	polje: matematika grana: algebra projekt: Izgradnja hrvatskoga nazivlja u matematici – temeljni pojmovi

Figure 4: Entry *broj* (number) in the terminological database *Struna*.

However, during the compilation of *Mrežnik*, it turned out that there are a number of other resources of the Institute of Croatian Language and Linguistics that could be connected to *Mrežnik*: *Croatian School Grammar* (<http://gramatika.hr/>), *Croatian Orthography Manual* (<http://pravopis.hr/>), language games (<http://hrvatski.hr/igre/>), and some papers from the popular journal *Hrvatski jezik* (<https://hrcak.srce.hr/hrjezik>), and this would present useful information. These resources were included only if we thought that they would be useful to potential users.

In the above-mentioned entry *broj* (number), in addition to mathematical meanings there are also grammatical meanings of the same word. These meanings are connected to the *Croatian School Grammar* (the chapter on number).

The entry *palatalization*¹⁶ is also connected to the *Croatian School Grammar*. The entry palatalization has the structure shown in Table 11.

	Entry structure
palatalizacija palatalizacija im. ž. (G palatalizáciĵe, DL palatalizáciĵi, A palatalizáciĵu, I palatalizáciĵom; mn. N palatalizáciĵe, G palatalizáciĵa, DLI palatalizáciĵama, A palatalizáciĵe)	grammatical block
<i>gram.</i> Palatalizacija je glasovna promjena u kojoj nenepčanicima <i>k, g, h, c</i> zamjenjuju nepčanicima <i>č, ž, š</i> ispred <i>e</i> i <i>i</i> .	field label and definition
– Palatalizacija dolazi od latinske riječi <i>palatum</i> , što znači nepce. – U slav. jezicima tzv. prva palatalizacija, provedena u praslav. jeziku, izmijenila je grlene suglasnike <i>k, g, h</i> u <i>č, ž, š</i> (kod nas ispred <i>e</i> : <i>junak</i> – <i>junače</i> ; ispred <i>i</i> : <i>noga</i> – <i>nožica</i> ; ispred nepostojanog <i>a</i> nastalog od poluglasa <i>ɔ</i> : <i>prah</i> – <i>prašak</i>).	examples
Kakva je palatalizacija? prva, druga, treća Što se s palatalizacijom može? ne provesti je, provesti je	collocations
mrtve tvorenice: palatalizacijski	word formation
Hrvatska školska gramatika: http://gramatika.hr/pravilo/palatalizacija/8/#pravilo	external link to Croatian School Grammar

Table 11: The entry *palatalization* connected with the *Croatian School Grammar*.

This is connected to the chapter on palatalization in *Croatian School Grammar* (Hudeček & Mihaljević, 2017b):

¹⁶ Palatalization refers to the process of change in sound in which a non-palatal consonant *k, g, h, c* changes to a palatal consonant *č, ž, š* in front of *e* and *i*.

PALATALIZACIJA

Nenepčanici (nepalatali) *k, g, h, c* zamjenjuju se nepčanicima (palatalima) *č, ž, š* ispred *e i i*.

nenepčanik	k	g	h	c
nepčanik	č	ž	š	č
primjer	majka – majčin	Bog – Bože	duh – duše	učiteljica – učiteljičin

Pri palatalizaciji suglasnici *k i c* uvijek daju *č*, a ne *ć*: *peko**h* – *peče*, *ujak* – *ujače*, *zec* – *zeče*.



Figure 5: The chapter on *palatalizatin* in the *Croatian School Grammar*.

In a similar way entries *imenica*, *glagol*, *pridjev* (*noun*, *verb*, *adjective*), etc. are connected to chapters of the *Croatian School Grammar*.

As can be seen above, the *Croatian School Grammar* has illustrations which can facilitate learning and make it more fun. To the same end, certain dictionary elements have been gamified. This is especially true for the module for non-native speakers and schoolchildren, but even some entries in the basic module have gamification elements.¹⁷ This will be illustrated by the entry *glagolica* (Glagolitic script) shown in Table 12, which is connected to the table of Glagolitic letters from the *Croatian Orthographic Manual* (Jozić et al., 2013) and contain links to games for learning the Glagolitic script.

¹⁷ For more on the gamification of language content see in Mihaljević, J., 2016a, 2016b, 2017, and Mihaljević, A. & Mihaljević, J. (2019).

	Entry structure
glagòljica glagoljica im. ž. (G glagòljicē, DL glagòljici, A glagòljicu, I glagòljicōm)	grammatical block
<i>ling.</i> Glagoljica je najstarije slavensko pismo nastalo polovicom 9 st., koje je poslije 12. stoljeća u stalnoj uporabi samo u Hrvatskoj.	field label and definition
<p>– U Jurandvoru se održavaju radionice u kojima se može naučiti pisati uglatu glagoljicu.</p> <p>– Mogu se pohvaliti da uz ćirilicu i latinicu pišem i čitam glagoljicu.</p> <p>– Ploča je pisana latinicom i glagoljicom te na dva jezika: latinskim i starohrvatskim.</p>	examples
<p>Kakva je glagoljica? ispisana, tiskana, uklesana; kurzivna, obla, uglata</p> <p>Što se s glagoljicom može? pisati njome, uklesati je</p> <p>Koordinacija: bosančica i glagoljica, ćirilica i glagoljica, latinica i glagoljica</p>	collocations
tvorenica: glagoljični	word formation
<p>Hrvatski u školi: http://hrvatski.hr/igra/7/</p> <p>Hrvatski pravopis: http://pravopis.hr/uploads/slova-2.pdf</p>	external links to language games and the <i>Croatian Orthography Manual</i>

Table 12: The entry *glagoljica* (Glagolitic script) from *Mrežnik*.

The entry *glagoljica* is connected to the table of Glagolitic letters from the *Croatian Orthography Manual* shown in Figure 6.

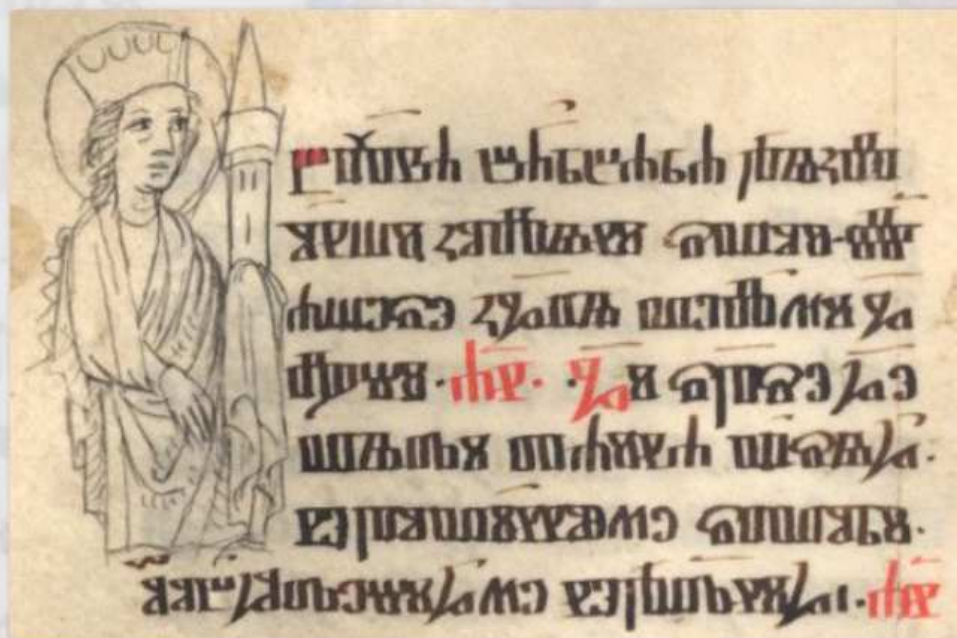
slovo	brojevena vrijednost	transliteracija	transkripcija (čitanje)
ⱦ	1	a	a
Ɱ	2	b	b
Ⱳ	3	v	v
ⱴ	4	g	g
ⱶ	5	d	d
ⱸ	6	e	e, je
ⱺ	7	ž	ž
ⱼ	8	í	z
ⱼ	9	z	z
ⱼ, ⱼ	10	ī	i
ⱼ	20	i	i, ji, j
ⱼ	30	j	j
ⱼ	40	k	k

Figure 6: The Glagolitic script from the *Croatian Orthographic Manual*.

The entry *glagoljica* is also connected to educational games for learning the Glagolitic script. One of these games is shown in Figure 7.

Znam glagoljicu

Nauči uglatu ili oblu glagoljicu s pomoću natjecateljskoga kviza. Za svako pitanje imaš 10 sekunda za odgovor.



započni kviz s ѧѡѣѥѧѧ glagoljicom

započni kviz s **ႤႬႬႬႬႬ** glagoljicom

Figure 7: A quiz for learning the Glagolitic script made by Josip Mihaljević.

The entries and subentries denoting punctuation marks: *zareza*, *točka*, *točka sa zarezom*, etc. (comma, period, semicolon, etc.) are connected to chapters of the *Croatian Orthography Manual*, where rules on how to use these marks are explained. This is illustrated by the subentry *točka sa zarezom* (semicolon) of the entry *točka* (period).

	Subentry structure
točka sa zarezom	subentry
pravop. Točka sa zarezom pravopisni je znak (;) koji se piše pri jačemu odvajanju od onoga koje označuje zarez, a slabijemu od onoga koje označuje točka	definition
<p>– Točka sa zarezom (;) razgodak je koji ima vrijednost između točke i zareza.</p> <p>– Točka sa zarezom na zapešću: Što označava novi trend među korisnicima Instagrama?</p>	examples
U hrvatskome pravopisnom nazivlju u istome se značenju upotrebljavaju nazivi točka-zarez i točka sa zarezom. Budući da u nazivlju istoznačenice nisu poželjne, a polusloženice se ne uklapaju u strukturu hrvatskoga jezika te ih je, kad je to moguće, bolje zamijeniti istoznačnim nazivom drukčije strukture, prednost se daje nazivu točka sa zarezom.	normative advice
mrtvi sinonim: točka-zarez	word formation
Hrvatski pravopis: http://pravopis.hr/pravilo/tocka-sa-zarezom/62/	external link to the <i>Croatian Orthography Manual</i>

Table 13: The subentry *točka sa zarezom* (semicolon) connected with the *Croatian Orthography Manual*.

Točka sa zarezom

Točka sa zarezom piše se:

a) kao znak jačega odvajanja od onoga koje označuje zarez, a slabijega od onoga koje označuje točka: Kad ga vidim, radujem se; ne dođe li, tugujem; teško je reći što će koji dan donijeti.

b) pri nabrananju ako je u cjelinama koje se nabranaju već upotrijebljen zarez:

- Veznici nezavisnosloženih rečenica dijele se na: sastavne – i, pa, te, ni, niti; suprotne – a, ali, no, nego, već; rastavni – ili...
- Tijekom dana toplomjer je pokazivao: 36,7; 37,2; 38,2.

Figure 8: The chapter *Točka sa zarezom* (semicolon) from the *Croatian Orthography Manual*.

Although *Mrežnik* is linked to many language resources from the Institute of Croatian Language and Linguistics, it does not as yet contain a link to an etymological dictionary as such a Croatian dictionary does not exist online yet. Hopefully, the near future will witness a link to an etymological dictionary (the *Croatian Etymological Dictionary* is compiled at the Institute).¹⁸ However, at the moment some entries are linked to short etymological articles from the Institute's journal for the popularization of the Croatian language, *Hrvatski jezik* (*Croatian Language*), which is available online. For example, the entry *ministar* (minister) shown in Table 14 from *Mrežnik* is connected to a short text on the etymology of the words *minister* and *magister* from the etymological section (*Odakle nam riječi?* – Where do words come from?) of the journal *Hrvatski jezik* (Ivšić, 2014), as shown in Figure 9.

¹⁸ Matasović et al., 2016.

	Entry structure
mìnistar ministar im. m. (GA mìnistra, DL mìnistru, V mìnistre, I mìnistrom; mn. NV mìnistri, G mìnistārā, DLI mìnistrima, A mìnistre)	grammatical block
Ministar je osoba ili muškarac koji je na čelu kojega ministarstva.	definition
<p>– Ministar je najavio i novi sustav ocjenjivanja liječnika primarne zdravstvene zaštite na principu od jedan do pet zvjezdica.</p> <p>– Ministar je također naglasio da je trenutna dužnička kriza prilika za stvaranje čvršće fiskalne unije unutar eurozone.</p>	examples
<p>Kakav je ministar? bivši, nekadašnji, resorni, tadašnji; britanski, francuski, njemački, slovenski; HDZ-ov</p> <p>Čega je tko ministar? financija, gospodarstva, obrane, policije, prometa, zdravstva, znanosti</p> <p>Što ministar može? istaknuti (ulogu novih naraštaja, da...), naglasiti (da..., kako ...), najaviti (nastavak suradnje, podiizanje trošarina, sustav ocjenjivanja liječnika, potpisati (memorandum, rješenje, ugovor)</p> <p>Što se s ministrom može? optužiti ga, obavijestiti ga, pitati ga, pozvati ga, smijeniti ga, upoznati ga, zadužiti ga</p> <p>Koordinacija: ministar i potpredsjednik Vlade, ministar i premijer; ministri i saborski zastupnici, odnosi se samo na muškarce: ministri i ministrice</p> <p>U vezi s ministrom spominje se: dužnost, izjava, ostavka, pomoćnik, sastanak, zamjenik</p>	collocations
tvorenice: ministarski, ministarstvo, ministrica, ministrov	word formation
<p>Hrvatski jezik:</p> <p>https://hrcak.srce.hr/index.php?show=clanak&id_clanak_jezik=245998</p>	external links to a paper in the journal <i>Hrvatski jezik</i>

Table 14: Some elements from the entry *ministar* (minister) in *Mrežnik*.

DUBRAVKA IVŠIĆ

Tko je veći, magistar ili ministar?

Etimološke veze hrvatskih riječi *magistar* i *ministar*

Riječ *magistar* u hrvatskome je jeziku razmjerno nova, potvrđena je od 19. stoljeća i ima dva osnovna značenja: 'akademska titula' i 'ljekarnik'. Posuđena je iz latinskoga *magister* 'učitelj, nadređeni, vođa, glava', uz posredovanje njemačkoga jezika. Latinska je riječ postala od kontrastivnoga pridjeva **magis-tero-*, koji u osnovi ima isti indoeuropski korijen **mg'h₂-* 'velik' kao i npr. latinska riječ *magnus* 'velik', grčka *mégas* (μέγας) 'velik', sanskrtska *mahā-* 'velik' (odatle *maharadža*, doslovno 'velik kralj'). Izvorno bi značenje latinske riječi *magister* bilo 'velik u odnosu na druge'.

Latinska riječ *magister* u talijanskome se razvila u *maestro* sa značenjem 'vođa orkestra, dirigent, skladatelj', a iz talijanskoga je posuđena i u hrvatski.

Figure 9: The explanation of the etymology of the word *ministar* (minister) from *Hrvatski jezik*.

6. Conclusion

One of the most important characteristics of *Mrežnik* is that it is a web-born dictionary and thus is created as a hypertext document. This means that it has numerous internal and external links to other entries in *Mrežnik* and to other language resources. Table 15 shows internal and external links in *Mrežnik* and their prototypal linking to the sense or to the lemma. Compounds and derivatives are mostly linked to the lemma, but in some cases they have to be linked to a particular sense, e.g. the entry *bilježnica* means 'notebook' and 'female notary'. The adjective *bilježničin* 'belonging to the female notary' can be derived only from the second meaning of the word *bilježnica*. Terminology is usually linked to the lemma as the specific terminological meaning does not correspond to any sense of the entry. However, in some cases as in the example of the entry *broj* above, it can be linked to a particular sense.

LINKS			
internal		external	
to the sense (prototypical)	to the lemma	to the lemma (prototypical)	to the sense
synonyms, antonyms, hyponyms, meronyms, male/female pair, co-hyponyms	compounds and derivatives	verbal valence, collocations, metaphors, terminology, language games	terminology, <i>Croatian Grammar</i> , <i>Croatian Orthography Manual</i> , journal <i>Hrvatski jezik</i>

Table 15: Links in *Mrežnik*.

However, we constantly posed ourselves the question: Is the dictionary linked to a

certain repository because it is possible or because it is useful? The answer to this question is reflected in each *Mrežnik* entry, and many possible links were not made because we did not consider them useful enough. Table 16 shows the position of different language resources in the structure of *Mrežnik*:

EXTERNAL RESOURCES			
incorporated into the entry		links	
developed in parallel with the <i>Mrežnik</i> project	developed independently	developed in parallel with the <i>Mrežnik</i> project	developed independently
language advice for children, idioms, pragmatic note	language advice, Croatian equivalence for Anglicisms	ethnics and ktetics, conjunctions, language games	verbal valence, collocations, terminology, metaphors, <i>Croatian Grammar</i> , <i>Croatian Orthography Manual</i> , journal <i>Hrvatski jezik</i>

Table 16: External resources in *Mrežnik*.

There are also plans for future linking of *Mrežnik* content to other Croatian dictionaries (e.g. dialectal dictionaries, jargon dictionaries) and also dictionaries of other languages, a sign language dictionary, speech synthesizer, sentiment analyser¹⁹, etc. Ensuring stability of links will not create a problem, as all the linked resources are those created and maintained by the Institute of Croatian language and linguistics in which *Mrežnik* is being compiled.

7. Acknowledgments

This paper is written within the research project *Croatian Web Dictionary—MREŽNIK* (IP-2016-06-2141), financed by the Croatian Science Foundation.

8. References

- Baza hrvatskih glagolskih valencija*. Accessed at: <http://ihjj.hr/projekt/baza-hrvatskih-glagolskih-valencija/27/>. (30 May 2019)
- Blagus Bartolec, G. et al. (2016). *555 jezičnih savjeta*. Zagreb: Institut za hrvatski jezik i jezikoslovlje. Available at: <http://jezicni-savjetnik.hr/>.
- Bolje je hrvatski*. Accessed at: <http://bolje.hr/>. (30 May 2019)
- Štrkalj Despot, K., Hudeček, L., Stojanov, T., & Ljubešić, N. (2019). State-of-the-art

¹⁹ See Mihaljević, J. in print.

- on monolingual lexicography for Croatia (Croatian). *Slovenščina 2.0: Empirical, Applied and Interdisciplinary Research*, 7(1), pp. 65–77.
<https://doi.org/10.4312/slo2.0.2019.1.65-77>.
- Etnici i ktetici*. Accessed at: <http://hrvatski.hr/etnici-i-ktetici/>. (6 June 2019)
- Flinz, C. (2011). The microstructure of online linguistic dictionaries: obligatory and facultative elements. In I. Kosem & K. Kosem (eds.) *Proceedings of eLex 2011, 10-12 November 2011*, Ljubljana: Trojina, pp. 83–88.
 Available at: <http://elex2011.trojina.si/Vsebine/proceedings/eLex2011-10.pdf>.
- Frazemi*. Accessed at: <http://hrvatski.hr/frazemi/>. (6 June 2019)
- Gronvik, O. & Smith Ore, C.-E. (2013). What should the electronic dictionary do for you – and how? In I. Kosem et al (eds.) *Proceedings of eLex 2013 conference*, 17-19 October 2013, Tallinn, Estonia. Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut, pp. 243–260. Available at: http://eki.ee/elex2013/proceedings/eLex2013_17_Gronvik+Ore.pdf.
- Hrvatski u igri*. Accessed at: <http://hrvatski.hr/igri/>. (6 June 2019)
- Hrvatski u školi*. Accessed at: <http://hrvatski.hr/>. (6 June 2019)
- Hrvatsko strukovno nazivlje – Struna*. Accessed at: <http://struna.ihjj.hr/>. (30 May 2019)
- Hudeček, L. (2018). Izazovi leksikografske obrade u jednojezičnome mrežnom rječniku (na primjeru Hrvatskoga mrežnog rječnika – *Mrežnika*). Вісник Львівського університету. Серія філологічна, 69. Львівський національний університет імені Івана Франка. Львів, pp. 29–38.
- Hudeček, L. et al. (2017). Radionica na Croaticumu – provjera rječničke koncepcije modula za strance na terenu. *Hrvatski jezik*, 4(4), pp. 9–12. Available at: https://hrcak.srce.hr/index.php?show=clanak&id_clanak_jezik=304942.
- Hudeček, L. & Mihaljević, M. (2017a). A New Project – Croatian Web Dictionary *MREŽNIK*. In I. Atanassova et al. (eds.) *The Future of Information Sciences. INFuture2017, Integrating ICT in Society*. Zagreb: Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences, pp. 205–213.
- Hudeček, L. & Mihaljević, M. (2017b). *Školska gramatika hrvatskoga jezika*. Zagreb: Institut za hrvatski jezik i jezikoslovlje. Available at: <http://gramatika.hr/>.
- Hudeček, L. & Mihaljević, M. (2017c). Hrvatski mrežni rječnik – *Mrežnik*. *Hrvatski jezik* 4(4). pp. 1–7. Available at: https://hrcak.srce.hr/index.php?show=clanak&id_clanak_jezik=286083.
- Hudeček, L. & Mihaljević, M. (2017d). The Croatian Web Dictionary Project – *Mrežnik*. In I. Kosem et al. (eds.) *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference*. Brno – Leiden: Lexical Computing CZ s.r.o, pp. 172–192.
- Hudeček, L. & Mihaljević, M. (2018a). Croatian Web Dictionary *Mrežnik*: One year later - What is different? In D. Fišer & A. Pančur (eds.) *Proceedings of the Conference on Language Technologies & Digital Humanities*. Ljubljana: Oddelek za prevajalstvo; Inštitut za novejšo zgodovino, pp. 106–113.

- Hudeček, L. & Mihaljević, M. (2018b). *Hrvatski mrežni rječnik – Mrežnik. Upute za obrađivače. Radna inačica*. Available at: <http://ihjj.hr/mreznik/uploads/upute.pdf>.
- Ivšić, D. (2014). Tko je veći, magistar ili ministar? *Hrvatski jezik*, 1(3), pp. 43–44. Available at: <https://hrcak.srce.hr/166798>.
- Jozić, Ž. et al. (2013). *Hrvatski pravopis*. Zagreb: Institut za hrvatski jezik i jezikoslovlje. Available at: <http://pravopis.hr/>.
- Lew, R. (2013). Online dictionary skills. In: I. Kosem et al. (eds.) *Proceedings of eLex 2013 conference, 17-19 October 2013*. Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut, pp. 16–31. Available at: http://eki.ee/elex2013/proceedings/eLex2013_02_Lew.pdf.
- Mihaljević, A. & Mihaljević, J. (2019). Mrežne igre u poučavanju i učenju hrvatskoga jezika. *Dijete i jezik – zbornik radova s Međunarodnoga znanstvenog skupa Dijete i jezik*. Osijek, pp. 113–137.
- Matasović, R. et al. (2016). *Etimološki rječnik hrvatskoga jezika. 1. svezak. A – Nj*. Zagreb: Institut za hrvatski jezik i jezikoslovlje.
- Mihaljević, J. (2016a). Elektroničke mrežne igre za učenje glagoljice. *Baščina*. 17, pp. 35–35.
- Mihaljević, J. (2016b). E-učenje i hrvatski jezik. *Hrvatski jezik*. 3(3), pp. 24–27. Available at: https://hrcak.srce.hr/index.php?show=clanak&id_clanak_jezik=252878.
- Mihaljević, J. (2017). Nove mrežne igre za učenje glagoljice. *Baščina*. 18, pp. 42–43.
- Mihaljević, M. (2018). Hrvatski mrežni izvori za djecu i strance. *Вісник Львівського університету. Серія філологічна*, 69. Львівський національний університет імені Івана Франка. Львів, pp. 75–89.
- Möhrs, C. & Töpel, A. (2011). The “Online Bibliography of Electronic Lexicography” (OBELIX). In I. Kosem & K. Kosem (eds.) *Proceedings of eLex 2011, 10-12 November 2011*, Ljubljana: Trojina, Institute of Applied Slovene Studies, pp. 199–202.
- Nesi, H. (1999). The specification of dictionary reference skills in higher education. In R.R.K. Hartmann (ed.) *Dictionaries in language learning. Recommendations, national reports and thematic reports from the Thematic Network Project in the Area of Languages, sub-project: dictionaries*. Berlin: Freie Universität Berlin, pp. 53–67.
- Repozitorij metafora*. Accessed at: <http://ihjj.hr/metafore/>. (20 February 2019)
- Volim glagoljicu*. Accessed at: <http://hrvatski.hr/volim-glagoljicu/>. (6 June 2019)

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



Representation and Classification of Polyfunctional Synsemantic Words in Monolingual Dictionaries and Language Corpora: The Case of the Croatian Lexeme *Dakle*

Virna Karlič¹, Petra Bago²

¹ Department of South Slavic Languages and Literatures

² Department of Information and Communication Sciences, Faculty of Humanities and Social
Sciences, University of Zagreb, Ivana Lučića 3, HR-10000

Email: {vkarlic, pbago}@ffzg.hr

Abstract

The paper will discuss the central issues concerning lexicographic descriptions of synsemantic words, with special regard to those with multiple syntactic and pragmatic functions. This topic will be exemplified through a description of a representative example, the Croatian lexeme *dakle* (Eng. *well, now; consequently; accordingly, so, then, therefore, thus*). We will focus on the shortcomings of lexicographic descriptions of such words in four contemporary monolingual dictionaries of the Croatian (standard) language. We pay particular attention to the inconsistent part of speech classification in these dictionaries, as well as to the type and content of their definitions, which generally do not take into account multiple syntactic and pragmatic functions of the word. This paper will analyse the functions and the use of lexeme *dakle*, an analysis based on language material extracted from the Croatian web corpus hrWaC, and processed by two independent annotators. We have attained fair agreement between annotators for the first task of determining the (supra)syntactic function (Cohen's κ is 0.4332), and poor agreement for the second task of determining the semantic-pragmatic function (Cohen's κ is 0.2908). Ultimately, the data collected, when compared to dictionary content, can serve as a starting point for a general discussion of an adequate methodology for lexicographic description of polyfunctional synsemantic words.

Keywords: monolingual lexicography; language corpora; pragmatics; synsemantic words; polyfunctionality; Croatian language; lexeme *dakle*

1. Introduction

Lexicographic descriptions of polyfunctional synsemantic (functional / grammatical / closed class) words are often problematic, particularly since they have numerous syntactic and pragmatic functions. Contemporary (but theoretically and methodologically traditional) monolingual dictionaries of the Croatian language reduce the description of this kind of lexeme to its main syntactic-semantic function. However, these lexemes have important and frequently employed pragmatic roles in written and spoken discourse, roles that are generally left out of dictionary definitions. The shortcomings of such descriptions are especially salient in the annotation process of language corpora, resulting in an overly generic categorization of polyfunctional synsemantic words in these annotations. This problem becomes exacerbated as new

dictionaries are compiled based on inaccurately annotated language corpora. We believe this vicious cycle can only be broken by the application of a pragmatic approach to dictionary descriptions of such words.

These issues become clear when specific lexemes are examined. This analysis will focus on the use and syntactic/pragmatic functions of the lexeme *dakle* (Eng. conj. *well, now; consequently* [Bujas, 1999] / *accordingly, so, then, therefore, thus* [Bujas, 2005]). In Croatian monolingual dictionaries¹ the lexeme is categorized as a conjunction or an adverb, while in the Croatian web corpus hrWaC (Ljubešić & Klubička, 2014) over 99% of the occurrences of the word *dakle* are annotated as conjunctions, which is inconsistent with previous linguistic research, as well as our analysis of hrWaC. According to Dedaić (2010), the lexeme has developed four predominant functions in discourse: conclusional, reformulational, argumentative/rhetorical, and attitudinal. In spoken language, especially in scientific discourse, *dakle* is also frequently used as a filler word (Pintarić, 2002; Silić & Pranjković, 2005).

Our research was conducted on a random sample of 400 KWIC examples of the word *dakle* extracted from hrWaC. Every example has been annotated by two annotators on two levels. The first level contains five distinct labels: sentence connective (conjunction), textual (discourse) connective, modifier (particle/adverb), filler word, or “other”. The second, discourse function level also contains five distinct labels, as identified by Dedaić (2010): conclusional, reformulational, argumentative/rhetorical, attitudinal, or “other”. We analysed and compared the distribution of the labels with the descriptions and categorizations of the word in Croatian monolingual dictionaries and web corpora.

The paper is structured as follows: Section 2 discusses general issues observed in lexicographic descriptions of synsemantic words, with emphasis on the contemporary Croatian monolingual (standard) language dictionaries. Section 3 analyses dictionary entries (the types and content of definitions, and part of speech classification) of the lexeme *dakle* as an example of a synsemantic polyfunctional word. Lexicographic descriptions are also compared to the features described in contemporary linguistic studies and grammar textbooks, as well as with the classification applied in the Croatian web corpus hrWaC. Section 4 focuses on the experimental methodology and the annotation results of labelling grammatical/discourse and pragmatic functions on corpus examples of the lexeme *dakle*, followed by Section 5 with a discussion and conclusion.

¹ *Hrvatski jezični portal / Croatian Language Portal* [HJP] (1991–2004), *Rječnik hrvatskoga jezika / Croatian Language Dictionary* [RHJ] (1998), *Školski rječnik hrvatskoga jezika / School Dictionary of Croatian Language* [SRHJ] (2012); *Veliki rječnik hrvatskoga standardnog jezika / Comprehensive Dictionary of Croatian Standard Language* [VRH] (2015).

2. Synsemantic words in (Croatian) dictionaries

On the semantic level, words are classified into two major classes: autosemantic (content / lexical / open-class) and synsemantic (empty / grammatical / functional / closed-class) word-forms. While autosemantic words have lexical meaning and refer to the extralinguistic world independent of their use, synsemantic words serve as functional units with grammatical (operational) meaning; they are used to mark the relations between the language units at a syntactic, semantic, and pragmatic level (Kunzmann-Müller, 1998: 239). In some cases, it is difficult to determine the border between autosemantic and synsemantic words, which is why Kordić (2002) introduces the intermediate category of *words on the border of lexicon and grammar*. The description of words in this intermediate category is a difficult task due to their oscillation between lexical and grammatical status, an alternation which can be observed in dictionaries and grammars of the Croatian language.

Based on an analysis of the descriptions of synsemantic words in Croatian dictionaries, Kunzmann-Müller (1998) concludes that the Croatian lexicography of synsemantic words is just beginning to develop. These language units have so far received fairly little attention as a result of the absence of an adequate theoretical and methodological apparatus, although they have always been included in Croatian dictionaries (ibid. 241-242). For this reason, Hoekstra (2010: 1009) points to the importance of implementing contemporary linguistic insights into lexicographic practice:

To sum, it is important that lexicography stays in touch with the advances that are made in the disciplines of phonology, morphology, syntax and semantics as these disciplines may provide tools for structuring the encyclopedic information about words and collocations that is presented to the laymen who are the primary target group of dictionaries.

The example of the lexeme *dakle* allows us to present the problem of determining how part of speech makes lexicographic analysis and corpus annotation more difficult, and to identify the possible causes for problems with further classification.

While Croatian lexicography currently does not give much attention to synsemantic words, dictionaries specialized for particular synsemantic word classes do exist for some languages.² These approach the subject differently – while some merely list synsemantic words, others describe them in detail, across all language levels. The level of analysis here is, in large part, determined by dictionary type (e.g. a language learning dictionary vs. a monolingual dictionary). For example, Kobozeva and Zakharov (2004) note that a dictionary of discourse markers should include graphic, phonetic, syntactic, semantic, communicative, pragmatic, paralinguistic and derivational information in order to serve

² As an example we list only a few particle dictionaries: *Lexikon deutscher Partikeln* by Helbig (1988); *Dictionary of Slovenian Particles* by Žele (2015); Shimchuk & Shchur: *Slovar' russkix chastic* (1999); *A Dictionary of Japanese Particles* by Kawashima (2000); *A Dictionary of the Chinese Particles* by Dobson (1974) etc. It is worth emphasizing that the lexicographic analysis of individual types of synsemantic words varies greatly, according to their specific grammatical, semantic and functional features.

as a source of study for Russian language learners, but also as a source of further linguistic study. In a discussion about the definition of a lexeme, Hoekstra (2010)—calling upon the work of Bergenholtz (1985) and Coffey (2006)—states that an intentional definition (a paraphrased meaning) is not a suitable solution for synsemantic words, and calls for detailed syntactic descriptions followed by relevant examples of the word’s use.

Osswald (2015) emphasizes that the lexicographic analysis of synsemantic words in monolingual dictionaries is especially problematic, because the definition cannot rely on a denotative meaning. He also explains that such dictionaries usually do not include the syntactic features (or functions) of synsemantic words because “the user is expected to have some basic knowledge of the respective language, and mastering the use function words is considered part of general grammatical competence” (ibid. 7). However, the author points to the “duty of documentation” in monolingual reference dictionaries and calling upon the work of Lang (1989), he concludes that lexicographic descriptions of synsemantic words should “[...] follow grammatical insights; syntactic constructions and their constraints should be part of the entry; and building the entry should consist of two stages, first, recording the relevant facts and, second, designing the final entry presentation” (Osswald, 2015: 7).

A lexicographic entry, thus, needs to mark the non-denotative meaning of the word; that is, according to Adamska-Sałaciak (2012), it needs to define the word “without describing the thing behind the word”. She claims such metalinguistic definitions that describe usage and function have been in use for a long time:

Thus, instead of defining an expression by describing its referent (i.e. the thing or situation named), a metalinguistic definition focuses on how the expression is used. It starts with a phrase such as: “(is) used to/for...”, “when you/people say...”, “you call sb a...”, and proceeds to specify the function(s) which the expression serves in communication.

An analysis of synsemantic words in Croatian monolingual dictionaries (HJP, RHJ, ŠRHJ, VRH) reveals that metalinguistic definitions are, in most cases, absent. Observed definitions do not contain detailed information on the words’ syntactic features, language use, and pragmatic functions. Grammatical descriptions are, in large part, reduced to part of speech classification, and this classification is inconsistent among observed dictionaries.

Synsemantic lexemes with multiple syntactic and pragmatic functions introduce additional problems. Descriptions of such words in Croatian dictionaries generally only partly describe their polyfunctionality. Thus, we will demonstrate this tendency in the following sections using the lexeme *dakle* as a case study.

3. An example of the polyfunctional synsemantic lexeme *dakle*

Entry ‘dakle’ ³	HJP/RHJ	ŠRHJ	VRH
Part of speech categorization	conjunction	adverb	adverb
Lexicographic definitions’ content			
Syntactic function	-	connective function in a compound sentence	connective function in a compound sentence
Semantic-pragmatic function	conclusional function	conclusional function	conclusional function
Synonym(s)	+	-	+

Table 1: The description of dictionary entries for the lexeme *dakle* within contemporary monolingual dictionaries of the Croatian (standard) language

The analysis of dictionary entries for the lexeme *dakle* (Eng. conj. *well, now; consequently* [Bujas, 1999] / *accordingly, so, then, therefore, thus* [Bujas, 2005]) within contemporary monolingual dictionaries of Croatian (standard) language (see Table 1) lead us to the following conclusions:

(1) Definitions of this lexeme in the analysed dictionaries are metalinguistic (followed by examples, and, in some cases, synonyms), but point to just one or two semantic-pragmatic functions: introducing a conclusion and/or a consequence. The function of introducing a conclusion is featured in relevant examples in all of the analysed dictionaries. An exception can be found in VRH, which lists *Što, dakle, ja tu mogu!?* (Eng. *So what can I do!?*), as an example for introducing a conclusion, an example we deem inappropriate, as it primarily represents the rhetorical and/or expressive function of the word. VRH is also the only dictionary to feature an example for introducing a consequence, although such a decision is questionable as well, as the function it serves better illustrates the function of introducing a conclusion (*Uzeo je stvari, dakle odlazi na put* / Eng. *He took his stuff; therefore, he is going on a trip*).

³ (1) **däklē** *vezn.* – označuje zaključak ili posljedicu [*dakle, to smo se dogovorili; dakle, stigao si*]; prema tome, onda, i zato, pa zato [HJP, RHJ]; (2) **däklē** *pril.* 1. uvodi zaključak [*Ti, dakle, odlaziš.*] 2. ima vezničku funkciju u nezavisnosloženoj zaključnoj rečenici [*Uzeo je stvari, dakle odlazi na put.*] [ŠRHJ]; (3) **däkle** *pril.* 1. uvodi zaključak [*Ti, dakle, odlaziš.; Alkohol šteti, dakle valja ga izbjegavati.; Što, dakle, ja tu mogu?!;*] 2. <u *vezn. službi*> u nezavisnosloženoj zaključnoj rečenici označuje posljedicu [*Uzeo je stvari, dakle odlazi na put.*]; *Sin.* elem, ergo, prema tome [VRH].

According to a pragmatic study by Mirjana N. Dedaić (2010)⁴, the lexeme *dakle*, when observed as a discourse particle, accomplishes multiple functions: “*Dakle* seems to have developed four principal functions in discourse: (a) conclusional; (b) reformulational; (c) argumentative/rhetorical; and (d) attitudinal” (ibid. 129). Considering the first two functions, the author states:

Dakle occurs by and large in two environments roughly defined as environment (1), in which *dakle* marks a **causative-resultative relationship** [sic] between S1 and S2, and (2) in which it marks S2 to be a reformulation of S1, with consequential inferences. (ibid.)

The author additionally states that these two functions (conclusional and reformulational) are not necessarily mutually exclusive. The other two functions of the lexeme *dakle* (argumentative/rhetorical and attitudinal) are listed as secondary and originate from its conclusional function, “[...] which allows for occasional manipulation in recipient’s reasoning. It also incites attitudes towards unfulfilled expectations, allowing for attitude-revealing explicatures” (ibid. 110).

Considering that the analysed entries of the lexeme *dakle* capture only one of its four listed functions (*cf.* Dedaić, 2010), namely the conclusional function, the representation of other functions (reformulational, argumentative/rhetorical, and attitudinal) is a matter requiring further investigation and inclusion in the lexicographic descriptions of the word.

(2) Part of speech classification of the lexeme *dakle* is inconsistent among the analysed dictionaries. While in two dictionaries (HJP, RHJ) it is categorized as a conjunction, the other two (ŠRHJ, VRH) categorize it as an adverb, wherein the lexicographic definition contains information of its connective function. Such inconsistencies are likewise consistent in Croatian language grammar textbooks and linguistic studies, in which the lexeme is listed as a conjunction, a textual connector, a particle, a modal word, a modifier, a discourse marker/particle, an adverb, or a filler word. This can be seen as a reflection of the polyfunctionality of the lexeme, but also a consequence of applying different approaches to uninflected words. The origin of the observed methodological problems include the following: (1) difficulties with differentiating the traditional part of speech categories—in this case, conjunctions, particles, and adverbs; (2) limitations of traditional grammar focused only on the sentence level; and (3) more recent application of contemporary linguistic (text/discourse oriented) approaches, an application which opens new issues (notably terminological inconsistencies and diverse interpretations of “new” terms)⁵.

While the lexeme *dakle* is classified as either an adverb or a conjunction in the four

⁴ The study is based on an analysis conducted on the examples “collected from conversation events, media talk shows and reports, various written material (Internet, newspapers, and books), and the Croatian National Corpus, which includes journalistic texts, essays, and fiction—more than three thousand occurrences in total” (Dedaić, 2010: 210-112).

⁵ More discussion on this topic is available in works of Badurina (2009) and Glušac (2012).

analysed dictionaries, in the online language corpus hrWaC it is labelled as a conjunction in over 99% of instantiations.

For these reasons, we conducted a corpus-based study to investigate the (supra)syntactic and pragmatic polyfunctionality of the lexeme *dakle* in order to identify the correlation between the existing linguistic/lexicographic descriptions and its (written) language use.

4. Polyfunctionality of lexeme *dakle*: a corpus-based experiment

4.1 Methodology

We conducted a corpus-based experiment on two different annotation tasks to investigate polyfunctionality of lexeme *dakle*. We calculated the sample size needed for the experiment taking into account the total size of the population (the size of hrWaC containing over 1.3 billion tokens), a margin of error of 5%, and a confidence level of 95%. The number of 385 was rounded up to 400 random KWIC examples from hrWaC⁶.

Step 1

Conjunction ⁷	Conjunctions are uninflected words which connect words, word groups, or clauses within complex sentences.
Textual connector ⁸	Connectors organize and signal relations between the text/discourse components.
Modifier (particle or adverb) ⁹	Syntactically independent words that modify the sentence meaning.
Filler words ¹⁰	Syntactically independent words used unconsciously/automatically, without any connection to their meaning.
Other	

Table 2: Annotation scheme for determining the (supra)syntactic function

⁶ We believe that hrWaC is an adequate Croatian corpus for pragmatic research, as it contains documents from varied sources, and not only documents written in standard language like newspaper articles and literary texts (e.g. Croatian National Corpus and Croatian Language Corpus).

⁷ An example from instructions for annotators: *Danas ne mogu doći na košarku, dakle igrat ćete bez mene.* (Eng. *Today I cannot come to a basketball practice so you'll play without me.*)

⁸ An example from instructions for annotators: *Dakle, na temelju svega što je u članku izneseno proizlaze sljedeći zaključci ...* (Eng. *Therefore, based on everything in the article, the following conclusions are ...*)

⁹ An example from instructions for annotators: *To, dakle, stvarno nije bilo lijepo od tebe.* (Eng. *Well, that was really not nice of you.*)

¹⁰ From instructions for annotators: *Although the filler words are a feature primarily of oral language production, they are listed here as a possible category. If, in the examples presented, the annotators notice an unnecessary accumulation of the lexeme dakle, it is possible to categorize it as a filler word.*

Step 2

Conclusion ¹¹	Introducing the conclusion which logically stems from the previous discourse, but is not explicitly stated.
Reformulation ¹²	Reformulating a statement which has previously been explicitly stated in the discourse. The reformulation can include: (a) expansion of the previous statement (b) summary of the previous statement
Argumentative / rhetorical function ¹³	(a) discourse organization (initiating the act of communication, changing the subject, returning to the subject etc.) (b) rhetorical questions (c) enticing the collocutor (d) persuading the collocutor
Attitudinal function ¹⁴	Expressing the locutor's emotions, attitudes, or states in reference to the collocutor or the contents of the utterance.
Other	

Table 3: Annotation scheme for determining the semantic-pragmatic function

Annotation of the examples from the corpus was undertaken in two steps: (1) determining the (supra)syntactic function; (2) determining the semantic-pragmatic function of the word in discourse. In both steps, the annotators were required to choose one of the five possible categories (see Tables 2 and 3). In determining the (supra)syntactic function, annotators had the option of labelling the lexeme *dakle* as a conjunction, a textual connector, a modifier or a filler word. The fifth category was the option “other” if the annotators could not decide on one of the offered possibilities. In the second step, to determine the semantic-pragmatic function of the word in discourse, we followed Dedaić's classification (2010). The annotators had the option of choosing if the word functioned as a conclusion, a reformulation, had an argumentative/rhetorical or an attitudinal function. As in the first step, the final category was the option “other” if the annotators could not decide on one of the offered possibilities.

The two annotators had a high level of education in linguistics. They were remotely trained and given precise instructions containing definitions and illustrative examples for each of the categories offered. They had no prior experience in corpus annotation, worked separately during the annotation tasks, and had no restriction on time.

In order to evaluate the annotated examples we used accuracy as well as Cohen's κ

¹¹ An example from instructions for annotators: *A: Spremi se, doći ćemo po tebe u sedam. B: Dakle, na večeru idemo poslije predstave.* (Eng. *A: Get ready, we'll pick you up at seven. B: So, we're going to dinner after the show.*)

¹² An example from instructions for annotators: *Takvu ružnu stvar si rekla mom najboljem prijatelju, dakle, Ivanu.* (Eng. *You said this ugly thing to my best friend, [dakle] to John.*)

¹³ An example from instructions for annotators: *Dakle, zovem se Andrej i imam 16 godina.* (Eng. *So, my name is Andrej and I am 16 years old.*)

¹⁴ An example from instructions for annotators: *Mislim, dakle, stvarno si neodgovoran.* (Eng. *I mean, [dakle], you are really irresponsible.*)

(Cohen 1960), as it is the predominant reliability measure of corpus annotation used in NLP due to the work of Carletta (1996). Cohen’s κ was developed for two annotators and nominal data, as is the case with our experiment. We considered using Krippendorff’s α , but Antoine et al. (2014) concluded that there is no benefit in using this measure on nominal data. Additionally, we would like to point out that we are aware the annotation process in the domain of pragmatics is highly affected by the annotators’ subjectivity. In the next section we present the results of our research.

4.2 Results

4.2.1 Distribution of the annotation categories

Table 4 presents the distribution of the annotation categories for determining the (supra)syntactic function of the lexeme *dakle*.

	Annotator A	Annotator B	Total
Conjunction	150 (37.5%)	69 (17.25%)	219 (27.38%)
Textual connector	246 (61.5%)	211 (52.75%)	457 (57.13%)
Modifier (particle or adverb)	3 (0.75%)	117 (29.25%)	120 (15%)
Filler words	1 (0.25%)	3 (0.75%)	4 (0.5%)
Other	0 (0%)	0 (0%)	0 (0%)
Total	400	400	800

Table 4: Distribution of annotation categories for determining the (supra)syntactic function

It is obvious that the categories are not balanced, as the textual connector accounts for more than half (57.13%) of all labels. The next two categories vary between annotators. While annotator A’s second most frequent choice was conjunction (37.5%), for annotator B it was the third most frequent choice (17.25%). Modifier is a category with the most drastic difference between annotators: while annotator A chose it in only 0.75% of the cases, annotator B chose it in 29.25% of the cases. Both annotators agreed that the lexeme *dakle* was rarely a filler word (0.5%), and none of them selected the option “other”.

Table 5 presents the distribution of the annotation categories for determining the semantic-pragmatic function of the word in discourse. From the data we can conclude that the distribution for the second step is overall more balanced between three categories (the argumentative/rhetorical function 40.5%, reformulation 32.38%, conclusion 26%). However, when examining each annotator separately, we can observe that each annotator has a different category prevailing. Annotator A chose the rhetorical and interactional function 56% of the time, while annotator B chose reformulation 43.5% of the time. As with the first step, both annotators agree that the lexeme *dakle* rarely serves as an attitudinal marker (1.13%), and none of them selected the option “other”.

	Annotator A	Annotator B	Total
Conclusion	88 (22%)	120 (30%)	208 (26%)
Reformulation	85 (21.25%)	174 (43.5%)	259 (32.38%)
Argumentative / rhetorical function	224 (56%)	100 (25%)	324 (40.5%)
Attitudinal function	3 (0.75%)	6 (1.5%)	9 (1.13%)
Other	0 (0%)	0 (0%)	0 (0%)
Total	400	400	800

Table 5: Distribution of annotation categories for determining the semantic-pragmatic function

4.2.2 Data reliability

We used accuracy as well as Cohen’s κ to measure data reliability for both steps, since it considers the possibility of the agreement occurring by chance. The results are shown in Table 6. The accuracy for determining the (supra)syntactic function is 0.655, while for the semantic-pragmatic function it is 0.5025. Before interpreting the results, we calculated Cohen’s κ for both annotation tasks. For the first task of determining the (supra)syntactic function, the result is 0.4332, while for the second task of determining the semantic-pragmatic function it is 0.2908. It is still not agreed upon as to what constitutes a good agreement, i.e. how to interpret Cohen’s κ . According to Landis and Koch (1977)¹⁵, for the (supra)syntactic function we have a moderate agreement, while for the semantic-pragmatic function we have a fair agreement. Altman (1990) proposed a slightly modified interpretation¹⁶, but the interpretation of our results stays the same (moderate and fair agreement, respectively). On the other hand, Fleiss et al. (2013) proposed another interpretation¹⁷. According to them, for the (supra)syntactic function we have fair to good agreement, but for the semantic-pragmatic function we have poor agreement.

We tend to agree with Fleiss et al.’s (2013) interpretation of Cohen’s κ . We believe that we have attained a fair agreement between annotators for the first task of determining the (supra)syntactic function of the lexeme *dakle*. However, we are aware of the disproportionate distribution of categories for this task, which might skew the results in our favour. For the second task of determining the semantic-pragmatic

¹⁵ Landis and Koch (1977) proposed the following interpretation of Cohen’s κ : < 0.0 poor agreement; 0.00 – 0.20 slight agreement; 0.21 – 0.40 fair agreement; 0.41 – 0.60 moderate agreement; 0.61 – 0.80 substantial agreement; 0.81 – 1.00 almost perfect agreement.

¹⁶ Altman (1990) proposed the following interpretation of Cohen’s κ : 0.00 – 0.20 poor agreement; 0.21 – 0.40 fair agreement; 0.41 – 0.60 moderate agreement; 0.61 – 0.80 good agreement; 0.81 – 1.00 very good agreement.

¹⁷ Fleiss et al. (2013) proposed the following interpretation of Cohen’s κ : < 0.40 poor agreement; 0.40 – 0.75 fair to good agreement; > 0.75 excellent agreement.

function of the word in discourse, we attained poor agreement between annotators. We believe the reason for this is that the categories in this task are not mutually exclusive, as Dedaić (2010) pointed out. In order to investigate this matter further, in following sections we analyse in more detail: (1) agreements and disagreements between annotators for each task, and (2) the combination of categories between annotation tasks.

	Accuracy	Cohen's κ
(Supra)syntactic function	0.655	0.4332
Semantic-pragmatic function	0.5025	0.2908

Table 6: Reliability measures

4.2.3 Analysis of agreements and disagreements between annotators

The next step was to analyse how many times the annotators agreed and on what categories, as well as how many times they disagreed and what were the categories that could be interpreted as “interchangeable”. Table 7 presents the frequency distribution of the agreements and disagreements for the first task of determining the (supra)syntactic function.

Agreements		Disagreements	
Categories	Frequency	Categories	Frequency
Textual connector	204 (51%)	Conjunction (for annotator A) and Modifier (for annotator B)	87 (21.75%)
Conjunction	56 (14%)	Textual connector and Modifier	29 (7.25%)
Modifier	3 (0.75%)	Conjunction and Textual connector	18 (4.5%)
		Conjunction and Filler words	2 (0.5%)
		Textual connector and Filler words	2 (0.5%)

Table 7: Distribution of agreements and disagreements for determining the (supra)syntactic function

We will first focus on agreements, and then on disagreements. The annotators agreed the most on when the lexeme *dakle* had the function of a textual connector (51%), which is expected since over half of the labels for this task were annotated with this

category. The annotators agreed 14% of the time the lexeme had the function of a conjunction and only 0.75% of the time that it had a modifier function. It is worth mentioning that none of the annotators chose the option “other”, which is also considered an agreement.

Analysing disagreements, we found an anomaly in that annotator A labelled an example as a conjunction, while annotator B labelled the same example as a modifier. However, there is not one instance of a vice versa case (annotator A labelling an example as a modifier and annotator B labelling it as a conjunction). We find this result very peculiar and one that needs to be investigated further, possibly by increasing the number of annotators. Other cases of disagreements had instances of a vice versa case (e.g. annotator A choosing X and annotator B choosing Y, as well as annotator A choosing Y and annotator B choosing X). In 7.25% of the instances, the annotators interchanged the labels of a textual connector with a modifier, and in 4.5% of the instances interchanged a conjunction and a textual connector.

Table 8 presents the frequency distribution of the agreements and disagreements for the second task of determining the semantic-pragmatic function of the word in discourse.

Agreements		Disagreements	
Categories	Frequency	Categories	Frequency
Argumentative/rhetorical function	87 (21.75%)	Conclusion and Argumentative/rhetorical function	89 (22.25%)
Reformulation	79 (19.75%)	Reformulation and Argumentative/rhetorical function	53 (13.25%)
Conclusion	35 (8.75%)	Conclusion and Reformulation	48 (12%)
		Argumentative/rhetorical function and Attitudinal function	8 (2%)
		Conclusion and Attitudinal function	1 (0.25%)

Table 8: Distribution of agreements and disagreements for determining the semantic-pragmatic function

As with the previous task, we will first focus on agreements, and then on disagreements. The annotators agreed the most on when the lexeme *dakle* had the argumentative/rhetorical function (21.75%). Similarly, the annotators agreed 19.75% of the time the lexeme was used for reformulation. Only 8.75% of the agreements were on the conclusional function. Analogous to the first task, none of the annotators chose

the option “other”, which we also consider an agreement. When analysing disagreements, the annotators mostly disagreed between the conclusional function and the argumentative/rhetorical function (22.25%). In 13.25% of instances the annotators interchanged the reformulational and the argumentative/rhetorical functions, while disagreement between the conclusional and the reformulational functions occurred 12% of the time.

4.2.4 Analysis of combination of categories between annotation tasks

In this section, for each annotator we analyse combinations of categories between the two annotation tasks, i.e. what category they selected for the first task and what category they selected for the second. The detailed results for annotator A and annotator B are presented in Table 9.

From the data it is evident that annotator A has more stable combinations of categories than annotator B. For example, annotator A covers 97% of all annotations with the top 5 combinations or 98.5% with the top 6. On the other side, annotator B has more combinations. With the top 5 combinations they cover 84% of all annotations, while with the top 6 they cover 88.5%. It takes the top 9 combinations for annotator B to cover 98% of all annotations. This data shows that every time annotator A selects a certain category in the first task, they are more likely to consistently select the same category in the second. On the other hand, every time annotator B selects a certain category in the first task, they are more likely to change categories for the second task.

5. Discussion and conclusion

The analysis of the functions and the use of the lexeme *dakle*, based on language material extracted from the Croatian web corpus hrWaC, has shown discrepancies between corpus data and dictionary descriptions. In Croatian monolingual dictionaries the lexeme is categorized as either a conjunction or an adverb, while in the corpus over 99% of occurrences are labelled as a conjunction. Our experiment has shown that in most cases (57.13%) the annotators have labelled the lexeme as a textual connector, while in considerably fewer cases they labelled it as a conjunction (27.38%) or a modifier (particle or adverb) (15%). However, we are aware of the great imbalance between annotators regarding the modifier category: while annotator A selected this category in only 0.75% of cases, annotator B selected it in 29.25%.

The disagreement between annotators regarding the first task is expected, due to the already mentioned issues with part of speech categorizations and grammatical descriptions of synsemantic (poly)functional words (presented in Section 3). It is also expected that the function of the filler word is confirmed in only 0.5% of the cases, due to hrWaC not containing spoken language material.

Annotator A		Annotator B	
Category of task 1 and Category of task 2	Frequency	Category of task 1 and Category of task 2	Frequency
Textual connector and Argumentative/rhetorical function	192 (48%)	Textual connector and Conclusion	92 (23%)
Conjunction and Reformulation	79 (19.74%)	Modifier and Reformulation	87 (21.75%)
Textual connector and Conclusion	46 (11.5%)	Textual connector and Argumentative/rhetorical function	71 (17.75%)
Conjunction and Conclusion	42 (10.5%)	Textual connector and Reformulation	44 (11%)
Conjunction and Argumentative/rhetorical function	29 (7.25%)	Conjunction and Reformulation	42 (10.5%)
Textual connector and Reformulation	6 (1.5%)	Modifier and Argumentative/ rhetorical function	18 (4.5%)
Modifier and Argumentative/rhetorical function	2 (0.5%)	Conjunction and Conclusion	17 (4.25%)
Textual connector and Attitudinal function	2 (0.5%)	Modifier and Conclusion	11 (2.75%)
Modifier and Attitudinal function	1 (0.25%)	Conjunction and Argumentative/rhetorical function	10 (2.5%)
Filler words and Argumentative/rhetorical function	1 (0.25%)	Textual connector and Attitudinal function	4 (1%)
		Modifier and Attitudinal function	1 (0.25%)
		Filler words and Argumentative/rhetorical function	1 (0.25%)
		Filler words and Attitudinal function	1 (0.25%)
		Filler words and Reformulation	1 (0.25%)

Table 9: Distribution of combination of categories between annotation tasks for annotator A and annotator B

We would like to point out one unexpected result regarding a disagreement between a textual connector and a modifier. The traditional grammar focused only on the sentence level includes the textual connectors within adverbs. Therefore, we expected the disagreement between these two categories to be larger than our data confirmed (only 7.25%). The analysis of semantic-pragmatic function of the lexeme *dakle* confirmed its polyfunctionality. In Croatian monolingual dictionaries, the definitions point to just one or two semantic-pragmatic functions: introducing a conclusion and/or a consequence. Our experiment has shown that in most cases (40.5%) the annotators labelled the lexeme with the argumentative/rhetorical function. Unexpectedly, even the reformulation is more frequent (30.38%) than the conclusional function (26%). Since Dedaić (2010) stated that the conclusional and the reformulational functions are not necessarily mutually exclusive, we expected these two categories to be interchangeable among annotators. However, our data demonstrates that the annotators disagree on these two categories in only 12% of the cases. A larger disagreement is confirmed between the argumentative/rhetorical function and the conclusional function (22.25%), while a similar disagreement is confirmed between the argumentative/rhetorical function and the reformulational function (13.25%). According to Dedaić (2010), the argumentative/rhetorical function originates from the conclusional function, which explains the aforementioned disagreement. We deduce that the lexeme *dakle* simultaneously performs more than one of these three functions proposed by Dedaić (2010). Our experiment hardly found the fourth attitudinal function (1.13%).

We find the combination of categories between annotation tasks very intriguing, as we are not certain if the (in)consistency of an annotator is indicative of their quality (due to the highly subjective annotation task in the field of corpus pragmatics research). As both annotation tasks are performed simultaneously, we cannot be sure of how one task influenced the other. In future work it would be beneficial to perform the annotation tasks separately.

We believe the experiment proves: (1) the polyfunctionality of the lexeme *dakle*, (2) the simultaneous multiple functionality of the lexeme, and (3) vague boundaries between (supra)syntactic and the semantic-pragmatic categories. It is our opinion that monolingual dictionaries for native speakers, like the ones analysed in our study, should contain lexicographic descriptions of all (or at least most frequent) functions of synsemantic words. Our pilot study has indicated that the functions of the lexeme *dakle* are not equally distributed. However, to identify a more precise frequency distribution of its functions, it is necessary to conduct a more extensive study that would include more annotators and, possibly, more corpus examples. With such information lexicographers can define and apply the criteria for structuring dictionary entries (e.g. the order or selection of functions defined). Dictionary entries of polyfunctional synsemantic words should contain metalinguistic definitions and usage descriptions, supported by illustrative examples based on language corpora. The analysis of language corpus data can improve linguistic (and thereby lexicographic) descriptions of such words, which will become a much-needed form of reciprocal feedback for adequate

processing of language corpora. Since Croatian monolingual dictionaries do not offer a methodical, exhaustive, and thorough lexicographic descriptions of polyfunctional synsemantic words, our pilot study offers an insight into developing an accepted procedure of their corpus-based processing and presentation.

6. Acknowledgements

We would like to thank the two annotators Sanja Šakić and Mateja Zirdum.

7. References

- Adamska-Salaciak, A. (2012). Dictionary definitions: problems and solutions. *Studia Linguistica Universitatis Iagellonicae Cracoviensis*, 2012(4), 323-339.
- Altman, D. G. (1990). Practical statistics for medical research. CRC press.
- Anić, V. (1998). *Rječnik hrvatskoga jezika*. Novi liber. [RHJ]
- Antoine, J. Y., Villaneau, J., & Lefeuvre, A. (2014, April). Weighted Krippendorff's alpha is a more reliable metrics for multi-coders ordinal annotations: experimental studies on emotion, opinion and coreference annotation. In *EACL 2014* (10 p).
- Badurina, L. (2008). *Između redaka: studije o tekstu i diskursu*. Hrvatska sveučilišna naklada, Zagreb.
- Bergenholtz, H. (1985). Vom wissenschaftlichen Wörterbuch zum Lernerwörterbuch. In *Lexikographie Und Grammatik. Akten Des Essener Kolloquiums Zur Grammatik Im Wörterbuch 28.-30.6. 1984*. Max Niemeyer Verlag.
- Birtić, M., Blagus Bartolec, G., Hudeček, L., Jojić, L., Kovačević, B., Lewis, K., & Vidović, D. (2012). *Školski rječnik hrvatskoga jezika*. Školska knjiga: Institut za hrvatski jezik i jezikoslovlje, Zagreb. [ŠRHJ]
- Bujas, Ž. (1999). *Veliki hrvatsko-engleski rječnik*. Globus, Zagreb.
- Bujas, Ž. (2005). *Veliki englesko-hrvatski rječnik*. Globus, Zagreb.
- Carletta, J. (1996). Assessing agreement on classification tasks: the kappa statistic. *Computational linguistics*, 22(2), pp. 249-254.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1), pp. 37-46.
- Coffey, S. (2006). High-frequency grammatical lexis in advanced-level English learners' dictionaries: From language description to pedagogical usefulness. *International Journal of Lexicography*, 19(2), pp. 157-173.
- Dedaić, M. (2010). Reformulating and concluding: The pragmatics of the Croatian discourse marker *dakle*. In M. Dedaić & M. Mišković-Luković (eds.) *South Slavic Discourse Particles*. Amsterdam: John Benjamins Publishing Company, pp. 107-131.
- Dobson, W. A. (1974). *A Dictionary of the Chinese Particles: With a Prolegomenon in which the Problems of the Particles are Considered and They are Classified by the Grammatical Functions*. University of Toronto Press.
- Fleiss, J. L., Levin, B., & Paik, M. C. (2013). *Statistical methods for rates and proportions*. John Wiley & Sons.
- Glušac, M. (2012). Prilozi kao vrsta riječi u hrvatskoj jezikoslovnoj literaturi. In M. Turk & I. Srdoč-Konestra (eds.): *Proceedings of the Fifth Slavistic Congress*.

- Rijeka: Filozofski fakultet, pp. 405-413.
- Helbig, G. (1988). *Lexikon deutscher Partikeln*. Verlag Enzyklopädie, Leipzig.
- Hoekstra, E. (2010). Grammatical information in dictionaries. In A. Dykstra & T. Schoonheim (eds.) *Proceedings of the XIV Euralex International Congress*. Afûk, Ljouwert: Fryske Akademy, pp. 1007-1012.
- Hrvatski jezični portal / Croatian Language Portal: <http://hjp.znanje.hr>. [HJP]
- Jojić, L. (Ed.). (2015). *Veliki rječnik hrvatskoga standardnog jezika*. Školska knjiga, Zagreb. [VRH]
- Kawashima, S. A. (1999). *A Dictionary of Japanese Particles*. Tokyo: Kodansha International.
- Kobozeva, I. M., & Zakharov, L. M. (2004). Types of information for the multimedia dictionary of Russian discourse markers. In *9th Conference Speech and Computer*.
- Kordić, S. (2002). *Riječi na granici punoznačnosti*. Hrvatska sveučilišna naklada, Zagreb.
- Kunzmann-Müller, B. (1998). Opis sinsemantičkih riječi u rječniku-izazov leksikologiji i leksikografiji. *Filologija*, (30-31), pp. 239-248.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, pp. 159-174.
- Lang, E. (1989). Probleme der Beschreibung von Konjunktionen im allgemeinen einsprachigen Wörterbuch. In F. J. Hausmann et al. (eds.). *Wörterbücher, dictionaries, dictionnaires. Ein internationales Handbuch zur Lexikographie*, 1, pp. 862-868.
- Ljubešić, N., & Klubička, F. (2014). {bs, hr, sr}WaC-web corpora of Bosnian, Croatian and Serbian. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, pp. 29-35.
- Osswald, R. (2015). Syntax and Lexicography. In Alexiadou, A. & Kiss, T. (eds.). *Syntax – Theory and Analysis. Volume 3, Handbooks of Linguistics and Communication Science, 1963–2000*. De Gruyter. (preprint)
- Pintarić, N. (2002). *Pragmemi u komunikaciji*. Zavod za lingvistiku filozofskog fakulteta Sveučilišta u Zagrebu.
- Shimchuk, E. G., & Shchur, M. G. (1999). *Slovar'russskikh chastits* [*Dictionary of Russian particles*]. Peter Lang - Europäische Verlag der Wissenschaften: Frankfurt am Main.
- Silić, J., & Pranjković, I. (2005). *Gramatika hrvatskoga jezika: za gimnazije i visoka učilišta*. Školska knjiga, Zagreb.
- Žele, A. (2015). *Dictionary of Slovenian Particles*. Slovenian language resource repository CLARIN.SI, <http://hdl.handle.net/11356/1128>.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



Reengineering an Online Historical Dictionary for Readers of Specific Texts

Tarrin Wills, Ellert Þór Jóhannsson

Dictionary of Old Norse Prose, Njalsgade 136, DK-2300 Copenhagen S, University of
Copenhagen

E-mail: tarrin@hum.ku.dk, ellert@hum.ku.dk

Abstract

This paper presents an example of how a digital historical dictionary can be reengineered for new uses and new audiences, without changing the underlying data and editing processes. We start from the premise that a large proportion of users of historical dictionaries will be using them to read specific old texts as part of their studies or research in fields that use the texts as source material (literature, history, religion, etc.). *Ordbog over det norrøne prosasprog / A Dictionary of Old Norse Prose* (ONP) has a vast archive of digitized texts, together with detailed referencing sufficient, in theory, to generate a glossary for each page and line of the texts. For the feature demonstrated here we reverse the normal dynamic dictionary-generation process. Instead of generating dictionary entries, the application searches for citations on an edition page and generates a running glossary to the edition, displaying it alongside the edition text. In this paper we present the new public interface to the dictionary (currently at onp.ku.dk) and the contextual glossaries that are generated from the dictionary's data. These have been developed using adaptive web technologies for use on a range of devices, including tablets and phones.

Keywords: Old Norse; lexicography; reading aids

1. Background

Comprehensive historical dictionaries such as *Ordbog over det norrøne prosasprog / A Dictionary of Old Norse Prose* (ONP) are major long-term research projects whose output includes tools which assist researchers in understanding the language and literature under investigation. Modern historical dictionaries use a range of digital methods to help compile and publish dictionaries, but very few lexicographic decisions are automated, with experts making all decisions about word categorization and semantics, for example. This is partly because the researchers who use such dictionaries expect extremely high levels of accuracy.

Many, if not most, users of dictionaries of written languages use them primarily to understand texts which they may be reading as objects of study or research in literature, history, history of religion and so on. A great deal of effort has been made in recent years towards making these dictionaries digital and therefore easy to search as a reference tool.

Anyone used to using such dictionaries will know that when they consult the dictionary in order to understand a specific text, they will not only find the word and the appropriate sense, but also, in a good proportion of cases, the specific passage they are reading cited in the dictionary. This is due to the fact that such dictionaries are remarkably comprehensive in their excerption of the corpora upon which they are based, with a strong tendency to cite passages that may be difficult or of interest for other reasons.

Post-1900 historical dictionaries also tend to be very detailed in their references, citing not only edition pages but also line numbers. This dense excerption and detailed referencing, when combined with digital texts, means that the lexicographic material can potentially be combined in complex ways with the original corpus. The present paper demonstrates that dictionaries can exploit the detailed and accurate referencing in digital historical dictionaries to turn the dictionary around, making a lexical glossary to the texts themselves.

1.1 History of ONP

The dictionary which later became known as ONP was established in 1939. Originally, the objective of the project was to supplement the renowned Old Norse dictionary, Johan Fritzner's *Ordbog over det gamle norske Sprog* ('Dictionary of the Old Norwegian language'; 1883-96), as many new scholarly text editions had been published in the early 20th century, which had not been excerpted for lexicographic purposes. However, it soon became clear that a new comprehensive lexical description of Old Norse was warranted, and so work began on an entirely new and extensive scholarly dictionary. The primary focus of this new lexicographic work was to be the vocabulary of prose texts, as a thorough overview of the vocabulary of the poetic language had then recently appeared with the publication of the revised *Lexicon Poeticum* (Jónsson, 1931). The project has from its inception been funded by the Arnamagnæan Commission and hosted by the University of Copenhagen.

The scope of the new dictionary was further defined by the time period for the textual source material. The dictionary was to account for the vocabulary of Icelandic and Norwegian medieval texts, from about 1150 to 1370 (for Norway) and from 1150 to 1540 (for Iceland). All the source texts are found in manuscripts of various qualities, many of which have been edited and published in scholarly editions. The dictionary was not to be limited to material from text editions, but could also cite medieval manuscripts.

In the early days of ONP, the staff were mostly concerned with collecting and organizing citations through extensive excerpting of all known Old Norse prose genres. Text citations were copied onto slips, which then were filed in alphabetical order by lemma. The citation archive was intended to contain examples illustrating the range in meaning of every word. A few key works were comprehensively excerpted, i.e. every single word

was written down in context on a slip and filed in the citation archive. With the increasing availability of text editions and ongoing excerption work, the citation archive continued to grow.

The ultimate aim of the excerption work was to build a foundation for a print publication. The initial plan was to publish a twelve-volume dictionary over a period of 25 years, with the first volume to appear in the mid-sixties (Widding, 1964: 21). This plan was not realized for various reasons, and the publication of the dictionary was delayed until 1989 when a volume of indices (ONP Registre) finally appeared. The print publication continued over the next 15 years with three additional volumes of dictionary entries (ONP 1-3, covering the alphabet from *a-* to *em-*). The rate of publication indicated that it would take around 45-50 years to publish the remaining nine volumes, so it was decided in 2005 to put the print publication on hold in order to explore alternative means of publishing the dictionary material. As technological advances were starting to fundamentally change the lexicographical world, a new publication plan was conceived, according to which ONP was to be published on a digital platform and made available online.

In 2010, the first version of ONP Online was published on the web, containing both entries from the printed volumes as well as all the citation slips from *h-* through the rest of the alphabet (for an overview see Johannsson, 2019). The shift from print publication to a digital publication entailed some changes in the editorial process. The traditional alphabetical approach was abandoned in favour of focusing on specific word types. The remaining headwords were divided into twelve different groups based on part of speech and morphological complexity. These groups were: simplex (uncompounded) nouns (with fewer than ten citations), simplex nouns (with ten or more citations), compound nouns, verbs, simplex adjectives, compound adjectives, simplex adverbs, compound adverbs, pronouns, numerals, conjunctions and prepositions. The editing work continued according to new editing procedures with edited entries published directly online.

In the digital ONP there is a distinction between semantic and structural editing. Nouns, adverbs and adjectives are being edited both structurally and semantically, whereas verbs and prepositions have been organized according to structure. The different types of entries are compared and discussed in some detail in Johannsson and Battista (2014: 173-174). Today the dictionary consists of approximately 65,000 headwords and over 800,000 citations. Around 30,000 headwords have been edited in some form: semantically, structurally or both, with 500,000 citations within the entries' semantic / grammatical trees. There are around 60,000 senses identified, of which 16,000 are defined in both English and Danish (mainly words starting A-E). A further 15,000 are only in Danish and 1,000 in English.

In addition to the dictionary itself, the digital resources include an index of approximately 5,000 manuscripts and other documents, and a bibliography with around

5,000 items. Every citation in the dictionary is linked to the manuscript which it originally comes from, providing a link between every word, its semantics, and the material record from which it derives. Every citation also has a reference to the page and line of the edition or manuscript from which it is excerpted, with over 17 citations recorded on average for each page of an edition.

The total corpus of Old Norse prose is difficult to quantify, but based on samples of excerpted texts the authors estimate it to be around 10 million words, including lexical variants but excluding texts which are otherwise substantially the same as other included texts. This means that for most texts the dictionary will have excerpted, and thus eventually will have defined, 5-10% of words in the entire text. ONP will thus by its completion have semantically categorized and defined a significant proportion of the entire corpus of Old Norse.

1.2 The data model and software used to generate the dictionary

ONP's data is managed through an Oracle database and edited via a desktop application which will eventually be replaced by the web applications described here. The data structures were largely developed in the 1980s (before TEI/XML was available as a possible digital standard for a historical dictionary), and can be represented by the schema in Figure 1. Each dictionary entry (article) consists of a headword with the semantic tree built through two linked tables, effectively allowing the semantic tree to be up to six levels deep when internal references are used in the tables.

Each citation is linked to the semantic tree with a corresponding scanned citation slip held on the filesystem in about two thirds of cases. The citation is linked to an edition or manuscript page (with page and line numbers) and through the indexes of the bibliography, texts and manuscripts it is identified as belonging to a particular edition and manuscript. The dating of the manuscript is used to sort the citations within each part of the semantic tree.

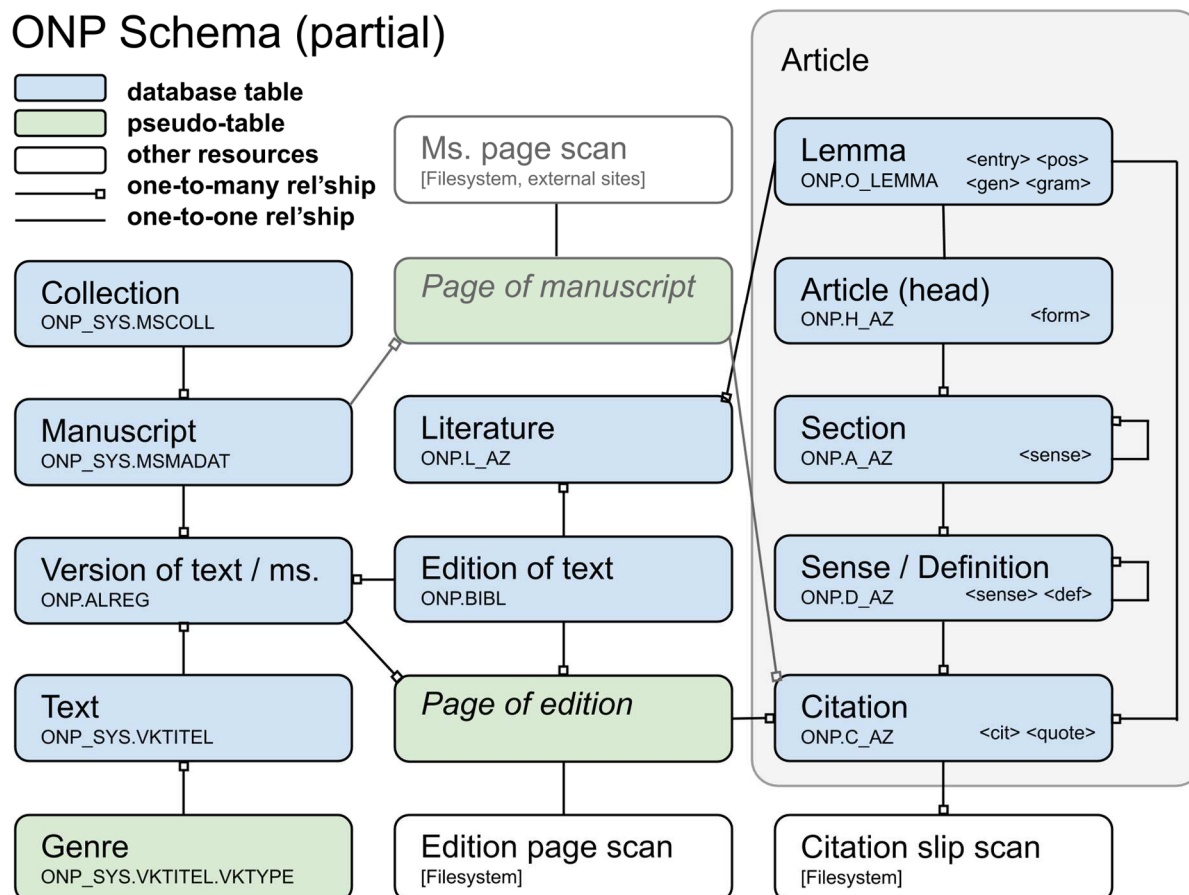


Figure 1: Simplified schema of ONP's database.

1.3 Comparisons with other digital historical dictionaries

Although ONP's methodology and data structure developed independently, it can be compared with other historical dictionary projects of its era and later. For example, the Middle English Dictionary (MED) covers a similar period to ONP (1100-1500) with comparable challenges, including a very a large number of potential manuscripts for each text and highly variable orthography. It also started in the interwar period (1925) and is larger but comparable in scope to ONP (3,000,000 citations¹ compared with ONP's 800,000).

MED was digitized from 1997 and has subsequently been updated. The framework of the online dictionary is similar to ONP and reflects a comparable underlying data model, with entries structured semantically and each citation including linked

¹ <https://quod.lib.umich.edu/m/middle-english-dictionary/about>. Accessed 22 May 2019.

information about its source — both the printed edition and the original main manuscript from which it derives.

The University of Michigan Library, which publishes MED, also supplies a searchable digital corpus in parallel, but users can neither access the corpus from the dictionary, nor access the dictionary from the corpus, although the references between the two are detailed enough to potentially allow this.

The Dictionary of Old English (DOE) is a more modern dictionary which started in the 1970s. It is now based on a fully-digitized corpus, with both the dictionary and corpus available online by subscription. At around 3,000,000 words² the corpus is smaller than both MED and ONP. Separate subscriptions are provided for the dictionary and corpus, and it is perhaps for this reason that the user cannot navigate digitally between the dictionary and corpus, despite their close connections in both referencing and digital methodology.

Similar historical dictionaries in Scandinavia are not as developed digitally. Some are incomplete, such as the Gammeldansk Ordbog (Dictionary of Old Danish), which includes a searchable headword list and access to scanned citation slips. Others belong to the print era and have not been digitized beyond OCR of the content and indexing of the headwords.

The concept demonstrated in this paper is perhaps most closely implemented in the Anglo-Norman Dictionary (AND). The digital AND was undertaken in a similar era to ONP's digital development and has a similar detail of referencing, although manuscript identification is not comprehensive. AND provides, in addition to citation source information for each citation in each entry, links to the citation's textual context in the corresponding corpus. It also generates a full alphabetical glossary for each text, with each citation linking to the corresponding headword. AND's web release, now over a decade ago, was well-received.³

Unlike ONP, which mostly relies on printed editions and provides scanned pages of them, the Anglo-Norman Dictionary has a full digital corpus. It does not, however, provide a parallel glossary to the corpus text itself, nor does it link the citations in the text's citation listing to the semantic tree of the dictionary entry, only to the complete entry.

² <https://tapor.library.utoronto.ca/doecorpus/wordcount.html> - Accessed 18 May 2019.

³ "One reviewer, after remarking that "the online AND permits an ease, speed and depth of consultation that a printed dictionary could never rival", concluded that it "represents the future of lexicography, in a freely available form that surpasses in every respect the commercial electronic versions of other dictionaries in the field" [D. Burrows in *Medium Aevum* Vol 26 (2007)]." <http://www.anglo-norman.net/dissemin/data/page2.htm> - Accessed 18 May 2019.

1.4 End-users and the digital historical dictionary

The major, comprehensive historical dictionaries that have been produced over the last century have researchers as their primary end-users. The dictionaries give a detailed semantic analysis of all words and, perhaps most usefully, a fairly complete concordance of the word's occurrence across all texts and genres in their corpora. These dictionaries do not normally aim to be comprehensive in their coverage of high-frequency words, but tend to be fairly comprehensive in citations of lower-frequency words, and include citations as evidence for every identifiable sense.

Many editions of the texts in the corpora which these dictionaries cover provide their own glossaries in alphabetical order, and some editions (for example, editions of Middle English poetry for students) have marginal word glosses. The end-users of editions of older texts are frequently students of language and literature, or those working in related fields — such as history, comparative literature or history of religion — which use the texts as their sources. For many individual editions and in some editing traditions, accompanying glossaries are not provided. There are often students or junior researchers who wish to understand these texts but who have a less advanced understanding of the language than the researchers who are the primary audience for historical dictionaries. In these cases the readers can understand much of the text but must make recourse to a dictionary.

Probably the most common type of dictionary used in these cases are abridged versions based on nineteenth-century historical dictionaries (e.g. for Old Norse: Zoëga, 1926; Heggstad, 2008; for Old English: Hall, 1960). These dictionaries normally remove almost all citations and are often not comprehensive, usually focusing on the (then) higher status texts. They are additionally limited by the original historical dictionary upon which they are based (Zoëga is based on Cleasby & Vigfusson (1957), Heggstad on Fritzner (1886-96) and Hall on Bosworth & Toller (1898)), which tend to be less comprehensive than their modern equivalents. Readers of such texts can also use online dictionaries, including digitized versions of the shorter or longer dictionaries, and others like the Oxford English Dictionary, which include information about earlier forms of the language.

For the first two examples above (MED and DOE), the digital resources appear to have sufficient linked information to be able to generate, for example, a glossary of a particular text, either in alphabetic order, or in the order which the words appear in the text. AND, as mentioned above, implements this capability, in alphabetical order. ONP also implements this capability, providing both alphabetical and text-order glossaries for each text in the corpus. Figure 2 shows the glossary for the saga of St. Agatha, sorted according to the order in which the citations appear in the text. Clicking on a citation will show the full citation detail, including definition and citation slip, in the same format as shown in Figure 4.

UNIVERSITY OF COPENHAGEN

ONP: Dictionary of Old Norse Prose

Home Words Manuscripts Works Bibliography Citations look up word

Agðu saga: Agat¹

Ms.: Holm perg 2 fol (c1425-1445) 82vb-84ra

Bibl.: Unger 1877¹ 1-6

Pages: 1 2 3 4 5 6

Alphabetical Text order

Filter 81 citations..

sikileyjarjarl <i>sb. m.</i> — 1¹	
jarl <i>sb. m.</i> — 1¹	Qvincianus Sikileyjar jarl
þyrja <i>vb.</i> — 1¹	Her þyrjar upp sögu heilagrar Agathe meyjar
tíginn <i>adj.</i> — 1⁴	hann leitadi mægra rada, at hann mætti fa hennar ok hugdiz mundu tignaz af því, at hann teingdiz vit tigna menn, þar er hann var sialfr lagr at burdum
lágr <i>adj.</i> — 1⁴	hann var sialfr lagr at burdum
saurlifr <i>adj.</i> — 1⁵	
taumr <i>sb. m.</i> — 1⁶	þa leysti hann tauma fegirni sinnar til aura hennar
girndarauga <i>sb. n.</i> — 1⁶	iarl leit girndar-augum a þessa ena göfðu mey
djöflablót <i>sb. n.</i> — 1⁷	hann ... eggjadz svo a djöfla blót af illsku sinni, at hann mætti eigi Kristz nafn heyra
eindómi <i>sb. n.</i> — 1⁸	Slikum endimun eggjadz iarll þa, er hann let gods mey taka ok seldi hana
óráðvandr <i>adj.</i> — 1¹⁰	
telja <i>vb.</i> — 1¹¹	
ógja <i>vb.</i> — 1¹²	

Figure 2: Glossary to a text in ONP (<https://onp.ku.dk/r24>).

For incomplete dictionaries such as DOE, AND and ONP, such a glossary will only include the words that have been edited and / or excerpted for the dictionary. Many users of the dictionary would be simply using it to trying to read an old text. It would therefore be useful to have a glossary to the texts they are using, even where the dictionary is incomplete. In many cases for these older and highly inflected languages, having the linked headword and word class information can potentially help a reader understand a text, even when a word is not defined.

There are potential issues in reproducing a large proportion of a copyrighted work in the form of collected citations from a particular work. A glossary could nevertheless be provided without including the full citations, for example, which would not raise any copyright issues. For editions that are out of copyright, appropriately licensed, or open access, the whole text can potentially be provided alongside the glossary.

From a technical point of view, producing such a glossary requires turning the dictionary inside out, so to speak — starting with the innermost detail of the dictionary entry and finding references to the same text throughout the entire dictionary in order to assemble a glossary of a particular text. This includes traversing the semantic tree backwards. This is potentially technologically complicated and slow for XML and NoSQL-type systems. Some data management technologies, however, are very efficient at this kind of operation, especially SQL-based RDBMS systems, and can work seamlessly with the existing and (in some cases) evolving lexicographic data.

The real utility in the technique presented here is in providing glossaries to texts that do not appear in user-friendly or student editions. These digital historical dictionaries are highly comprehensive with regard to their corpora, meaning that they can provide

a very useful service to users who wish to understand more obscure texts, or ones that have not been of particular interest in the past but nevertheless may be of increasing interest.

2. Method

2.1 The web application

Two web applications have been built in the last year for ONP, with different aims: one as an integrated web publishing and editing application, the other as a fast and archivable public interface to the dictionary. Both interfaces include a version of the feature described here, but the focus here will be on the public interface. They both retain the dictionary's Oracle RDBMS back-end and build an interface using PHP to interact with the database and generate HTML and/or JSON output. User interaction is coded in JavaScript and both applications use Bootstrap as the HTML framework.

A fundamental difference between the original print output of the dictionary (via TeX) and earlier versions of the web output (as largely static HTML) is that these earlier versions generated the output procedurally (as Windows applications written in Delphi), with the data tables queried separately. No table joins were used in the database queries in the earlier applications, possibly to reduce load on the database server. These applications treated the Oracle server as essentially a 'NoSQL' system. The new applications, however, make extensive use of the possibilities in SQL of joining multiple tables in complex queries. With modern hardware and software these operations are very quick, despite joining data from several tables containing hundreds of thousands of rows. This means that entries can be built from queries starting with the headword and linking the semantic tree and citation, or pages can be generated from locating information in the citation table itself, such as references to particular texts, and then linking the semantic trees and headwords back to the citations.

As the dictionary is constantly being updated, with individual entries now published as soon as they are reviewed and corrected, the web interfaces retrieve data directly from the evolving database. This means that as new entries are finalized they are available instantly in all parts of the application that use them, including the text reader described here. Corrections can therefore also be made instantly to the online dictionary.

The new web applications are written with Adaptive Web Design principles. The pages are designed to show all useful information laid out in one layer on larger devices, with smaller devices reflowing elements into a vertical scrolling page and making more use of tabs and pop-ups to access details about the entries and indexes. They also include print-friendly output.

2.2 Linking back from the editions

Through the web interface the user can navigate to the reader view either through the indices (by text, manuscript or bibliographic item – if the edition is available publicly), or through the entry and citation, to see other citations in the vicinity of the same text. Opening the reader view runs a query in which the database searches the citation table for citations that occur on the same page of the edition and links the corresponding headwords and definitions, if available. At this stage the semantic tree is traversed upwards one level, which is sufficient to give the full sense of the word in the vast majority of cases. The resulting information is formatted for the reader.

Figure 3 shows a sample view from the reader feature of the web application. Most of the unused space in the browser window is removed so that the page and gloss can fill the window. In order to effectively use the glossary as an aid to reading the text, it is helpful to have the relevant information available without requiring further interaction.

The server load on both the web and database servers to generate this output is negligible. For the entire operation of querying the database server and formatting the output as HTML, the web server takes around 0.3 real seconds. Subsequent views, which take advantage of the database server's query optimizer cache, take around 0.1 second. This means that, despite joining six tables, one of which contains 800,000 rows, the application server can generate 3-10 page views per second. This is much more than the anticipated real-world load on the application, even when search engine robots are taken into consideration.

The screenshot displays the ONP Reader interface. The top navigation bar includes links for Home, Words, Manuscripts, Works, Bibliography, and Citations, along with a search box labeled 'look up word'. The main header identifies the text as 'EgM(2001) (Egils saga Skalla-Grimssonar)' by 'in Bjarni Einarsson 2001 [EA A 19]', with page 52 of 30 shown. The left pane contains the text of the manuscript, with line numbers 15 through 42 visible. The right pane provides a glossary for the text, listing words and their meanings in Old Norse and modern Icelandic, along with their grammatical forms and semantic information. The glossary entries are numbered 3 through 25, corresponding to the words in the text.

Figure 3: The ONP reader (<https://onp.ku.dk/r11194-52>).

The scanned page of the edition or manuscript is shown in over 99% of cases. Where the edition is out of copyright, open access, or rights held by the Arnamagnæan Commission (as in the example in Figure 3), the scanned image of the page is shown together with buttons to browse through the work. Where the scans of the edition are covered by agreement with the Danish Copyright Agency (Copydan), no browsing buttons are supplied, as browsing access to the scanned editions is not covered by ONP's agreement with Copydan. In some cases the citations are linked to the original manuscript page, in which case the manuscript image is shown. In other cases a digital text is available.

On the right hand side of the reader view is the glossary generated from ONP's database. Each excerpted word is shown with:

- Line number (grey). Sometimes words within a particular line will appear in a different order because the database does not have information about ordering within an edition line.
- Word form in the text (underlined) where this information is in the database (around 75% of citations).
- Parallels from the source or related texts in Old Norse or other languages (italics), if available (around 5% of citations).
- Headword (bold) with word class information (italics), plus citation count (in brackets), which, when compared with other words, approximates the headword's frequency in the corpus (these are available for all citations in the corpus).
- Semantic tree node (for 69% of excerpted words), including the syntax of the word in the particular sense for the excerpted word (square brackets); the main definition in Danish (40% of words), English (19%), both (or neither); and if applicable the phrasal use of the word. If there is a higher-level definition then that is also shown, separated by an arrow.

The minimum information available for an excerpted word is the corresponding headword, its word class and citation count. Even this basic information can be useful to a reader who is less familiar in the language, as it allows the reader to look up a word with an unusual orthography in another dictionary, and helps them to understand the grammar of the sentence in which the word occurs. The majority of glossed words, however, include much more information than this.

Clicking or tapping on a gloss (citation) brings up a popup with the full citation and scanned slip if available, plus more detailed information about the manuscript (including linked images of the manuscript pages), as shown in Figure 4.

Much of the dictionary's definitions are at this stage only in Danish. This text appears in a different colour. Clicking / tapping such text will automatically translate it via the Google Translate API. In most cases this is fairly accurate, but still not ideal. Eventually all definitions will be in both Danish and English.

The fact that most editions appear as scanned images produces a small challenge in laying out the page in an adaptive way, because the edition text cannot be reflowed as the screen narrows. However, as the references to the corpus are by page and line of the editions, it is helpful to retain the edition layout in any case, where reflowing might cause confusion.

UNIVERSITY OF COPENHAGEN

Home Words Manuscripts

ONP Reader

¹karfi in EgM(2001) - 52³

AM 132 fol (c1330-1370) 62va-99rb – *Egils saga Skalla-Grimssonar*
 (Bjarni Einarsson 2001 [EA A 19], 3-28^a, 28^a-31²⁰ (Finnur Jónsson 1886-1888 [STUAGNL 17] 61^a-65^a), 31²⁰-32¹⁵, 32¹⁵-21 (Finnur Jónsson 1886-1888 [STUAGNL 17] 67¹¹⁻²⁰), 32²¹-62^a, 69^a-95^{1a}, 104¹³-186 (88-95^{1a}: nederste tekst, ^r151, 158-159, 173-186: øverste tekst, 160-172: næstøverste el. midterste tekst))

ms. photo folder: 61e-65v 66e-70v 71e-75v 76e-80v 81e-85v 86e-90v 91e-95v 96e-100v

handrit images: 62v 63e 63v 64e 64v 65e 65v 66e 66v 67e 67v 68e 68v 69e 69v 70e 70v 71e 71v 72e 72v 73e 73v 74e 74v 75e 75v 76e 76v 77e 77v 78e 78v 79e 79v 80e 80v 81e 81v 82e 82v 83e 83v 84e 84v 85e 85v 86e 86v 87e 87v 88e 88v 89e 89v 90e 90v 91e 91v 92e 92v 93e 93v 94e 94v 95e 95v 96e 96v 97e 97v 98e 98v 99e 99v

middelstort skip (mindre og slankere end knørr, sb. m.) som primært drives frem med årer // middle size ship (smaller and narrower than knørr, sb. m) which is primarily powered by oars

●●● þeir höfðu karfa þann er reru a borð .xij. menn eða .xiii. ok höfðu nærr .xxx. manna

þeir höfðu karfa þann, er reru á borð .xij. menn eða .xiii., ok höfðu nærr .xxx. manna.

Eg. 113.

Figure 4: Additional information on a glossed word.

As the scanned images cannot be altered, the gloss is instead modified to fit. Citations are spread out vertically on most devices so that they can best approximate the position on the corresponding page. The text size also scales on different device sizes so that the full glossary in most cases can be viewed as a whole alongside the edition page, with the glosses more or less in line with the word in the corresponding text. On the narrowest devices the text and gloss appear as separate tabs which the user can easily switch between.

Eventually the output will be made print-friendly, so that hard copies of the gloss can be printed, again, where copyright and licensing permits.

3. Discussion

The ONP Reader application demonstrates that using standard web application technologies a complex historical dictionary can be repurposed with a focus on the

individual texts in the corpus it covers. This broadens the utility of the dictionary to assist users who are primarily interested in the texts rather than the language itself, as well as those learning the language.

The utility of the system has been tested informally by giving beta access to members of the Arnamagnæan Collection's Old Norse reading group. Members of the group used a range of devices (various smartphones and laptop computers) to access the Reader, as well as direct hard copy print-outs from the web pages, and feedback was very positive. We anticipate further real-world feedback and have provided a user-feedback form for all pages.

The feature demonstrated here can also be integrated with other developments at ONP, including the incorporation of fully digitized corpora (see Wills, Jóhannsson & Battista, 2018). Using the TEI/XML texts published through the Menota project, ONP can potentially provide glosses to more simplified forms of the original texts, for example, if a normalized text is embedded in the digital edition.

This output of the dictionary as glosses to texts is more than just a means to make access to the dictionary easier. Research into second language acquisition suggests that glosses assist in language acquisition and text comprehension. There is long-standing evidence which demonstrates that glossed texts improve text comprehension and aid in vocabulary acquisition (Lomicka, 1998). This applies to comparable cases with digital glosses and 'authentic' texts (Abraham, 2007). Although the teaching and learning methods in acquiring written-only languages are different from those used with the acquisition of living languages, it is likely that tools such as the one presented here may also assist learners of languages such as Old Norse.

4. Acknowledgements

This research was funded by the Arnamagnæan Commission.

5. References

- Abraham, L. B. (2008). Computer-mediated glosses in second language reading comprehension and vocabulary learning: A meta-analysis. *Computer Assisted Language Learning*, 21(3), pp. 199-226.
- Anglo-Norman Dictionary*. <http://www.anglo-norman.net>. Accessed 22 May 2019.
- Bosworth, J. & Toller, T. N. (1898). *An Anglo-Saxon Dictionary*. Oxford: Oxford University Press.
- Cleasby, R., Vigfusson, G. & Craigie, W. A. (1957). *An Icelandic-English Dictionary*. 2nd edn. Oxford: Clarendon.
- Dictionary of Old English*. <https://www.doe.utoronto.ca>. Accessed 22 May 2019.
- Dictionary of Old Norse Prose*. <https://onp.ku.dk>. Accessed 22 May 2019.
- Fritzner, J. (1883-96). *Ordbog over det gamle norske sprog*. Kristiania (Oslo): Den norske forlagsforening.

- Hall, J.R.C. (1960). *A concise Anglo-Saxon dictionary*. 4th edition. Cambridge: Cambridge University Press.
- Heggstad, L., Hødnebo, F. & Simensen, E. (2008). *Norrøn Ordbok*. 5th edition. Oslo: Det norske samlaget.
- Jónsson, F. (1931). *Lexicon poeticum antiquæ linguæ septentrionalis: Ordbog over det norsk-islandske skjaldesprog oprindelig forfattet af Sveinbjörn Egilsson*. 2nd ed. Copenhagen: Møller.
- Johannsson, E. & Battista, S. (2014). A Dictionary of Old Norse Prose and its Users – Paper vs. Web-based Edition. In A. Abel, C. Vettori & N. Ralli (eds.) *Proceedings of the XVI EURALEX International Congress: The User in Focus*. Bolzano/Bozen: Institute for Specialised Communication and Multilingualism, pp. 169-179.
- Johannsson, E. (2019). Integrating analog citations into an online dictionary. In C. Navarretta, M. Agirrezabal & B. Maegaard (eds.) *Proceedings of the Digital Humanities in the Nordic Countries 4th Conference*, pp. 250-258.
- Lomicka, L. (1998). To gloss or not to gloss: An investigation of reading comprehension online. *Language learning & technology*, 1(2), pp. 41-50.
- Middle English Dictionary*. Robert E. Lewis, et al. (eds.). Ann Arbor: University of Michigan Press, 1952-2001. *Online edition in Middle English Compendium*. Frances McSparran, et al. (eds.) Ann Arbor: University of Michigan Library, 2000-2018. <http://quod.lib.umich.edu/m/middle-english-dictionary/>. Accessed 14 May 2019.
- ONP = Degnbol, H., Jacobsen B., Helgadóttir, T., Knirk, J., Rode, E., & Sanders, C. (eds.). *Ordbog over det norrøne prosasprog / A Dictionary of Old Norse Prose*. ONP Registre (1989). ONP 1: *a-bam* (1994). ONP 2: *ban-da* (2000). ONP 3: *de-em* (2004). Copenhagen: Den Arnamagnæanske Kommission.
- Oxford English Dictionary*. <https://www.oed.com>. Accessed 22 May 2019.
- Widding, O. (1964). *Den Arnamagnæanske Kommissions Ordbog, 1939-1964: Rapport og plan*, Copenhagen: G.E.C.GADS Forlag.
- Wills, T., Jóhannsson, E., & Battista, S. (2018). Linking Corpus Data to an Excerpt-based Historical Dictionary. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (eds.) *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*. Ljubljana: Ljubljana University Press, Faculty of Arts, pp. 979-987.
- Zoëga, G.T. (1926). *A Concise Dictionary of Old Icelandic*. Oxford: Clarendon.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



Assessing EcoLexiCAT: Terminology Enhancement and Post-editing

Pilar León-Araúz, Arianne Reimerink, Pamela Faber

Department of Translation and Interpreting
University of Granada
E-mail: {pleon, arianne, pfaber}@ugr.es

Abstract

EcoLexiCAT is a freely available online application, which integrates all features of the professional translation workflow in a stand-alone interface where a source text is interactively enriched with terminological information (i.e. definitions, translations, images, compound terms, corpus access, etc.) from different external resources. EcoLexiCAT is powered by MateCat and the external sources include EcoLexicon, BabelNet, the EcoLexicon English Corpus (powered by Sketch Engine) and IATE, as well as other common resources (e.g. Wordreference, Wikipedia, Linguee, etc.). Machine translation (MT) can also be optionally added. In order to evaluate the functionalities and performance of the tool, two experiments were carried out. In the first, one subject group used EcoLexiCAT and the other used MateCat, acting as the control group. In the second, both subject groups used EcoLexiCAT and only one used MT. Both experiments shed interesting light on user behaviour, performance and satisfaction while using EcoLexiCAT.

Keywords: EcoLexiCAT; CAT tools; terminology management; MT post-editing

1. Introduction: EcoLexiCAT

Today, machine translation (MT) and computer-assisted translation (CAT) are a crucial part of the professional translation workflow. Nevertheless, the post-editing of MT output has only recently started to become more widely accepted, and terminology management is often not seamlessly integrated into the translation process. As a possible solution to this problem in the field of environmental translation we developed EcoLexiCAT, a terminology-enhanced CAT tool that provides easy access to domain-specific terminological knowledge in context and MT (León-Araúz, Reimerink & Faber, 2017; León-Araúz & Reimerink, 2018; León-Araúz, Reimerink & Faber, 2019).

The integration of MT post-editing and terminology enhancement in a CAT environment constitutes the core of what has recently been termed “augmented translation” (De Palm & Lommel, 2017; Lommel, 2018, 2017). Augmented translation is a technological approach that leverages various technologies to support and augment translators’ mental processes while translating. Such technologies include translation memories, terminology management, adaptive machine translation, and automatic content enrichment (ACE). EcoLexiCAT can thus be regarded as an augmented

translation system for the environmental domain, since it combines to a certain extent all of the above, especially in terms of ACE, which is the whole idea underlying terminology enhancement. Similar approaches can be found in TaaS¹ (Terminology as a Service), SCATE (Smart-Computer-Aided Translation Environment) and the Ocelot plug-in developed in the project FREME².

EcoLexiCAT is freely available for any user interested in translating English or Spanish environmental texts³. It integrates all features of the professional translation workflow in a stand-alone interface where a source text is interactively enriched with terminological information (i.e. definitions, translations, images, compound terms, corpus access, etc.) from different external resources: (1) EcoLexicon, a multimodal and multilingual terminological knowledge base (TKB) on the environment (Faber, León-Araúz & Reimerink, 2014, 2016); (2) BabelNet, an automatically constructed multilingual encyclopaedic dictionary and semantic network (Navigli & Ponzetto, 2012); (3) the EcoLexicon English Corpus (EEC), powered by Sketch Engine, the well-known corpus query system (Kilgariff et al., 2004); (4) IATE, the multilingual terminological database of the European Union; and (5) other external resources that can be customized by users (i.e. Wikipedia, Wordreference, Linguee, etc.).

EcoLexiCAT is powered by MateCat⁴, which runs as a web server and communicates with other services through open APIs. It allows communication with pre-existing TMs, terminological databases, concordance searches within the TMs and machine translation (MT) engines, from which the MT provider MyMemory (a combination of Google Translate and Microsoft Translator) is freely available⁵.

The main interface (Figure 1) is divided into two main sections. The left-hand section is where the four external resources (i.e. EcoLexicon, BabelNet/Babelfy, Sketch Engine and IATE) provide the terminological enhancement of the translation process (text comprehension). The right-hand section, which is where the target text is produced, is an editor where the source text appears split into different segments (text production).

Figure 2 shows a segment within the editor. First of all, the source segment is enriched with information from EcoLexicon. This is done by lemmatizing all the words in the segment and matching them against the term entries in the TKB.

¹ <http://www.taas-project.eu/>

² <http://www.freme-project.eu/>

³ Temporarily hosted at <http://manila.ugr.es:9966>

⁴ <https://www.matecat.com/open-source>

⁵ <https://www.matecat.com/support/managing-language-resources/machine-translation-engines/>

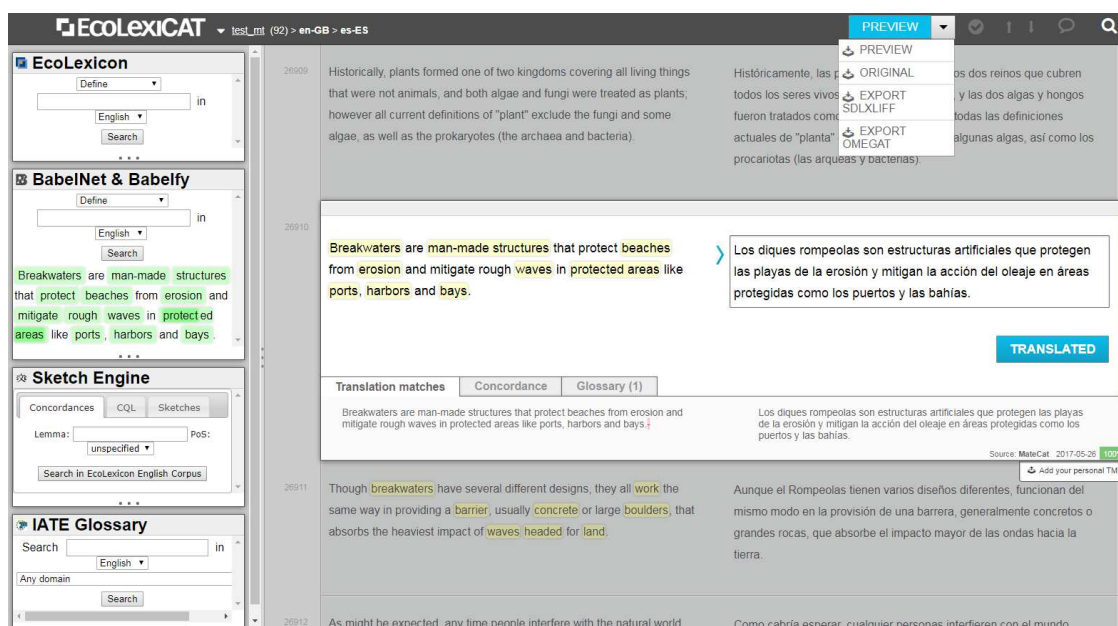


Figure 1: Main user interface of EcoLexiCAT.

All matching terms are highlighted in yellow. In the BabelNet box, the source text is matched against the contents of the KB. After applying the Babelify algorithm for disambiguation, matches are marked in green. If users right-click on any of them, a scroll-down menu gives access to all the different options provided by each of the resources of the left-hand section. In the case of EcoLexicon, these options correspond to the data categories in the TKB that are useful for text comprehension: translations, synonyms, definitions, semantic relations and images. The data categories of BabelNet included in EcoLexiCAT are definitions, translations, compound words, semantic relations, and images.

In the Sketch Engine box, the behaviour of a term selected in the source or target segments can be analysed in the EcoLexicon English Corpus (EEC; León-Araúz et al., 2018) hosted in Sketch Engine Open Corpora. Three different query modes are provided: lemma-based concordances, word sketches, and CQL (Corpus Query Language). In the IATE box, the set of English and Spanish terms downloaded from the database interacts with EcoLexiCAT as a fourth external resource.

Finally, other common language resources (e.g. Wikipedia, Wordreference, Linguee, etc.) are integrated as a pop-up box right under the active segment. Their results are shown as they appear online, since these resources are integrated as embedded websites.

In turn, the target segment is enriched with a predictive typing feature based on the matches from EcoLexicon. In addition, as in the source segment, users can right-click on any term typed in the target segment and send queries to all resources in the

opposite language directionality⁶.

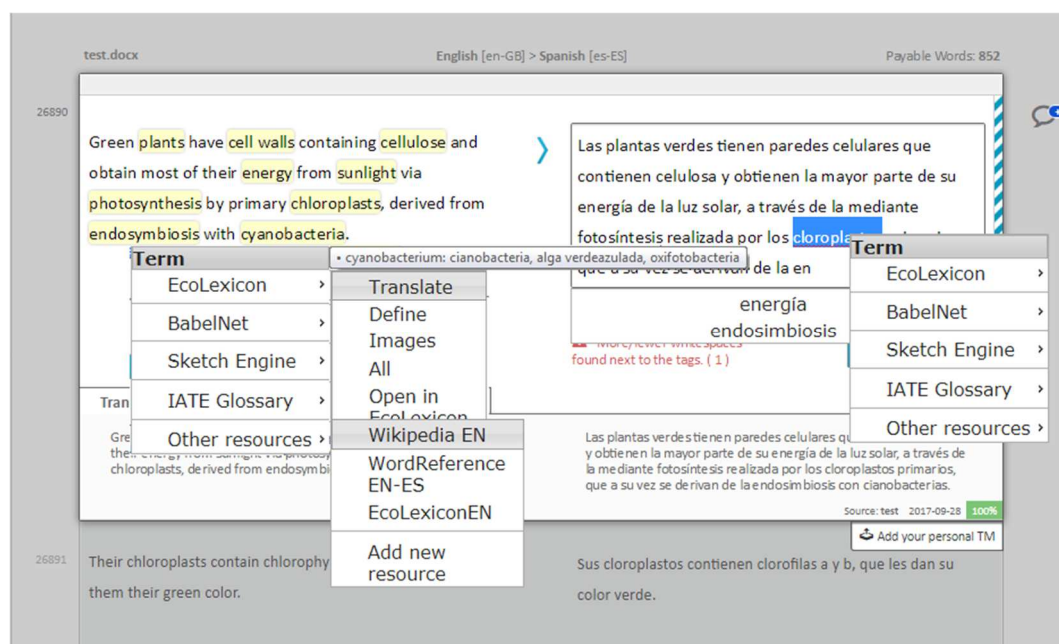


Figure 2: EcoLexiCAT editor.

After designing, creating and testing EcoLexiCAT, the next logical step was to evaluate the functionalities and performance of the tool based on the experience of prospective users in order to assess whether it meets the expectations of translators.

In the remainder of this paper, we present the experimental setup (Section 2) and the results of two experiments carried out to evaluate the tool, focusing on user expectations (Section 3), user behaviour (Section 4), user performance (Section 5) and user satisfaction (Section 6). In the first experiment (León-Araúz, Reimerink & Faber, 2019), one subject group used EcoLexiCAT and the other used MateCat, acting as the control group. In the second, both subject groups used EcoLexiCAT, but only one used MT. Accordingly, in the first experiment we studied the benefits of terminology enhancement, whereas in the second we focused on the benefits of MT post-editing. Finally, Section 7 presents the conclusions derived from this research.

2. Experimental setup

EcoLexiCAT was evaluated in two experiments conducted one year apart. This means that during the second experiment the tool had already been improved based on the results of the first.

⁶ For a more detailed account of the functioning of EcoLexiCAT, consult León-Araúz, Reimerink & Faber (2017), León-Araúz & Reimerink (2018) and León-Araúz, Reimerink & Faber (2019).

Prior to the translation task, participants of both groups were asked to fill out a brief questionnaire in order to collect data about their professional/training background, their expectations of terminological resources and CAT tools, and their habits regarding the use of dictionaries, corpora, terminological resources, etc. when confronted with a translation assignment.

The subject groups of the first experiment (EcoLexiCAT translators vs. MateCat translators) were students from the master's degree in Professional Translation of the Faculty of Translation and Interpreting of the University of Granada (Spain). In contrast, the subject groups of the second experiment (EcoLexiCAT translators vs. EcoLexiCAT post-editors) were students from both the master's degree and the final year of the Undergraduate Programme in Translation of the same faculty.

In the first experiment a total of 19 students, 22 to 37 years of age, were included in the evaluation: 10 EcoLexiCAT translators and nine MateCat translators. All subjects except for one were native speakers of Spanish; 11 subjects had English as their first foreign language, and five as their second foreign language. One subject was a native speaker of both English and Spanish, and two did not include English as one of their official working languages during their undergraduate degree, but had sufficient proficiency. The majority had a translation degree (84%); the others had degrees in modern languages or related areas. Only four subjects mentioned previous professional translation experience.

In the second experiment a total of 20 students, 20 to 54 years of age, participated in the evaluation: 10 EcoLexiCAT translators and 10 EcoLexiCAT post-editors. All subjects were native speakers of Spanish, 16 subjects had English as their first foreign language, and four as their second foreign language. Among the master's students, 90% had a translation degree and 70% had previous professional translation experience. In both experiments these characteristics were evenly divided over both groups.

In both experiments the subjects were presented with the same translation task. It consisted of two short, specialized translation assignments, one English-Spanish (EN-ES) and the other Spanish-English (ES-EN). The texts were extracts of scientific papers on the topic of Coastal Engineering, a domain widely covered in EcoLexicon. The reason for having chosen both directionalities was first to see whether behaviour and results varied according to directionality, and second, because the only corpus available so far is the EEC and usage examples are usually requested during the text production phase.

Subjects were required to deliver publishable texts in two hours. Therefore, the length of each source text was less than 200 words (EN-ES 194 and ES-EN 168 words). Other features of the source texts were high term density, syntactically complex sentences and collocational specificities that called for a deep understanding of both domain knowledge and written expression. Subjects were thus confronted with various

challenges during the comprehension and production phases of the translation workflow.

Moreover, in the two experiments both groups were asked to list all the problems encountered and the resources that helped them solve each problem. EcoLexiCAT translators and post-editors were allowed to use resources other than those in EcoLexiCAT only if they did not find the answer within the tool.

Finally, after finishing the assignments, EcoLexiCAT users filled out another anonymous questionnaire on the tool's usability, functionality and efficiency, which are three parameters established by the ISO 9126 (2001) standard for software product evaluation. They were also asked to highlight any issues related to the functioning of the tool and to propose possible improvements.

Apart from discovering the expectations of our prospective users, the purpose of this evaluation was threefold. We were not only able to assess user satisfaction but also user behaviour and performance. The first parameter was assessed based on the answers given by EcoLexiCAT translators and post-editors in the last questionnaire. The second parameter was based on the analysis of the subjects' behaviour according to Google Analytics. The third parameter was assessed by comparing the time employed and the average quality of the target texts delivered by all groups. Quality assessment was based on a scale where both translation and linguistic errors and accurate choices were accounted for. The editing logs of EcoLexiCAT and MateCat were used to see how long subjects took to translate each text.

3. User expectations

In the first questionnaire, the participants were asked to classify the following features in CAT tools as essential, desirable or unnecessary: access to MT engines, access to corpora, interoperable file formats, access to terminological resources, access to terminological resources defined by users, and QA and revision options. The results in Figure 3 for both experiments show that the most important features were found to be format interoperability, terminological resources, and QA and revision options. Access to corpora was regarded as slightly more essential than desirable, whereas access to MT engines was only desirable. This might be due to the fact that post-editing of MT is still not widely accepted by the translation community.

When asked about other features not included in the above list, most subjects could not identify anything else that they considered to be relevant in CAT tools. Exceptions were image editors and customizable QA rules and target text preview.

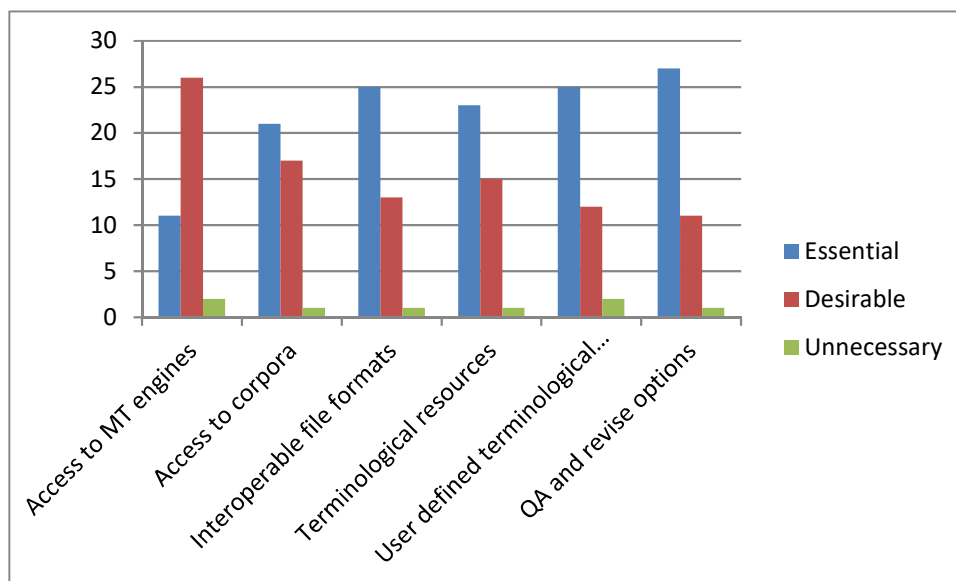


Figure 3: User expectations about CAT tools.

The participants were also asked to do the same with a set of data categories usually included in terminological resources. The data categories were: definitions, translations, synonyms and variants, context and usage examples, conceptual relations, register, images, phraseological and collocational information, etymology, pronunciation, compounds and derivatives, part-of-speech, pragmatic information on term usage, and access to corpora.

The results in Figure 4, also merged from both experiments, show that definitions, translations, synonyms and variants, context and usage examples, phraseology and collocations and access to corpora are the most relevant data categories. Desirable categories include conceptual relations, register, images, etymology and compounds and derivatives. Pronunciation is the category most often regarded as unnecessary.

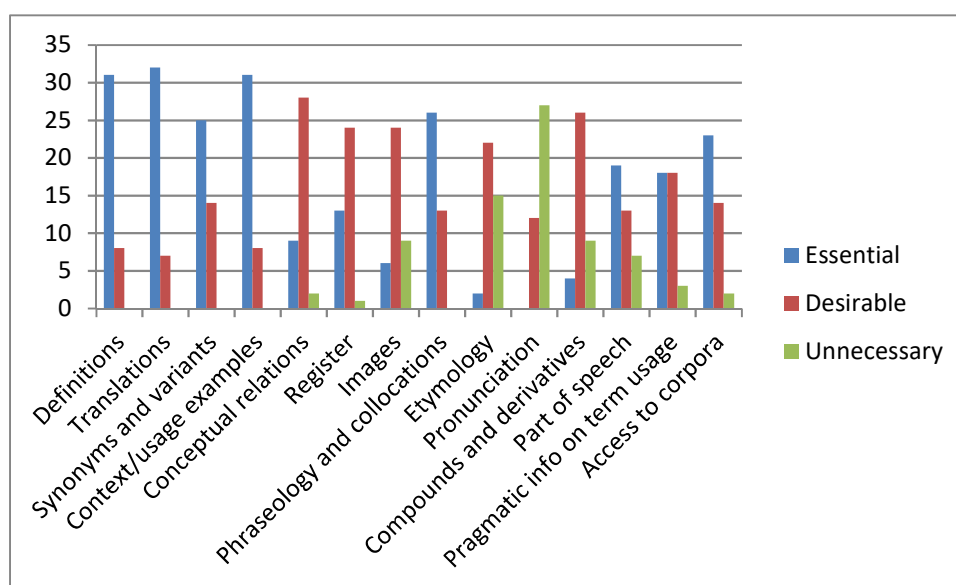


Figure 4: User expectations about terminological resources.

When asked about other features not included in the above list, most subjects could not identify any other that they regarded as relevant for terminological resources. Exceptions were specialized reference works and term use frequency, connotations, and false friends. The resources that subjects used the most for their translation assignments were as follows: Wordreference, Linguee, Reverso Context, IATE, Merriam-Webster, Oxford dictionaries, Collins, Cambridge Dictionary, RAE, esTenTen and enTenTen corpora in Sketch Engine, the BNC, CREA, the web as a corpus, CORPES XXI, Pons and Termium Plus, Glosbe, DeepL, ProZ forum, WIPO Pearl, and Medline Plus.

The subjects' answers indicated that EcoLexiCAT meets most user needs and expectations, but they also highlight how to improve the tool as well as EcoLexicon. For instance, currently there is a phraseology module (essential for most subjects) under construction in EcoLexicon that will be linked to EcoLexiCAT in the future. Part-of-speech is currently included as a data category in EcoLexicon but not in EcoLexiCAT. Therefore, based on the fact that most users considered it essential or desirable, it will be included in the next version. Furthermore, some of the resources reported by users had already been included based on the feedback received after experiment 1. However, it was impossible to include others because they do not allow embedding.

4. User behaviour

While completing their assignments, EcoLexiCAT subjects were monitored through Google Analytics. Prior to the evaluation task, we defined a series of "Events" based on the kind of actions that we wished to monitor. These "Events" in Google Analytics can be tracked according to a three-level structure consisting of Category (e.g. EcoLexicon), Action (e.g. definition by clicking on the terms) and Label (e.g. breakwater), which would mean that when users search for the definition of breakwater in EcoLexicon by clicking in the editor, the event is stored as such. This allowed us to compare the real use of each resource and the kind of queries that subjects make through a certain kind of action (e.g. definitions, translations, images, etc. from the right-click menu, by clicking in the editor, in the search form of each left-hand box, etc.). Table 1 shows a summary of the main actions tracked within each resource.

In experiment 1, a total of 5,693 events were stored during the completion of the assignments. Obviously, most of them took place within MateCat (4,874), but of the other resources, EcoLexicon stands out with 473 events (58%). EcoLexicon is followed by BabelNet, with 262 events (32%); other resources, with 47 (6%); IATE, with 27 (3%); and Sketch Engine with 10 (1%) (Figure 5).

In experiment 2, a total of 8,650 events were stored, and this higher number makes sense since both subject groups worked with EcoLexiCAT. Again, most of them took

place within MateCat (7,694). EcoLexicon, with 695 events (74%), was followed by other resources, with 88 events (9%); Sketch Engine, with 72 (8%); BabelNet, with 60 (6%); and IATE, with 41 (4%) (Figure 6). The number of events for other resources is higher than in experiment 1 (from 6% to 9%), and this is probably because new resources were added after the first experiment. The use of Sketch Engine is much higher than in experiment 1 (from 1% to 8%), which is undoubtedly an indication of the subjects' competence in corpus analysis. What is surprising is that the use of BabelNet dropped dramatically (from 32% to 6%).

Matecat	Insert-text, open-segment, delete-text, translate, search, download-original, download-translation
EcoLexicon	Definitions-click, definitions-menu, definitions-form, translations-click, translations-menu, translations-form, showAll-menu, showAll-form, images-form, images-menu, open-menu (EcoLexicon in a browser), relations-form, relations-menu
BabelNet	Definitions-click, definitions-menu, definitions-form, translations-click, translations-menu, translations-form, compound_words-menu, compound_words-form, images-menu, images-form, relations-form, showAll-menu, showAll-form, relations-menu
Sketch Engine	Concordance-menu, concordance-form, sketches-menu, sketches-forms, CQL-form
IATE	Search-menu, search-form
Other resources	Load-Linguee EN-ES, Load-Linguee ES-EN, load-WordReference EN-ES, load-WordReference ES-EN, load-Cambridge EN-ES, load-Cambridge ES-EN, load-EcoLexicon EN, load-EcoLexicon ES, load-MetaGlossary, load-TermiumPlus, load-Wikipedia EN, load-Wikipedia ES, load-Onelook, load-Majstro, load-RAE

Table 1: Main actions tracked within each resource in EcoLexiCAT.

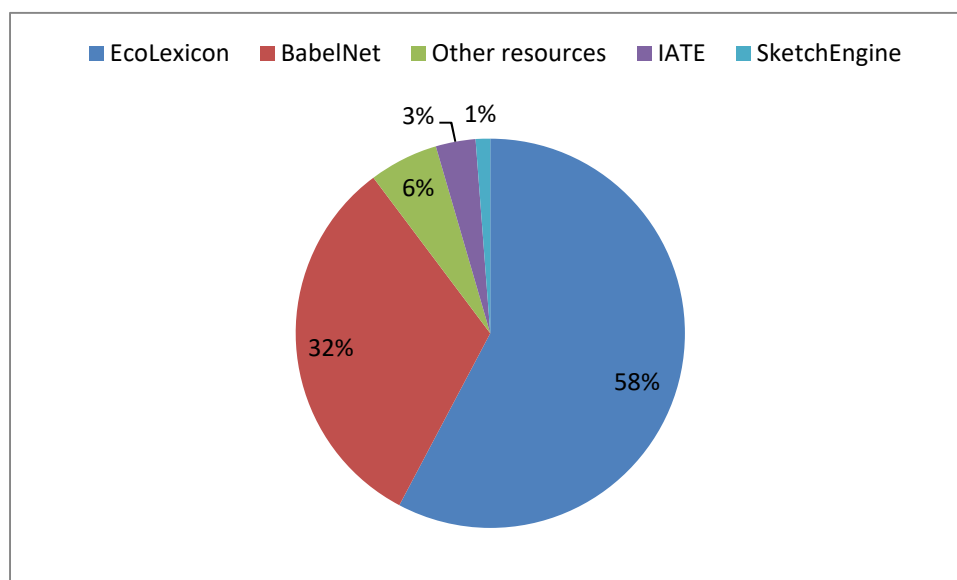


Figure 5: Events per resource in experiment 1.

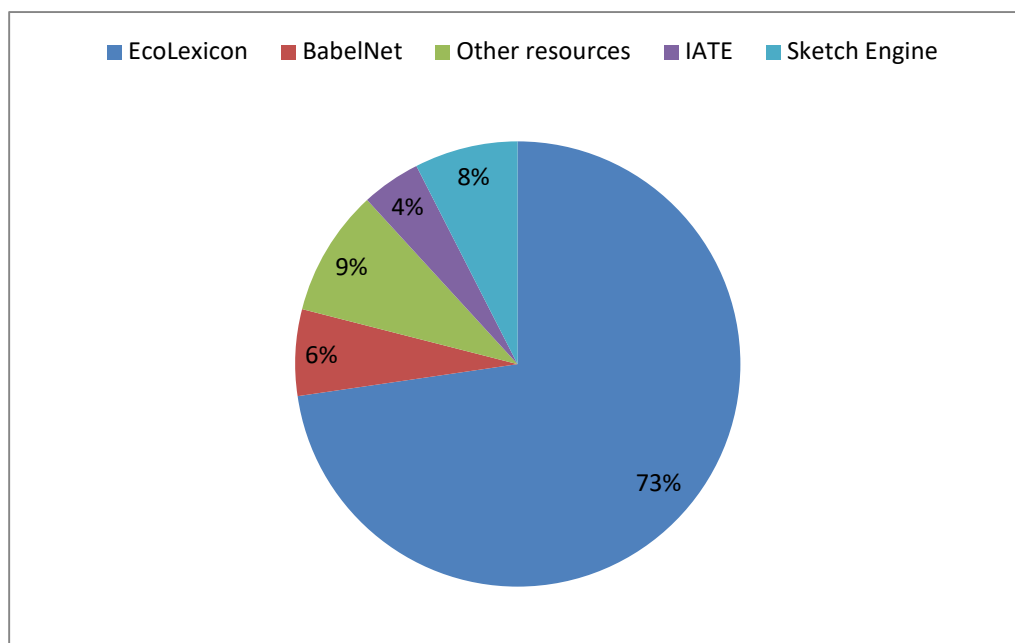


Figure 6: Events per resource in experiment 2.

From a quantitative point of view, the following figures (7-16) show the number and type of actions performed within each of the resources. This illustrates the usefulness of both the data categories of each resource (e.g. definitions, translations, images, etc.) and the way in which each category can be accessed (e.g. clicking, from the menu, writing the query in the box, etc.). For instance, in EcoLexicon (Figures 7-8) and BabelNet (Figures 9-10), definitions and translations are the preferred data categories. Clicking in the editor is clearly the preferred action in EcoLexicon in both experiments. However, in experiment 2, translations-form (writing the query in the box) and translations-menu (selecting from the right-click menu) were clearly preferred over definitions-form, definitions-click and translations-click in BabelNet. The number of actions for definitions-click and translations-click are the same, because when users clicked on one of the highlighted terms in the source segment of the editor, both kinds of information were deployed in the EcoLexicon and BabelNet boxes at the same time.

In experiment 1 in EcoLexicon, the subjects preferred to consult definitions and translations through the form in the box rather than the right-click menu access, whereas in BabelNet the opposite occurred. In experiment 2, where new events were added for new functionalities (e.g. semantic relations), in EcoLexicon subjects clearly preferred the definitions-menu option over the definitions-form option. Images were rarely consulted in either resource in both experiments. The open-menu option of EcoLexicon was used only once in experiment 1. From the EcoLexicon right-click menu, users have the possibility of opening EcoLexicon in a browser for a more detailed view of the conceptual networks. After experiment 1, the decision was made to add related concepts as a new data category in the EcoLexicon box to encourage users to explore the semantics contained in EcoLexicon. The relations-menu option

was used seven times in experiment 2, and the relations-form option was used four times.

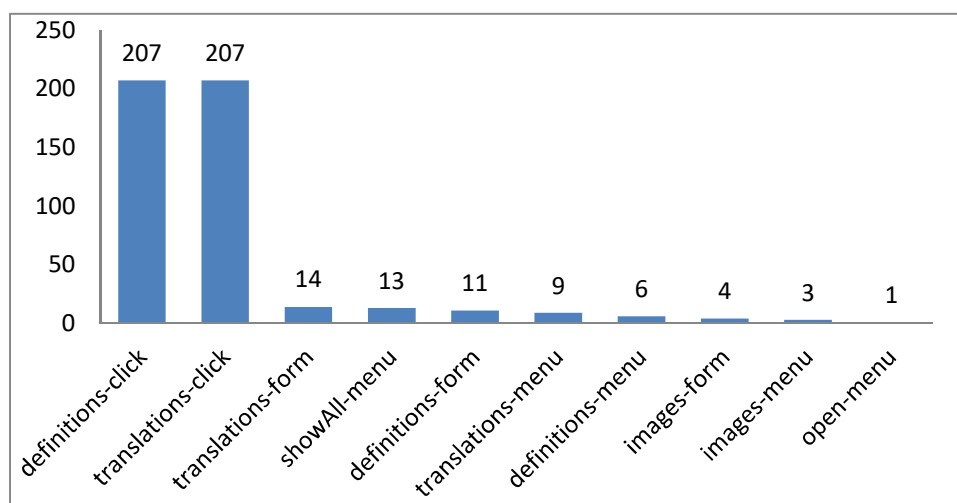


Figure 7: Actions performed within EcoLexicon – experiment 1.

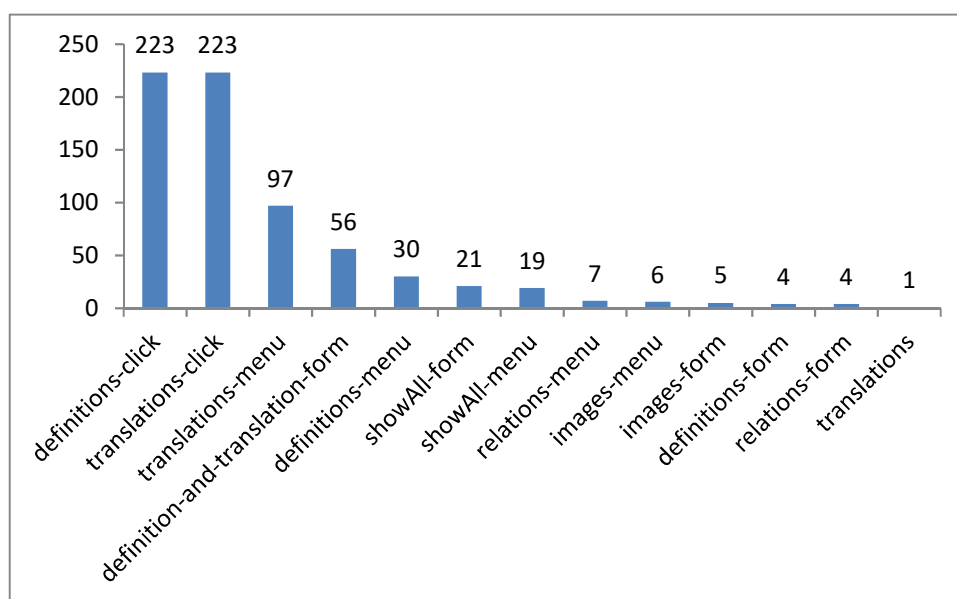


Figure 8: Actions performed within EcoLexicon – experiment 2.

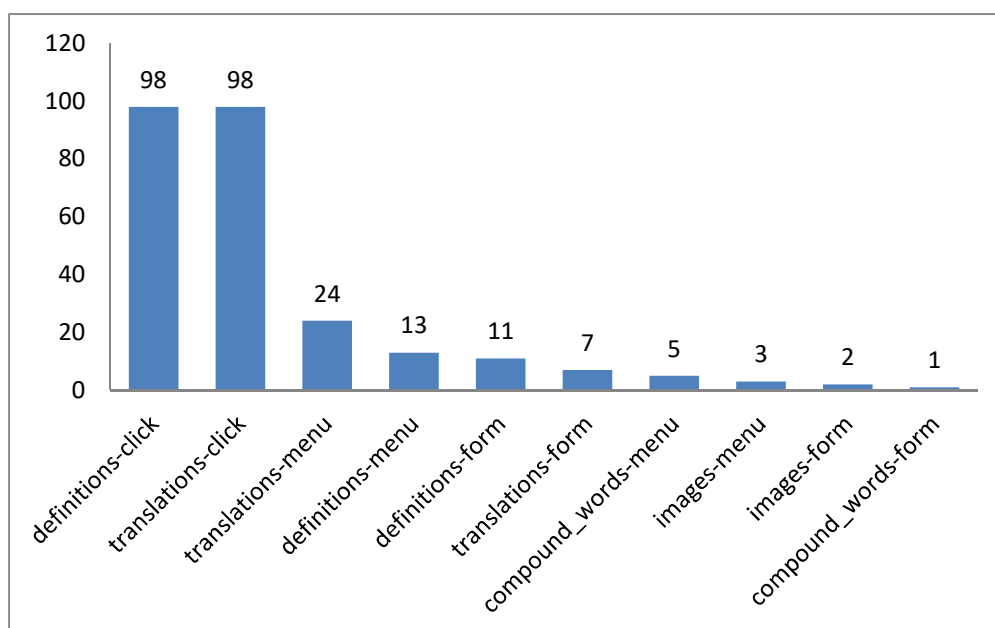


Figure 9: Actions performed within BabelNet – experiment 1.

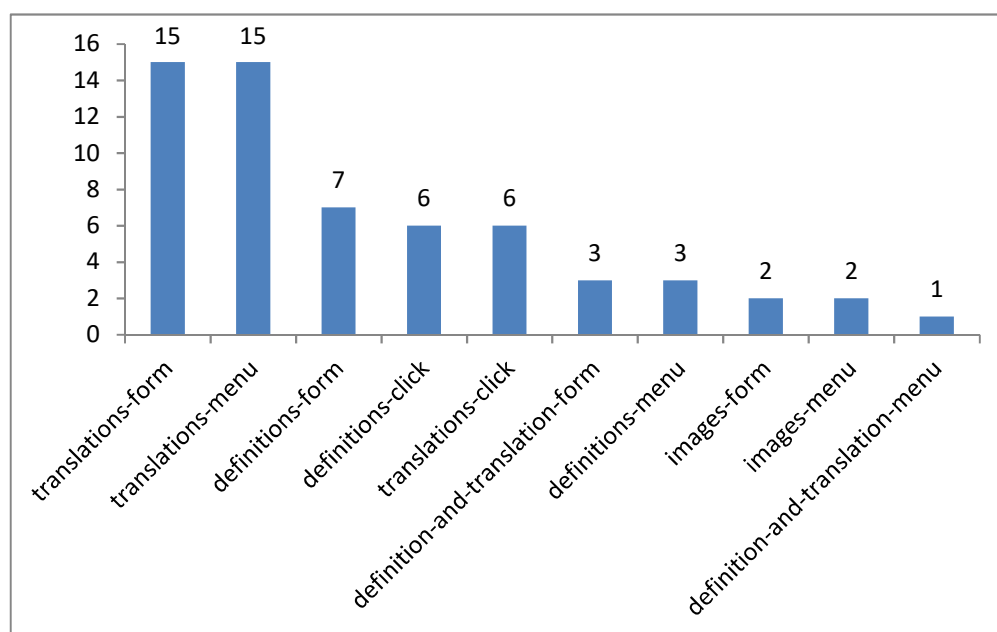


Figure 10: Actions performed within BabelNet – experiment 2.

The low number of actions carried out in Sketch Engine (Figure 11) in experiment 1 shows that the subjects were either not aware of the kind of information that can be extracted from a corpus, or did not know how to build meaningful queries. The latter is shown by the fact that seven of the 10 actions were simple concordance searches from the menu, where only the term needs to be selected in the editor. The subjects did not seem to be familiar with the basic syntax for more complex searches that would have provided more useful information, and they did not use the more advanced functionalities of corpus analysis, such as word sketches.

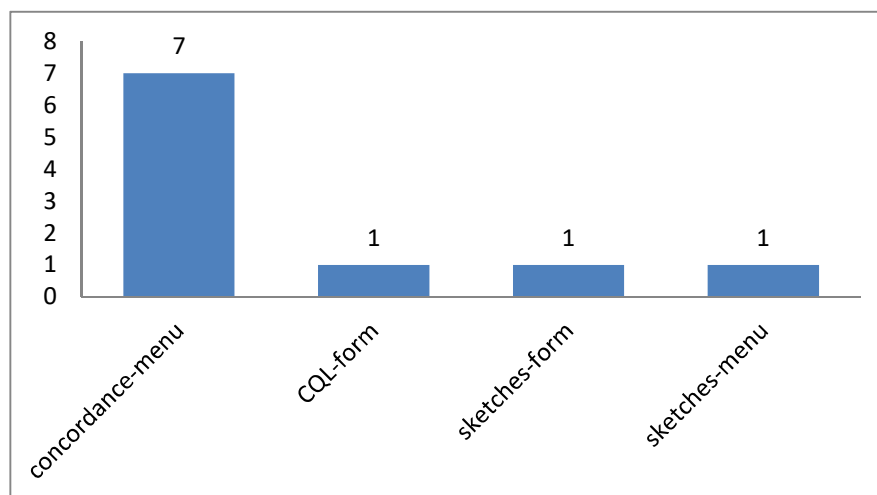


Figure 11: Actions performed within Sketch Engine – experiment 1.

In experiment 2, the subjects used Sketch Engine a great deal more (Figure 12), even if we take into account that the number of subjects working with Sketch Engine in EcoLexiCAT doubled in comparison with experiment 1. The concordance-menu was still clearly the preferred search option, but the other options were used as well, especially the word sketches, as opposed to the behaviour in experiment 1. The subjects in experiment 2 seem to be better versed in corpus analysis than those of experiment 1, although the more advanced option of CQL was only used twice.

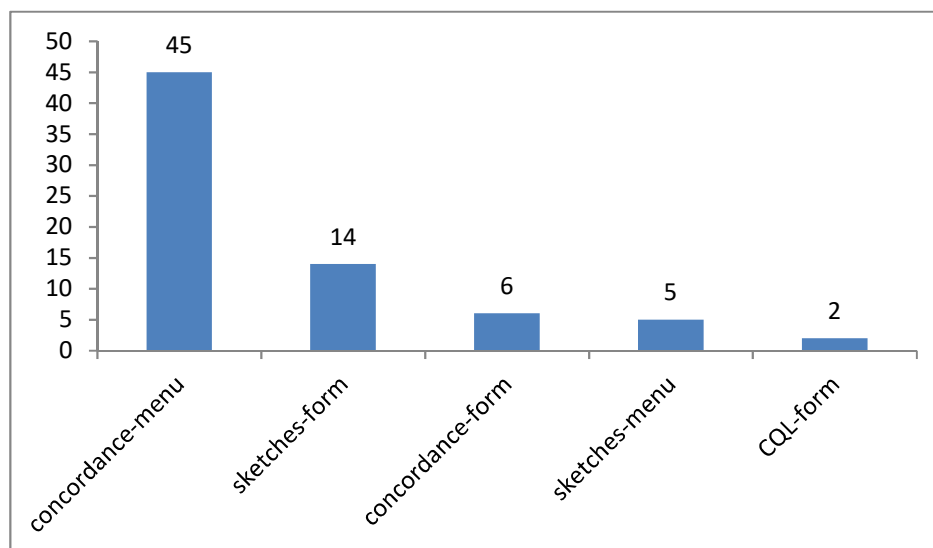


Figure 12: Actions performed within Sketch Engine – experiment 2.

The subjects in both experiments indicated the importance of having access to corpora in CAT tools, as most of them chose the essential or desirable options in the initial questionnaire (Section 3). In all likelihood, students are taught in their classes that corpus analysis is essential in the translation process, but not enough time is devoted to showing them how to actually obtain such information. In a study by Durán Muñoz

(2012), professional translators did not include access to corpora in their preferences when asked about terminological resources, probably because of lack of skills in corpus analysis and user-unfriendly search engines. Therefore, a user manual for EcoLexiCAT would have to provide easy-to-follow instructions on how to use the corpus options.

In IATE (Figures 13-14), 27 and 41 actions were carried out in experiments 1 and 2, respectively, with a slight preference for the right-click menu over the use of the form in the box in both experiments.

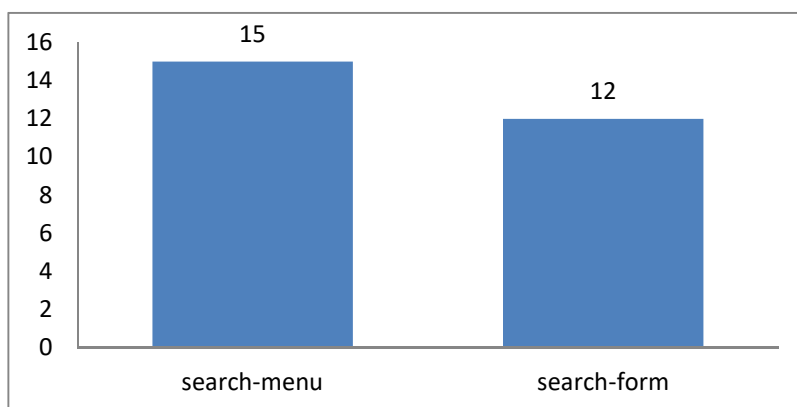


Figure 13: Actions performed within IATE – experiment 1.

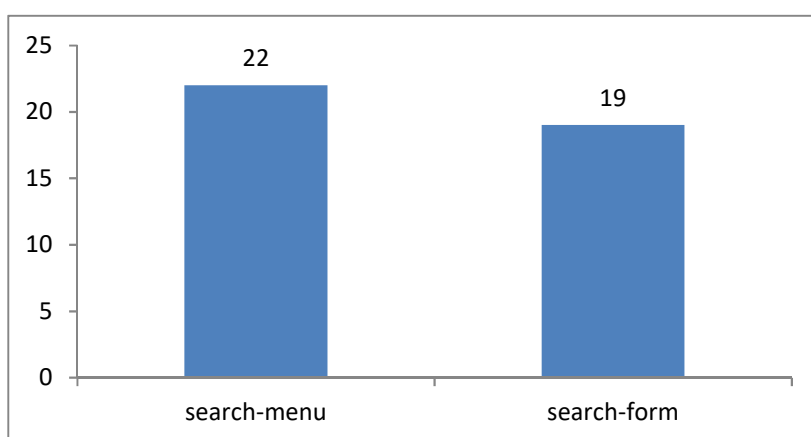


Figure 14: Actions performed in IATE – experiment 2.

With regard to other resources (Figures 15-16), the subjects in both experiments mostly used Linguee to find translation equivalents and terms in context, primarily during the first EN-ES translation task.

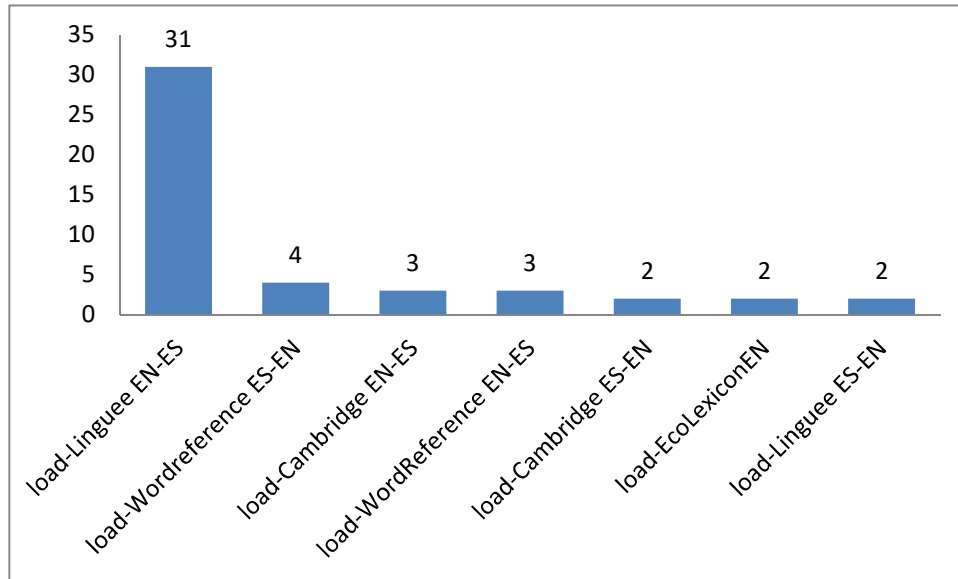


Figure 15: Actions performed within other resources – experiment 1.

However, in experiment 2 the subjects used more resources such as Wikipedia, TermiumPlus and Metaglossary, some of which were new resources added after experiment 1.

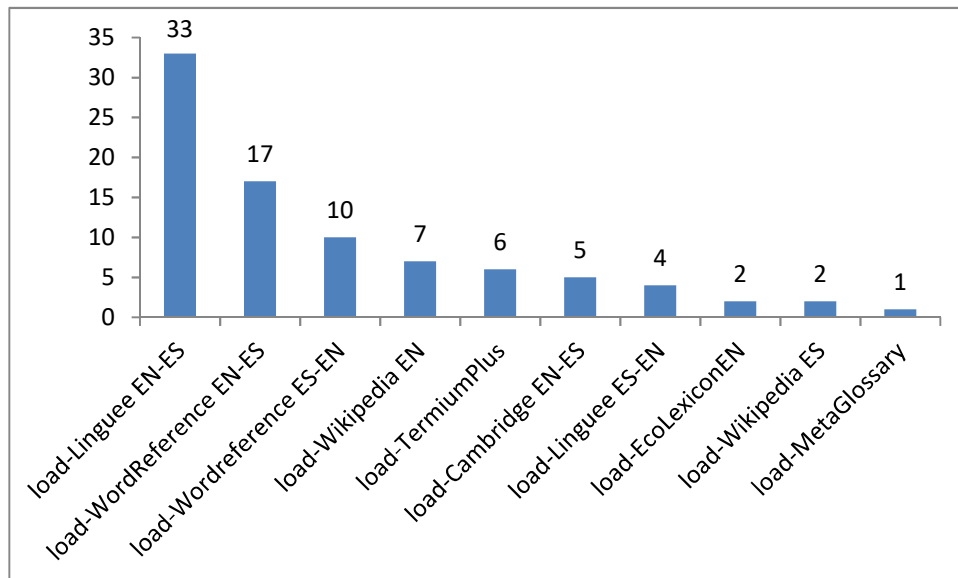


Figure 16: Actions performed within other resources – experiment 2.

From a qualitative point of view, the terms or text chains searched (labels) through each action within each resource were analysed and compared. In both experiments, there were many more searches with an English term or text chain as a starting point than a Spanish one. This is probably because the EN-ES task was performed first. Both tasks were on the same subject matter, and the subjects may have carried out most of the necessary research during the EN-ES task and already be familiar with many of the terms in both languages and the underlying domain knowledge.

Most users seem to have initially looked at the options provided in EcoLexicon for the terms marked in yellow in the source text (e.g. *detached breakwaters* and *hard coastal structures*). Then, when no option was given in EcoLexicon, subjects viewed the options marked in green in BabelNet, since the terms searched are clearly different at least in the most frequent searches. The order in this process was clearly influenced by the subject matter of the tasks as well as by the order and hierarchical structure of the terminological enhancement provided by EcoLexiCAT.

Regarding the kind of terms and chains searched for, multiword terms such as *hard coastal structure*, *detached breakwater*, and *artificial submerged reefs* were most extensively researched in nearly all resources. The search terms also matched the translation difficulties reported in the questionnaire that all subject groups filled in while translating in both experiments. Almost all difficulties reported were related to the lack of previous domain knowledge, which would impair the understanding of certain concepts, and to the lack of equivalences in the resources checked. Most of the resources that helped them solve their difficulties were the ones included in EcoLexiCAT, with the exception of some general language dictionaries and parallel texts found online. A few students also reported phraseological issues, which explains the queries of chains like *storm-induced*, *system* or *subject to*.

Curiously, EcoLexicon was searched for certain terms that initially seemed easy to translate, such as *erosion* (19 in experiment 1, 11 in experiment 2) and *cliffs* (10 and five, respectively). However, when working in a subject domain for the first time, researching more general terms and finding out how these concepts are related to others often helps to construct an initial mental representation of the domain.

What seems strange is that in experiment 1 some students looked for general language expressions, such as *continuamente* (continuously) and *significantly* in specialized resources such as EcoLexicon or BabelNet, instead of using the other resources menu. This indicates that maybe these resources should also be included on the left-hand side of the screen as a fifth box instead of as a pop-up window. On the other hand, this did not happen in experiment 2, which may again indicate that these subjects were better translators. However, in experiment 2 some subjects used the definition-and-translation-form to search for *define* and *remedy*. Furthermore, in both experiments the subjects looked for specialized terms in general resources such as Cambridge dictionary (*estuary* and *storm-induced* in experiment 1, and *detached breakwaters* and *soft cliffs* in experiment 2). Apart from that, some subjects in experiment 1 and fewer in experiment 2 used the definitions box in EcoLexicon (action: definitions-form) to find terms already marked in yellow in the text editor, such as *coastal structure*. This apparently strange behaviour can be explained by the fact that the subjects in our study were students with hardly any professional experience, although most students had a previous or almost finished translation degree, were students of a master's degree in translation, or both.

5. User performance

All target texts were evaluated by one reviser to ensure that the same criteria were applied in all cases. To assess the quality of the target texts of all groups, ten translation problems were identified for both the EN-ES and ES-EN assignments. The problems identified were based on those that the subjects mentioned repeatedly and on the reviser’s expertise in the text type and domain. Depending on how well the subjects solved these problems, they could obtain up to 10 translation points. On the other hand, the language errors in both Spanish and English were subtracted from a maximum grade of 10. The final grade was then the average between the translation points obtained and the linguistic quality of the target text.

For example, one translation problem of the English-Spanish assignment was finding the correct terminological equivalent in Spanish for the different types of current (longshore, tidal and rip current). Another problem was understanding the exact location of a *groyne* in “perpendicular or slightly oblique to the shoreline extending into the surf zone (generally slightly beyond the low water line)”. An example of a translation problem in the Spanish-English assignment was understanding that *bocana* and *desembocadura* are synonyms, and can both be translated as *river mouth*.

In experiment 1 (Figure 17), the EcoLexiCAT translators outperformed the MateCat translators in both directionalities, although only slightly in the ES-EN assignment. The average quality of the target texts of both groups was not very high. This is understandable because most subjects in both groups did not have any professional translation experience or previous knowledge of the environmental domain. The results were promising though, as EcoLexiCAT helped to obtain a better target text in less time.

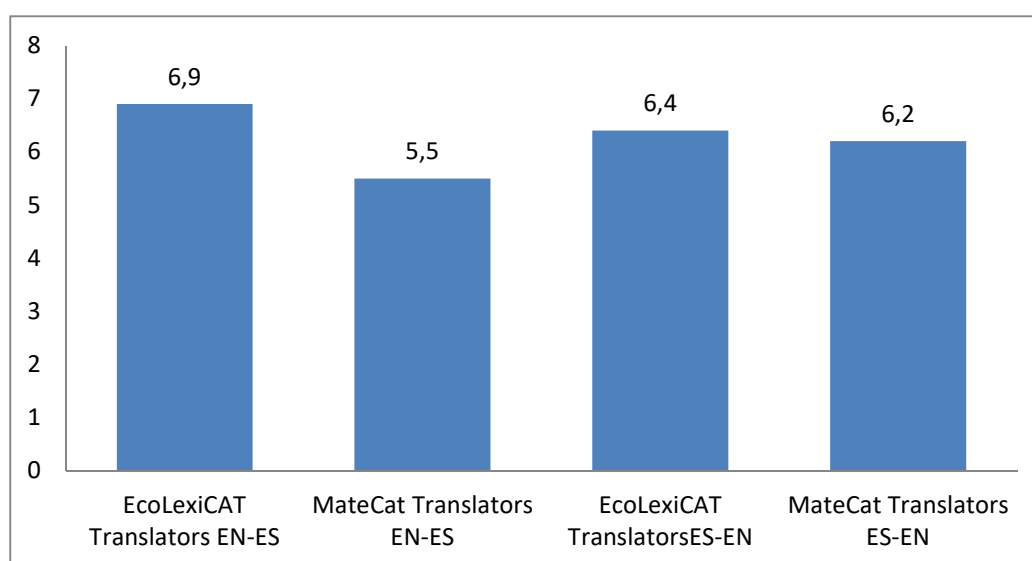


Figure 17: User performance in experiment 1 – quality.

It is also interesting that the control group used very similar resources to solve the translation problems: EcoLexicon, BabelNet, Wordreference, IATE, Linguee, and Wikipedia.

In terms of the time invested (Figure 18), in both directionalities EcoLexiCAT translators outperformed the control group. Surprisingly, the EcoLexiCAT group took longer in the ES-EN assignment than in the EN-ES one, whereas the control group took longer in the EN-ES assignment. This is striking because even though it was a shorter source text, the assignment involved translating into a non-mother tongue of most of the subjects.

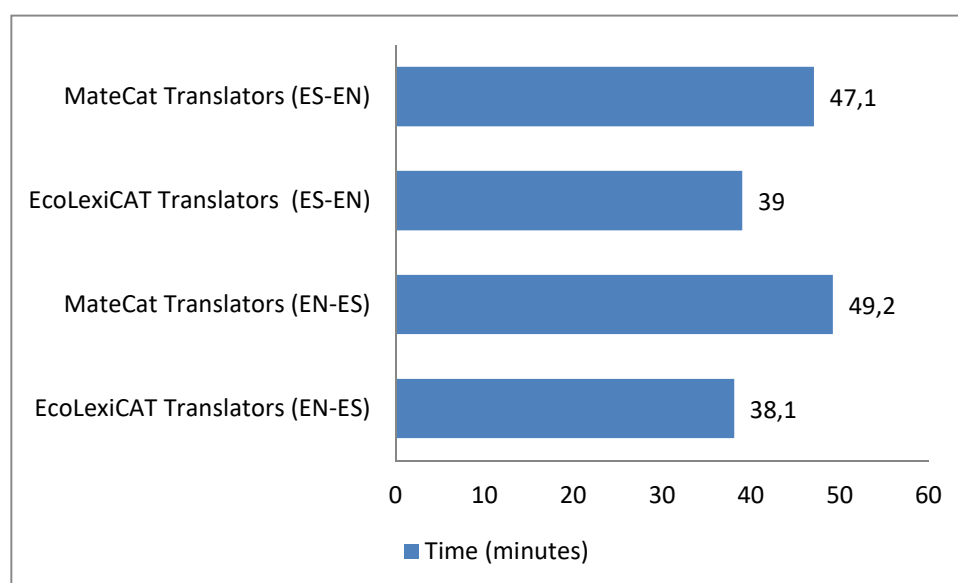


Figure 18: User performance in experiment 1 – time invested.

In experiment 2, however, the average quality of the target texts of both groups was higher than the average quality of both groups in experiment 1. This is surprising, as half of the subjects were undergraduate students in experiment 2, whereas in experiment 1 all of them were master's students. If we look at the translators group in experiment 2 (Figure 19, EN-ES: 8 and ES-EN: 7.3) and the group that translated with EcoLexiCAT in experiment 1 (Figure 17: EN-ES: 6.9 and ES-EN: 6.4), the improvement is clear, approximately one point more in both cases. In fact, in the ES-EN assignment, there was an average 9.3-minute time gain (Figure 20). This may be due to the fact that in experiment 2 better students were recruited, or that the improvements in EcoLexiCAT after experiment 1 had an impact on user performance.

As for the comparison between translators and post-editors in experiment 2, in terms of quality (Figure 19) the translators outperformed the post-editors in the EN-ES task, whereas in the ES-EN assignment the opposite occurred. In terms of the time invested (Figure 20), in both assignments post-editors outperformed translators, with a difference of 14.4 minutes for the EN-ES assignment and 10.6 in the ES-EN task.

This means that post-editing definitely reduces the average time spent on translation tasks, but it does not necessarily entail any improvement in quality.

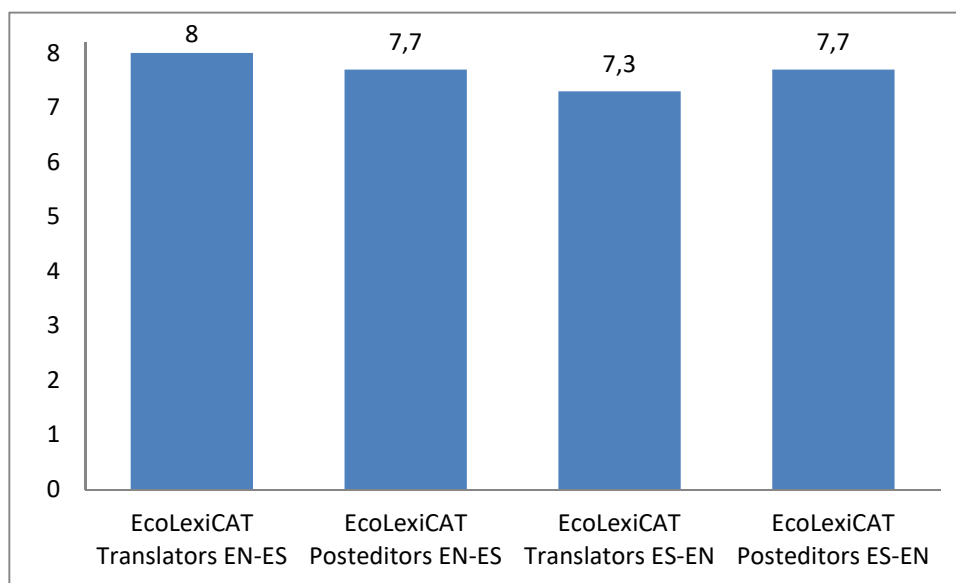


Figure 19: User performance in experiment 2 – quality.

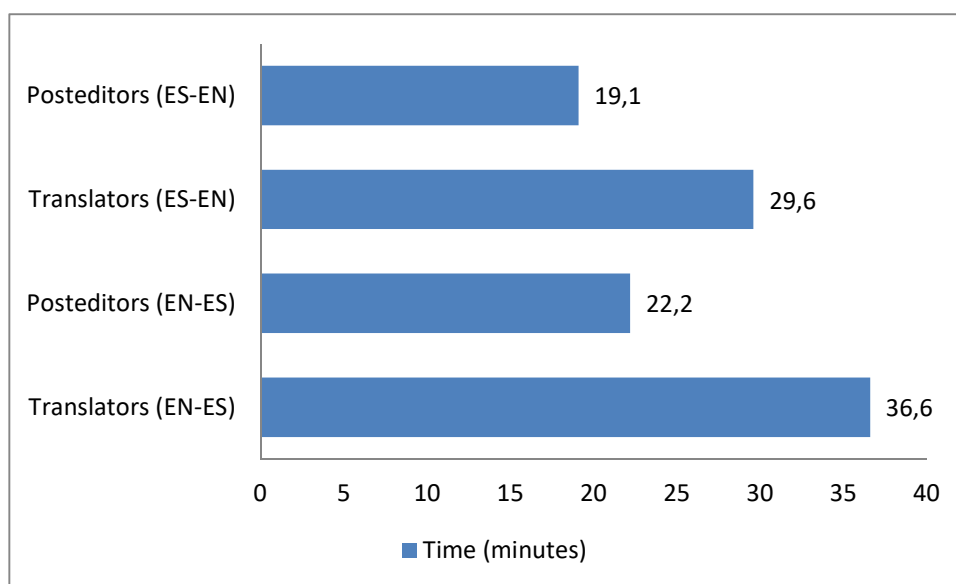


Figure 20: User performance in experiment 2 – time invested.

A qualitative comparison can be made when looking at the errors of the three groups (EcoLexiCAT translators in experiment 1; EcoLexiCAT translators in experiment 2; and EcoLexiCAT post-editors in experiment 2). Tables 2 and 3 show a collection of some of these errors accompanied by their frequency in all three subject groups.

When comparing the type of error made in the ES-EN task (Table 3), it seems that post-editing greatly reduced basic grammar and spelling problems in English, as all

subjects were translating into a non-mother tongue. However, when post-editing into their mother tongue (Table 2), subjects seemed to be more indulgent with the MT outputs.

In the EN-ES task, post-editors sometimes agreed too easily with MT options, for example when giving very literal translations of *are no remedy* and when translating *soft cliff* as *acantilado suave*, which in this context should be *blando*, since *soft* in this context refers to easily eroded cliff material. *Soft cliff* was a problem for all three groups, but the translators at least avoided the *suave* option. On the other hand, MT seemed helpful for the translation of the terms *inner surfzone* and *beach fills*. MT was also very helpful with the construction *to stabilize relatively deep tidal channels*, as the translators did not seem to understand that *relatively* affected the adjective *deep* and not the verb *stabilize*.

Translation problem EN-ES	Experiment 1 Translation	Experiment 2 Translation	Experiment 2 Post-editing
Term: hard coastal structure	1	2	1
Term: groyne, breakwater	2	2	4
Exp: are no remedy	5	1	6
Term: soft cliff	6	9	8
Term: high surge levels	2	5	4
Term: surf zone, inner surf zone	6	6	2
Comp: "Groyne...inner surf zone..."	6	2	1
Term: beach fills	3	5	2
Term: tidal, longshore and rip currents	2	0	1
Term: straight groynes, T-head, L-shaped and Y-shaped groynes	3	2	0
MT: at a more offshore position	1	2	1
MT: artificial submerged reefs	0	4	0
MT: mean sea level	0	4	3
MT: to stabilize relatively deep tidal channels	6	5	0
MT: diminish the generation of rip currents	0	1	0
MT: storm-induced erosion of sandy dunes and soft cliffs	3	0	4
MT: near the groyne heads	1	1	2

Table 2. Translation problems in EN-ES assignment.

However, because of the complicated word order of the sentence *storm-induced erosion of sandy dunes and soft cliffs during conditions with relatively high surge levels*, MT was not helpful in this case, whereas the translators in experiment 2 were capable of understanding the content. There are various indicators that the students of experiment 2 were generally better than those of experiment 1. For example, there were comprehension problems with the sentence: *Groynes are long, narrow structures perpendicular or slightly oblique to the shoreline extending into the surf zone (generally slightly beyond the low water line)*. Nevertheless, the subjects in experiment 2 tended

to be less precise, since they omitted *submerged* in the phrase *artificial submerged reefs*, and *mean* in *mean sea level* in their translations. In the first case, the post-editors did not show this problem, which was probably solved by the MT option.

In the ES-EN task the differences are not as clear, possibly because all the students were translating into a foreign language. However, MT again led to a more literal translation (e.g. in *remodelados*). In addition, in all cases where post-editors had problems with the term *ambientes mesomareales*, this was due to the fact that MT omitted *mesomareal*, and the post-editors did not correct this. Some results again show that the subjects of experiment 1 did not perform as well as those of experiment 2, as they had problems with expressions such as *están sujetos a* and comprehension problems with the sentence *para los casos de desembocaduras sin diques y con diques de encauzamiento*.

Translation problem ES-EN	Experiment 1 Translation	Experiment 2 Translation	Experiment 2 Post-editing
Term: sistemas abiertos	1	1	0
Exp: remodelados	4	4	9
Exp: están sujetos a	5	0	0
Term: acreción	2	1	1
Term: ambientes mesomareales	1	3	6
Term: un canal formado por una flecha	3	4	5
Term: bocana, desembocadura	9	9	8
Comp: relación existente... bocana	2	1	0
Term: diques de encauzamiento	9	2	1
Comp: Los procesos sedimentarios...barras.	4	0	1
MT exp: se aceleran	5	3	6
MT corrientes inducidas por el oleaje	1	1	0
MT condiciones de cierre	2	1	0
MT estuarios mareales	1	1	0
MT con una energía moderada del oleaje	0	1	0

Table 3. Translation problems in ES-EN assignment.

6. User satisfaction

User satisfaction was measured in three subject groups of 10 members each: EcoLexiCAT translators in experiment 1; EcoLexiCAT translators in experiment 2; post-editors in experiment 2. When asked about the general usefulness of the tool for the translation of environmental texts, the subjects in the first experiment said that the tool was very useful (60%) or useful (40%). Likewise, in the second experiment the subjects stated that the tool was very useful (70% EcoLexiCAT translators and 80% EcoLexiCAT post-editors) or useful (30% EcoLexiCAT translators and 20% EcoLexiCAT post-editors). No subjects answered “not very useful” or “useless”. These figures indicate that the tool had improved from the first to the second experiment,

and also that post-editors found it even more useful than the translators.

The parameters of functionality, usability and efficiency were evaluated, based on the rating of different items on a 1-to-5 Likert scale, where 1 was the lowest rating and 5 the highest. After that, subjects could fill out a free-text field to report problems, make suggestions for improvement, and/or note the tool's strengths.

Regarding functionality (Figures 21-23), the subjects were asked whether the tool contained suitable features for: (1) the translation of environmental texts; (2) the comprehension phase of an environmental text; and (3) the production phase of an environmental text.

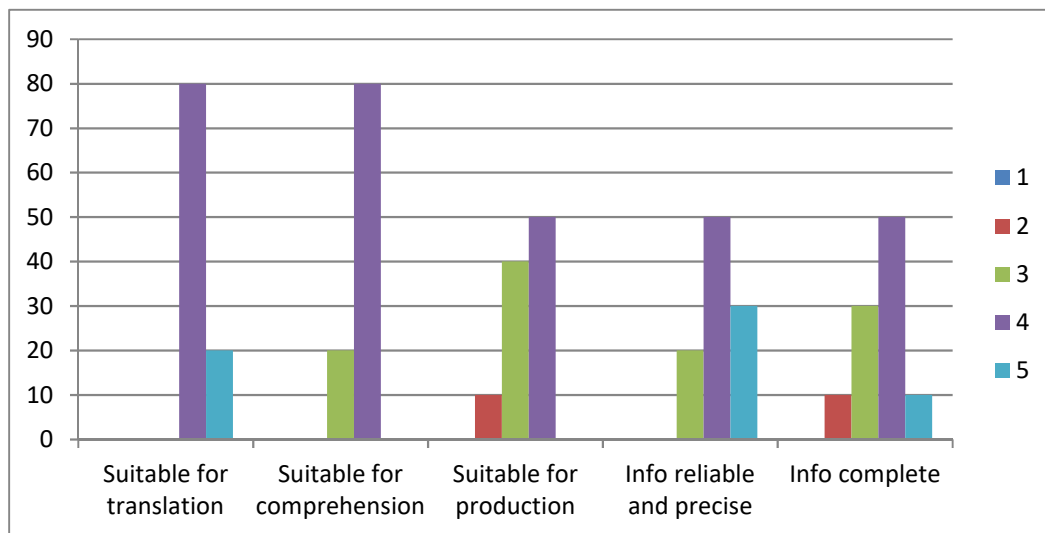


Figure 21: Functionality of EcoLexiCAT – translators in experiment 1.

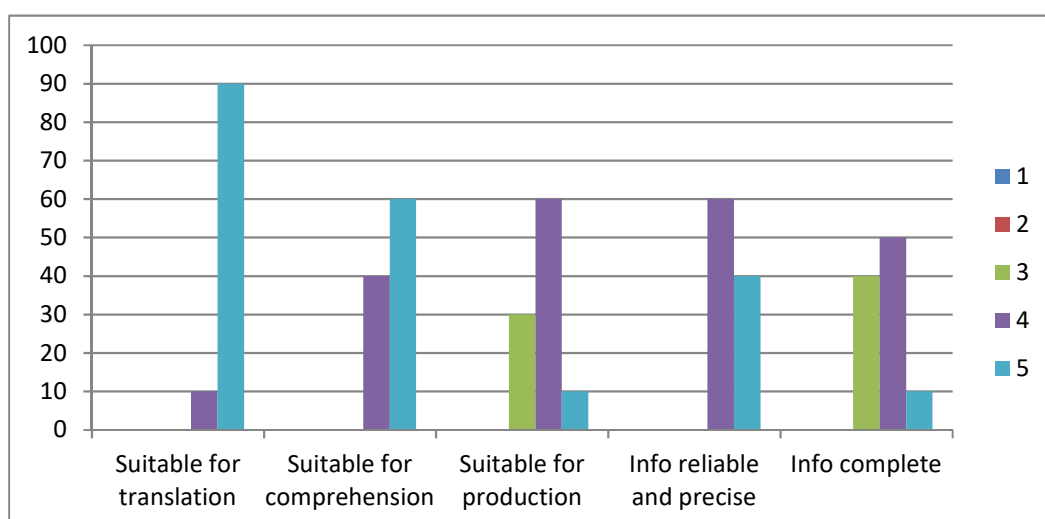


Figure 22: Functionality of EcoLexiCAT – translators in experiment 2.

Generally speaking, from experiment 1 to experiment 2, the tool was rated better, since its suitability for translation was given a score of 5 by 20% of translators in experiment 1, as compared to 90% of translators and 80% of post-editors in experiment 2. Its suitability for the comprehension phase was rated better than that for the production phase in all three groups, though the upward trend continued from experiment 1 to experiment 2. Comprehension was rated with a score of 4 by 80% of translators in experiment 1, but with a 5 by 60% of translators and post-editors in experiment 2. Production received a somewhat lower score, which means that EcoLexiCAT is currently more comprehension-oriented, and that future improvements should focus on increasing assistance in production-oriented tasks. However, a slight upward trend was still evident from experiment 1 to experiment 2. The minimum score in experiment 2 is 3, and the percentage of 4 rose from 50% in experiment 1 to 60% in experiment 2. Not surprisingly, 40% of the post-editors rated production with a 5 and 60% with a 4, which is only natural, since in the case of obtaining highly reusable MT output the text production phase was obviously enhanced.

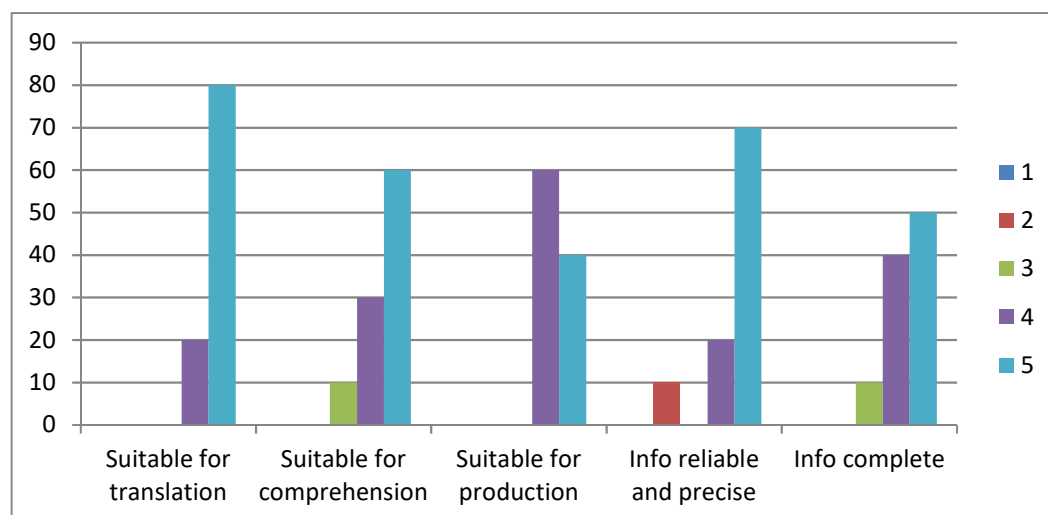


Figure 23: Functionality of EcoLexiCAT – post-editors in experiment 2.

Regarding the reliability, precision, and completeness of the information provided, the scores given by translators in both experiments were practically the same, whereas those given by post-editors were much higher. For instance, reliability and precision were given a 4 by 50% and 60% of translators but received a 5 from 70% of post-editors. These figures call for the continuous extension, improvement and maintenance of terminological resources. Similarly, in experiment 1 one of the subjects stated that the only improvement necessary was to expand the knowledge currently contained in EcoLexicon. Moreover, a translation difficulty reported by a few subjects was the fact that in all resources synonyms and term variants are listed with no clues on how to choose one or another.

When asked to rate the usefulness of external resources during their assignments (Figures 24-26), EcoLexicon, Sketch Engine, Linguee and Wikipedia were rated best in experiment 1. However, this did not exactly correspond to user behaviour (Section 4),

since Sketch Engine was rarely consulted, and Wikipedia was not consulted at all. This shows how users' introspection cannot be the only method used to evaluate a tool. In experiment 2, where new resources were added as other resources (TermiumPlus, Metaglossary, OneLook, and Majstro), EcoLexicon was again the best rated resource (rated 5 by 90% of translators and post-editors), followed by Linguee, Wordreference, Cambridge, and Wikipedia. Again, these results do not exactly correspond to the figures reported in Section 4. For example, Sketch Engine was not reported among the best resources even though it was used more often than individual other resources. Among the worst rated resources (because they were not useful or were not needed), Termium Plus, Metaglossary, OneLook, and Majstro were mentioned. These resources were among those integrated after experiment 1.

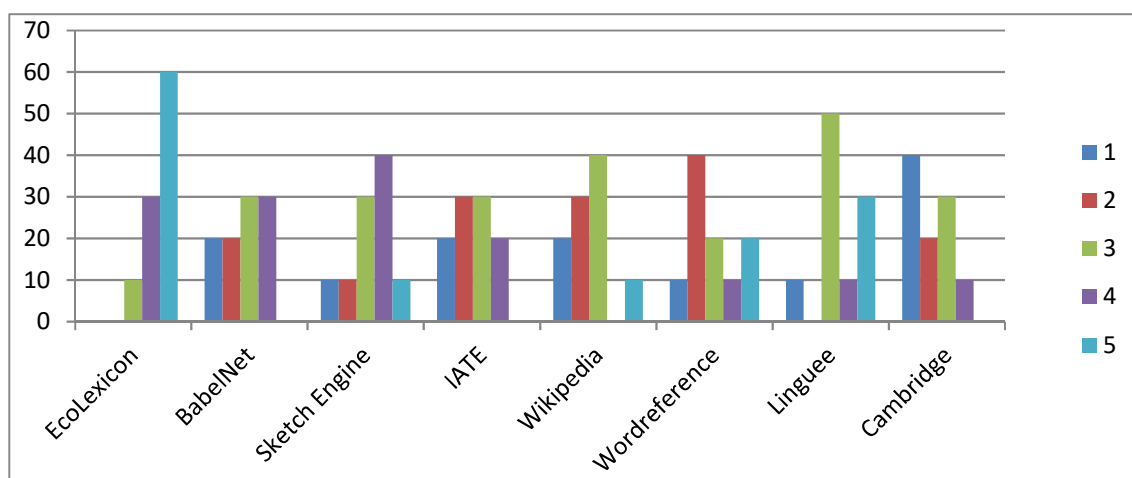


Figure 24: Usefulness of external resources – translators in experiment 1.

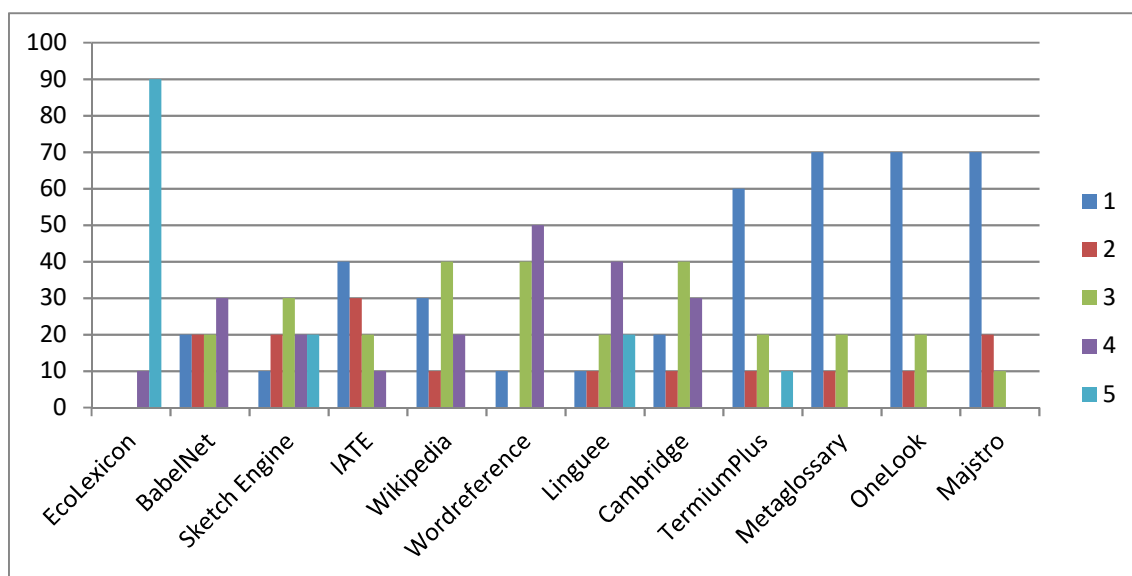


Figure 25: Usefulness of external resources – translators in experiment 2.

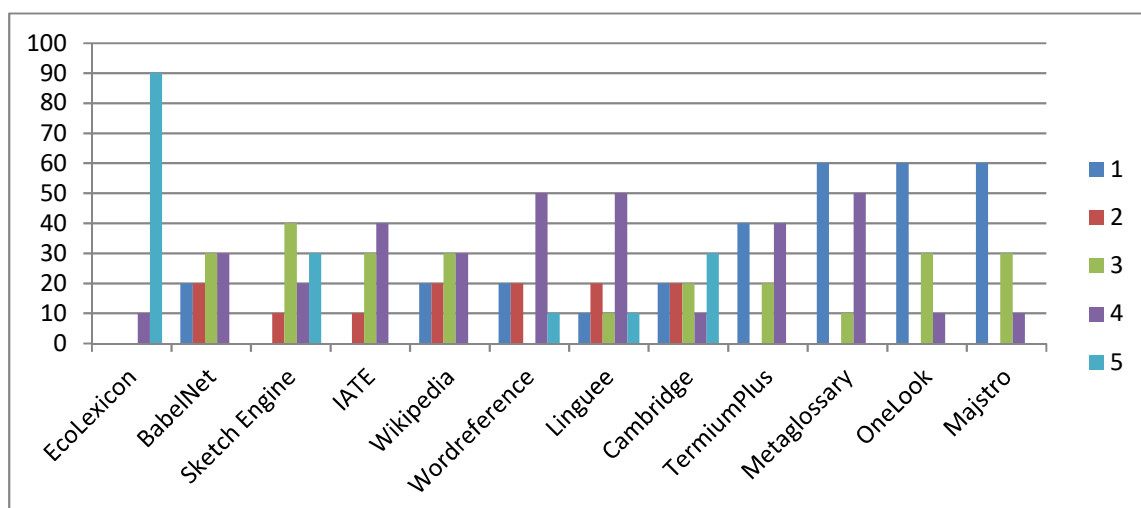


Figure 26: Usefulness of external resources – post-editors in experiment 2.

As for usability (Figures 27-29), the subjects were asked the following about EcoLexiCAT: (1) if it was intuitive and easy to use; (2) if it had a functional design; and (3) if it provided an adequate interaction with the layout (e.g. resizing of the windows).

In both experiments the interaction with the layout was rated the worst. Thus, future improvements should head in this direction, although some of them were already integrated after experiment 1. The score of the design remained stable in the translators groups (in both experiments 40% of the translators rated it with a 5, and 50% with a 4), although the post-editors rated it higher (70% with a 5 and 30% with a 4). Regarding ease of use, this was the parameter that improved the most, since 40% of translators in experiment 1, 70% of the translators in experiment 2, and 100% of the post-editors in experiment 2, rated it with 5.

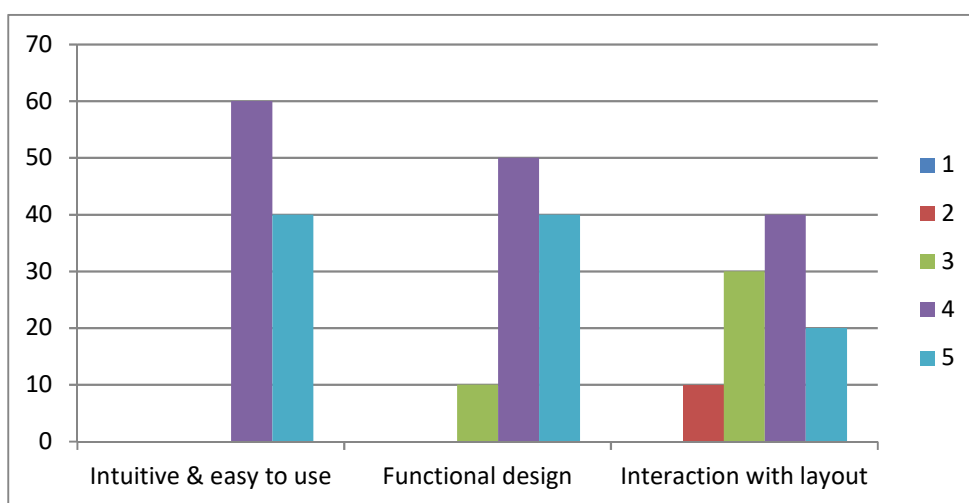


Figure 27: Usability of EcoLexiCAT – translators in experiment 1.

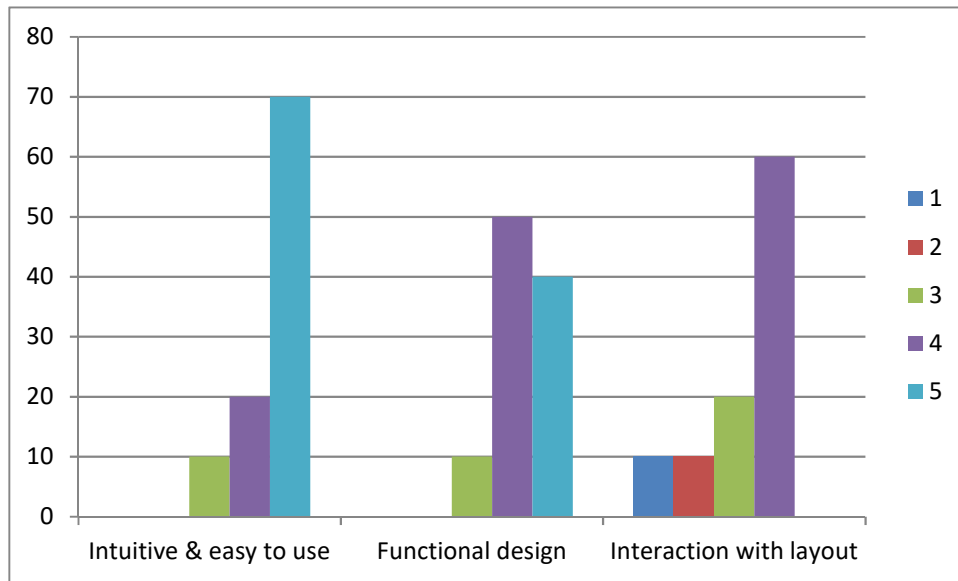


Figure 28: Usability of EcoLexiCAT – translators in experiment 2.

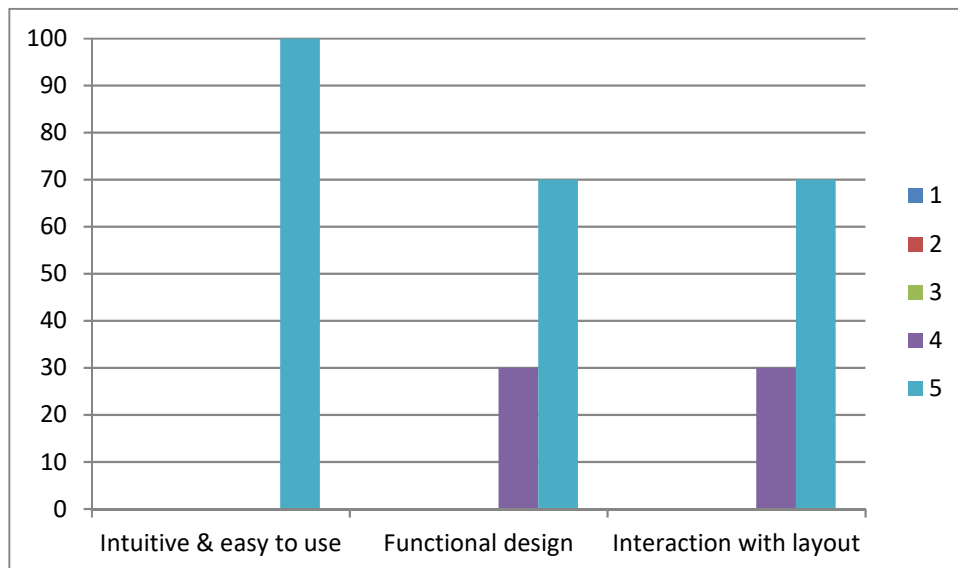


Figure 29: Usability of EcoLexiCAT – post-editors in experiment 2.

Finally, efficiency (Figures 30-32) was assessed based on whether the information loaded at the right speed and fluidity: (1) user interaction with the editor; (2) interaction of the editor with external resources; and (3) user interaction with external resources. All parameters improved from experiment 1 to 2. In experiment 1, they were mostly rated with a 4, whereas in experiment 2 they were mostly rated with a 5. In experiment 1, user-editor and user-resources interaction scored worse than information loading speed and editor-resources interaction. In experiment 2, user-editor interaction and information loading speed improved significantly, but user-resources and editor-resources interaction showed some flaws, even if the general trend was positive. Comparing translators' and post-editors' assessments in experiment 2, post-editors clearly gave higher scores to all parameters.

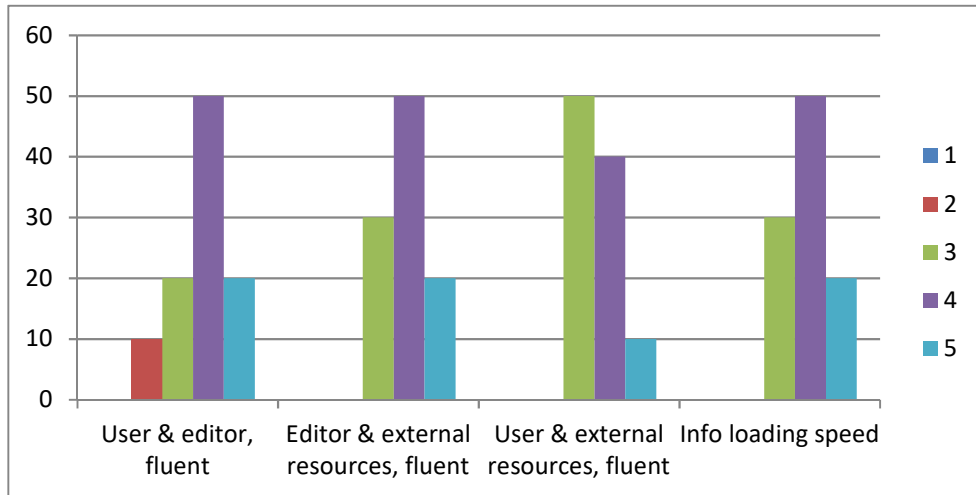


Figure 30: Efficiency of EcoLexiCAT – translators in experiment 1.

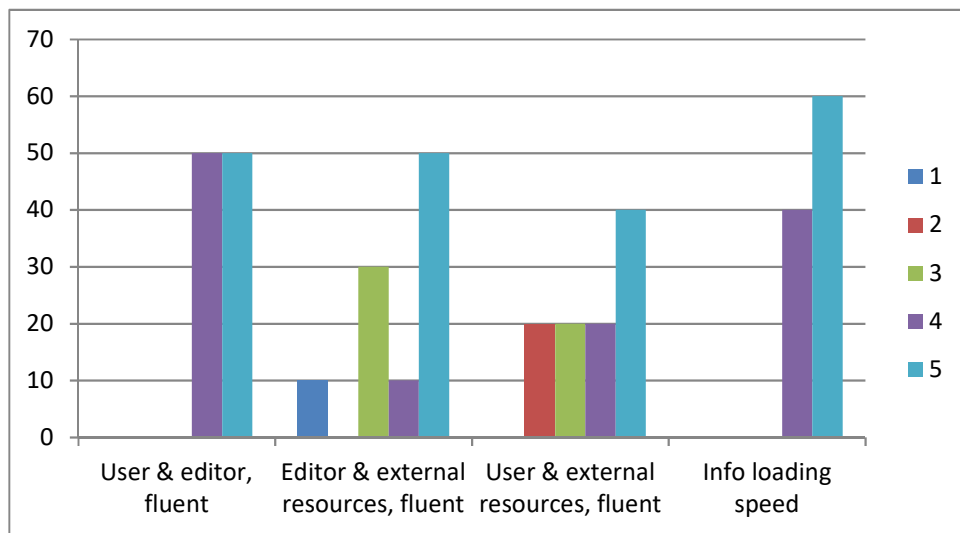


Figure 31: Efficiency of EcoLexiCAT – translators in experiment 2.

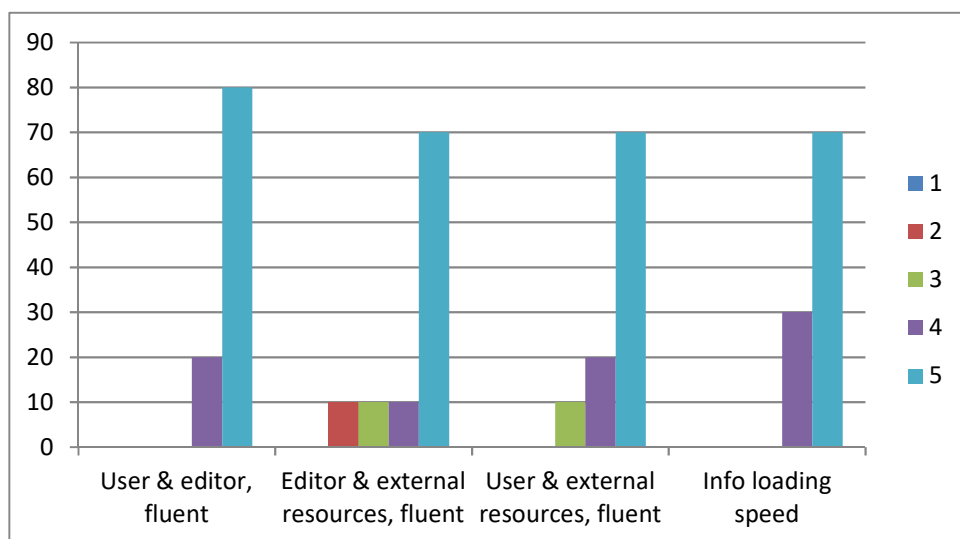


Figure 32: Efficiency of EcoLexiCAT – post-editors in experiment 2.

The post-editor group also answered a question regarding MT efficiency. They were asked to assess on a 1-5 Likert scale the frequency with which they encountered common issues in MT (i.e. inadequate terminology, literal translation, problems with numbers and figures, omissions, additions, etc., with the results shown in Figure 33). These results, together with those related to the time invested, show that the reusability of MT output was significantly high. Unintelligible segments were rare, as well as omissions, additions and issues related to spelling, gender and number, punctuation and capitalization, and words that should be kept in the source language. In contrast, word order, literal translations, and inadequate terminology were the issues that were most often encountered, and on which the post-editing process had to focus. Most users acknowledged that MT was of great help.

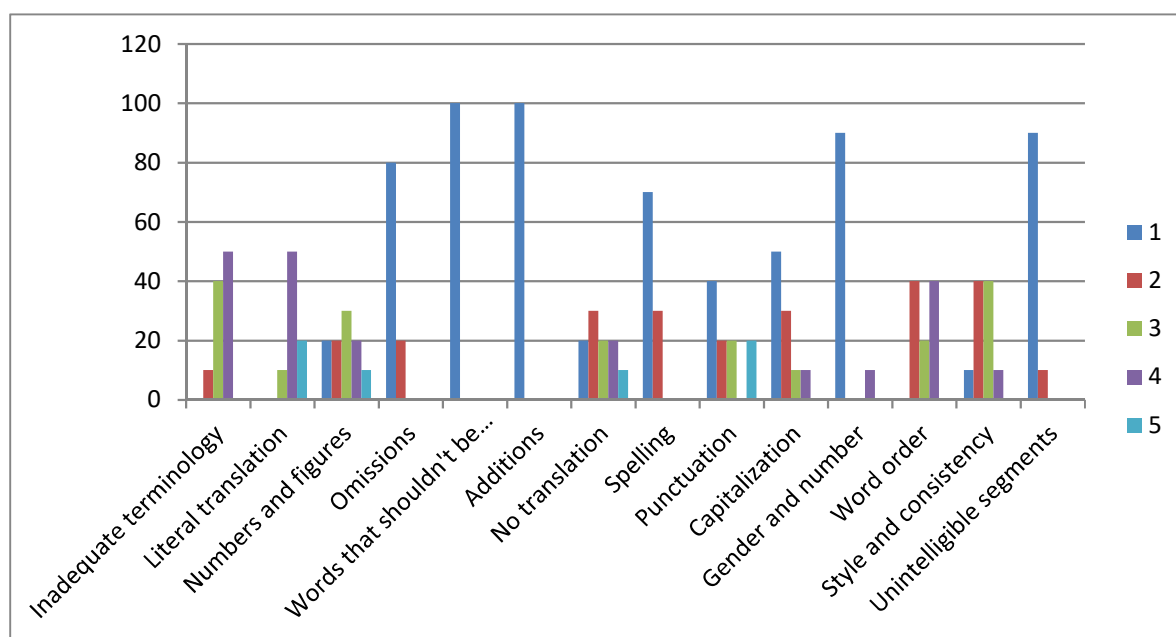


Figure 33: MT issues – post-editors in experiment 2.

When asked about the tool's flaws and possible improvements, in experiment 1 several subjects reported some bugs and efficiency issues regarding the other resources pop-up window – it could not be resized or moved, making things difficult to see – and the predictive typing feature in the target segment, which did not work well in the case of multiword terms. Moreover, certain plural multiword terms in Spanish were not lemmatized properly, and thus not recognized as terms in EcoLexicon. These issues were addressed before experiment 2, but again users reported other problems related to both issues: sometimes the other resources window would disappear until the browser was refreshed, and the predictive typing feature added a line break in the target segment.

One subject in experiment 1 suggested adding the other resources window to the left-hand side of the screen, as already inferred from the analysis of user behaviour (Section 4). However, in experiment 2 the users seemed to be happy having the general

language resources in that window, instead of placing them with the terminology resources.

Among other suggestions for improvement, the subjects in experiment 1 proposed the addition of the resources added before experiment 2. In experiment 2, the subjects proposed the inclusion of an environmental corpus in Spanish, part-of-speech information, style guides, reliability rates for terms usage, possibility of having shortcuts for the different searches, and a better integrated quality assessment tool.

When asked about the positive aspects of the tool, many subjects in both experiments pointed out that the quick and easy access to so many resources in the same interface, as well as the fact that the search terms do not need to be typed, is the main strength of the tool, which is the whole idea behind our concept of terminology enhancement. However, there were also several users that felt overwhelmed by the amount of information shown. They proposed making the layout more flexible so that users could customize the order, amount, and position of resource boxes. Users also highlighted the usefulness of Sketch Engine and EcoLexicon, especially its definitions, term equivalents, and images.

7. Conclusions and future work

Based on user expectations, EcoLexiCAT can be regarded as a tool specifically tailored to user needs and conceived in line with the augmented translation approach. According to user performance, the results of the experiments indicate that integrating terminology enhancement in the translation workflow in a stand-alone interface improves the quality of the translation and reduces the time spent on the task. MT post-editing, however, reduces the time spent on the task but does not necessarily raise the quality. With regard to user behaviour, we can conclude that the most useful resource in EcoLexiCAT is EcoLexicon, which is hardly surprising, since the tool is specifically conceived for environmental translation. The increased use of Sketch Engine was observed in experiment 2. Definitions and term equivalents were the data categories most often consulted in all resources. Users also showed a clear preference in the way they accessed information. In this sense, clicking was the preferred mode, followed by the right-click menu option, and finally by typing the search in the form.

Regarding user satisfaction, the three parameters point to a favourable evaluation of EcoLexicon, although efficiency will be the first aspect to be improved in the future. Post-editors tended to rate the tool better as a whole, since all parameters showed higher figures in this subject group. Comparing translators' general assessments in experiments 1 and 2, those in experiment 2 were slightly better. We can thus conclude that both improvements from experiment 1 to 2 and the MT feature had a positive impact on the evaluation of EcoLexiCAT.

Based on these studies, EcoLexiCAT thus seems to be on the right path. However, it

still needs to be assessed by more prospective users. Wider studies with larger samples, including professional translators, will be carried out in the future. Other features and resources will also need to be added to the tool, especially those related to text production tasks, such as phraseological information and access to the EcoLexicon Spanish corpus. All flaws and bugs reported will also be fixed. Moreover, if EcoLexiCAT were extensively used, it would be possible to draw meaningful conclusions about the kind of terms/concepts most researched through each of the resources and data categories. This would provide valuable insights into how to build and improve augmented translation tools.

8. Acknowledgements

This research was carried out as part of the project FFI2017-89127-P, Translation-oriented Terminology Tools for Environmental Texts (TOTEM), funded by the Spanish Ministry of Economy and Competitiveness. The authors would like to thank the students who participated in the evaluation of EcoLexiCAT.

9. References

- De Palma, D.A. & Lommel, A. (2017). *Augmented Translation Powers up Language Services*. Common Sense Advisory.
- Durán Muñoz, I. (2012). Meeting translators' needs: translation-oriented terminological management and applications. *The Journal of Specialised Translation*, 18, pp. 77–92.
- Faber, P., León-Araúz, P. & Reimerink, A. (2014). Representing environmental knowledge in EcoLexicon. In E. Bárcena, T. Read & J. Arús (eds.) *Languages for Specific Purposes in the Digital Era. Educational Linguistics*, 19. Springer, pp. 267–301.
- Faber, P., León-Araúz, P. & Reimerink, A. (2016). EcoLexicon: new features and challenges. In I. Kernerman, I. Kosem, S. Krek, & L. Trap-Jensen (eds.) *GLOBALEX 2016: Lexicographic Resources for Human Language Technology in conjunction with the 10th edition of the Language Resources and Evaluation Conference*, Portorož, pp. 73–80.
- Kilgariff, A, Rychly, P., Smrz, P. & Tugwell, D. (2004). The Sketch Engine. In *Proceedings of the 11th EURALEX International Congress*. Lorient: EURALEX, pp. 105–116.
- ISO/IEC 9126-1 (2001) Software engineering -- Product quality. International Organization for Standardization, Geneva, Switzerland.
- León-Araúz, P., Reimerink, A. & Faber, P. (2017). EcoLexiCAT: a Terminology-enhanced Translation Tool for Texts on the Environment. In I. Kosem, J. Kallas, C. Tiberius, S. Krek, M. Jakubíček & V. Baisa (eds.) *Electronic lexicography in the 21st century. Proceedings of eLex 2017*. Brno: Lexical Computing CZ s.r.o, pp. 321–341.

- León-Araúz, P., Reimerink, A. & Faber, P. (2019). Translating environmental texts with EcoLexiCAT. In M. Ji (ed.) *Translating and communicating environmental cultures*. London: Routledge.
- León-Araúz, P., San Martín, A. & Reimerink, A. (2018) The EcoLexicon English Corpus as an open corpus in Sketch Engine. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (eds.) *Proceedings of the 18th EURALEX International Congress*, Ljubljana: Faculty of Arts, pp. 893-901.
- León-Araúz, P. & Reimerink, A. (2018). Evaluating EcoLexiCAT: a Terminology-Enhanced CAT Tool. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis & T. Tokunaga (eds.) *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France: European Language Resources Association (ELRA).
- Lommel, A. (2018). Augmented translation: A new approach to combining human and machine capabilities. *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas* (Volume 2: User Papers), Boston, March 17 - 21, 2018.
- Lommel, A. (2017). Augmented Translation. In Brown-Hoekstra (ed.) *The Language of Localization*. XML Press.
- Navigli, R. & Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193, pp. 217–250.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



Lexical Tools for Low-Resource Languages:

A Livonian Case-Study

Valts Ernštreits

The University of Latvia Livonian Institute, Kronvalda 4-220, Riga LV1010, Latvia

E-mail: valts.ernstreits@lu.lv

Abstract

This article focuses on the empirical experience and conclusions, resulting from the creation of language research and acquisition tools for Livonian – one of the smallest languages in Europe.

A cluster was created for Livonian containing three interconnected databases, each with distinct types of data – lexical, morphological, and a corpus. The lexical database contains the lemmas and their data, the morphological database stores morphological forms, while all textual material, including the dictionary examples, is in the corpus. When indexing the corpus, every word refers to a lemma in the lexical database and its morphological information (new lemmas are added prior to indexation), ensuring consistency of the language data, and from each database the full data set of the other databases can be accessed.

The function of each cluster is to extract the maximum amount of information from limited data sources. While technologies designed for languages with a large number of speakers focus on using quantitative methods and automation to extract qualitative information from a large and constantly expanding amount of linguistic data, the main function of technologies designed for small languages is to extract the same type of information from a limited and largely static data set.

This article also examines a string of problems faced when working with a small amount of resources (inadequate language data, insufficient personnel, lack of rules for automating processes, etc.) and methods for resolving these problems in the case of Livonian.

Keywords: Livonian; low-resource languages; lexicography; corpora; data collection

1. Introduction

Livonian, one of the smallest languages in Europe, at present is spoken fluently by ~20 people¹. Although currently listed in UNESCO's Atlas of the World's Languages in Danger as critically endangered (unesco.org), historically, the Livonians have had a

¹ According to the most recent census, 250 Livonians live in Latvia (csb.gov.lv); however, the majority of them do not speak Livonian and a reliable estimate of the number of speakers cannot be made. Due to the scattered nature of the Livonian population and the complex language situation, since the most recent Livonian speaker census in 1935–1937 (Blumberga, 2006), no other attempts have been made at assessing the total number of Livonian speakers. However, in Summer 2019, there are plans for a sociolinguistic pilot project to be carried out by UL Livonian Institute researchers to determine the number of Livonian speakers and their level of proficiency.

significant role in the development of modern-day Latvia and the entire Baltic Sea region. For this reason, Livonian language resources must be accessible not only to the Livonian community for language revitalization work, but also to society at large. Though the number of Livonian speakers is extremely small, Livonian requires the same opportunities and language tools as any other language. This article is focused on the experience and conclusions resulting from the design of technical language support tools for Livonian over the course of the last five years. Some aspects are also discussed in previous studies (e.g., Ernštreits 2019).

A seemingly small user base, associated limitations in being able to access financial resources as well as institutional lack of interest are not the only problems one encounters when creating modern language tools for exceptionally small languages like Livonian. The small amount of speakers also places a limit on the number of people who could potentially be involved in creating these language tools, which means that every potential language tool must be evaluated based on the actual possibilities for creating it and also its effectiveness. This same problem likewise affects the accessibility of the end product; for these tools to be usable by a wider audience for the purposes of language research and learning, one has to already consider the fact that these tools will need to be equipped with translations into one or more other languages (usually: Latvian, Estonian, English).

An added challenge faced by Livonian is that following the Second World War and the Soviet occupation, Livonian speakers were scattered across Latvia and the world. The same is true for Livonian language sources and researchers, which are located at various different institutions (Ernštreits, 2012). This means that when creating or using any resource for Livonian, people from very different backgrounds are involved and are working from different platforms and locations around the globe.

Another important aspect is that the grammars of small languages are often insufficiently studied. As a result, there are many processes which cannot be automated due to lack of knowledge of grammatical rules, while other alternatives, such as solutions based on neural networks, do not function well due to insufficient data. Additionally, sources of language data have been recorded at different times and so they are often written using different transcriptions², which limits the possibilities for people without existing specialized knowledge from using these sources and makes it difficult to process these texts electronically. However, the primary problem is that small languages consistently suffer from a lack of sufficient institutional interest, as well as inadequate data sources resulting from insufficient documentation.

The abundance of available resources is also the primary distinguishing factor when

² The problem faced by Livonian is the wide-ranging use of phonetic transcription, which, moreover, is not used for its basic function – accurately depicting pronunciation – but rather as a systematic means for writing down sources, including lexical sources (Ernštreits, 2011).

creating technologies for different languages. Technologies designed for languages with a large number of speakers focus primarily on using quantitative methods and automation to extract qualitative information from a large and continuously expanding amount of linguistic data. The primary function of technologies designed for small languages is to acquire that same information from a limited and largely static data set, primarily using qualitative methods in an effort to extract the maximum amount of information in circumstances where the available human resources and opportunities for automating this entire process are also limited.

A database cluster containing lexical morphological databases as well as a Livonian language corpus was created to resolve all of the aforementioned problems faced by Livonian. Its function is to ensure information acquisition from the limited Livonian language sources, while simultaneously optimizing the tasks carried out by the personnel working with the database and creating a base for further expanded use of both existing and future databases.

2. Creating the database clusters

2.1 Earlier Livonian language dictionaries and databases

The history of Livonian language dictionaries is relatively long. The first Livonian dictionary (Livonian-German-Livonian) was published in 1861 (~9,000 lemmas; SW); it was followed in 1938 by a Livonian-German dictionary (~13,000 lemmas; LW). Both of these publications were primarily intended for researchers, and the Livonian entries were written using phonetic transcription (Ernštreits, 2011).

The first Livonian dictionary (Livonian-Latvian-Livonian) intended for general use, and in which all Livonian entries were written using the orthography of the Livonian literary language, was only published in 1999 (~5,000 lemmas; LLLS 1999). This was also the first collection of Livonian vocabulary compiled using electronic tools. It was assembled, beginning in 1995, from entries in the 1938 dictionary using the Filemaker database software, though due to various reasons the primitive system used for compiling this dictionary was not further developed. However, for its time it was somewhat advanced. One of the first Livonian fonts and also lemma-sorting algorithms were designed for this dictionary, as well as the first Livonian keyboard drivers, the principles of which continue to be used up to the present day.

After a lengthy hiatus, in 2012, the most extensive lexicographic publication in the Livonian literary language – the Livonian-Estonian-Latvian Dictionary (13,000 lemmas; LELD) was published. The basis for this dictionary is the nearly 40 years of work by Estonian researcher Tiit-Rein Viitso, who collected and compiled Livonian vocabulary, language examples, and morphology. For understandable reasons, the basis of the dictionary was prepared using analogue methods. In the project's final phase, the information from the card index was transferred to MS Word format.

During the next year, the dictionary was transformed from its original text format into a database and published online (murre.ut.ee). Following that, in 2015, the indexing tool *Liivike* was created, which used this database as a lemma reference source and enabled the creation of a corpus of Livonian texts in phonetic transcription within the Archive of Estonian Dialects and Kindred Languages at the University of Tartu. The aforementioned electronic dictionary and the tables of morphological patterns published in that dictionary were also used in the University of Helsinki project “Morphological Parsers for Minority Finno-Ugrian Languages” (2013–2014).

All of the aforementioned linguistic tools, however, had their problems, e.g., the web version of the dictionary was created as a static database, and therefore was difficult to update and correct. The dialect corpus utilized Uralic phonetic transcription and so was suitable only for research purposes (rather than, for example, language acquisition), its indexing system also allowed only for fully indexed texts to be uploaded or edited. This led in many cases to “forced indexation”, especially for unclear cases, and sometimes indexation errors due to the poor Livonian language skills of the people doing the indexing. Also, due to the structure of the workflow, later corrections of various inadequate indexations were extremely difficult to correct, e.g., systematic indexation mistakes could be corrected in isolated textual units, but not across the entire corpus, etc.

The morphological analyser and other tools created by the University of Helsinki project worked well, but were made using an existing set of morphological rules and were therefore static and sometimes incorrect. As further developments have clearly shown, morphological rules for Livonian remain at a hypothetical stage in many cases, as they still need to be further clarified and/or adjusted based on information gained from the corpus. However, the most severe flaw of all these previously existing systems and linguistic tools was the fact that they used the same initial source (the database based on LELD digital data), but were also isolated, not providing any feedback with updates or corrections, requiring all efforts to keep the databases updated to be fully manual, and thus being quite ineffective and never performing consistently. As a result, the understanding that a new approach to linguistic tools was needed gradually began to form.

2.2 The precursor of the Livonian language database cluster – the

Estonian-Latvian dictionary

It could be said the events outlined above happened led to the creation of the cluster and its databases. In 2013, a working group was formed with professionals from Latvia and Estonia in order to compile both the print and electronic versions of the Estonian-Latvian and Latvian-Estonian dictionaries. These dictionaries, containing 40,000 lemmas each, had to be compiled from scratch and published as a joint effort of

the Latvian Language Agency (Estonian-Latvian; ELD) and the Estonian Language Institute (Latvian-Estonian; LED) within a timeframe of two and a half years.

Originally, the Estonian side was to use its own lexicographic working environment EELex for compiling the dictionaries; however, once testing began, the Latvian side concluded that this system was outdated and worked too slowly. For example, an average of 1.5 minutes was needed to open an entry, make an edit, and close the entry in this system, which meant that a compiler, who needed to compile at least 30 entries per day in order to meet the project deadline, would lose at least 45 minutes of work time per day. In addition, though this system could be used online, it was possible to use it with only one type of operating system and one type of browser.

When there were no results after almost half a year of attempts to resolve the issues with the EELEX system relating to the speed of operation and other aspects of compiling entries, the Latvian side decided to begin immediate work on a solution. Within 48 hours they had constructed a temporary online system not connected with any particular operating system, which decreased the opening/closing time for each entry to two seconds and permitted the user to see the entry with its final formatting as information was added to it, to move examples and entire definitions between groups without difficulty, search for entries, view the completion status of each entry as well as use data from the Estonian Language Frequency Dictionary (EKSS), the unified corpus (cl.ut.ee), and the Glossary of Estonian Basic Vocabulary (EKPS) for selecting lemmas and also print out any part of the dictionary or print out the full dictionary in its final formatting. Unexpectedly, this system proved to be so productive that the decision was made to continue work on the Latvian side using this system. During the course of the project it was supplemented with a string of other tools necessary to ensure the quality of the final product – a reverse dictionary, compound word inspection, and other tools.

ŠKIRKĻA LABOŠANA

☒ Saglabāt
 ☐ gatavs
 ☒ Saglabāt

Šķirkļa pamatdati

lugema

⌵

lugeda, `loen

Tips

⌵

v

⌵

Saistītie vārdi `loe

☐ pie saglabāšanas ģenerēt gramatiskā lauka vārdus

Šķirkļa nozīmes pievienot jaunus

1. ⌵ ()

lasīt

pievienot salikteni

pievienot piemēru

• raamatut lugema

lasīt grāmatu

• kaarti lugema

lasīt karti

• ta loeb mu mōtteid

viņš lasa manas domas

• ta loeb ūlikoolis filosoofia ajalug

viņš lasa filozofijas vēsturi univ

2. ⌵ ()

skaitīt

pievienot salikteni

pievienot piemēru

• ta loeb raha

viņš skaita naudu

☒ automātiska saglabāšana

lugema v{\lugeda, `loen} 1. lasīt • raamatut lugema lasīt grāmatu; kaarti lugema lasīt karti; ta loeb mu mōtteid viņš lasa manas domas; ta loeb ūlikoolis filosoofia ajalugu viņš lasa filozofijas vēsturi universitātē 2. skaitīt • ta loeb raha viņš skaita naudu; laps oskab juba sajani lugeda bērs jau prot skaitīt līdz simtam 3. skaitīties • see ei loe tas neskaitās; jārjest loe! pēc kārtas skaitīties! 4. uzskatīt • kūsimus loeti otsustatuks jautājumu uzskatja par izlemtu 5. būt svarīgam • aega on vāhe, iga minut loeb laika ir maz, katra minūte ir svarīga

lugema V 542 205 337

valts 2015-02-12 11:49:53

status uz validēts

marika 2013-08-02 13:29:05

status uz pabeigts

import 2013-01-26 17:53:05

Imports no pamatvārdu faila

Figure 1: An inside view of the Estonian-Latvian dictionary compiling module.

This system, which was built using our own resources in conjunction with corpus data for lemma selection, proved to be one of the main steps in compiling the 1,096-page-long Estonian-Latvian dictionary. This work was done over an incredibly short time period without sacrificing the quality expected from lexicographic sources.

A logical question that might emerge from this is whether it is possible to speed up this work even more by utilizing parallel corpora to find correspondences. The Estonian side proceeded down exactly this experimental path. They worked in cooperation with the Latvian language resource company “Tilde” to create the basis for a dictionary compiled in an automated manner utilizing parallel corpora; however, this work did not take into account the aspects discussed in this article

Thus, in addition to the term “low-resource languages”, an additional term should be used – “low-resource language combinations”, i.e., those language combinations without parallel corpora or corpora formed from a limited number of sources, translators, documents, and text genres. This leads to the problems noted in the introduction, namely that in circumstances characterized by insufficient information (as is the case for the Estonian and Latvian language combination) automatic methods cannot be used due to inadequate data. The aforementioned experimental Latvian-English dictionary, which was essentially a structured word list collected from parallel corpora without any word use examples, was criticized for its low lexicographic quality (Bušs, 2015).

2.3 The formation of the Livonian language database cluster

Following the successful completion of the Estonian-Latvian dictionary project in 2015, the decision was made to adapt the lexicographic system created for compiling the Estonian-Latvian dictionary to the needs of the LELD. In 2017, it was supplemented with fields for correspondences in a second language, adjusted to be used with the Livonian writing system and Livonian alphabetic sorting, fields were added for the supplementary information found in the LELD database, but not included in the print version – the sources for the lemmas, correspondences, and examples – along with other necessary additions.

Following the beginning of work on the new dictionary and its publication online (livones.net), it was concluded that from the perspective of language acquisition it was still vital to resolve one of the most troublesome Livonian language problems faced by everyday users – the method for displaying the inflectional morphology of words in the dictionary.

Livonian morphology is relatively complicated. In order to show word inflection, Livonian follows the Estonian example of using word types (usually these are noted in each entry with a numeral following the lemma), which are a model used to show the changes that occur for all words within a particular word group. Livonian has 256

declension types and 68 conjugation types. It is impossible for a user to remember all of these, and it is also complicated to form the inflectional forms of other words by analogy. In order to simplify looking up forms and to free users from needing to constantly use a word type table, a morphological database was created, which utilized the templates included in the LELD and in a partially automated manner generated a template of declination or conjugation forms corresponding to each lemma. As a result, users can see all the forms of that word by clicking on the numeral corresponding to the word type of that word.

After the creation of the morphological database and the active use of the databases, subsequent research showed that the dictionary examples contain lemmas which are not found in the dictionary itself, as these examples had been used to illustrate the use of other lemmas. As a result, the idea arose to supplement both existing databases with a corpus, which would extract vocabulary from texts – using the examples in the LELD as its first text – and collect associated morphological data for the morphological database. This morphological data would be used to test the accuracy of the morphological form template and to gather information about the morphological forms not included in the templates. Since the first part of the cluster was the lexical database, it was logical to connect all subsequent databases to it. However, it was not possible to fully gauge the effectiveness of this solution until the third part of the cluster – the corpus – was completed.

Currently, these three databases are accessible for linguistic research purposes through registered-access modules (lingua.livones.net). Their public parts – mainly targeted towards language acquisition – are currently fully accessible in a separate section of the Livonian culture and language web portal Livones.

3. Cluster operating principles

The cluster is composed of three Livonian language databases – the lexical database, morphological database, and corpus – and consists of interconnected data archives, which have been compiled using the Livonian literary language and are completely editable and usable online. The lexical database forms the backbone of the cluster and each section contains a different type of data. All databases are built with a relational database structure and JSON objects, and the dictionary engine is powered with a PHP application for the backend and simple API calls on the frontend.

The general working principles within all the databases are based on simplified approaches – all necessary work is performed mainly by dragging, clicking, entering search criteria, or completing necessary fields. Workflow is made intuitive and no programming skills whatsoever are required by personnel involved in any of the processes. User controls are eased with visual attribution (e.g., colour-indexed statuses, book-ready lemma articles, etc.).

3.1 Lexical database

The lexical database contains only information about the lemmas, parallel forms, semantics, representations in other languages, and the source of the lemmas. The only grammatical information it contains is the word class, word type, and a reference to the use of the word in singular and plural. The other grammatical information concerning the lemma – the word form template created by generating the forms using the word type as a model – are stored in the morphological database, with the lexical database only containing links to these forms.

All the example texts and the references to their source in the lexical database are shown as data from the corpus. The lexical database only contains a link to the respective sentence in the corpus. The lemma is also linked to every indexed use of the lemma in the corpus. The original examples used in the LELD were also transformed into a separate part of the corpus and their separation from the lexical database was one of the main changes undertaken in the process of connecting all of the parts into a single cluster. This was done to prevent duplication of data and ensure the consistency of the data across the entire cluster.

The lexical database also includes various statuses that allow one to identify the status of work performed (e.g., finalized, missing grammar, etc.) or to limit public access (e.g., technical lemmas from the corpus, such as Latvian-like personal names or casual new borrowings). These may also be used for language standardization purposes. This module also has several additional functions, such as various search and selection options, a reverse dictionary, and also options for printing search results in the form of a pre-formatted dictionary.

3.2 Morphological database

The morphological database contains fields for all known word forms and parallel forms (the set of forms depends on the word class). They are partially filled with word form templates, which are generated according to the word type example, a process which is partially automated with the help of simplified formulas. These formulas are also used in generating form template sets for new lemmas to be included in the lexical database. The database also contains empty fields for rarely encountered forms or those not included in the morphological examples found in the LELD, as well as those with formation principles that remain unclear and also parallel forms.

The morphological database is used for corpus-indexing purposes, offering possible matches for indexation, and – after indexation – for collecting morphological data from the corpus in order to verify word form templates statistically or point out differences in declination principles. Although morphological paradigms are linked to lemmas and have been collected over decades of field research, this statistical verification is done

due to the fact that these paradigms still remain hypothetical to some extent, since there are many specific forms that are quite rare and may appear differently than initially assumed or may be statistically not dominant. This is the gap that feedback from the corpus can fill.

The result is accessible in matrix form, offering an overview of all forms of words included in the corresponding paradigm, and the automatic generation process also helps to reveal inconsistencies and subsequently to create new sub-paradigms. Moreover, based on this database, an overview of paradigm patterns is available for further methodological grouping. It is also possible to change a word's type within a word class with the same morphological principles without losing existing data which had already been generated or links to the lexical database or corpus.

3.3 The Corpus of Written Livonian

The Corpus of Written Livonian contains a variety of indexed and unindexed Livonian texts and serves as a base for obtaining new lemmas for the dictionary as well as forms for the morphological database via the indexing process.

The corpus has a dual purpose – it serves as a linguistic source for research on Livonian, but also as a tool for researching other areas, e.g., folklore or ethnography, as it also simultaneously serves as a repository of written texts in Livonian. Sources used in the corpus are, therefore, quite varied. Although initially it mostly contained texts in literary Livonian (books, manuscripts, etc.), other written texts (folklore, texts in dialects, etc.) have been gradually added. The corpus also contains lots of metadata about the added texts, including their origin, dialect (if applicable), compiler or author, historical background, and other references. This data may also be used for narrowing searches – e.g., texts from a particular village, author, etc.

When texts are uploaded, they are split into subsections (e.g., chapters), paragraphs, sentences, and separate words, and then joined back together when the entire text is presented. Previously uploaded texts are normalized so that they are represented using the unified contemporary Livonian orthography. Normalization mostly affects only orthographical representation, leaving things such as dialectal peculiarities intact. The same applies to texts written in phonetic transcription.

LIV **EE** **LI** **Libiešu-igauņu-latviešu vārdnīca** Parādīgam Korpus Korpus meklēšana Vārdnīca Lietotāja Cron Izeja

Teksti Atgriezies pie teikumiem

MARKĒŠANA

Teksta detaļas B-Äböd-36

Pieejamība: (Publiskai lietošanai)

Mēģ ievietēt šom Sūmō-ugrōd rovsugt, neliz kui sūomlizt, ēstlizt, ungārd ja munt.

ee: _____

lv: _____

Mēģ <small>ma pñ NomPl</small>	Ivlizt <small>Ivli z NomPl</small>	šom <small>šōda va PñPl</small>	Sūmō-ugrōd	rovsugt	, neliz	kui ² <small>kui² adv</small>	sūomlizt	ēstlizt <small>ēstli z NomPl</small>	ungārd	ja <small>ja¹ conj</small>
-----------------------------------	---------------------------------------	------------------------------------	------------	---------	---------	--	----------	---	--------	--

munt
munt pñ NomSg

munt *pñ NomSg*

☐ alternatīvais markējums

Korpus:

- munt *pñ* *46 munt *NomSg* 2
- munt *pñ* *46 munt *GenSg* 1

munt meklēt

Vārdnīca [Pref.Pass]:

- munt *pñ* *46 munt *NomSg* ?
- munt *pñ* *46 munt *GenSg* ?
- rō¹ *pñ* *3 munt *NomPl* ?
- rō¹ *pñ* *3 munt *GenPl* ?

Pievienot jaunu vārdu

Meklēt: Meklēt Notīrīt

1 Livliet.

2 Mēģ Ivlizt šom Sūmō-ugrōd rovsugt, neliz kui sūomlizt, ēstlizt, ungārd ja munt. Livlietōn amā ležgli sugrov āt ēstlizt, ja slepietlizt, ku mēģ nei ležgliuzt sugrovōd šom, pidābbōd ēstlizt mēģi ka lđōlluggōd miešō.

3 Jēgi āpgust ne kaimōbbōd livliet lipušon tulzpidāv andōdi. Ēstlizt āt uzstādōnd otmip logātibōmōtōr rindākieš, nei mēģi lđōlluggōd miešō pidās.

4 5

Figure 2: An inside view of the indexation module.

During the indexation process a mandatory reference is made to the lemma (lexical database) and its particular form (morphological database). In the case of new lemmas or deviations from prior indexation, new records are generated in the lexical database and subsequently in the morphological database directly from the indexation module, using the default lemma form, reference to the form, and its source.

Indexation itself is performed by selecting lexemes and their forms, and the lemma article view from the dictionary is available for the purposes of checking every form selected. For every word to be indexed, possible versions are offered based on either previous corpora statistics or the morphological database, and in most cases indexation can be performed by simply clicking to accept the offered combination or choosing a form from the list offered. It is also possible to search for a lexeme in the lexical database on the spot, choose a different, unlisted morphological form, or add a completely new lexeme. Indexed words and sentences are marked with colour indicators in order to distinguish fully indexed, partially indexed, and unindexed parts.

All texts are available for searching as soon as they are uploaded and do not have to be fully or even partially indexed. While indexing, it is possible to leave an indexed word completely unindexed or marked as questionable, which does not limit the availability of texts for research. Since indexing languages with unclear grammatical rules involves a lot of interpretation, it is also possible to add a completely independent second indexing interpretation (e.g., *pin̄kōks* ‘with a dog’: noun, singular, instrumental ~ noun, singular, comitative) or a reference to a completely different lemma and form (*kōrandōl* ‘in the yard’: adverb ~ noun, singular, allative).

Indexation sources	
Corpora	Primary indexation source – candidates from previously indexed forms, statistics, sentence translations, etc.
Morphology database	Secondary indexation source – candidates from forms listed in the database
Lexical database	Lemma information (semantics, grammar information, etc.)
Indexation process	
General principles	The indexation process attributes a lemma and a form to every indexed word. Indexation is performed by clicking (all steps), picking from a drop-down menu (step 4), entering search criteria (step 5), correcting the lemma form (step 6). When clicking any candidate, the lemma article data is displayed. When indexation is completed, the module automatically jumps to the next word.
Step 1	Primary choice – click to accept the most popular indexation match from corpora statistics (the choice offered inside the yellow box at the sentence level) or indicate it as not to be indexed (e.g., number).
Step 2	If not, click to choose an alternative indexation match from the corpora statistics.
Step 3	If not, choose an alternative matching lemma and form from the morphology/lexical database
Step 4	If not, choose an alternative matching lemma from the morphology/lexical database and choose an alternative form from the drop-down menu.
Step 5	If the lemma is not found (e.g., a very different form), manually search for an alternative lemma within the lexical database, then return to step 4.
Step 6	If the lemma is still not found, add a new lemma (semi-manual, when adding a new lemma to the lexical database the word in the form as found in the sentence and the source reference are taken together), correct the lemma form and add the word category, then return to step 4.
Step 7	If necessary, add a second alternative indexation by clicking the checkbox and repeating steps 2 to 6.
Step 8	If in doubt about the indexation outcome, set the status to yellow (unfinished) or red (clear indexation).
Indexation output	
Corpora	Statistics and indexation candidates
Morphology database	Actual forms, statistics
Lexical database	References to the existing lemmas / new lemmas added; source references; example data – references to the sentences; semantics – references to the sentences and meanings in the lemma article (planned)
Other options	
Translations	Translations into several languages may be added.
Corrections	Corrections may be made in the sentence itself, in any indexed items, in translations, etc.
Sentence management	Public access may be restricted, if needed; a sentence can be added to one or multiple lemma articles as an example.

Table 1: Indexation scheme and options available in the indexation module.

It is possible to edit every sentence separately in order to eliminate possible mistakes in the original text, to add translations in several languages, and to set limitations for

sentences, text portions, or entire texts with regard to public use for language standardization purposes. At every stage it is also possible to index texts or their parts, or to make corrections to existing indexations on the spot. This option is also available dynamically when entering the corpus from the search module.

4. Problems and solutions

A cluster consisting of three interconnected databases has allowed Livonian to turn a lack of resources into an advantage. The general trend for large languages is that different institutions control and develop their own type of database – lexical and morphological databases, corpora, and tools for language acquisition. The benefits of sharing these resources remain untapped not only due to differing interests, but also often due to the incompatibility of these resources. The reasons for this are not always exclusively technical. At the same time, in the Livonian case, the lack of institutional resources has resulted in a solution which ultimately ensures the compatibility of various linguistic data, data consistency, avoiding data duplication, and provides high quality data processing. It also makes it possible to use various types of complementary data from a single resource for research as well as language acquisition, with the option of combining linguistic data with other types of data.

The Livonian experience shows manual work is inevitable that when developing linguistic tools for small linguistic communities, and only some processes can be fully entrusted to automated solutions, at least in their initial phases. For example, even automated text recognition would not be effective since most of the texts are handwritten or printed at a poor level of quality. Also, as only a small proportion of them are available electronically, automated indexing does not work because of a lack of clear and verified grammar rules, limited data, etc.

Thus, one of the main sources for improving efficiency can be found in maximizing the efficiency of all areas of manual work, supporting semi-automated solutions instead of fully automated approaches, which – due to insufficient or occasionally incorrect input data – may lead in the long run to completely undermining the entire effort by, for example, creating a large number of misinterpretations. Also, since there are significantly fewer linguistic sources for small languages anyway, the creation of fully automated solutions may also be questionable from the perspective of the effort necessary to create them versus the actual benefits gained from their creation.

Increasing productivity is one of the main approaches for compensating for insufficient personnel with an adequate level of relevant linguistic and language knowledge. A considerable lack of human resources affects not only the Livonians, but nearly any small language. In the case of Livonian, this has been addressed using two different approaches.

The first approach is to simplify work methods and technical solutions, bringing them

down to the level of simple, familiar everyday actions such as clicking, choosing from drop-down menus, dragging, etc., which also helps to limit possible mistakes.

The second approach addresses the overall principles of database performance and workflow, which are organized so that personnel only complete the tasks for which they are objectively qualified. This means that people with lesser skills only perform actions matching their skill level. For example, they transcribe texts from manuscripts following a set of normalization rules, but final normalization prior to adding the texts to the database is performed by more skilled scholars. This approach is also integrated into the corpus-indexing principles, where less-skilled personnel only index simple items of which they are completely certain (such items also happen to make up most of the texts to be indexed), leaving complicated cases for more skilled personnel. Ultimately, this saves time and effort for everyone involved.

Another means for increasing effectiveness is to ensure that when the system is being created, it is coded so as to allow many types of uses as well as dynamic and creative options for adapting the databases and their contents to serve different uses. The databases created for Livonian, for example, also allow one to simultaneously perform linguistic research on the language while dynamically setting the language standard, which is relevant for many insufficiently studied and standardized languages (e.g., adjusting morphological templates, suggesting better vocabulary, excluding poor quality texts from public view, etc.). In addition, language materials can also hold significant cultural value, so it is possible to keep them available as textual units for research and other uses unrelated to linguistics.

Incidentally, in the Livonian case something that has been important and seems elementary by current standards is that technical independence is built into the foundations of this system. This relates not only to being able to access all functions online, which makes it possible to work with the data in the database and expand the database regardless of one's actual physical location, but also that it functions independent of any operating system, browser, or other programs. Likewise, the user does not need to have language support (fonts, keyboard drivers) for Livonian or any other language used in the system, which in the past had turned out to be a significant barrier to, for example, Livonian language acquisition.

Surprisingly, one of the most important factors in developing and using the language database cluster for Livonian has turned out to be that it is left unfinalized. In most cases, the content of databases is usually completely prepared and finalized before making it available for further use. However, such finalization of content tends to be quite complicated, due to a lack of sufficient people or time to perform the necessary work, though mostly due to inadequate knowledge of clear rules and relevant studies. With regard to a mandatory requirement that corpora content be finalized, in many cases this leads to "forced indexation", which is a significant source of misinterpretations and leads to additional work later on involving the elimination of

incorrect indexations. Moreover, waiting for completion and finalization of content – e.g., lemma articles, morphological standards, etc. – may limit or significantly postpone its use for research or language acquisition.

In the Livonian case, this is addressed by making all content available immediately, e.g., texts are fully searchable right after they are uploaded and there is no requirement for them to be indexed at all. During indexation it is also possible to index the entire text, index it partially, mark it as questionable, or add different interpretations. At the same time, all actions (indexation, adding lemmas, etc.) can be performed at any stage of working with the databases, even while researching some other subject (though indicators are used for marking completed workflows).

This means that all resources are fully usable, each to a certain extent depending on readiness, of course, and at the same time unclear cases can be left unclear until they can be resolved at a future point or indexed purely as an interpretation, leaving them for final attention at a later time.

Leaving the database unfinalized also prevents the work from stalling – for example, if it is not possible to precisely define a word or place it in a specific morphological category. This makes it possible to work with other, achievable tasks, as there is always much to be done when it comes to working with small languages.

This also allows for the application of the open-contribution principle, where every researcher using these databases is able to contribute little by little in the areas on which they are working within a particular study, by adding indexations or corrections, resolving unclear cases, contributing translations, etc.

The combination of all of these efforts makes it possible to extract the maximum amount of data from limited sources with minimal effort. In a sense, it is reminiscent of Livonian Rabbit Soup, which has nothing to do with rabbits and is made as an extra dish by simply not throwing out the water left over from boiling potatoes for dinner.

5. Future plans

Though initially this system was created as a Livonian language data archive and a tool for language research, standardization, and acquisition, its principles can also be adjusted to suit other types of studies by supplementing it with other digital archives (containing images, audio recordings, video, 3D scans, data from other databases) as well as other information. In this way, the synergy and coordination among various archives can create a rich, high-quality tool suitable for multi-faceted studies in other fields or for interdisciplinary research, for the effective use of data and research results for the preservation, maintenance, and development of any low-resource language and cultural community existing in circumstances characterized by limitations on data, personnel, financing, and other resources.

One of the projects the UL Livonian Institute will undertake in the near future that will further develop this platform is the creation of a Livonian place name database. This database will link Livonian place names found in the corpus texts with their corresponding Latvian place name cartographic and geospatial data. This will be followed by linking the existing databases geographically with their sources, using existing metadata in the corpus and lexical database relating to the language informants and the data-recording location. This will make it possible to have a completely new perspective on the use of Livonian vocabulary and Livonian dialects.

Taking all this into account, it becomes clear that the opportunities and technologies offered by the modern electronic world, when used wisely, can be a positive support for the preservation and development of all low-resource languages; and they are already helping to close the gap in resources between large and small languages.

6. Acknowledgements

This study was supported by the Latvian Ministry of Education and Science research program “Latvian language” sub-project “Livonian Language” (VPP-IZM-2018/2-0002).

7. References

- Blumberga, R. (2006). *Lībieši dokumentos un vēstulēs*. Rīga: Latvijas vēstures institūta apgāds.
- Bušs, O. (2015). Sōnaraamat vōi/ja eksperiment. *Keel ja kirjandus*, 10, pp. 744–746. cl.ut.ee. *Tasakaalus korpus*. Accessed at: <https://www.cl.ut.ee/korpused/grammatikakorpus/> (10 June 2019)
- csb.gov.lv. *Centrālās statistikas pārvaldes datubāzes*. Accessed at: https://data1.csb.gov.lv/pxweb/lv/iedz/iedz__tautassk__taut__tsk2011/TSG11-06.px/table/tableViewLayout1/ (10 June 2019)
- EKPS: *Eesti keele põhisõnavara sõnastik*. (2014). Tallinn: Eesti Keele Sihtasutus.
- EKSS: *Eesti kirjakeele sagedussõnastik*. (2002). Tartu: Tartu Ülikooli Kirjastus.
- ELD: *Eesti-läti sõnaraamat. Igaunu-latviešu vārdnīca*. (2015). Tallinn: Eesti Keele Sihtasutus.
- Ernštreits, V. (2011). *Lībiešu rakstu valoda*. Rīga: Latviešu Valodas aģentūra, Līvõ kultūr sidām.
- Ernštreits, V. (2012). Lībiešu valodas situācijas attīstība Latvijā. In I. Druviete (ed.) *Valodas situācija Latvijā: 2004-2010*. Rīga: Latviešu Valodas aģentūra, pp. 142–166.
- Ernštreits, V. (2019). Electronical resources for Livonian. In *Proceedings of the Fifth International Workshop on Computational Linguistics for Uralic Languages. Tartu: ACL SIGUR, the special interest group for Uralic Languages*. Accessed at: <https://www.aclweb.org/anthology/W19-03> (30 July 2019)
- LED: *Läti-eesti sõnaraamat. Latviešu-igauņu vārdnīca*. (2015). Tallinn: Eesti Keele

Sihtasutus.

LELD: *Līvõkīel-ēstikīel-leṭkīel sōnārōntōz. Liivi-eesti-lāti sōnaraamat. Lībiešu-igauņu-latviešu vārdnīca.* (2012). Tartu, Rīga: Tartu Ülikool, Latviešu valodas aģentūra.

LLLS: *Līvõkīel-leṭkīel-līvõkīel sōnārōntōz. Lībiešu-latviešu-lībiešu vārdnīca.* (1999). Rīga: Līvõ kultūr sidām.

LW: *Livisches Wörterbuch mit grammatischer Einleitung.* (1938). Helsinki: Suomalais-Ugrilainen Seura.

lingua.livones.net. Accessed at: <http://lingua.livones.net/lv/module/login> (14 June 2019)

livones.net. *Lībiešu valodas vārdnīca.* Accessed at: <http://www.livones.net/lili/lv/vardnica/> (10 June 2019)

murre.ut.ee. *Līvõkīel-ēstikīel-leṭkīel sōnārōntōz. Liivi-eesti-lāti sōnaraamat. Lībiešu-igauņu-latviešu vārdnīca.* Accessed at: <http://www.murre.ut.ee/liivi/> (10 June 2019)

SW: *Joh. Andreas Sjögren's Livisch-deutsches und deutsch-livisches Wörterbuch.* (1861). St. Petersburg: Kaiserlichen Akademie der Wissenschaften.

unesco.org. *UNESCO Atlas of the World's Languages in Danger.* Accessed at: <http://www.unesco.org/languages-atlas/> (13 June 2019)

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



Ontological Knowledge Enhancement in EcoLexicon

Juan Carlos Gil-Berrozpe, Pilar León-Araúz, Pamela Faber

University of Granada

Department of Translation and Interpreting, Buensuceso 11, 18071 Granada, Spain

E-mail: jcgilberrozpe@ugr.es, pleon@ugr.es, pfaber@ugr.es

Abstract

Contemporary research has focused on how concepts are represented and organized in the mind, leading to neurocognitive theories such as grounded cognition or embodied cognition. These theories have greatly influenced further studies in linguistics and terminology. In this way, conceptualization, categorization, and knowledge organization are the foundation of cognitive-oriented terminology theories which highlight the relevance of situated knowledge structures, such as Frame-based Terminology. Accordingly, the practical application of Frame-based Terminology is EcoLexicon, a dynamic terminological knowledge base on environmental science. Concepts in this terminological resource are domain-specific within the Environmental Event, a model that interrelates concepts by assigning them different roles. However, the Environmental Event does not include specific category types to annotate these concepts ontologically. Therefore, this paper presents a process of ontological knowledge enhancement in EcoLexicon. This process was mainly based on the categorization of its concepts in semantic classes with a multidimensional approach. As a result, EcoLexicon was ontologically enhanced not only in terms of this categorization, but also through a redesign of the conceptual categories module, which involved modifying the existing category hierarchy and implementing new features focused on describing the combinatorial potential of concepts and categories (i.e. the conceptual combinations function and the ontological view).

Keywords: conceptual categories; conceptualization; categorization; ontology; environmental knowledge

1. Introduction

According to classic theories of cognition, knowledge representations are amodal data structures located in a semantic memory that is completely isolated from the modal systems of the brain for perception, action, and introspection (Barsalou, 2008). However, contemporary theories of cognition, including grounded cognition (Barsalou, 2010; Kiefer & Barsalou, 2013) and embodied cognition (Gallese & Lakoff, 2005; Martin, 2007; Meteyard et al., 2012), propose a more interrelated depiction of knowledge in our minds.

Grounded cognition considers that factors such as the environment, situations, the body and simulations are essential for contextualizing the cognitive representations in the brain's modal systems (Barsalou, 2010). Likewise, embodied cognition implies that the body is the main grounding mechanism and that all cognitive processes depend on perception and action (Meteyard et al., 2012). In line with this, concepts are embodied in the sense that their conceptual features are represented in sensory and motor brain areas based on experience (Kiefer & Pulvermüller, 2012). Not surprisingly, every discipline with a cognitive perspective pays attention to how concepts are represented

and organized in the mind (Mahon & Caramazza, 2009) or, in other words, to how conceptual information is categorized.

These grounded or embodied approaches to conceptualization are particularly relevant to the fields of linguistics and terminology because of the cognitive shift (Faber, 2009) in these disciplines over the last decade. This cognitive shift has specifically affected the study of terminology in relation to specialized knowledge representation, category organization and conceptual description. Not surprisingly, terminology is a discipline that combines linguistic and cognitive facets, since terms are linguistic elements which carry conceptual meaning within the framework of specialized knowledge texts (Faber, 2009). As such, lexicographic and terminological resources should draw on various aspects or details coming from psychological studies.

Accordingly, cognitive-based theories of terminology are also inspired in contemporary theories of cognition. Thus, they claim that specialized concepts are not activated in isolation, but are typically contextualized in background situations and events (Faber & San Martín, 2010). For instance, when perceiving an entity, people also perceive the space where it is located, including the agents, patients or events affecting it. Moreover, brain-imaging experiments have confirmed that simulations of potential actions are greatly involved in the conceptualization of entities and events, even including those which are mentioned in specialized language texts (Faber et al., 2014).

Because of the influence of cognition in terminology, it is necessary to develop or enhance the ontological information displayed in terminological resources so as to offer more accurate representations of concepts and their descriptions. This would lead to a more expressive formal ontology, which would not only benefit human users by facilitating knowledge representation and acquisition, but also non-human users by offering a higher degree of interoperability and usefulness. In most cases, this process starts by structuring the knowledge contained in the resource in a given manner, and this is the point where categorization plays a key role. In fact, classifying knowledge through categorization is inevitable, because any concept can be included in a set of hierarchically-organized categories (Murphy & Lassaline, 1997), which can range from general to specific levels.

In this context, this paper addresses a process of ontological knowledge enhancement in EcoLexicon¹, a terminological knowledge base on environmental science. This process was mainly based on the categorization of its concepts in semantic classes with a multidimensional approach. As a result, EcoLexicon was ontologically enhanced not only in terms of this categorization, but also through the redesign of the previous conceptual categories module, which involved modifying the category hierarchy and implementing new features (i.e. the conceptual combinations function and the ontological view).

¹ <http://ecolexicon.ugr.es/en/index.htm>

2. Conceptual categorization of environmental knowledge

Neurological characteristics such as conceptualization, categorization, and knowledge organization are the foundation of Frame-based Terminology (FBT), a cognitive-oriented terminology theory which highlights the relevance of situated knowledge structures represented as frames (Faber, 2015). FBT combines specialized knowledge representation with cognitive linguistics and semantics, taking aspects from both psychological and linguistic models. Frames are the cornerstone of FBT, and they are usually defined as the knowledge structures which contain information about the conceptual level and which relate entities and events associated with a particular scene or situation from human experience (Faber, 2015). Accordingly, any scientific or technical text contains specialized knowledge units that activate domain-specific semantic frames that are linked to the domain and to the user's background knowledge.

FBT has its main practical application in the form of a terminological resource: EcoLexicon (Faber et al., 2016). EcoLexicon is a dynamic terminological knowledge base on environmental science that provides a wide range of information about each of its entries, including conceptual, linguistic, phraseological, and multimodal aspects. EcoLexicon currently contains approximately 4,500 environmental concepts and 23,500 terms distributed in seven languages (English, Spanish, German, French, Dutch, Modern Greek, and Russian), with plans to include terms in Chinese, Portuguese, and Arabic. In addition, one of the most important functionalities in EcoLexicon is its general view (Figure 1), where conceptual networks are displayed and show how concepts are interrelated through different semantic relations (generic-specific, part-whole, and non-hierarchical relations).

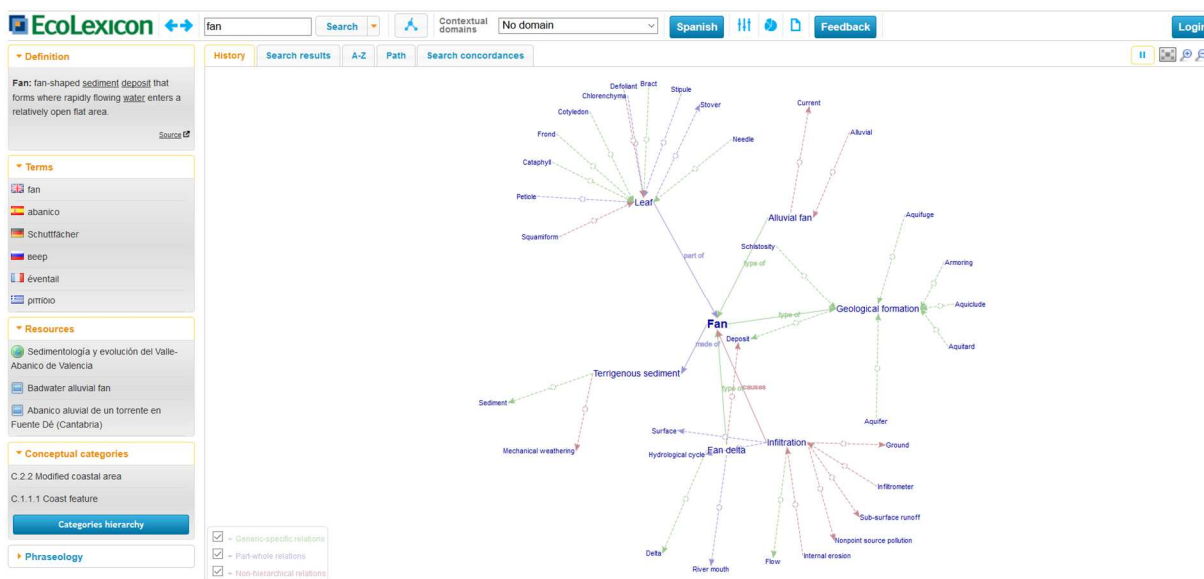


Figure 1: General view of EcoLexicon.

2.1 Environmental Event

According to FBT, conceptual networks are based on an underlying domain and on a closed inventory of both hierarchical and non-hierarchical semantic relations (Faber et al., 2009). These were the main premises used when building EcoLexicon, and the targets were conceptual relations and the combinatorial potential of concepts, extracted from corpus analysis.

In EcoLexicon, knowledge can be accessed from general to more specific relational structures. The most basic level is the Environmental Event (EE). In this frame, general categories of environmental entities are linked by predicates codifying the states, processes, and events in which the entities can take part (Faber, 2015). As stated by León-Araúz et al. (2012), the EE contains basic meanings that relate concepts, roles, and categories pertaining to general environmental knowledge. Moreover, the EE also links generic categories at the superordinate level and provides the basis for subframes that can be used to restrict contextual information to what is most relevant.

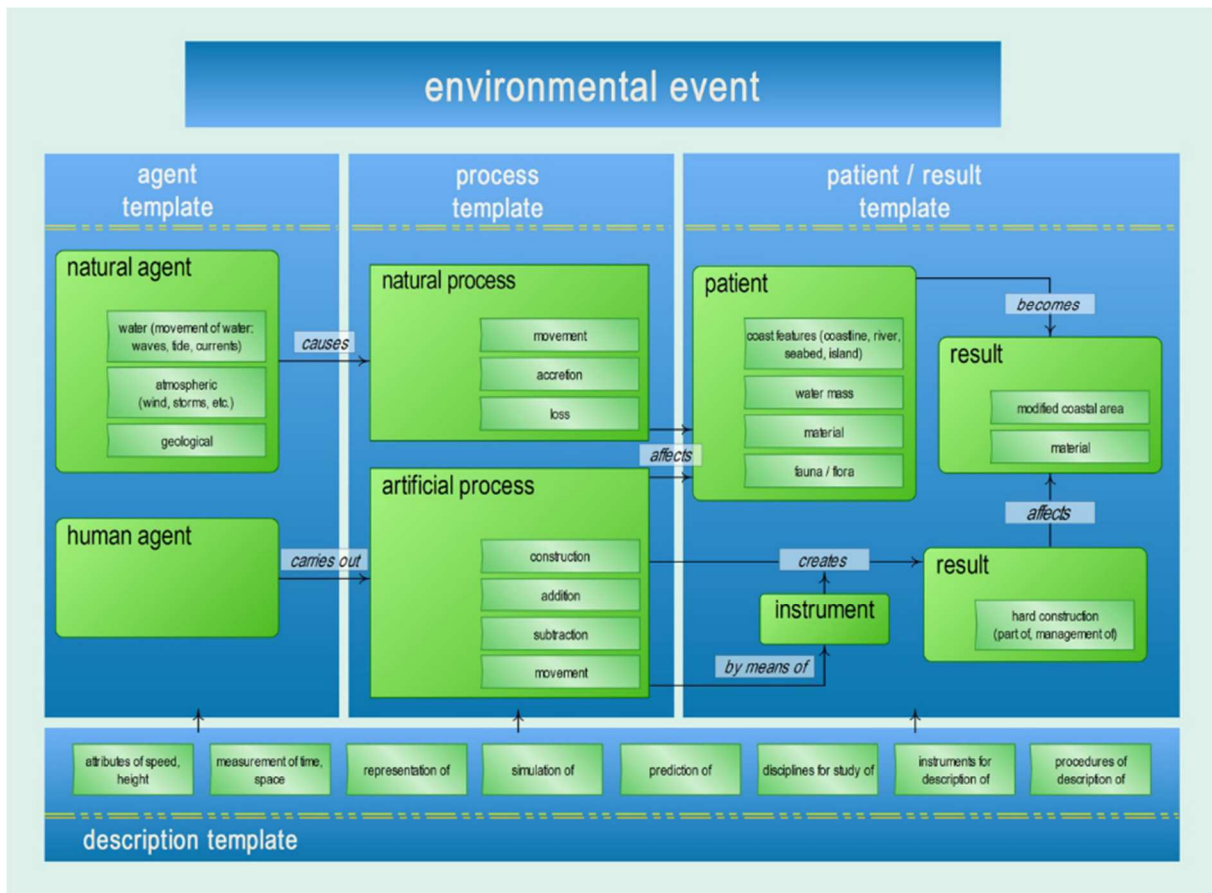


Figure 2: Environmental Event.

As shown in Figure 2, the Environmental Event has two types of AGENT that can initiate processes, i.e. NATURAL AGENTS (inanimate) and HUMAN AGENTS (animate).

On the one hand, natural forces (e.g. water movement) cause NATURAL PROCESSES (e.g. river erosion) in specific locations, commonly regarded as PATIENTS (e.g. riverbed) which, as a RESULT, may suffer alterations (e.g. deterioration, modification of size or shape). On the other hand, humans can also carry out ARTIFICIAL PROCESSES (e.g. construction) to alter the EFFECTS normally caused by natural processes (e.g. protection), or to create new effects through the use of certain INSTRUMENTS (e.g. defence structures).

Nevertheless, the conceptual representation of environmental knowledge cannot be achieved simply by assigning these generic semantic roles to concepts as if all of them would belong to a universal type of event (León-Araúz et al., 2012). In fact, contextualization has to be taken into account, because the way in which a concept interacts with other concepts can influence its categorization (Evans & Green, 2006). For this reason, the EE was originally used as a macrostructure for the further design of context-dependent microstructures (e.g. coastal engineering, meteorology, oceanography).

In recent years, the content of EcoLexicon has widely expanded, including a large quantity of conceptual and semantic information that has allowed us to interrelate all of its content, and thus go beyond the specific cases observed in the original EE. Because of this expansion in conceptual meaning, the need for an enhanced ontology of environmental categories has become apparent, since the EE does not include specific category types to annotate all environmental concepts ontologically, but only semantic roles. For this reason, we carried out an in-depth categorization process of all concepts in the database, a revision of the ontology underlying EcoLexicon, and the implementation of new features to its conceptual categories module, which will be explained in the following sections.

2.2 Conceptual categorization process

An ontology is usually regarded as a database describing the concepts of a knowledge field, their properties or characteristics, and how concepts are related to each other (Weigand, 1997). Moreover, ontologies are often organized as classification hierarchies and tend to be as universal as possible so that they can be used and reused for different applications. Such hierarchies tend to position the three most basic ontological categories at the top level: ENTITIES or OBJECTS, PROCESSES or EVENTS, and ATTRIBUTES or PROPERTIES (Mahesh & Nirenburg, 1995; Moreno-Ortiz & Pérez-Hernández, 2000).

In this context, various ontology-based projects for categorizing environmental knowledge have already been carried out, such as the Environmental Ontology²

² <http://www.obofoundry.org/ontology/envo.html>

(ENVO) (Buttigieg et al., 2013, 2016). More specifically, ENVO defines itself as “a community-led, open project which seeks to provide an ontology for specifying a wide range of environments relevant to multiple life science disciplines and, through an open participation model, to accommodate the terminological requirements of all those needing to annotate data using ontology classes” (Buttigieg et al., 2013). Although this project was initially focused on the representation of biomes, environmental features, and environmental materials, it has been continuously expanding to include ontological information related to a multitude of interrelated fields (Buttigieg et al., 2016).

In a similar way, the conceptual categorization process in EcoLexicon followed the premises behind ENVO’s ontological reasoning by adapting the conceptual categories and hierarchies to the specific needs of the environmental knowledge contained in EcoLexicon. Because of the dynamism of environmental sciences (León-Araúz et al., 2012), it was essential to take into account the multifaceted nature of concepts, as they can belong to more than one category depending on their salient features (Kageura, 1997). For this reason, the conceptual categorization process was carried out from a multidimensional perspective.

A series of semantic classes belonging to different top-down categorization levels was established to determine degrees of specificity (Murphy & Lassaline, 1997) and conceptual similarity (Hahn & Chater, 1997), so that every concept could be tagged with a category showing its interrelation with ontologically-similar elements. These semantic classes were mainly based on concept definitions and on the contextual information in the EcoLexicon corpus, but they were also contrasted with the ontological classes found in ENVO (Buttigieg et al., 2013, 2016). Consequently, an enhanced category system for EcoLexicon was established and hierarchically organized (Figure 3).

Process	Loss		
Process	Method		
Process	Movement		
Process	Movement	Earth / Soil movement	
Process	Movement	Energy movement	
Process	Movement	Fluid movement	
Process	Movement	Fluid movement	Water movement
Process	Movement	Transport	
Process	Movement	Wave	
Process	Movement	Wind movement	
Process	Phase		
Process	Phase	Phase of cycle	
Process	Phase	Phase of treatment	
Process	Phenomenon		
Process	Phenomenon	Atmospheric phenomenon	
Process	Phenomenon	Atmospheric phenomenon	Precipitation
Process	Phenomenon	Optical phenomenon	

Figure 3: Example of the category hierarchy.

In this way, the 4,500 concepts in EcoLexicon were classified in 152 categories, distributed in five categorization levels. To begin with, the most general level is composed of the three starter ontological categories (Mahesh & Nirenburg, 1995; Moreno-Ortiz & Pérez-Hernández, 2000):

A: ATTRIBUTE – properties of entities and processes

E: ENTITY – physical and mental objects

P: PROCESS – events extending over time and involving different participants

However, depending on the ontological nature of concepts, they can be subclassified in up to five levels of specificity, as can be seen in the category hierarchy involving CREATION concepts:

E: ENTITY

E-1: CREATION

E-1.1: ARTIFACT (e.g. *dc bus*)

E-1.1.1: CONDUIT (e.g. *duct*)

E-1.1.2: CONTAINER (e.g. *sedimentation tank*)

E-1.1.3: INSTRUMENT (e.g. *centrifugal pump*)

E-1.1.3.1: MEASURING INSTRUMENT (e.g. *accelerometer*)

E-1.1.3.2: RECORDING INSTRUMENT (e.g. *albedograph*)

E-1.1.3.3: SAMPLING INSTRUMENT (e.g. *automatic sampler*)

E-1.1.3.4: TRANSFORMING INSTRUMENT (e.g. *solar cell*)

E-1.1.4: VEHICLE (e.g. *dredger*)

E-1.2: SOFTWARE (e.g. *computer application*)

E-1.3: STRUCTURE (e.g. *pier*)

E-1.3.1: BUILDING (e.g. *oil refinery*)

E-1.3.2: DEFENSE STRUCTURE (e.g. *reef breakwater*)

Additionally, those concepts with a multidimensional nature (Kageura, 1997) were classified in as many categorization hierarchies as necessary, depending on the salient features observed in their definitions and in the corpus. For instance, one of the most multifaceted concepts is *port*, which was classified according to four categories:

- **Concept:** *port*
- **Definition (from EcoLexicon):** place along a river or seacoast that gives ships and boats protection from storms and rough water, and where ships can load and unload cargo. It can be natural or artificial.
- **Conceptual category:**
 - E-1.3: STRUCTURE
 - E-4.1: ARTIFICIAL GEOGRAPHIC FEATURE
 - E-4.2: NATURAL GEOGRAPHIC FEATURE
 - E-12.1.2: FACILITY

Figure 4 shows a fragment of the categorization table that was used to summarize the classification process. The first column contains the concept analyzed; the second column indicates whether the concept is multidimensional; the third column describes the number of categories applied to a single concept; and the remaining columns contain the top-down categories applied to each concept.

sheet pile	NO	FIRST CATEGORY:	Entity	Part	Part of structure
taiga	NO	FIRST CATEGORY:	Entity	Geographic feature	Natural geographic feature
stem	NO	FIRST CATEGORY:	Entity	Part	Part of lifeform
thallus	YES	FIRST CATEGORY:	Entity	Part	Part of lifeform
		SECOND CATEGORY:	Entity	Part	Part of lifeform
slope	NO	FIRST CATEGORY:	Entity	Part	Part of landform
continental slope	YES	FIRST CATEGORY:	Entity	Space	Layer
		SECOND CATEGORY:	Entity	Part	Part of landform
grain size	NO	FIRST CATEGORY:	Attribute	Physical attribute	Size
drum	YES	FIRST CATEGORY:	Entity	Creation	Structure
		SECOND CATEGORY:	Entity	Creation	Artifact
sieve	NO	FIRST CATEGORY:	Entity	Creation	Artifact
sieving	YES	FIRST CATEGORY:	Process	Elimination	
		SECOND CATEGORY:	Process	Phase	Phase of treatment
tank	YES	FIRST CATEGORY:	Entity	Creation	Structure
		SECOND CATEGORY:	Entity	Creation	Artifact
aeration tank	YES	FIRST CATEGORY:	Entity	Creation	Structure
		SECOND CATEGORY:	Entity	Creation	Artifact
calibration tank	YES	FIRST CATEGORY:	Entity	Creation	Structure
		SECOND CATEGORY:	Entity	Creation	Artifact

Figure 4: Example of the categorization table.

From an ontological point of view, 16 categories were associated with attributes, 93 with entities, and 43 with processes. (For a full list of the conceptual category hierarchy in EcoLexicon and some examples of each category, see Appendix A.)

3. Ontological perspective in EcoLexicon

The ontological enhancement process in EcoLexicon was mainly based on the categorization of its concepts in semantic classes with a multidimensional approach. As a result, not only was it possible to improve the structuration and organization of all the environmental knowledge it contained, but also to offer new practical applications and functionalities so that the end user could make the most of the ontological information. Essentially, the ontologically-enhanced functions that were implemented in EcoLexicon are the following: (i) the ontological view, an optional addition to the conceptual networks displayed in the general view; and (ii) a new conceptual categories module, including the revised category hierarchy and a conceptual combinations function.

3.1 Ontological view

The general view of EcoLexicon includes a series of elements that show all the information contained in the database in a user-friendly interface that facilitates access to the different types of data. The main information about each entry is broken down into five modules: (i) definition module, with a terminological definition based on the explication of the *genus* and the *differentiae*; (ii) term module, with the lexical denominations for a concept in the different languages available and linguistic information; (iii) resource module, with multimodal resources such as images, videos and hyperlinks; (iv) conceptual categories module, with the list of categories to which the concept belongs; (v) phraseology module, with the phraseological pattern and the collocational information about the concepts and terms. Furthermore, this terminological knowledge base also offers more functionalities, including the possibility of searching specific concordances in the EcoLexicon corpus and extracting statistics about the information in the database.

The most prominent feature of EcoLexicon is its dynamic visual display of conceptual networks, where concepts are surrounded by their multilingual denominations and related to each other through semantic relations. In EcoLexicon, three different types of semantic relations are distinguished: generic-specific relations (*type_of*), part-whole relations (*part_of*, *made_of*, *delimited_by*, *located_at*, *takes_place_in*, *phase_of*), and non-hierarchical relations (*affects*, *causes*, *attribute_of*, *opposite_of*, *studies*, *measures*, *represents*, *result_of*, *effected_by*, *has_function*).

In relation to the ontological enhancement process in EcoLexicon, this visual display of conceptual networks was improved through the implementation of an optional feature known as the ontological view (Figure 5). As a result of the conceptual categorization, each concept in EcoLexicon is tagged with one or more of the 152 categories, which allows for including this information so that the end user can observe the combinatorial potential of concepts according to their ontological nature.

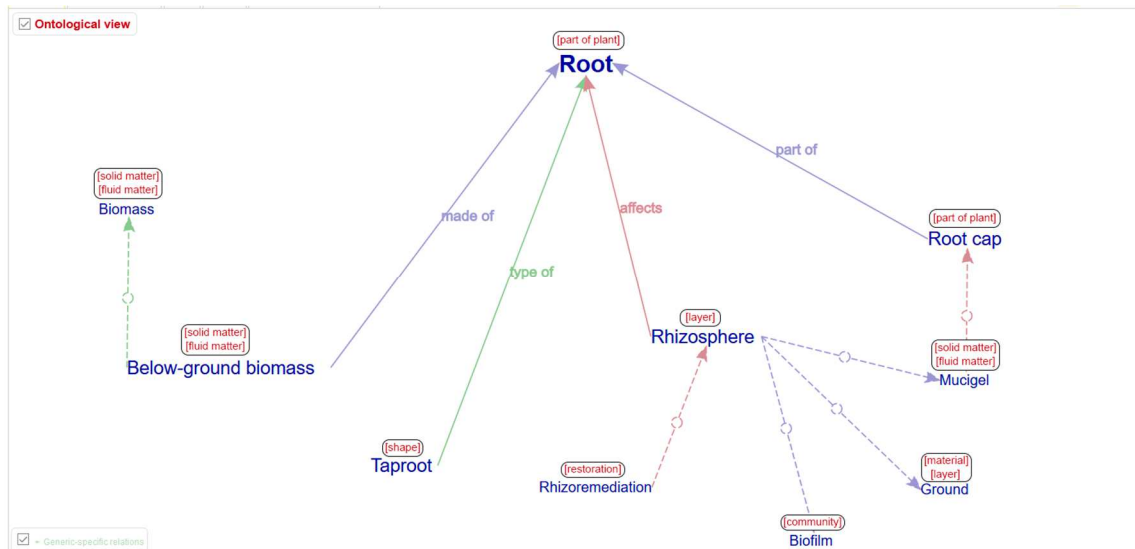


Figure 5: Ontological view (concept: *root*).

In Figure 6, the ontological view feature has been activated, so that a series of bubbles pop up over each concept (in blue) and indicate the conceptual categories to which each concept belongs (in red). Thanks to this functionality, there is a series of observations that can be made regarding the combinatorial potential of the chosen concept. For instance, it is interesting to confirm that *solar cell* (TRANSFORMING INSTRUMENT & PART OF INSTRUMENT) shares exactly the same categories with the other concepts to which it is related through a generic-specific relation: *amorphous cell* (TRANSFORMING INSTRUMENT & PART OF INSTRUMENT); *crystalline solar cell* (TRANSFORMING INSTRUMENT & PART OF INSTRUMENT); and *thin-film solar cell* (TRANSFORMING INSTRUMENT & PART OF INSTRUMENT). In the same way, since *solar cell* is categorized as a PART OF INSTRUMENT, its membership in larger conceptual categories is expressed through part-whole relations: *photovoltaic system* (TRANSFORMING INSTRUMENT & SYSTEM) and *solar panel* (TRANSFORMING INSTRUMENT). Finally, the concepts that are linked to *solar cell* through

non-hierarchical relations are indeed related to the nature of this concept as a TRANSFORMING INSTRUMENT: *energy* (ENERGY & MEASUREMENT) and *solar radiation* (ENERGY MOVEMENT).

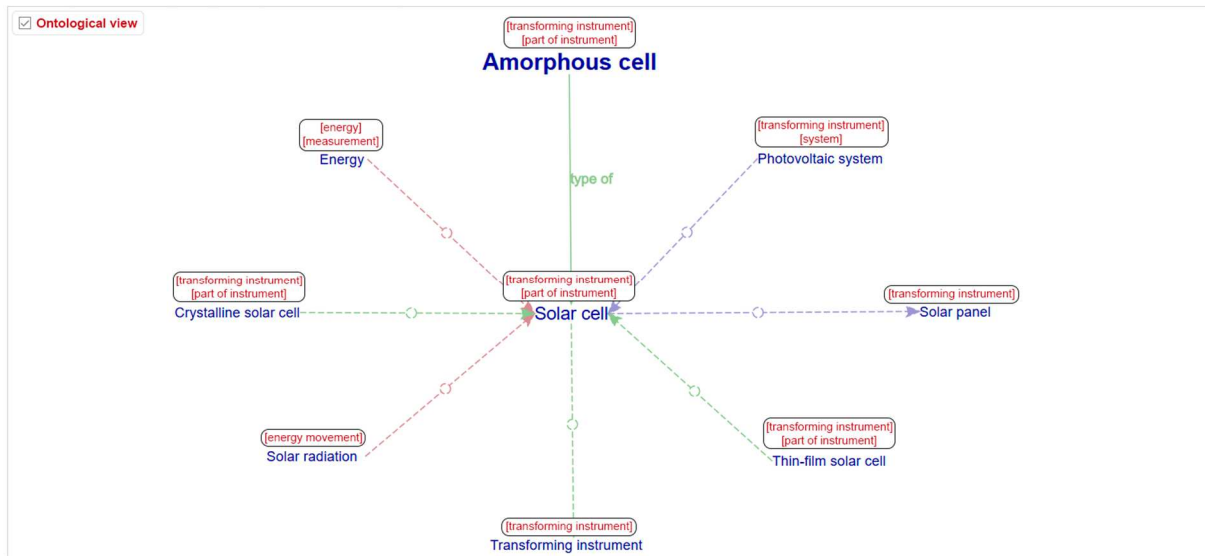


Figure 6: Ontological view (concept: *amorphous cell*).

3.2 Conceptual categories module

The original conceptual categories module in EcoLexicon only classified concepts according to the semantic roles designated in the Environmental Event (Faber, 2015; León-Araúz et al., 2012). For this reason, after performing the conceptual categorization process it was necessary to redesign this module. This involved two major changes: (i) the modification and update of the category hierarchy function; and (ii) the implementation of the conceptual combinations function. Figure 7 shows the conceptual categories module when selecting the concept *port*. Four conceptual categories (E-1.3: STRUCTURE, E-4.1: ARTIFICIAL GEOGRAPHIC FEATURE, E-4.2: NATURAL GEOGRAPHIC FEATURE, and E-12.1.2: FACILITY) are showcased, as well as the buttons for category hierarchy and conceptual combinations.



Figure 7: Conceptual categories module (concept: *port*).

3.2.1 Category hierarchy

The enhanced conceptual category hierarchy function of this new module contains a hierarchically-organized list of all 152 semantic classes (for a full list of the conceptual categories, see Appendix A). The members of each category can be accessed by clicking on the triangle to the left, enlarging the list to view the more specific subcategories (Figure 8). When a category is selected, a new window pops up with all the concepts belonging to it. This provides easy access to each entry, its information, and its ontologically-interrelated concepts in EcoLexicon (Figure 9). For example, in Figure 9 the concepts belonging to the DEFENSE STRUCTURE category are listed alphabetically, and clicking on any of them (e.g. *cofferdam*, *dike*) would lead EcoLexicon to its full entry with all the information.

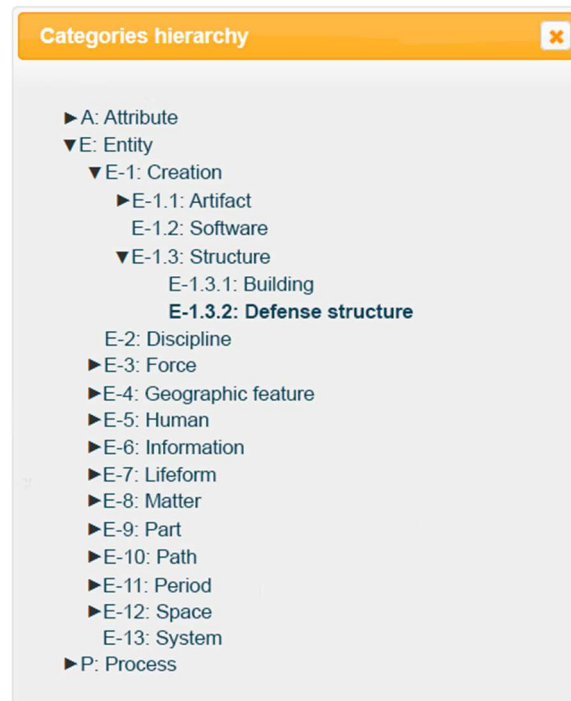


Figure 8: Category hierarchy function (category: DEFENSE STRUCTURE).

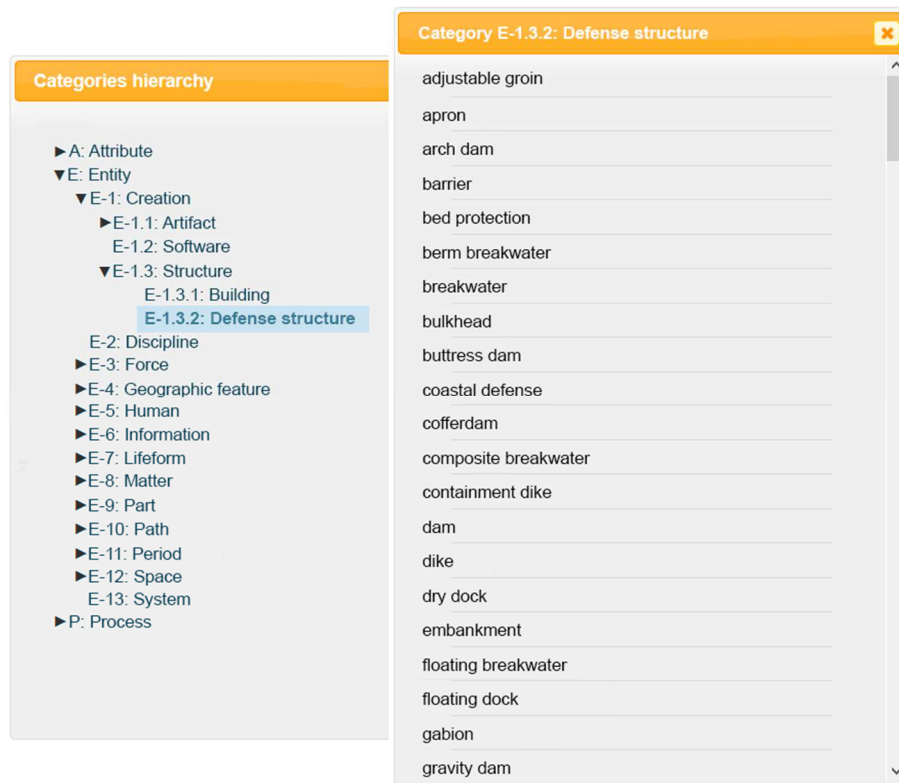


Figure 9: Category hierarchy function with examples (category: DEFENSE STRUCTURE).

3.2.2 Conceptual combinations

In the conceptual combinations function of the new conceptual categories module, users can perform a simple or advanced query. Figure 10 shows the query screen and the results screen of the simple query “hard structure”. The simple query box can be used to perform a proximity search, since it then autocompletes with the available concepts as the user writes different letters. As shown in the results screen, the system automatically converts the user’s search into a query expression (“hard structure [CONCEPT]”) and displays a list of results in EcoLexicon that shows the combinatorial potential of the queried concept with other concepts through specific semantic relations. These results are, by default, collected under conceptual propositions made of conceptual categories (in black) linked through semantic relations (in orange). For instance, the fourth result in Figure 10 is listed as “[Defense structure] *made of* [Material]”, but in order to see the specific concepts codified under those categories, it is necessary to click on the “+ Show specific results” option (in blue) next to this conceptual proposition, and thus the actual results of the query will appear: “HARD STRUCTURE *made of* CONCRETE”, “HARD STRUCTURE *made of* STEEL”, “HARD STRUCTURE *made of* QUARRY STONE”, etc.

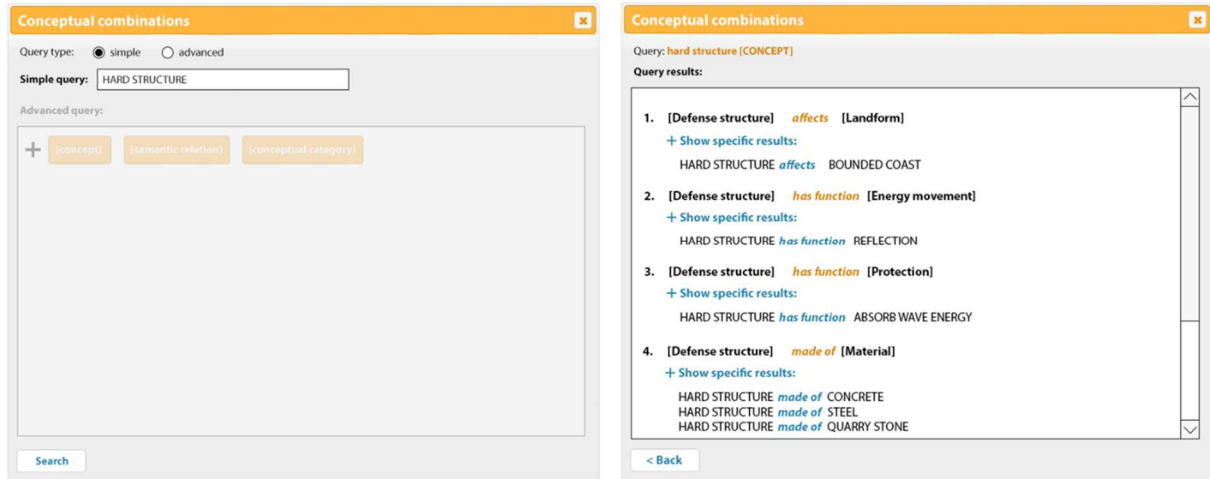


Figure 10: Simple query (left side) and results (right side) in the conceptual combinations function using the expression “hard structure [CONCEPT]”

On the other hand, the advanced query presents a series of particularities that allow users to perform more complicated searches. As shown in Figure 11, the advanced query is based on three elements: (i) concepts; (ii) semantic relations; (iii) conceptual categories. By clicking on the orange bubbles next to the “+” symbol, users can add as many elements to the query as they want in any order, since this query allows for free element combination (e.g. “category + relation”, “concept + relation + category”, “category + relation + category”, etc.). Similarly, any element can also be deleted. The concept bubble has a free text box to type anything, whilst the semantic relation and the conceptual category bubbles display a picklist showing all the relations or categories contained in EcoLexicon. However, it is also possible to choose the option “ANY” in the semantic relation and conceptual categories bubbles. In fact, displaying all the possibilities with a picklist is the simplest way for users to find and choose the most suitable option for their query. In addition, each bubble contains “AND” and “OR” buttons, which are useful if users want to look for more than one concept, relation and/or category found in the same position.

Figure 12 shows the query screen and the results screen of the advanced query “Water movement [CATEGORY] + any [SEMANTIC RELATION] + Natural water body [CATEGORY]”. In order to perform this search, users must select the option “advanced” next to “Query type”, and this will activate the advanced query box, where the user will then create a conceptual category bubble in order to select “Water movement”, a semantic relation bubble in order to select “ANY”, and a conceptual category bubble in order to select “Natural water body”. As a consequence, this expression displays a series of results that include conceptual propositions linking concepts belonging to the WATER MOVEMENT category and the NATURAL WATER BODY category through any semantic relation. For instance, the first case is the conceptual proposition “[Water movement] *affects* [Natural water body]”, including examples such as “FLOOD CURRENT *affects* BAY”, “TIDE *affects* TIDAL RIVER”, and “REGRESSION *affects* SEA”.

Figure 11: Advanced query in the conceptual combinations function

Query: Water movement [CATEGORY] + any [SEMANTIC RELATION] + Natural water body [CATEGORY]

Query results:

1. [Water movement] **affects** [Natural water body]
 + Show specific results:
 FLOOD CURRENT **affects** BAY
 FLOOD CURRENT **affects** ESTUARY
 OCEAN GYRE **affects** OCEAN
 FLOOD CURRENT **affects** ESTUARY
 TIDE **affects** TIDAL RIVER
 FLOOD CURRENT **affects** ESTUARY
 REGRESSION **affects** SEA
 SEICHE **affects** LAKE
2. [Water movement] **result of** [Natural water body]
 + Show specific results:
 INFLOW **result of** SPRING
 WAVE **result of** RIFFLE
3. [Water movement] **takes place in** [Natural water body]
 + Show specific results:

Figure 12: Advanced query (left side) and results (right side) in the conceptual combinations function using the expression “Water movement [CATEGORY] + any [SEMANTIC RELATION] + Natural water body [CATEGORY]”

4. Conclusion

Contemporary theories of cognition have greatly influenced the most recent approaches to linguistics and terminology. Since terms are linguistic units that convey conceptual information dependent on the context, they cannot be analyzed in isolation, but rather as part of a situated environment where different brain modal systems interact. In the specific case of the development of terminological resources, it is essential to focus on how concepts are represented and organized in the mind or, in other words, on how conceptual information is categorized.

In addition, the influence of cognition on terminology has led to an enhancement of the ontological information displayed in linguistic and terminological resources, since it is necessary to portray more accurate representations of concepts and their information. Accordingly, more expressive formal ontologies benefit both human and

non-human users by facilitating knowledge acquisition and offering a higher degree of interoperability, respectively. In this sense, EcoLexicon has experienced a process of ontological knowledge enhancement, mainly based on the categorization of its 4,500 concepts in 152 semantic categories. Thus, these top-down semantic categories distributed in up to five categorization levels were established to determine degrees of specificity and conceptual similarity, so that every concept could be tagged with a category showing its interrelation with other ontologically-related concepts.

As a result, not only it was possible to improve the structure and organization of the environmental knowledge contained in EcoLexicon, but also to offer new conceptual applications and functionalities, which benefitted from the ontological information that was implemented. Two new features derived from the conceptual categorization process were put in place: (i) the ontological view, an optional enhancement to the conceptual networks displayed in the general view that shows the combinatorial potential of concepts; and (ii) a revised conceptual categories module, including the modification and update of the category hierarchy function, and the inclusion of a new conceptual combinations function. This last feature is particularly useful for end users, since it allows them to perform simple and advanced queries regarding specific combinations of conceptual propositions (focusing on concepts, conceptual categories, and semantic relations).

In conclusion, this process of ontological enhancement in EcoLexicon will be useful not only for the improvements presented here in relation to the conceptual categories module, but also for the development of complementary features, such as the new phraseological module. More specifically, this last module would benefit from the integration of the category hierarchy into its functionalities, since it would make it possible to analyse phraseological units from an ontological approach.

Further research would require a series of users (experts and non-experts) to assess the main ontological features presented in this paper so as to validate their actual usefulness. Finally, since the future is based on interoperability among resources, it will be necessary to explore how the conceptual categorization can be implemented in the resources derived from EcoLexicon: the EcoLexicon corpus and EcoLexiCAT. Therefore, we plan to implement category annotation to enrich the EcoLexicon corpus, and ontological information derived from the conceptual categories module will be displayed in the EcoLexiCAT interface. Future work will also focus on how the ontological knowledge in EcoLexicon can be shared with external resources through Linked Data (León-Araúz et al., 2011a; León-Araúz et al., 2011b).

5. Acknowledgements

This research was carried out as part of project FFI2017-89127-P, Translation-Oriented Terminology Tools for Environmental Texts (TOTEM), funded by the Spanish Ministry of Economy and Competitiveness.

Funding was also provided by the FPU grant given by the Spanish Ministry of Education and Professional Training to the first author of the article (ref. FPU16/02194).

6. References

- Barsalou, L. W. (2008). Grounded Cognition. *Annual Review of Psychology*, 59(1), pp. 617–645.
- Barsalou, L. W. (2010). Grounded Cognition: Past, Present, and Future. *Topics in Cognitive Science*, 2(4), pp. 716–724.
- Buttigieg, P. L., Morrison, N., Smith, B., Mungall, C. J. & Lewis, S. E. (2013). The Environment Ontology: Contextualising Biological and Biomedical Entities. *Journal of Biomedical Semantics*, 4(43).
- Buttigieg, P. L., Pafilis, E., Lewis, S. E., Schildhauer, M. P., Walls, R. L. & Mungall, C.J. (2016). The Environment Ontology in 2016: Bridging Domains with Increased Scope, Semantic Density, and Interoperation. *Journal of Biomedical Semantics*, 7(57).
- Evans, V. & Green, M. (2006). *Cognitive Linguistics: An Introduction*. Edinburgh: Edinburgh University Press.
- Faber, P. (2015). Frames as a Framework for Terminology. In H. J. Kockaert & F. Steurs (eds.) *Handbook of Terminology*, 1. Amsterdam/Philadelphia: John Benjamins, pp. 14–33.
- Faber, P. & San Martín, A. (2010). Conceptual Modeling in Specialized Knowledge Resources. *Information Technologies and Knowledge*, 4(2), pp. 110–121.
- Faber, P., León-Araúz, P. & Prieto-Velasco, J. A. (2009). Semantic Relations, Dynamicity, and Terminological Knowledge Bases. *Current Issues in Language Studies*, 1(1), pp. 1–23.
- Faber, P., León-Araúz, P. & Reimerink, A. (2014). Representing Environmental Knowledge in EcoLexicon. In E. Bárcena, T. Read & J. Arús (eds.) *Languages for Specific Purposes in the Digital Era*, Educational Linguistics, 19. Cham: Springer, pp. 267–301.
- Faber, P., León-Araúz, P. & Reimerink, A. (2016). EcoLexicon: New Features and Challenges. In I. Kernerman, I. Kosem, S. Krek & L. Trap-Jensen (eds.) *Proceedings of GLOBALEX 2016: Lexicographic Resources for Human Language Technology in conjunction with the 10th edition of the Language Resources and Evaluation Conference*. Portorož: ELRA, pp. 73–80.
- Gallese, V. & Lakoff, G. (2005). The Brain's Concepts: The Role of the Sensory-Motor System in Conceptual Knowledge. *Cognitive Neuropsychology*, 22(3/4), pp. 455–479.
- Hahn, U. & Chater, N. (1997). Concepts and Similarity. In K. Lamberts & D. Shanks (eds.) *Knowledge, Concepts, and Categories*. Cambridge (MA)/London: MIT Press, pp. 93–131.
- Kageura, K. (1997). Multifaceted/Multidimensional Concept Systems. In S.E. Wright

- & G. Budin (eds.) *Handbook of Terminology Management: Basic Aspects of Terminology Management*. Amsterdam/Philadelphia: John Benjamins, pp. 119–132.
- Kiefer, M., & Barsalou, L.W. (2013). Grounding the Human Conceptual System in Perception, Action, and Internal States. In W. Prinz, M. Beisert & A. Herwig (eds.) *Action Science: Foundations of an Emerging Discipline*. Cambridge (MA)/London: MIT Press, pp. 381–407.
- Kiefer, M. & Pulvermüller, F. (2012). Conceptual Representations in Mind and Brain: Theoretical Developments, Current Evidence and Future Directions. *Cortex*, 48(7), pp. 805–825.
- León-Araúz, P., Faber, P. & Magaña-Redondo, P. J. (2011a). Linking Domain-Specific Knowledge to Encyclopedic Knowledge: An Initial Approach to Linked Data. In *Proceedings of the 2nd Workshop on the Multilingual Semantic Web (The 10th International Semantic Web Conference)*. Bonn, Germany, pp. 68–73.
- León-Araúz, P., Faber, P. & Montero-Martínez, S. (2012). Specialized Language Semantics. In P. Faber (ed.) *A Cognitive Linguistics View of Terminology and Specialized Language*. Berlin: De Gruyter Mouton, pp. 95–176.
- León-Araúz, P., Magaña-Redondo, P. J. & Faber, P. (2011b). Integrating Environment into the Linked Data Cloud. In W. Pillman, S. Schade & P. Smits (eds.) *Proceedings of the 25th International Conference Environmental Informatics (Enviroinfo Ispra 2011)*. Aachen: Shaker Verlag, pp. 370–379.
- Mahesh, K. & Nirenburg, S. (1995). A Situated Ontology for Practical NLP. In *Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing, International Joint Conference on Artificial Intelligence (IJCAI-1995)*. Montreal, Canada, pp. 1–10.
- Mahon, B.Z. & Caramazza, A. (2009). Concepts and Categories: A Cognitive Neuropsychological Perspective. *Annual Review of Psychology*, 60(1), pp. 27–51.
- Martin, A. (2007). The Representation of Object Concepts in the Brain. *Annual Review of Psychology*, 58, pp. 25–45.
- Meteyard, L., Rodríguez Cuadrado, S., Bahrami, B. & Vigliocco, G. (2012). Coming of Age: A Review of Embodiment and the Neuroscience of Semantics. *Cortex*, 48(7), pp. 788–804.
- Moreno-Ortiz, A. & Pérez-Hernández, C. (2000). Reusing the Mikrokosmos Ontology for Concept-based Multilingual Terminology Databases. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000)*. Athens, Greece, pp. 1061–1067.
- Murphy, G. L. & Lassaline, M. E. (1997). Hierarchical Structure in Concepts and Basic Level of Categorization. In K. Lamberts & D. Shanks (eds.) *Knowledge, Concepts, and Categories*. Cambridge (MA)/London: MIT Press, pp. 93–131.
- Weigand, H. (1997). Multilingual Ontology-Based Lexicon for News Filtering – The TREVI Project. In K. Mahesh (ed.) *Ontologies and Multilingual NLP: Proceedings of the 15th International Joint Conference on Artificial Intelligence (IJCAI)*. Nagoya, Japan, pp. 138–159.

Appendix A: Full conceptual category hierarchy in EcoLexicon

A: Attribute

- A-1: Ability [ex. AUTOTROPHIC, PERMEABILITY, TSUNAMIGENIC]
- A-2: Direction [ex. DOWNSTREAM, WINDWARD, ONSHORE]
- A-3: Location [ex. HADOPELAGIC, MESOTIDAL, SUBAQUEOUS]
- A-4: Measurement [ex. QUANTITY, SPECIFIC HEAT CAPACITY, NUTRIENT CONCENTRATION]
 - A-4.1: Magnitude [ex. ALTITUDE, RADICULAR ZONE DEPTH, AMBIENT TEMPERATURE]
 - A-4.1.1: Level [ex. MAXIMUM FLOW, HIGHEST ASTRONOMICAL TIDE, FREEZING POINT]
 - A-4.1.1.1: Mean [ex. MEAN FLOW, MEAN TIDE LEVEL, AVERAGE PRECIPITATION]
- A-5: Origin [ex. ARTIFICIAL, AEOLIAN, LITHOLOGIC]
- A-6: Physical attribute [ex. COLOR, SOIL TEXTURE, XERICITY]
 - A-6.1: Composition [ex. BIOCLASTIC, WOODY, MONOLITHIC]
 - A-6.2: Shape [ex. BACCIFORM, EUHEDRAL, HOOK-SHAPED]
 - A-6.3: Size [ex. BIG, SMALL, GRAIN SIZE]
 - A-6.4: State [ex. CARBONATE EQUILIBRIUM, SLOPE INSTABILITY, UNCONSOLIDATED]
 - A-6.4.1: Climate [ex. BIOCLIMATE, SAVANNA CLIMATE, PERIGLACIALISM]
- A-7: Time [ex. APERIODIC, SEMIDIURNAL, TEMPORARY]

E: Entity

- E-1: Creation [ex. WIND TURBINE GENERATOR SYSTEM, COLLECTOR, SEPTIC SYSTEM]
 - E-1.1: Artifact [ex. CULVERT, DC BUS, STATOSCOPE]
 - E-1.1.1: Conduit [ex. DRAINAGE DITCH, PIPELINE, DUCT]
 - E-1.1.2: Container [ex. CLOUD CHAMBER, SEDIMENTATION TANK, RETENTION BASIN]
 - E-1.1.3: Instrument [ex. CENTRIFUGAL PUMP, FISHING NET, WEATHER SATELLITE]
 - E-1.1.3.1: Measuring instrument [ex. ACCELEROMETER, BAROMETER, SOUNDING MACHINE]
 - E-1.1.3.2: Recording instrument [ex. ALBEDOGRAPH, MARIGRAPH, WATER-LEVEL RECORDER]
 - E-1.1.3.3: Sampling instrument [ex. COLLECTOR, AUTOMATIC SAMPLER, VAN DORN BOTTLE]
 - E-1.1.3.4: Transforming instrument [ex. UPWIND TURBINE, CONVERTER, SOLAR CELL]
 - E-1.1.4: Vehicle [ex. BOAT, DREDGER, ELECTRIC VEHICLE]
 - E-1.2: Software [ex. COMPUTER APPLICATION, CONTOUR GRIDDER, MODFLOW]
 - E-1.3: Structure [ex. SPILLWAY, PIER, ENGINEERING STRUCTURE]
 - E-1.3.1: Building [ex. GEOTHERMAL POWER PLANT, TIDE STATION, OIL REFINERY]
 - E-1.3.2: Defense structure [ex. REEF BREAKWATER, HIGH GROUYNE, RETAINING WALL]
- E-2: Discipline [ex. BIOCLIMATOLOGY, HUMAN ECOLOGY, PHYTOPATHOLOGY]
- E-3: Force [ex. TRACTIVE FORCE, TECTONIC FORCE, GRAVITY]
 - E-3.1: Dynamics [ex. ATMOSPHERIC DYNAMICS, SLOPE DYNAMICS, COASTAL DYNAMICS]
 - E-3.2: Energy [ex. ELECTRICITY, WIND ENERGY, SOLAR ENERGY]
 - E-3.3: Stress [ex. FRICTION, DYNAMIC PRESSURE, TENSION]
- E-4: Geographic feature [ex. ENTRY CHANNEL, AQUIFER, BIOME]
 - E-4.1: Artificial geographic feature [ex. GROUYNE BAY, QUARRY, PORT]
 - E-4.1.1: Artificial water body [ex. POOL, POND, RESERVOIR]
 - E-4.2: Natural geographic feature [ex. ABYSS, HIGH PLATEAU, BAY]
 - E-4.2.1: Landform [ex. FAN DELTA, RIVER GORGE, EMERGENT COAST]

- E-4.2.1.1: Natural water body [ex. SEA CHANNEL, KARST SPRING, LAGOON]
- E-4.2.2: Landscape [ex. TIDAL SHOAL, MONSOON FOREST, MANGROVE SWAMP]
- E-5: Human [ex. PORT AUTHORITY, HUMAN BEING, SOCIAL AGENT]
 - E-5.1: Institution [ex. METEOROLOGICAL SERVICE, CITY COUNCIL, PUBLIC INSTITUTION]
 - E-5.2: Specialist [ex. GEOGRAPHER, GEOLOGIST, OCEANOGRAPHER]
- E-6: Information [ex. PIECE OF DATA, CARTOGRAPHIC INFORMATION, HYDROLOGIC DATA]
 - E-6.1: Classification [ex. CLIMATE CLASSIFICATION, CLADE, URBAN HIERARCHY]
 - E-6.1.1: Scale [ex. BEAUFORT SCALE, STATE-OF-SEA SCALE, SPECTRUM]
 - E-6.2: Document [ex. PLAN, PROTOCOL, TIDE TABLE]
 - E-6.2.1: Law [ex. LEGISLATION, WILDLIFE LAW, PRINCIPLE OF ENVIRONMENTAL LAW]
 - E-6.3: Parameter [ex. STRUCTURAL CRITERION, QUALITY INDICATOR, K FACTOR]
 - E-6.4: Record [ex. BASELINE CARTOGRAPHY, ECHOGRAM, METEOROLOGICAL SERIES]
 - E-6.5: Representation [ex. GEODATABASE, AURORAL OVAL, SOIL PROFILE]
 - E-6.5.1: Graph [ex. ADIABATIC CHART, STRATIGRAPHIC COLUMN, COMPOUND HYDROGRAPH]
 - E-6.5.2: Line [ex. RATING CURVE, ISOHALINE, MERIDIAN]
 - E-6.5.3: Map [ex. NAUTICAL CHART, ORIENTATION MAP, ORTHOPHOTOMAP]
 - E-6.5.4: Mathematical expression [ex. COEFFICIENT, STANDARD DEVIATION, WAVE EQUATION]
 - E-6.5.5: Model [ex. EKMAN SPIRAL, EROSION MODEL, SIMULATION]
 - E-6.5.6: Picture [ex. PHOTOMOSAIC, SATELLITE IMAGE, ORTHOPHOTO]
 - E-6.5.7: Unit [ex. STERADIAN, FARADAY, MILLIMETER]
 - E-6.6: Theory [ex. PLATE TECTONICS, EQUILIBRIUM THEORY, STATIONARY WAVE THEORY]
- E-7: Lifeform [ex. DETRITIVORE, NATIVE SPECIES, ORGANISM]
 - E-7.1: Animal [ex. AMPHIBIAN, LIVESTOCK, CRUSTACEAN]
 - E-7.2: Community [ex. BENTHOS, BIOCENOSIS, BIOLOGICAL COMMUNITY]
 - E-7.2.1: Animal community [ex. STYGOFAUNA, COHORT, ZOOPLANKTON]
 - E-7.2.2: Plant community [ex. PHYTOBENTOS, FLORA, PHYTOPLANKTON]
 - E-7.3: Fungus [ex. BASIDIOMYCOTA, MYCOBIONT, FACULTATIVE PARASITE]
 - E-7.4: Microorganism [ex. BACTERIA, FACULTATIVE AEROBE, ENTERIC VIRUS]
 - E-7.5: Plant [ex. CHAMAEPHYTE, PHYCOBIONT, MANGROVE]
- E-8: Matter [ex. GREYBODY, ORGANIC MATERIAL, SUBSTANCE]
 - E-8.1: Chemical substance [ex. CARBONIC ACID, ARSENIC, NITROGEN DIOXIDE]
 - E-8.2: Fluid matter [ex. TAR, LAVA FLOW, MUD]
 - E-8.2.1: Fluid astronomical body [ex. HEAVENLY BODY, STAR, SUN]
 - E-8.2.2: Gas [ex. POLAR AIR, EXHAUST GAS, SMOG]
 - E-8.2.3: Water [ex. RUNOFF WATER, DRINKING WATER, RAINWATER]
 - E-8.2.3.1: Cloud [ex. ALTOSTRATUS, STRATOCUMULUS, FRONTAL FOG]
 - E-8.3: Particle [ex. VOLCANIC ASH, INTERLEUKIN, ULTRAFINE PARTICLE]
 - E-8.4: Solid matter [ex. SOLID FUEL, SOLID WASTE, SOLUTE]
 - E-8.4.1: Deposit [ex. ALLUVIUM, SEDIMENT FLOW, AEOLIAN DEPOSIT]
 - E-8.4.2: Material [ex. CEMENT, REINFORCED CONCRETE, SEMICONDUCTOR]
 - E-8.4.2.1: Mineral [ex. ANTHRACITE, COARSE SAND, ZEOLITE]
 - E-8.4.2.2: Rock [ex. LIMESTONE, QUARTZ DIORITE, CLASTIC SEDIMENTARY ROCK]
 - E-8.4.2.3: Soil [ex. LEPTOSOL, MOLLISOL, SATURATED SOIL]

- E-8.4.3: Snow/ice [ex. AVALANCHE, SNOWFLAKE, ANCHOR ICE]
- E-8.4.4: Solid astronomical body [ex. ASTEROID, PLANET, SATELLITE]
- E-9: Part [ex. DISCARDS, SECTION, STATOR]
 - E-9.1: Part of instrument [ex. ANEMOMETER MAST, WIND TURBINE ROTOR, FLAP]
 - E-9.2: Part of landform [ex. BEACH HEAD, BERM CREST, SOIL PROPERTIES]
 - E-9.3: Part of lifeform [ex. ALLELE, CELL WALL, TISSUE]
 - E-9.3.1: Part of animal [ex. EOSINOPHIL, OTOLITH, VALVE]
 - E-9.3.2: Part of fungus [ex. ASCOSPORE, SPOROCARP, PARAPLECTENCHYMA]
 - E-9.3.3: Part of plant [ex. BRACTEOLE, CHLOROPLAST, DEHISCENT FRUIT]
 - E-9.4: Part of structure [ex. HARBOUR MOUTH, SPILLWAY CREST, GROUYNE HEAD]
 - E-9.5: Part of vehicle [ex. GUNWALE, HULL, KEEL]
 - E-9.6: Part of water body [ex. DOWNSTREAM, APHYTAL ZONE, SEA FLOOR]
- E-10: Path [ex. ROAD, GULLY, VIADUCT]
 - E-10.1: Imaginary path [ex. PLANETARY ORBIT, ECLIPTIC PLANE, EARTH'S ELLIPTIC ORBIT]
- E-11: Period [ex. LUNAR DAY, AUTUMN, USEFUL LIFE]
 - E-11.1: Era [ex. DEVONIAN, MESOZOIC ERA, PLEISTOCENE EPOCH]
- E-12: Space [ex. CAPILLARY INTERSTICE, MEDIUM, ECOLOGICAL NICHE]
 - E-12.1: Area [ex. SEDIMENTARY ENVIRONMENT, PROTECTED AREA, ECOREGION]
 - E-12.1.1: Administrative area [ex. CITY, MUNICIPAL BOUNDARY, THE UNITED STATES OF AMERICA]
 - E-12.1.2: Facility [ex. BIOMASS POWER PLANT, MEASURING STATION, GAUGING SITE]
 - E-12.1.3: Land [ex. BASIN SLOPE, MEADOW, AREA OF LAND]
 - E-12.2: Layer [ex. ATMOSPHERE, PLANETARY BOUNDARY LAYER, LOWER MANTLE]
 - E-12.3: Limit [ex. WAVE CREST, LIMIT OF UPRUSH, AMPHIDROMIC POINT]
 - E-12.4: Position [ex. BIFURCATION, DEPOCENTER, PERIGEE]
- E-13: System [ex. DETRITUS FOOD CHAIN, NETWORK, ISOLATED SYSTEM]
- P: Process
 - P-1: Action [ex. BIOLOGICAL ACTION, SPAWNING, ENVIRONMENTAL CRIME]
 - P-1.1: Analysis [ex. SEDIMENTOLOGICAL ANALYSIS, ENVIRONMENTAL IMPACT ASSESSMENT, WEATHER FORECAST]
 - P-1.2: Chemical reaction [ex. COMBUSTION, ANABOLISM, DEFLAGRATION]
 - P-1.3: Collection [ex. ENERGY STORAGE, SOIL WATER RETENTION, SAND TRAPPING]
 - P-1.4: Interaction [ex. INTERSPECIFIC COMPETITION, AIR-SEA INTERACTION, ENDOGENIC GEOLOGICAL PROCESS]
 - P-1.5: Management [ex. COASTAL MANAGEMENT, SUSTAINABLE WATER USE, WASTE MANAGEMENT]
 - P-1.6: Measurement [ex. STREAM GAUGING, DENSITOMETRY, STOCHASTIC PROCESS]
 - P-1.7: Protection [ex. ABSORB WAVE ENERGY, SOIL CONSERVATION, FLOOD PREVENTION]
 - P-2: Activity [ex. SUBSISTENCE AGRICULTURE, SHIFTING CULTIVATION, FACTORY FARMING]
 - P-3: Addition [ex. TECTONIC ACCRETION, ARTIFICIAL NOURISHMENT, PHOSPHATE FERTILIZATION]
 - P-4: Change [ex. CLIMATE CHANGE, ECOLOGICAL DEGRADATION, ENVIRONMENTAL IMPACT]
 - P-4.1: Change in size/intensity [ex. TIDE ACCELERATION, CYCLOGENESIS, ANTICYCLOLYSIS]
 - P-4.1.1: Decrease [ex. RETARD LITTORAL DRIFT, WAVE SETDOWN, REDUCTION IN LONGSHORE TRANSPORT]

- P-4.1.2: Increase [ex. SEA LEVEL RISE, ALGAL BLOOM, RISE OF THE WATER TABLE]
- P-4.2: Change of direction [ex. DEFLECTION, DENSITY STRATIFICATION, SECULAR VARIATION]
- P-4.3: Change of state [ex. CONDENSATION, SOIL LIQUEFACTION, SOLIDIFICATION]
- P-4.4: Disease [ex. BRONCHITIS, YELLOW BAND DISEASE, MONILIA DISEASE]
- P-4.5: Division [ex. CLEAVAGE, DISPERSION, BREAKING DROPS]
- P-4.6: Transformation [ex. ACIDIFICATION, METAMORPHISM, TERRITORIAL TRANSFORMATION]
 - P-4.6.1: Pollution [ex. ATMOSPHERIC POLLUTION, OZONE POLLUTION, OCEAN DUMPING]
 - P-4.6.2: Restoration [ex. BIOREMEDIATION, ENVIRONMENTAL RECOVERY, REVEGETATION]
- P-5: Cycle [ex. TIDAL CYCLE, CARBON CYCLE, HYDROLOGIC CYCLE]
- P-6: Elimination [ex. DEFORESTATION, MASS EXTINCTION, ELIMINATION OF SOLID WASTE]
- P-7: Emission [ex. PARTICULATE EMISSION, HYDROMAGMATIC ERUPTION, EVAPOTRANSPIRATION]
- P-8: Formation [ex. BRECCIA FORMATION, ATMOSPHERIC IONIZATION, PRIMARY PRODUCTION]
- P-9: LOSS [ex. COASTAL DEGRADATION, INTERNAL EROSION, MECHANICAL WEATHERING]
- P-10: Method [ex. AIR LAYERING, HODOGRAPH METHOD, POLYCULTURE]
- P-11: Movement [ex. DRIFT, OSMOSIS, TRAFFIC]
 - P-11.1: Earth/soil movement [ex. CONTINENTAL DRIFT, SLOPE MOVEMENT, TECTONIC EARTHQUAKE]
 - P-11.2: Energy movement [ex. FORCED CONVECTION, ATMOSPHERIC RADIATION, CLOUD ELECTRIFICATION]
 - P-11.3: Fluid movement [ex. CAPILLARITY, LAMINAR FLOW, MAGMA INTRUSION]
 - P-11.3.1: Water movement [ex. COASTAL CIRCULATION, DRIFT CURRENT, GRAVITY FLOW]
 - P-11.4: Transport [ex. TRANSFER, LONGSHORE TRANSPORT, UPWELL]
 - P-11.5: Wave [ex. REGULAR WAVE, ATMOSPHERIC WAVE, PROGRESSIVE WAVE]
 - P-11.6: Wind movement [ex. SEA BREEZE, ANTICYCLONIC CIRCULATION, WARM FRONT]
- P-12: Phase [ex. KARYOKINESIS, CYTOKINESIS, PRELIMINARY TREATMENT]
 - P-12.1: Phase of cycle [ex. TIDAL STAGE, LITHOGENESIS, OROGENY]
 - P-12.2: Phase of treatment [ex. PRIMARY SEDIMENTATION, THERMOPHILIC DIGESTION, PREAERATION]
- P-13: Phenomenon [ex. LUNAR ECLIPSE, EXTREME EVENT, ENVIRONMENTAL NOISE]
 - P-13.1: Atmospheric phenomenon [ex. SQUALL, ADVECTIVE THUNDERSTORM, TROPICAL CYCLONE]
 - P-13.1.1: Precipitation [ex. HYDROMETEOR, FREEZING RAIN, CONVECTIVE PRECIPITATION]
 - P-13.2: Optical phenomenon [ex. RAINBOW, AUROREAL STORM, LIGHTNING FLASH]

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



Smart Lexicography for Low-Resource Languages: Lessons Learned from Buddhist Sanskrit and Classical Tibetan

Ligeia Lugli

SOAS University of London, Thornhaugh Street, London WC1H 0XG, room 339
E-mail: ll34@soas.ac.uk

Abstract

Traditional lexicography requires titanic efforts and enormous resources. For many languages, such resources have never been available. As a result, they have received only limited lexicographic coverage. Today, these languages can take advantage of many of the same digital tools and strategies that have simplified and expedited dictionary-making for mainstream languages. However, the resource gap remains evident even in the digital era, with basic corpus processing tasks that lie at the foundation of contemporary ‘smart lexicography’ still constituting a challenge for many under-resourced languages.

Drawing on my own experience in Sanskrit and Tibetan lexicography, this paper aims to offer some guidance as to the advantages and limitations of the application of smart lexicography to under-resourced languages. In particular, this paper suggests that in order to optimize resources, it may be advisable to prioritize high-quality lexical annotation of the corpus over highly curated dictionary entries, and to let digital tools take care of the lexicographic representation of the annotated linguistic information.

Keywords: automated lexicography; GDEX; Buddhist Hybrid Sanskrit; Tibetan

1. Introduction

This paper serves two purposes. On the one hand it provides a progress report of two ongoing lexicographic projects, (1) a Buddhist Sanskrit lexical resource called The Buddhist Translators Workbench commissioned by the Mangalam Research Center (Berkeley, CA), and (2) a diachronic valency lexicon of Tibetan verbs, which is being developed at SOAS (University of London) within the AHRC-funded project Lexicography in Motion. On the other hand, this paper outlines strategies for applying smart lexicography to low-resource languages.

Smart lexicography is intended here as an optimally efficient cooperation between human lexicographers and machines, whereby all task that can be automated are delegated to computers, while lexicographers focus on points of curation that require human judgement. This includes re-using pre-existing dictionary content and ensuring that any new human-curated output can in turn be re-used by other projects or in subsequent iterations within the same project.

What constitutes a ‘low-resource language’ is more difficult to define. Low is a fundamentally relative concept, as it acquires meaning only relative to its antonym ‘high’. Languages can be considered low-resource only when compared with high-resource languages, like English or other major spoken languages that tend attract much study, funding and technological development. In this paper, I use the expression ‘low-resource languages’ to indicate those for which computational and human resources are insufficient to take full advantage of state-of-the-art automated or semi-automated lexicographic workflows.

Many reasons may limit the ability to apply automation to lexicographic tasks. For the projects discussed here, one crucial obstacle has been the difficulty of producing suitably annotated corpora quickly. Sadly, Rundell and Kilgariff’s (2011) assertion that “the timescale for creating a large lexicographic corpus has been reduced from years to weeks, and for a small corpus in a specialized domain, from months to minutes” does not apply to the languages considered here. The main problem for these languages has been generating sufficient manually annotated data to develop reliable NLP pipelines for corpus pre-processing. Few people have the adequate skills to create the amount of annotated data necessary to train Machine Learning-based models, or even to test rule-based systems. Moreover, these people are usually highly skilled, not easily amenable to the dull routine of corpus annotation and required for more sophisticated lexicographic tasks.¹

Fortunately, the unavailability of large amounts of training data needs not entirely preclude the application of automation to the lexicography of low-resource languages. It does however impose significant limitations on the scope of such application and the results that can be achieved through it.

A key to the adoption of smart lexicography for low-resource languages lies in the re-conceptualization of the dictionary product and of its core design principles. Good lexicographic practice dictates that entries are designed primarily to meet the needs of the dictionary prospective audience, or ‘market’ (Atkins & Rundell, 2008, Ch. 2; Landau, 2001: 343). While this is undoubtedly a commendable approach, when working with low resource languages much is to be gained if the needs of the lexicographic team take primacy over those of the audience. As this paper will show, ambitious microstructures designed to fulfil audience needs may slow down the progress of small teams working on low-resource languages to unsustainable levels. By contrast, investing the lexicographers’ linguistic expertise to create annotated data for use in the future can lead to faster and more rewarding results. This is because annotated data is inherently versatile. It can be immediately displayed to users in the form of a lexical

¹ This is critical issue for historical languages like Buddhist Sanskrit and Classical Tibetan, for which no active speakers are available. Contemporary low-resource languages may pose different challenges; cf. Nasiruddin 2013 who sees Machine Learning as promising for under-resourced languages for which crowd-sourcing solutions are available.

database or minimally curated ‘proto-dictionary’, it serves to develop NLP pipelines and can be later re-used to create full-fledged dictionaries (cf. Pajsz, 2009; Atkins & Rundell, 2011; Mianáin & Convery, 2014). This strategy fits the definition of smart lexicography given above insofar as it constitutes an optimally efficient cooperation between lexicographers and computers, given the available human and digital resources.

This is the general strategy we have adopted, to varying degrees and with different practical solutions, in the two projects discussed in this paper.

2. The Buddhist Translators Workbench

2.1 Project overview

The project was commissioned by the Mangalam Research Center in 2012, with an eye to providing translators with useful lexical information about key Sanskrit Buddhist vocabulary. The primary aim of the project was to help translators achieve a nuanced understanding of selected Buddhist vocabulary and, ideally, move away from the overly terminological renditions and calques that often characterize English translations of Buddhist Sanskrit Texts (Griffiths, 1981). Two features were deemed essential to achieve this goal.

First, the dictionary would have to be corpus-driven. Semantic descriptions and lexico-semantic relations should be derived from the corpus rather than from traditional interpretation. This decision was at odds with the perceived needs of a sizeable portion of our intended audience, which was primarily interested in historical normative lexicography and asked that we derive our content from traditional Buddhist definitions found in ancient treatises and present it in the form closer to encyclopaedic articles than dictionary entries (Lugli, 2019). Dauntingly, introducing corpus lexicography in the field of Buddhist Sanskrit also required building a suitable corpus from scratch. Buddhist Sanskrit is a non-classical variety of Sanskrit, sometimes referred to as ‘Buddhist Hybrid Sanskrit’ (Edgerton, 1953), which is especially difficult to segment and has hardly received any attention from the NLP community until very recently.² With no computational tools available to process Buddhist Sanskrit, we opted for working with a very small unprocessed corpus consisting of 33 Buddhist Sanskrit texts dating from the first half of the first millennium CE and belonging to various traditions and text-types. The choice of the texts was largely determined by the quality of the available digital editions and the availability of translations. Given the amount of manual labour involved in retrieving and analysing corpus examples for each lemma, starting on such a small corpus seemed a justifiable choice.

² See Lugli (2018 and forthcoming), as well as Handy (2019).

Second, detailed lexical analysis would be presented in narrative form together with sense-descriptions, examples and a short etymological overview. As a compromise between our intended mission and our audience’s requests, we decided to open our entry with a rather lengthy narrative description of the headword that would explain the relationship between its general and specialized uses in a format akin to a miniature essay. Great efforts were invested in the design and implementation of a granular microstructure that would provide users with the information necessary to gauge the semantic versatility of key Buddhist words in context, and appreciate their relationship with semantically and etymologically related words. Since our intended audience comprised both seasoned scholars and students we also took care of presenting the information in a way that would satisfy both user groups. The entry would provide our analysis of a lemma while at the same time also offering users the opportunity to conduct their own analysis based on an extensive range of examples extracted from the corpus. All the examples found in the corpus would be semantically categorized, but only those judged to be most illustrative of a sense or construction would be rendered in English.³ For each sense of a lemma, the entry would also provide a ‘contrastive section’ with examples illustrating the relationship between the lemma and semantically or etymologically related words in context.⁴

2.1.1 Problems

Several entries were produced using the microstructure outlined above. Work was progressing extremely slowly and it gradually became clear that the amount of labour required to prepare an entry was simply not sustainable. This was partly due to the large amount of curated information that each entry required. The translation of all the relevant examples alone typically took several days. Yet, what proved to be really unsustainable was the kind of workflow that the essay-like entry required—and its tolerance for lack of systematicity. Combined with the training background of our lexicographic team, this workflow led to catastrophic results.

People proficient in Buddhist Sanskrit tend to have a solid philological and philosophical training, but no training in lexicography and corpus linguistics. This affects their lexicographic output in several ways. First, they are not used at looking for patterns in data and find it difficult to abstract word senses from individual citations, or spot correlations between meaning and co-text. Second, they tend to focus on philosophically interesting examples where the lemma is used in a less than typical way.⁵ Third, and most important, they are used to a scholarly workflow that starts with taking notes and progresses by gradually refining these notes into a publishable

³ On the system of semantic categorization used in the project see Lugli (2015).

⁴ For more information of the principles informing the entry design, see Gomez and Lugli (2015).

⁵ Cf. Atkins and Rundell (2008: 52).

piece of writing. This workflow was initially encouraged as it was thought suitable to produce the verbose entries that the project required. This proved to be the single most problematic aspect of the early phases of the project. The unstructured workflow made it difficult to monitor progress, reproduce the lexical analysis that informs an entry, or hand over an unfinished entry to colleagues whenever a contributor left. Most importantly, unstructured note-taking was in no way re-usable and could not contribute to advancing the NLP infrastructure that we needed to build a lemmatized corpus.

After years of painfully slow progress, a costly lesson was learned: before starting lexicographic work (especially on a low resource language), it is advisable create a highly structured digital workflow designed to optimize resources. In our case, a good way to optimize resources was to ensure the re-usability of the lexicographers' output for both dictionary content and corpus creation.

To move from this realization to its implementation was not easy. The idea of adopting a rigid workflow met with significant resistance and was at first rejected on the grounds that it would be too mechanical a job for postdoctoral scholars, and junior students would not have sufficient proficiency in the language to perform it accurately. Both objections are valid. It proved difficult to find collaborators who are both capable and willing to annotate Buddhist Sanskrit using a systematic workflow. Still, the time invested in searching for these people and developing a computer-assisted workflow proved a good investment.

2.2 Towards smarter lexicography for Buddhist Sanskrit

In 2017 we developed a web-based annotation tool that requires lexicographers to record syntactic and semantic information for each citation (i.e. KWIC) they analyse.⁶ The corpus is still unprocessed, so the annotation tool requires lexicographers to manually segment and lemmatize the examples, mark all syntactic dependencies involving the lemma, semantically tag the lemma and its dependencies, and annotate conceptual relations between the lemma and other co-text items (e.g. cases where the concept expressed by a lemma is said to be caused by a concept expressed by another word in the sentence). Given the interpretive difficulties of the sources, lexicographers are also asked to record any uncertainty in the annotation using a four-fold typology that allows to distinguish between philological problems, textual ambiguity, disputed interpretation and personal uncertainty (Lugli, 2015). Finally, the annotation process involves aligning the Sanskrit examples with their published English translations.

Such detailed annotations are time consuming. However, switching to an annotation-based workflow has sped up lexicographic work by an order of magnitude compared to

⁶ <https://btw.mangalamresearch.org/en-us/meaning-mapper/>

the unsystematic workflow we initially had. It has improved the efficiency of our in-house lexicographic training phase, enabling our contributors to transition from a ‘humanities mindset’ to the adoption of corpus-linguistics methods. It has also made lexicographers’ analyses more transparent and easier to check, thus drastically reducing the time allocated to revisions. Most importantly, the new workflow has enabled us to adopt an iterative lexicographic cycle whereby proto-dictionary entries automatically derived from the annotations can be made accessible to our audience before fully curated entries become available.

2.2.1 A Visual Dictionary of Buddhist Sanskrit

With the new workflow, the immediate output of our lexicographers’ work on a headword is not a dictionary entry; it is a dataset containing annotated citations for that headword. This dataset can be exported from the annotation tool to several formats, including vertical, xml or CSV. Each format has its own uses. Here I will focus on the CSV format, which offers the advantage of easily lending itself to analysis through widely used statistical computing platforms, such as R.

The CSV files exported from our annotation tool have one row per citation and one column per annotation field. For example, there are columns containing semantic descriptors of the headword at various levels of granularity (e.g. semantic field, sense and subsense). There are also columns for grammatical details such as gender and number, as well as several columns devoted to syntactic information. The representation of syntactic dependencies over CSV columns is somewhat clumsy, especially if compared to CONLL formats, but is nonetheless effective. Each type of syntactic relation corresponds to a variable (e.g. ‘modifies’, or ‘isSubjectOf’) that takes as values the lemma forms of the words linked to the headword through the specified syntactic relation. The same applies to conceptual relations. The resulting CSV features 170 columns and is best explored through data visualizations.

These visualizations, which we currently generate using the popular R package *ggplot2*, are used internally to check the consistency of the annotations. They also serve to refine the lexicographers’ interpretation of a lemma in context, highlighting collocational trends and co-textual patterns that might have been overlooked while reading through the citations.

Once the dataset for a headword has been checked and the team agrees that the annotations it contains are reliable, it is merged with the datasets already created for other words and the information it contains can immediately be made available to the public via those very same data-visualizations we used internally to refine the annotations. To this end we currently use *Shiny*, an R package that allows users to create web-based interactive apps with minimal programming skills (Chang et al., 2018). *Shiny* is extremely versatile and supports data-visualizations as well as text sections, thus allowing the display of traditional dictionary content, such as definitions and

examples, as well as charts.

At present, our Shiny app is a rapidly evolving working prototype called (over-ambitiously) A Visual Dictionary of Buddhist Sanskrit.⁷ It opens with a shallow description of the senses and semantic domains covered by the lemma, which is automatically derived from the annotated dataset, followed by a series of data-visualizations that allow users to explore various aspects of the lemma. The top visualization can be configured to chart most of the information contained in the annotated dataset, including the distribution of a headword's senses, subsenses and semantic prosody across different genres, period, traditions and periods.

Below this graph, the app displays two corpus examples where the headword expresses the sense or subsense chosen by the user. The examples are accompanied by bibliographic references and, whenever possible, they are followed by a translation taken from a published translation of the relevant text. Currently the examples are randomly selected from among all the examples available for a word-sense combination. A GDEX-based system may be devised once we have a segmented and lemmatized corpus.



Figure 1: Example display in the Shiny app.

After the examples, the user is presented with a series of word clouds, illustrating the relative frequencies of various co-textual items that occur in the user-specified relation with the lemma. Further down, the user can visualize the distribution of word-senses in a specified text. This visualization addresses one of the primary concerns of the original Buddhist Translators Workbench project, that is helping translators gauge the degree of specialization that a lemma might have in a given text and appreciate the semantic continuity that often exists between the artificially created word senses. This feature is especially useful for students of Buddhist philosophy, as it helps identify cases where the inherent vagueness of a word was exploited for hermeneutical reasons.

⁷ <https://ligeialugli.shinyapps.io/VisualDictionaryOfBuddhistSanskrit>

Typically, the chart would highlight these cases by showing the deployment of different senses of the same word in close proximity. At the time of writing, the unit used to measure proximity is the page of the Sanskrit edition of the text. This is unhelpful, as the length of pages changes from text to text and thus impairs comparison of a lemma's semantic distribution across different sources, which is a desirable feature.⁸ We are in the process of switching to a sentence-based measure to enable such comparisons.⁹

The last visualization that our app currently offers is a chart that categorizes lemmata by semantic domain to identify near-synonyms. We will probably soon switch to a different modality of visualization for this chart, and as soon as we will have sufficient data we intend to move away from relying on semantic annotation for this feature and we will seek to use corpus data and collocational information to detect potential synonyms.

It is important to emphasize that this app is a work in progress and has not been developed by our professional engineer. It is conceived as a nimble tool to communicate our results to our audience in real time without incurring into additional software-development cost.

2.2.2 Future developments

We are currently creating datasets for headwords pertaining to the semantic fields of language and mental activity, with an emphasis on lemmata that cover both semantic fields. Once we complete datasets for all the words in these semantic fields, we will start a new iteration of the lexicographic process and craft human-curated descriptions of the words to replace the shallow, automatically generated summaries that currently open the entries. Once the curated descriptions are in place, our lexicographic team will move on to annotating citations for words related to a new semantic field, while contributors with no specialized knowledge of Buddhist Sanskrit will be tasked with filling in our original work-intensive microstructure with the data annotated by the lexicographers. This allows the 'real' dictionary to keep growing at reduced cost. Once the datasets for one semantic field are deemed complete, they will also be made available to the public in CSV, CONLL and xml formats for re-use in other projects.

This iterative model allows us to concentrate our very limited human resources on one task at the time, first annotation and then lexicographic curation, while simultaneously enabling our audience's access to lexical analysis at an early stage. It also allows us to work towards the development of a fully processed corpus. The manually segmented citations have been used to develop a rule-based segmenter and lemmatizer that is

⁸ I am grateful to Ammon Shea for suggesting this feature.

⁹ This is not without problems, as 'sentence' is not a straightforward concept in our sources, and some differences in the division of text into sentences may occur from text to text.

currently being used to automatically process our corpus (Lugli, forthcoming). The manually annotated dependencies are also being used to test a Sanskrit sketch grammar for use in Sketch Engine that has been developed by the present author. This sketch grammar is designed to infer syntactic relations from a segmented corpus, without the need for PoS tagging or dependency annotation. As it relies on morphology only, it cannot achieve the same level of delicacy as the manually annotated citations. However, the ability to infer even the most basic syntactic relations (e.g. verb’s subject and object) automatically would constitute a significant advance for Buddhist Sanskrit corpus linguistics. If the automatically inferred syntactic relations will prove sufficiently accurate, we shall be able to further streamline our lexicographic work by limiting annotation to semantic information. In the future, semantic tagging could also be automated, but this avenue has not been explored yet within the project.

3. Lexicography in Motion: a Tibetan verb valency lexicon

The context of the diachronic Tibetan verb lexicon project differs significantly from that of the Buddhist Translators Workbench. This project builds on extensive previous work on Tibetan NLP. It disposes of at least two PoS taggers and lemmatizers (Garrett, Hill & Zadoks, 2014; Meelen & Hill, 2017), as well as of a large tokenized, lemmatized and PoS-tagged corpus (Meelen, Hill & Handy, 2017). It also benefits from pre-existing high-quality dictionaries, including works devoted entirely to Tibetan verbs (Hackett, 2019; Hill, 2010). Moreover, the team possesses expertise not only in the Tibetan language, but also in professional lexicography and computational linguistics. Still, this project also faces some key difficulties characteristic of the lexicography of low-resource languages, especially for older diachronic strata – which are the focus of the present discussion. Even though pre-processed corpora for these strata of the language exist, they do not possess the layer of annotation required for our lexicographic purposes. The main research goal of the project is to shed light on verb argumentation patterns through corpus evidence. To this end, the lexicon relies on an annotation system for syntactic dependencies that distinguish between twelve types of arguments.¹⁰ Few researchers in our team possess the necessary level of language proficiency to carry out the dependency annotations or check the output of automatic parsers. They are the same people who were initially tasked with creating the dictionary content.

This creates intra-project competition for human resources, as the same team-members are needed for NLP and corpus development on the one hand, and for lexicographic curation on the other. We planned to address this problem by tasking these researchers with corpus annotation first, and with lexicographic editing later on. The idea was that once a critical mass of manually annotated data was achieved, dependency annotation could be automated. In the meantime, the rest of the team would prepare the microstructure of the dictionary and ready a dictionary writing schema for the

¹⁰ For details, see <https://tibetan-nlp.github.io/lim-annodoc/deprels>.

lexicographers to use as soon as the corpus was ready.

The theory behind this plan is sound. In practice, however, reaching a critical mass of manually annotated sentences and developing a reliable automated dependency annotation has been taking most of the team’s time and energy, leaving very little room for lexicographic curation of dictionary entries. As a result, the automation of lexicographic tasks has acquired a more prominent and pervasive role in the project than we initially envisioned.

A challenge in this project is that our corpus’ design is still in flux. The corpus is being built while we devise strategies for automatically extracting and displaying lexicographic information from it. Any trials and tests need to be run on the exiguous manually annotated corpus that we currently have, which amounts to around 100,000 words. However, the solutions we come up with through the trials need to be scalable to the full corpus once we have it. The size of our final corpus is not set, but will ideally include several hundred million words.¹¹ Size is not the only difference between the corpus we are using for trials and the one on which we intend to base our final lexicographic product. The final corpus will comprise three diachronic layers, while so far we have been working only on Classical Tibetan. The dependencies annotation will be enriched with morpho-syntactic information that is currently not available, and portions of the corpus will be aligned to English translation. In brief, our strategies for automating the project’s lexicographic output need to be adaptable to changes in the corpus.

3.1 Lexicographic automation for a diachronic Tibetan verb valency

lexicon

In collaboration with the Sketch Engine team, we have generated a sample dictionary draft from our small manually-annotated corpus of Classical Tibetan. It contains 774 entries, based on a headword list derived from existent Tibetan dictionaries. We also derived a headword list from the corpus, but this proved unsatisfactory, as it erroneously included nominalized verbal forms, due to PoS-tag ambiguity. When our full corpus will be ready, we will derive a new headword list from it and compare it with the list extracted from dictionaries to ensure that verbs not recorded in existing dictionaries but attested in the corpus will be included in our lexicon.

Our small test corpus is associated with a sketch grammar that allows verbs’ word sketches to be arranged by argument structure in Sketch Engine. The word sketch

¹¹ Ideally it would comprise a 300 million-word corpus of Tibetan that has been PoS tagged in recent years (Meelen, Hill & Handy, 2017), plus an additional corpus of contemporary Tibetan and a small corpus of Old Tibetan that we are creating from scratch within the project.

information is mapped onto our DWS entry template, which is arranged by argument structure. As DWS we are using Lexonomy, a free dictionary writing software closely connected with Sketch Engine. Lexonomy allows users easily to edit entry templates that can be auto-populated with information from a corpus hosted on Sketch Engine (Měchura, 2017). Lexonomy's out-of-the-box configuration allows lexicographers to pull dictionary examples from a Sketch Engine corpus from individual example slots in each entry. This practice requires lexicographers to manually select and add the examples to the entries, which is time consuming. To push all the examples from the corpus directly to the relevant slots in the entries seems more efficient; so we opted for this solution. This required the assistance of the Sketch Engine team and the payment of a (very reasonable) fee.

3.1.1 GDEX development for Classical Tibetan

In the dictionary draft, all examples are accompanied by full bibliographic and period metadata and are sorted using a GDEX formula that models an ideal good dictionary example (Kilgarriff et al., 2008). The main parameters of our GDEX are sentence length, absence of additional arguments beside the argument pattern to be illustrated by the example, and a reduced presence of pronouns, to avoid anaphoric references that may be difficult to interpret out of context. To filter out sentences that might be difficult to read, examples with many verbs are penalized, and so are those displaying lengthy strings of adjectives, determiners and adverbs.

Our GDEX formula was first intuitively developed on the basis of an ideal model of 'good Tibetan example sentence'. The output of the formula was then tested against 150 sentences manually rated by the lexicographers on a 0-2 scale, where 2 is a perfect example, 1 is an example that may need some manual editing, and 0 is a bad example. 70% of the examples were rated 0, and only 8% were rated 2. Given the limited time the lexicographers could spare for rating examples, only two iterations of the formula have been possible so far. The formula that we have developed through these iterations is successful in promoting good examples to the top of the example list; but given the paucity of 2-rated examples it was impossible to fine-tune the formula to distinguish between 1- and 2-rated examples. It also needs improvement in filtering out 0-rated sentences. Currently, while all good examples are among the top-rated sentences, almost one third of the top-rated sentences are bad examples.

The identification of complete sentences is one of the most challenging aspects of modelling good examples for Classical Tibetan. The corpus is divided into sentences according to Tibetan punctuation, but this does not follow the same principles as Western punctuation and is rarely indicative of sentence boundaries. Steps have been taken to include likely identifiers of final sentence boundaries in the GDEX formula. For instance, sentences ending with final particles are promoted, while sentences ending with case markers are penalized. However, more work remains to be done to identify initial sentence boundaries.

As it is often the case with GDEX, our current formula promotes simple sentences. These may well be user-friendly, but are not necessarily representative of the style employed in Classical Tibetan sources. For this reason, our entries will also contain examples sorted through an alternative GDEX formula that does not penalize multiple verbs, modifiers and determiners as much as the current one. It will be up to the user to choose which set of examples to peruse.

In an effort to promote to top of the example list the most representative sentences, we have also augmented GDEX sorting with argument-specific collocational information.¹² The highest GDEX-ranked example that features in the relevant argument slot the most frequent word for that argument slot is promoted to the top. Likewise, the top GDEX ranked example that has in the relevant argument slot the second most frequent word for that argument slot will occupy the second position in the example list, and so on. This is to ensure that at the top of the example list we will have typical sentences like ‘to drive a car’ and not idiosyncratic expressions like ‘to drive a gas guzzler’.

To be representative, the top examples also need to be drawn from a variety of sources. All else being equal, the sentences at the top of the example list will be taken from different texts.¹³

3.1.2 Future developments

To be useful, dictionary examples need not only to be ‘good’ and representative, but also easy to peruse. In the case of ancient languages such as Old and Classical Tibetan, adding a translation of the examples would help in this regard. It is unlikely that our lexicographers will have time to craft such translations; so our attention has turned to the possibility of using published translations of the sources. While it may save us time, this option is not without its problems. Only a fraction of our final corpus has been translated. This leaves us with the uncomfortable choice of either limiting the selection of our top examples to the few texts that we can align with published English translations, thus not taking full advantage of the power of the integrated GDEX workflow we have devised, or risk leaving the top examples untranslated, thus compromising the user-friendliness of our lexical resource. A solution would be to allow users to decide whether to restrict the selection of the examples to those accompanied by a translation. We have not yet investigated how to implement this feature within the Lexonomy infrastructure.

The most daunting challenge awaiting us is the addition of word senses to the entries. Currently, the entries are divided by argument pattern and not by sense. This allows us to auto-populate the entries purely on the basis of word sketches, without recourse

¹² Cf. Gantar et al. (2016: 214).

¹³ Cf. Cook et al. (2014: 320-321).

to automatic sense induction or sense discrimination. Senses feature in our Lexonomy entry schema as xml attributes of example elements, alongside bibliographic and period metadata. The original aim of this arrangement was to allow lexicographers to manually tag the top examples with sense labels while editing the automatically generated dictionary draft. It now seems unlikely that the lexicographers will have sufficient time to sense-tag the examples, as their linguistic expertise is still needed to develop the dependency parsed corpus. We will therefore explore avenues to automate this aspect of the lexicographic work, too.

4. Conclusions: lessons learned

Automated lexicographic solutions can only be as smart as the language resources they rely on. Languages that lack suitably processed and annotated corpora are at a disadvantage. Especially so if there is a paucity of people able to annotate those corpora and develop adequate NLP tools for them. Still, this is no excuse for reverting to entirely manual workflows. The lexicographers' work and output should be designed to serve more than one purpose, so that beside building dictionary content it also feeds into NLP research and contributes to the creation of better corpora, which will, in due course, enable faster lexicographic workflows.

Building the corpora and NLP infrastructure necessary for the automation of lexicographic tasks is a lengthy process. In the meantime, there is no reason to fall back to entirely manually curated dictionary entries, which would only divert the lexicographers' precious language-specific expertise from the task of corpus development. There is no need to wait until a fully processed corpus and perfect NLP pipeline are in place, either. While the corpus is being developed, manually annotated sentences can be displayed to the public, without extra curation, via *ad interim* lexical resources through free and easy to set up tools such as Shiny or Lexonomy.

5. Acknowledgements

The project Lexicography in Motion is funded by the British Arts and Humanities Research Council. The Buddhist Translators Workbench was started with funding from the US National Endowment for the Humanities and is currently funded by the Mangalam Research Center for Buddhist Languages.

6. References

- Atkins, S. & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: OUP.
- A Visual Dictionary of Buddhist Sanskrit. Accessed at: <https://ligeialugli.shinyapps.io/VisualDictionaryOfBuddhistSanskrit/> (June 15 2019).
- Buddhist Translators Workbench. Accessed at <https://btw.mangalamresearch.org/>

- (June 15 2019).
- Chang, W., Cheng J., Allaire, J.J., Xie, Y. & McPherson, J. (2018). *Shiny: Web Application Framework for R*. <https://CRAN.R-project.org/package=shiny>
- Cook P., Rundell, M., Lau, J. L. & Baldwin, T. (2014). Applying a Word-sense Induction System to the Automatic Extraction of Diverse Dictionary Examples. In A. Abel, C. Vettori & N. Ralli (eds.) *Proceedings of the 16th EURALEX International Congress*. Bolzano: EURAC research, pp. 319-328.
- Edgerton, F. (1953). *Buddhist Hybrid Sanskrit grammar and dictionary*, 2 vol. New Haven: Yale University Press.
- Gantar, P., Kosem, I. & Krek, S. (2016). Discovering Automated Lexicography: The Case of the Slovene Lexical Database, *International Journal of Lexicography* 29(2), pp. 200–225.
- Garrett, E., Hill, N., Zadoks, A. (2014). A Rule-based Part-of-speech Tagger for Classical Tibetan. *Himalayan Linguistics*, (13)1, pp. 9-57.
- Garrett, E., Hill, N., Kilgariff, A., Vadlapudi, R. & Zadoks, A. (2015). The contribution of corpus linguistics to lexicography and the future of Tibetan dictionaries. *Revue d'Etudes Tibétaines*, 32, pp. 51–86.
- Garrett, E. (2017). Lexicography in Motion: Documentation. <https://tibetan-nlp.github.io/lim-annodoc/>.
- Gomez, L. O. & Lugli, L. (2015). Buddhist Translators Workbench white paper. <http://dx.doi.org/10.17613/M6866Z>.
- Griffiths, P. (1981). Buddhist Hybrid English: Some notes on philology and hermeneutics for Buddhologists. *Journal of the International Association of Buddhist Studies* 4(2), pp. 17 -132.
- Hackett, P. (2019). *A Tibetan Verb Lexicon*. Second edition. Boston: Snow Lion. First ed. 2003.
- Handy, C. (2019). A context-free method for the computational analysis of Buddhist texts. In D. Veidlinger (ed.) *Digital Humanities and Buddhism: An Introduction*. Berlin: De Gruyter.
- Hill, N. (2010). *A Lexicon of Tibetan Verb Stems as Reported by the Grammatical Tradition*. Munich: Bayerische Akademie der Wissenschaften.
- Kilgariff, A., Husak, M., McAdam, K., Rundell, M. & Rychly, P. (2008). GDEX: Automatically Finding Good Dictionary Examples in a Corpus. In E. Bernal and J. DeCesaris (eds.) *Proceedings of the Thirteenth EURALEX International Congress*. Barcelona: Institut Universitari de Linguística Aplicada, Universitat Pompeu Fabra, pp. 425–432.
- Landau, S. I. (2001). *Dictionaries: The art and craft of lexicography*. Cambridge: CUP. First ed. 1984.
- Lugli, L. (2015). Mapping meaning across time and cultures: innovations in Sanskrit lexicography. In *Words Dictionaries and Corpora: Proceedings of the 9th International Conference of ASIALEX*.
- Lugli, L. (2018). Drifting in Timeless Polysemy: Problems of Chronology in Sanskrit Lexicography. *Dictionaries: Journal of the Dictionary Society of North America*,

- vol. 39(1), pp. 105–129.
- Lugli, L. (2019). Words or terms? Models of terminology and the translation of Buddhist Sanskrit vocabulary. In A. Collett (ed.) *Buddhism and Translation: Historical and Contextual Perspectives*, New York: SUNY.
- Lugli, L. (In preparation). Towards Buddhist Sanskrit Corpus Linguistics: advances in segmentation, lemmatization and syntactic inference for Buddhist Sanskrit.
- Meelen, M. & Hill, N. (2017). Segmenting and POS tagging Classical Tibetan using a memory-based tagger. *Himalayan Linguistics*, 16(2), pp. 64–89.
- Meelen, M., Hill, N. & Handy, C. (2017). The Annotated Corpus of Classical Tibetan (ACTib), Part II - POS-tagged version, based on the BDRC digitised text collection, tagged with the Memory-Based Tagger from TiMBL [Data set]. Zenodo. <http://doi.org/10.5281/zenodo.822537>.
- Měchura, M. B. (2017). Introducing Lexonomy: an open-source dictionary writing and publishing system. In I. Kosem et al. (eds.) *Electronic Lexicography in the 21st Century: Lexicography from Scratch. Proceedings of the eLex 2017 conference*, Leiden.
- Mianáin, P. O., Convery, C. (2014). From DANTE to Dictionary: The New English-Irish Dictionary. *Proceedings of the 16th EURALEX International Congress*, pp. 807-817.
- Nasiruddin, M. (2013). A State of the Art of Word Sense Induction: A Way Towards Word Sense Disambiguation for Under-Resourced Languages. arXiv:1310.1425.
- Pajzs, J. (2009). On the Possibility of Creating Multifunctional Lexicographical Databases. In H. Bergenholtz, S. Nielsen & S. Tarp (eds.) *Lexicography at a crossroads. Dictionaries and encyclopedias today, lexicographical tools tomorrow*. Bern: Lang, pp. 327-354.
- Rundell, M. & Atkins, S. (2011). The DANTE database: a User Guide. In I. Kosem & K. Kosem (eds.) *Proceedings of eLex 2011. Trojina: Institute for Applied Slovene Studies*, pp. 233–246.
- Rundell, M. & Kilgariff, A. (2011). Automating the creation of dictionaries: where will it all end? In F. Meunier, G. Gilquin & M. Paquot (eds.) *A Taste for Corpora: In Honour of Sylviane Granger*. Amsterdam: John Benjamins, pp. 257–282.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



The *Russian Academic Neography*

Information Retrieval Resource

Marina N. Priemysheva, Yulia S. Ridetskaya, Kira I. Kovalenko

Institute for Linguistic Studies, Russian Academy of Sciences,
199053, 9 Tuchkov pereulok, St. Petersburg, Russia

E-mail: mn.priemysheva@yandex.ru, vjs_neolex@mail.ru, kira.kovalenko@gmail.com

Abstract

Creation of electronic dictionaries and retrodigitalization are very popular trends in modern lexicography. The idea to use computer techniques in Russian neology appeared in 2013, but only recently has the *Russian Academic Neography* information retrieval resource been created. It represents both published dictionaries (annual, decadal and thirty-year dictionaries), which include about 116,000 words and collocations that had not been registered by normative explanatory dictionaries of the Russian language, and new materials that were not included in published volumes or that are being prepared for publication. Simple and advanced types of search give an opportunity to find words by various parameters (word, word component, year or time period, labels, etc.). It is also intended to include chronological and frequency parameters in the future. The aim of the *Russian Academic Neography* information retrieval resource is to represent the newest Russian vocabulary and to make it available for a wide spectrum of users.

Keywords: new-word dictionary; neography; electronic dictionary; *Russian Academic Neography*; information retrieval resource

1. Introduction

New scientific directions, such as corpus linguistics and computer lexicography, have allowed authors and publishers of dictionaries to go beyond traditional paper lexicography and discover new possibilities for creating and using vocabulary information. During the last two decades, lexicographers have searched for optimal forms and means of achieving the most convenient and productive ways of representing vocabulary, as well as going beyond the existing formats through the creation of new, interactive resources.

At the present stage, there are several trends in the presentation and use of dictionaries in electronic form. Nowadays, lexicography works can be roughly divided into electronic dictionaries (newly created computer dictionaries and online dictionaries) and retrodigitalized dictionaries (all forms of paper dictionaries converted into electronic format).

Electronic dictionaries are a very successful genre of modern computer lexicography, and appear in various forms and solutions. As a Russian language resource, they are

incredibly popular. Along with small projects, such as the online dictionary of jargon and slang (<http://www.slovonovo.ru>) and the popular dictionary of the Russian language (<http://slovoborg.su>), large electronic thesauruses catering for a variety of functional purposes and based on databases of various sizes are available: the dictionary of collocations based on the National Corpus of the Russian language (<http://www.ruscorpora.ru/obgrams.html>), “Database of pragmatically marked vocabulary” (<http://spml.ipmip.nspu.ru/?action=main>), *CrossLexica* (<https://www.xl.gelbukh.com>), open electronic thesauruses of the Russian language (<https://russianword.net>; <http://ruslex-encode.ru>) and many others. Of course, this trend in the development of modern computer lexicography still requires a long period of adaptation and crystallization of forms and tasks, and each project needs to find its place in the scientific paradigm: the creation and design of such dictionaries often resembles a lexicographic game, rather than a serious scientific project.

Retrodigitalized dictionaries are represented quite significantly on the Internet, but there is still a question of technical implementation. These dictionaries have various formats: from a database of a single edition (for example, <https://www.slovardalja.net>, <https://ushakovdiction.ru>, <http://orfo.ruslang.ru>) to databases of dictionaries of the same type (<http://etymolog.ruslang.ru/>) or a number of typologically diverse dictionaries (<https://www.slovari.ru>, <http://grammar.ru/SPR/?id=1.0>, <http://gramota.ru/slovari/>, etc.). Currently there is a tendency to create compilations of dictionaries, such as, for example, the “Historical Dictionary of the Russian Language” (<http://dic.feb-web.ru/rusdict/index.htm>) and “Academic Corpus of the Russian Language Vocabulary” (Lesnikov, 2019a, 2019b). The architecture, structure and interfaces of these databases are very diverse: each of the electronic lexicographic projects in Russia currently functions autonomously, and there is still an ongoing search for an optimal electronic lexicographic form.

2. General characteristics of the texts

in the *Russian Academic Neography* portal

When considering the tradition of lexicographic representation of dictionaries and dictionary resources in Russia, the *Russian Academic Neography* information-retrieval resource occupies an intermediate place and this, among other things, is its originality:

- it is both a professional resource for specialists of lexicology and lexicography, and a reference resource, designed for a wide spectrum of users;
- it is a joint database of dictionaries of the same type (annual, decadal and a thirty-year dictionaries);
- it is (in the near future) an online dictionary of new vocabulary.

At the same time, the textual database of the resource is quite specific, which is determined by the traditions of Russian neography. Russian academic neography as a separate theoretical and practical lexicographic trend has existed since the 1960s. Its theoretical basis is formed in numerous works of N. Z. Kotelova, E. A. Levashov and T. N. Butseva. As was determined by Kotelova, Russian neography is represented by three types of neologism dictionaries, work on which was conducted, and continues to be conducted, by the team of the New Words Dictionaries group of the Institute for Linguistic Studies of the Russian Academy of Sciences (Leningrad / St. Petersburg).

1. *New in Russian Vocabulary* annual dictionaries, recording all the new words of a given year, including innovations of particular authors and occasionalisms (18 issues were published: 1977–1994; work on the annual dictionaries of 2010–2019 has been resumed recently). This is a series of reference dictionaries that include absolutely all the innovations of Russian speech in the focal period. “The *New in Russian Vocabulary. Dictionary Materials...* annual dictionaries are an attempt to show the flow of spontaneous language life, to demonstrate the facts of birth, change, or entry into the language of words in all their diversity. They present everything new that occurred during daily examination in the texts of ten sources (constant from year to year) in four checked months (of a given year), including the words of short-term existence and one-time use. Each annual dictionary includes about 4,000 vocabulary units” (Kotelova, 2015: 367).

2. *New Words and Meanings* decadal dictionaries record only those lexical units that entered the Russian language in a given decade and were included in the language use. *New Words and Meanings* are explanatory dictionaries, which complement large explanatory dictionaries of the literary language (such as the *Big Academic Dictionary of the Russian Language*), as “decadal dictionaries show only facts that have become the property of the language, at least for a certain time” (Kotelova, 2015: 367). Decadal dictionaries of the 1960s, 70s, 80s and 90s have been published; the last one is a three-volume book (about 1,000 pages per volume), in which the linguistic elements of Russian life of the 1990s are clearly and visually represented.

3. *Dictionary of New Words* is a thirty-years dictionary and records only the words that entered into common usage and could be included in the dictionaries of the Russian literary language. *The Dictionary of New Words of the Russian Language of 1950–1980s* is the normative explanatory dictionary of neologisms of the post-war era, which was intended to complement the explanatory dictionaries of the Russian language.

Currently, the resource of all published new-word dictionaries is about 116,000 words and collocations that were used in the Russian language in 1960–2000, but which had not been registered by any of the explanatory dictionaries of the Russian language of the 19th and 20th centuries (in comparison, the *Dictionary of the Modern Russian Literary Language* in 17 volumes includes about 120,000 words). That means that these dictionaries significantly complement all available vocabulary resources of the Russian

language, containing as many words as had been registered by the lexicographic works before.

On the one hand, each of the dictionaries of the series has its own special scientific function but, on the other hand, it also has a complementary relationship, from a historical perspective, with another type of new-word dictionary. N.Z. Kotelova noticed that “Depending on the lexicographic situation, society needs one or another dictionary of neologisms. The need, for example, to create a normative dictionary of neologisms of a significant period can be considered to be less pressing in a situation of rapidly reprinted and updated general explanatory dictionaries. Dictionaries of new words are designed to facilitate knowledge of the language, giving a description of the innovations from the various points of view: they show their internal form (first of all, the producing word), supply stylistic labels, give forms of inflection, illustrate with good examples of usage, and help with mastering the best variant among competing options. This information is also needed for translators and authors of bilingual dictionaries” (Kotelova, 2015: 370).

Also at the disposal of the new-word dictionary compilers is a fourth resource, which is not available to a wide audience: it is a bank of Russian neologisms, “including three indices: 1) words, 2) word meanings, 3) collocations. It gives an opportunity to review the entire array of neologisms, see the development of pre-existing derivational, thematic nests and series of words, the formation and degree of filling of new ones, evaluate quantitatively innovations for a given attribute (derivational, partial, structural and phraseological, etc.), compare with innovations in other languages — in general or by ranks, to see the variation or synonyms, to observe projections into extralinguistic spheres. It fixes a point of reference for future work in the field of neology — it provides the possibility of automatic processing of neological material, the implementation of formal transformations (for example, the compilation of a reverse vocabulary of neologisms), etc.” (Kotelova, 2015: 370). In other words, the bank of Russian neologisms helps to find a new language unit and define its place in the language lexical system.

Reflecting the synchronous level of the Russian language, annual dictionaries form the basis of decadal dictionaries, and each of the types becomes the historical dictionary of the Russian language of the period being described. However, the main value of a series of new-word dictionaries lies not only in the combination of historical and synchronous approaches in the lexicographic description, but also in fairly accurate dating of one or another occurrence: it is the combination of these principles that makes up the peculiarity of Russian neography.

At the present stage of the collection, recording and description of new words, the work of lexicographers has become even more complex.

Before entering the dictionary, words and word meanings must pass a multistage selection process. First, the material from the source list for the primary search is

analysed. For 2018 the list included *Komsomolskaya Pravda*, *Kommersant*, *Gazeta.ru*, *Rossiyskaya Gazeta*, *Vedomosti*, *Lenta.ru*, *Izvestia*, *Rbc.ru*, *RBK* (magazine), *Metro* (newspaper), and *Novy Peterburg* (newspaper). The source list is created on the basis on IndEx — an indicator calculated by the *Integrum* information and analytical system, which assesses the resource rank in the media space. “The calculation takes into account the number of publications in the media, the visibility of mentioning the object in the media, the role of the object in the publication, emotional colour of the publication and the significance of the (cited) source... The higher the indicator, the more visible the analysed object is in the media space” (<https://www.integrum.ru/ratings/smi/media/jul18>).

The survey of sources also includes monitoring of social networks, news feeds, popular blogs and non-professional Internet dictionaries. The initially selected lexical material is rigorously tested for novelty using the internal databases of the Institute for Linguistic Studies, as well as authoritative normative, explanatory and special dictionaries, Russian National Corpus, corpus of the Russian texts in Google.books.com, and the *Integrum* corpus of texts — the largest in the Russian Archive of texts of Russian language media. Contextual queries in the *Integrum* information-analytical system help in selecting new vocabulary that is synonymous, antonymous, hyponymic, etc. for previously found neologisms.

Modern methodological principles for the selection of lexical units for academic dictionaries of neologisms and the formation of a new-word database were developed in the early 2010s. Thanks to corpus data, it became possible to clarify the first written record of a word in Russian language texts, that is, to find out the approximate time that a word appeared in the language.

Thus, at present, the following vocabulary is available for study (classified according to time and quality parameters):

- neologisms of 1990–1999;
- neologisms of 2000–2009;
- neologisms of the last decade;
- vocabulary dated to the period of the 1990s and missed in explanatory, orthographic, terminological and other authoritative dictionaries;
- occasionalisms, individual authorial innovations, i.e. neologisms, the written record of which is unique.

The new words, new meanings, compounds and collocations that are found enter a local neological database accompanied by technical and information marks. The neologisms of the last decade are distributed by year (2010–2019) in order to create the primary word lists of the *New in Russian Vocabulary. Lexical materials*.

3. The history of the creation of the *Russian Academic*

Neography information retrieval resource

New computer technologies have made it possible not only to expand the sources of new-word dictionaries, but to present the vocabulary data of academic neography in open access. Materials of all published dictionaries are represented on the website of the Russian Academy of Sciences <http://iling.spb.ru/dictionaries.html.ru>. The materials of the four decadal dictionaries can be found in Wiktionary <https://ru.wiktionary.org/wiki/>. Information about the neologisms of the last decades is being published on the web-page of the Academic Neography in the social network <https://www.instagram.com/neographia.spb>.

However, the needs of modern science have long dictated the transition to a new paradigm for the creation and use of dictionaries of new words: going out beyond the existing series of dictionaries makes it possible to create a resource of all neological publications and also to continue the work in the new online format. The future implementation of this lexicographic information retrieval resource will not adopt the existing principle of transition from paper format to electronic, but instead that of online format to paper, in which the paper format can be optional and diverse. This will allow us to speed up the introduction of new words into scientific circulation by representing them in the resource soon after their appearance in speech.

The idea of such a resource — the Neology Service of the Russian Language (neologia.ru) — came from the team leader T. N. Butseva (Butseva, 2013). The resource was technically developed at a very high level on the basis of a specially developed program with an original interactive interface (Dmitriev, 2013). However, the main obstacle in its creation and work was the incredible difficulty of marking up and converting 116,000 dictionary entries into the electronic database. The tasks set by the authors of the project, which were very important for Russian science, required enormous technical and human resources and have not been implemented.

Recently, the *Russian Academic Neography* electronic information retrieval resource (<https://neographia.iling.spb.ru>) has been developed, which continues and develops the concept of the previous resource. At present, it has reached the advanced stage of technical finalization and functions in its test mode (the main part of the vocabulary from 1977-1990s is going to be available by September 2019, in October and November it will be filled by new units for 2016-2017, and at the beginning of 2020 new materials for 2013-2015, 2018 and 2019 will be included).

4. Resource interface and functionality

The *Russian Academic Neography* resource includes a database of the published dictionaries and some unpublished materials, together with a query system. It is both a lexicographic resource and an information portal of Russian neology and neography as a whole.

The new-word database includes both the previously published lexicographical works and the new editions of annual dictionaries created by the team members. The database will be supplemented by new lexical units that were not included for one reason or another in published volumes or materials that are being prepared for publication.

In the final version the resource will include following subdivisions:

1. Information about Academic Neography.
2. Information about dictionaries and dictionary corpora.
3. Links to interesting neologisms of the current year (as a news feed); neologisms from dictionaries of previous years (period 1960–2010s); rare neologisms not represented in the dictionaries of the period before the 1960s (section “From the history of words”), as well as lexicographic and linguistic sketches and articles.

The technical implementation of the *Russian Academic Neography* information retrieval resource has been created by A. Andreev. The dictionaries are processed using a specially written program, which is based on the SWI-Prolog 7.6 development environment. Internally, the set of word entries is stored as a semantic network, with nodes corresponding to different fields in an entry, which had been identified by their formatting (font, size, etc.). TEI encoding is used as an intermediate representation between the textual source and the semantic network. The user query is processed by a set of heuristics in a DWIM fashion, so that the requested fields are automatically guessed in most cases. It is then transformed into a semantic graph template and eventually compiled as a Prolog goal, which is executed yielding the search results. The Web UI is based on the PWP suite (Prolog Well-formed Pages). The application code, the data and the UI elements are all packed together into a single portable executable file.

There are two search options available: simple and advanced. A simple search is performed:

1. On request (word, word component, year of approximate appearance of a word in Russian).
2. Alphabetically. There is a search in the Latin alphabet and numbers, since the dictionaries include neologisms that consist of numbers and letters, as well

as neologisms with foreign-language components (initial and final).

Advanced search is possible by the following parameters:

1. Labels (grammatical, stylistic, emotional; labels that indicate the language of borrowing). In the series of annual dictionaries of the current decade, thematic ones have been added to the listed labels.
2. Full-text search on request.
3. By chronological parameter. Temporal boundaries make it possible to find new words of a certain period.

The entire database is built on the material selected by lexicographers manually, which means that the new words are attributed by the time parameter, as well as new meanings, new morphs (affixoids), and new collocations.

The inclusion of materials into the database is preceded by long preliminary work carried out by a large number of professional researchers: published editions of dictionaries are marked up in a certain way; semantic disambiguation is removed; reference entries included in compounds and collocations are duplicated, which makes it possible to remove the problem of formal, meaningless references; to facilitate the search by time parameter, the year is set for each quotation and for each collocation; technical errors are removed.

The *Russian Academic Neography* resource currently does not take into account the usage parameter, since the dictionaries of this series rely on the non-linguistic *Integrum* corpus of texts, the materials of which, however, allow us to identify the number and dynamics of new words used in Russian texts from the mid-1980s until now. In comparison, in the German dictionary database *das Online-Wortschatz-Informationssystem Deutsch* (<https://www.owid.de/>), the chronological and frequency parameters are presented in the form of diagrams (see the neologic section of *Neologismenwörterbuch*). The diagrams are available for words which appeared in German texts from the 1990s to 2017). Nevertheless, the authors of the *Russian Academic Neography* use data from a number of Russian resources which will give us the option to incorporate this function later. For example, *The National Corpus of the Russian Language* has a section called “graphics” (<http://www.ruscorpora.ru/new/graphic.html>), charts built on a chronological-frequency principle in this section are based on the *Google Ngram Viewer* service. The *Google Books Ngram Viewer* online search service, which has its own corpus of Russian-language texts, allows you to search for words and compare their usage from 1800 to 2008.

Thus, the *Russian Academic Neography* resource is a set of lexical and phraseological units, reflecting changes in the Russian language over the past 60 years.

5. The scientific potential of the resource

The *Russian Academic Neography* information retrieval resource is intended not only for specialists in the field of Russian lexicology and lexicography, but for all linguists and the wider audience.

Thanks to the query system, the following data is going to be available:

- materials of all new-word dictionaries published since 1971, which are currently a bibliographic rarity;
- the lexical materials of the Russian language (1960–2020s), not recorded in other dictionaries;
- when requesting chronology, it becomes possible to establish the occurrence of a word in a particular period;
- when requesting derivational formants, it becomes possible to identify relevant derivational models;
- with the root query, it becomes possible to identify word-building nests and derivational schemes;
- when requesting a label, the trends of the functional and stylistic dynamics of the vocabulary of the Russian language are identified;
- when requesting a source language, it is possible to reveal all borrowed lexemes in one or another period of time, etc.

Due to the fact that the portal database contains about 116,000 professionally collected and processed new units of the Russian language, which is as many as the average vocabulary of the Russian language represented in explanatory dictionaries, the scientific potential of the *Russian Academic Neography* information retrieval resource cannot be overestimated. Introducing a huge lexical layer of the modern Russian language and the newest Russian vocabulary, it is expected to be of great interest to professional linguists and a wider audience.

6. References

- Butseva, T. N. (2013). Neologicheskaya sluzhba russkogo yazyka [Neology Service of the Russian Language]. In V. P. Zakharov & M. N. Priemysheva (eds.) *Leksikologiya, leksikografiya i korpusnaya lingvistika*. St. Petersburg: Nestor-Istoriya, pp. 93–98.
- Dmitriev, D.V. (2013). Neologia.ru: principy postroeniya internet-resursa dlya kollektivnoj leksikograficheskoy raboty [Neologia.ru: Principles of the Internet

- Resource Construction for Joint Lexicographic Work]. In V. P. Zakharov & M. N. Priemysheva (eds.) *Leksikologiya, leksikografiya i korpusnaya lingvistika*. St. Petersburg: Nestor-Istoriya, pp. 99–109.
- Kotelova, N. Z. (2015). Teoreticheskie aspekty opisaniya neologizmov [Theoretical aspects of the new words description]. In Kotelova N.Z. *Izbrannye raboty*. St. Petersburg: Nestor-Istoriya, pp. 254–269.
- Lesnikov, S. V. (2019a). Akademicheskie tolkovye slovari russkogo yazyka kak yadro akademicheskogo slovarnogo korpusa russkogo yazyka [Academic Explanatory Dictionaries as a Core of the Academic Corpus of the Russian Language Vocabulary]. In *Sbornik nauchnykh statej po itogam raboty Mezhdunarodnogo nauchnogo foruma “Nauka i innovacii: sovremennye koncepcii” (g. Moskva, 5 aprelya 2019 g.)*. Part 1. Moscow: Infiniti, pp. 38–47.
- Lesnikov, S. V. (2019b). Akademicheskij slovarnyj korpus russkogo yazyka [Academic Corpus of the Russian Language Vocabulary]. In: *XLVIII Mezhdunarodnaya filologicheskaya nauchnaya konferenciya SPbGU, 18–27 marta 2019* (<http://conference-spbu.ru/conference/40/reports/9649>).

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



***A Thesaurus of Old English* as Linguistic Linked Data: Using OntoLex, SKOS and *lemon-tree* to Bring Topical Thesauri to the Semantic Web**

Sander Stolk

Leiden University, Leiden, the Netherlands

E-mail: s.s.stolk@hum.leidenuniv.nl

Abstract

An increasing number of dictionaries are represented on the Web in the form of linguistic linked data, utilizing OntoLex-Lemon for this purpose. Lexicographic resources other than dictionaries, however, have thus far not been the main focus of efforts surrounding this model. In this paper, we discuss porting a topical thesaurus to the Web: *A Thesaurus of Old English*. By means of this case study, this paper discusses how this thesaurus – and topical thesauri in general – can be represented with OntoLex-Lemon, SKOS and *lemon-tree* through a fully automated process. Along with discussing the terminology required for expressing *A Thesaurus of Old English* as linguistic linked data, this paper indicates challenges encountered in the conversion process. These challenges range from material that is not meant to be made available to the general public to distinctions and relations that have been left implicit in the legacy form but are of much value and, indeed, required to be expressed explicitly in its linked data form. The aim of this paper, thus, is to provide recommendations for representing topical thesauri on the Web and to grant insight into aspects that may be encountered in porting similar lexicographic resources in the future.

Keywords: thesaurus; linguistic linked data; conversion; automation

1. Introduction

An increasing number of dictionaries are represented on the Web in the form of linguistic linked data using the OntoLex-Lemon vocabulary (Bosque-Gil et al., 2016; Khan, 2016). Such a representation is thought to facilitate interoperability across linguistic resources, have the potential to increase their visibility, and promote their reuse (Declerck et al., 2015; Klimek & Brümmer 2015). However, lexicographic resources other than dictionaries have thus far not been the main focus of efforts surrounding OntoLex-Lemon and its modules. In this paper, we discuss porting a topical thesaurus to the Semantic Web: *A Thesaurus of Old English*.

A Thesaurus of Old English captures the lexis of the early medieval variant of English, spoken between roughly 500 and 1100 by the Anglo-Saxons (Roberts et al., 2015). This lexicographic resource presents a feature common to topical thesauri but uncommon to dictionaries: its topical system (i.e., a hierarchy of categories) that organizes lexical senses according to their meaning (Kay & Alexander, 2016). Moreover, this thesaurus

also distinguishes conceptual levels within the topical system – a feature that was already present in the first modern thesaurus, *Roget's Thesaurus* (1852). By means of this case study, then, this paper presents areas problematic for representing *A Thesaurus of Old English* – and topical thesauri in general – in OntoLex-Lemon alone, and turns to the novel model *lemon-tree* for the needed expressivity. This model combines OntoLex-Lemon with the SKOS vocabulary, filling minor but important lacunae perceived for topical thesauri specifically, thereby increasing the portability and interoperability of these lexicographic resources (Stolk, 2019).

Next to treating the terminology required for porting *A Thesaurus of Old English* to a linguistic linked data form, this paper will indicate further challenges in this process. These range from material available in the legacy form that is not meant to be made available to the general public (e.g., notes purely editorial in nature) to distinctions and relations that have been left implicit in the legacy form but are of much value and, indeed, required to be expressed explicitly in its linked data form. The aim for this paper, thus, is to provide recommendations for representing topical thesauri on the Web and to grant insight into aspects that may be encountered in porting similar lexicographic resources in the future.

2. *A Thesaurus of Old English*

A Thesaurus of Old English (TOE) captures the lexis of Old English. The words and their senses of this historical variant of English, spoken roughly between 500 and 1100, are grouped together in sets of synonyms and placed in an overarching hierarchy of categories. In addition, TOE indicates the distribution of words in the surviving Old English texts. Thus, some are flagged as found only in poetic works or as glosses. As of May 2017, the thesaurus contains 51,483 senses that have been sorted and categorized manually in 22,451 categories¹. Accumulating and editing this wealth of information for the first publication of the thesaurus in 1995 took a team of scholars – led by Christian Kay, Jane Roberts, and Lynne Grundy – over fifteen years (Roberts, 1978). The fruit of their labour has certainly not gone unnoticed in the scholarly field concerning Old English.

Since its publication, TOE has been met with high praise. Rolf Bremmer Jr, for instance, states that the thesaurus fills a “voluminous gap [...] on the shelf of lexicographical tools” available for Old English (2002). Richard Dance, too, calls TOE “invaluable” for lexical studies and deems it an “impressive piece of scholarship” (1997). Manfred Görlach goes so far as to state that TOE is “the most important contribution to Old English studies for years”, as its content allows scholars to “investigate what distinctions Anglo-Saxons felt important enough to make in the lexicon” (1998). This historical thesaurus, then, is considered a valuable asset to many scholars. Opening up

¹ These numbers are based on an export of the TOE database provided on 26 May 2017.

the knowledge contained within – by providing the thesaurus in an appropriate form – is therefore an important aspect for its use in research.

Work on TOE continued after its first publication in 1995, resulting in further editions. None of these, however, was published in a linguistic linked data form. The benefits promised by such a form – e.g., interoperability and reuse – warrants looking into how such a lexicographic resource can be represented using the relevant standards. This paper therefore details the process of bringing TOE to the Semantic Web. This process, which converts the contents of the current TOE database into the desired linked data form is illustrated with *frēols* (in the sense of ‘free, not enslaved’, see DOE, s.v. ‘frēols adj.’) that is positioned in the TOE category “Freedom, being free”. This lexical sense and the category it belongs to are depicted in Figure 1 along with relevant context in the form of synonymous senses (cf. *frēot*) and superordinate categories from the topical system.

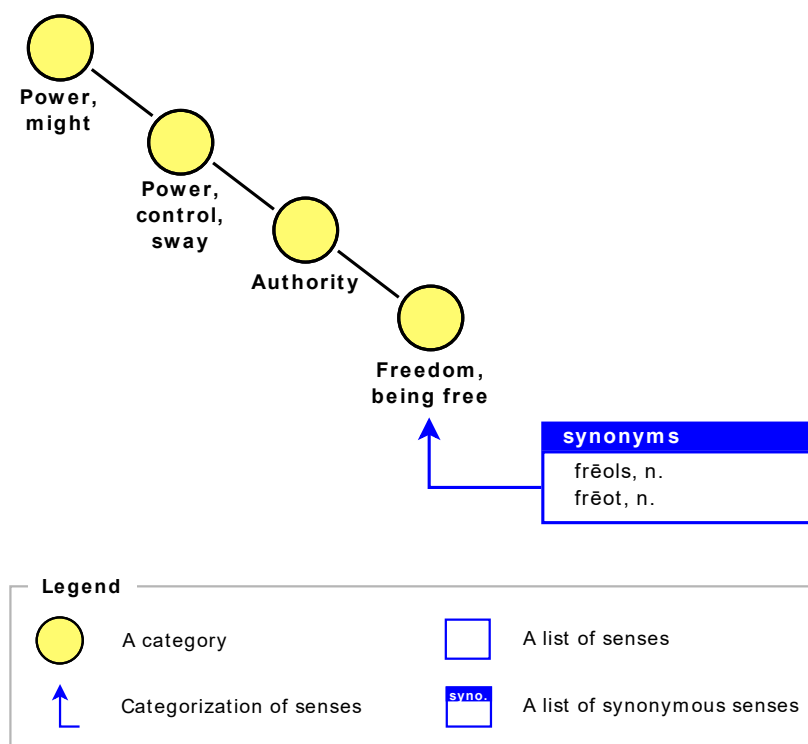


Figure 1: Sample of content from TOE.

In order to discuss the conversion process, we will first continue to describe the current digital form of the TOE database, referred to as its legacy form. The subsequent section provides a better insight into the desired, linguistic linked data form of TOE, which leverages the compact *lemon-tree* model for topical thesauri (Stolk, 2019) alongside the W3C standards OntoLex-Lemon and SKOS (*OntoLex*; *SKOS*). Finally, the conversion

process itself between these two forms is described, followed by the conclusion.

3. Legacy Form

The electronic edition of TOE hosted by the University of Glasgow employs a MySQL database to retrieve and display the thesaurus contents in webpages (TOE, ‘Creation of the *Thesaurus*’). The database format is a tabular one, which makes exports possible to other formats that can capture rows and columns (MySQL 5.7 Reference Manual, ‘What is MySQL?’). Such formats include Excel spreadsheets and CSV files (MySQL 5.7 Reference Manual, ‘Alternative Storage Engines’). In fact, the University of Glasgow provides licensees of the TOE database with a copy by means of such formats. The version of the database provided for this research dates from 26 May 2017.

The TOE database consists of three tables. Each of the tables start with a single row containing the column headings. The rows below it – also known as records – capture instances. The first table discussed here is the category table of TOE, of which the structure is illustrated by Table 1.

catid	t1	t2	t3	t4	t5	t6	t7	subcat	pos	heading	notes
1	1								N	Earth, world	
2	1							1	N	As God's creation	xr Religion
3	1							1.01	N	In the beginning	
...
17187	12	1	1	9				18	V	To accept as a slave	
17188	12	1	1	9				19	V	To bring into bondage	
17189	12	1	1	10					N	Freedom, being free	
17190	12	1	1	10				1	N	Citizenship	
17191	12	1	1	10				2	N	A free man	
17192	12	1	1	10				3	N	A free woman	
17193	12	1	1	10				4	N	Freeman of lowest class	

Table 1: Structure of the TOE category table
(the category “Freedom, being free” is highlighted).

The category table of TOE is used to capture information on categories, where each record represents a single category. The table contains twelve columns in total:

- **catid**: This column acts as primary key, which “uniquely identifies each record in a database table” (*W3Schools.com*, ‘SQL Primary Key’).
- **t1 to t7**: These columns capture the location in the taxonomy. Values in **t1** specify the position of the first main category compared to others at the same level, values in **t2** of the second tree level, and so on.
- **subcat**: This column indicates the location further down the taxonomy on a subcategory level (where applicable). Subcategories are distinguished from main TOE categories, which are indicated by **t1** through **t7**, in order to indicate a conceptual level in the taxonomy with smaller semantic differences than is the

case with main categories (TOE, ‘Classification’). The subcategory position is not stored separately per subordination step, as the case with τ_1 to τ_7 , but as a single concatenated string delimited by stops.

- **pos:** This column stores the part of speech associated with a category. An indicated part of speech applies to all lexemes and their senses that are positioned directly at the category (i.e., they are not assigned to subordinate categories). Such a group of lexemes and senses in TOE always shares a single part of speech. Possible values are “aj” for adjective, “av” for adverb, “cj” for conjunction, “in” for interjection, “n” for noun, “p” for preposition, “ph” for phrase, “pn” for pronoun, “v” for verb, “vi” for intransitive verb, and “vt” for transitive verb (which may be monotransitive or ditransitive).
- **heading:** This column contains the name of each category in present-day English.
- **notes:** This column contains notes that are mostly editorial in nature. These include adjustments that have taken effect, matters still to be discussed, and so on. Due to their nature, the notes have so far been left unpublished in both paper and electronic editions.

Table 1 is identified by the key value 17189, called “Freedom, being free”, expressed by nouns, and located in the taxonomy at position 12.01.01.10 – the 12th top category, followed by the 1st subordinate one, etc. Note that subordination relations applicable to given categories are not captured explicitly in this table but need to be deduced from the position in the taxonomy. Thus, the “Freedom, being free” category is understood to have the category located at 12.01.01 in the taxonomy as its direct superordinate category: “Authority” (catid 169410).

The TOE table discussed next is the category-xref table, of which a sample is shown in Table 2.

xid	catid	refid	tnum
1	18	588	01.03.01.05.01
2	18	9166	05.10.05.04.09
3	45	478	01.02.01.01.03
...
839	17189	16858	11.12.01
840	17189	18102	12.07.03

Table 2: Structure of TOE category-xref table
(the cross-references available at category “Freedom, being free” are highlighted).

Each record in the category-xref table represents a cross-reference in TOE from one category to another. Such a cross-reference indicates a related category that may be of interest to the user, too, but is found in another branch of the taxonomy. The table for

these cross-references contains four columns in total:

- `xid`: This column acts as primary key.
- `catid`: This column acts as foreign key. Such a key links one table to another by means of a reference to a primary key (*W3Schools.com*, ‘SQL Foreign Key’). In this case, the column values refer to the primary key of the TOE category table. The categories indicated here are those at which a cross-reference is made.
- `refid`: This column, too, acts as foreign key to the TOE category table. The categories indicated here are those to which a cross-reference is made.
- `tnum`: The values of this column capture the location in the taxonomy of the category referenced in the `refid` column. (Note that this information is superfluous, as it can already be retrieved from the TOE categories table.)

To illustrate, the category “Freedom, being free” (`catid` 17189) has two cross-references: one to category “Absence of restraint, freedom” (`refid` 16858) and one to “Abstinence/exemption (from)” (`refid` 18102). These two categories referred to are found in another branch of the taxonomy than “Freedom, being free”. In other words, there exists no subordinate/superordinate relation between them. Hence, the cross-referencing mechanism is employed to indicate that, nonetheless, these categories have a related topic according to the editors.

lid	catid	prefix	word	catorder	et	notes	oflag	pflag	gflag	qflag
1	1		brytengrundas	1		ChristA 355	Y	Y	N	N
2	1		brytenwagas	2		ChristA 380	Y	Y	N	N
3	1		eormengrund	3		Beo 859	Y	Y	N	N
...
39486	17187		hēafod niman	1			N	N	N	N
39487	17188	=	(ge)hæftan	1			N	N	N	N
39488	17189		frēols	1			N	N	N	N
39489	17189		frēot	2			N	N	N	N
39490	17190		burhræden	1			Y	N	Y	N
39491	17190		burhscipe	2			N	N	N	N
39492	17191		bonda	1	bond		N	N	N	N
39493	17191		ceorl	2	churl		N	N	N	N

Table 3: Structure of TOE lexeme table
(the lexeme *frēols* that is found at category “Freedom, being free” is highlighted)

From the data it appears that cross-references in TOE occur between main categories only. No cross-references exist from one subcategory to another, from a main category to a subcategory, or vice versa. Thus, although we find “Freedom, being free” is related to “Absence of restraint freedom”, no cross-reference is made at one of its subcategories. It is likely that the editors of TOE deemed using cross-references for subcategories to be too fine-grained to indicate and maintain, and therefore kept such references confined to the main categories of the thesaurus. The third and last table of the TOE

legacy form is the lexeme table, depicted in Table 3.

Each record of the lexeme table represents an Old English lexeme that has been categorized based on one of its senses. The table contains eleven columns:

- `lid`: This column acts as primary key.
- `catid`: This column acts as foreign key to the TOE category table and assigns a lexeme, or rather one of the senses of a lexeme, to the category indicated.
- `prefix`: Values in this column, if filled in, can be “+” or “=”. These signs correspond to + and ± in the second edition of the Old English dictionary by Clark Hall (CASD)². Its introduction states the following:

Words beginning with *ge-* have been distributed among the letters of the alphabet which follow that prefix, and the sign + has been employed instead of *ge-* in order to make the break in alphabetical continuity as little apparent to the eye as possible. The sign ± has been used where a word occurs both with and without the prefix.

This information on *ge-* prefixes has been superseded in TOE³. The current knowledge on prefix use can be deduced from the values in the `word` column.

- `word`: This column contains the head-form of each Old English lexeme. Optional segments of a word (which can be prefixes like *ge-*) are indicated between parentheses. See, for example, the lexeme with `lid` 39487 in Table 3.
- `catorder`: The values of this column indicate the order in which categorized lexemes are to be displayed that are located at the same category.
- `et`: This column contains etymological notes related to the lexeme. For instance, the Old English *ceorl* (`lid` 39493) developed into *churl* (OED, s.v. ‘churl, n.’).
- `notes`: This column contains notes. These typically mention how often or where a lexeme is found in the Old English corpus. Thus, the noun *eormengrund* (`lid` 3) is noted to be found on line 859 in the poem *Beowulf*.
- `oflag`: This column represents one of the distribution flags of TOE. When the value “Y” is recorded, the word form of the lexeme in question – not in any one specific sense – is marked as “very infrequent” in the Old English corpus.

² Information gained in personal correspondence with prof. Marc Alexander (6 August 2017).

³ One example of knowledge in the `prefix` column being outdated is found with the lexeme with `lid` 582. The `prefix` column suggests the *ge-* prefix of this lexeme is mandatory (+), but the `word` column indicates that is no longer considered to be the case: “(ge)mȳþe”.

- `pflag`: A distribution flag marking those word forms found only in poetry.
- `gflag`: A distribution flag marking those word forms found only in glosses.
- `qflag`: A flag marking word forms as “highly dubious” (TOE, ‘Distribution Flags’).

To illustrate, the lexeme *frēols* has a sense categorized as belonging to category 17189, “Freedom, being free” (see `lid` 39488). This lexical sense is meant to be displayed as the first one of this category, with the synonymous sense of *frēot* (`lid` 39489) as the second one. The word-forms of *frēols* are not marked as occurring very infrequently in the Old English corpus, in poetry only, in glosses only, or as questionable.

The lexeme table of TOE is rather inefficient for editorial purposes. Each record provides information for a lexeme (such as its head-form, and the distribution of its word forms) but also for a specific sense of that lexeme (such as its placement in the topical system). In fact, the `lid` value of each record is not unique per lexeme. Instead, it is unique per lexical sense. Information on a lexeme is therefore often recorded multiple times and in multiple locations – in a record for each of its senses. When a structure allows redundancy of information, consistency is more difficult to ensure. Contradictory statements are certainly present in the current dataset⁴. Such defects will not be magically mended by porting TOE to linguistic linked data. What the process will do, however, is make a clearer distinction between lexemes (or lexical entries) and lexical senses, which may improve detection of inconsistencies.

4. Linguistic Linked Data Form

A linguistic linked data form for topical thesauri should reuse standardized terminology in order to be interoperable. OntoLex-Lemon and SKOS are highly suitable to this end for capturing both lexical items and a hierarchy of concepts that represent the topical system of a thesaurus. Content from TOE can thus be published on the Web in a form that is machine-interpretable and understood in a wider community. Figure 2 charts, in a coarse manner, the relation between the content from the TOE sample and the linked data terminology from SKOS and OntoLex-Lemon. The relation `a` in this figure, and throughout this paper, is shorthand for `rdf:type` and can be read as “is a” or “is of type” (*RDF 1.1 Turtle*). As can be seen in Figure 2, a categorized lexeme corresponds with a `LexicalSense` in the `ontolex` module from OntoLex-Lemon. Similarly, a TOE category corresponds with a `LexicalConcept`. Thus, the Old English words *frēols* and *frēot* have lexical senses that lexicalize the concept “Freedom, being free”. Superordination between concepts, such as between “Power, control, sway” and “Power,

⁴ The noun *earfopsīþ*, for instance, has two categorized senses in TOE (`lid` 22631 and 32588). Their registered `pflag` values contradict one another – “Y” and “N” respectively – even though both senses share their word forms and the distribution of these forms.

might”, is indicated through the broader relation from SKOS. A more thorough list of linked data terminology and corresponding TOE content is available in Table 4. Most of the TOE table elements translate directly to linked data counterparts, although there are a few exceptions. These exceptions, discussed below, are taken into account in the linked data form that is proposed for the content of TOE.

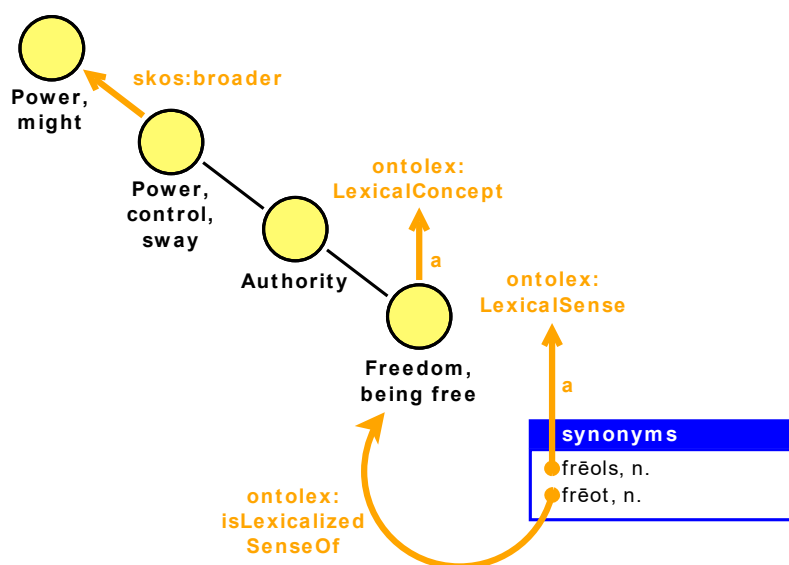


Figure 2: Sample of TOE content and its relation to linked data terminology from OntoLex-Lemon and SKOS

Firstly, some TOE content is not meant to be made available to the general public. Three elements are purely editorial in nature: the *notes* column from the category table and the *et* and *notes* columns from the lexeme table⁵. Various other elements are redundant or have been superseded. These bits may have been useful to the editors during the task of compiling the TOE dataset, but retaining them will likely prove detrimental or confusing. A case in point is the *catorder* column of the lexeme table. Although its values may aid in presenting synonymous senses in the desired order, they do not assist in determining the order for any given selection of senses. As the order of co-ordinate senses in TOE is a largely alphabetical one (with slight adjustments to take into account optional segments, length marks, and symbols specific to Old English), it would be possible – and preferable – to allow visualizations to determine the order of any selection of senses based on their head-forms. To this end, a label intended specifically for machines to order lexemes and their senses according to straightforward string comparison mechanisms (i.e., on ASCII characters only) would be easy to implement and utilize. The *prefix* column from the lexemes table, too, contains

⁵ Information gained in personal correspondence with Prof. Jane Roberts (30 August 2017).

information that may best be left unshared with users. Its values are no longer current and can, especially if juxtaposed with the prefix information encoded in the word column, confuse users by contradictory statements on whether word forms of a particular lexeme existed with or without the *ge-* prefix. The aforementioned bits of information that are not meant for public consumption should not be part of any publication – including one in a linguistic linked data form.

Secondly, the TOE dataset is in some places more explicit than needed and less explicit in others. The category table, for instance, does not contain a column that explicitly captures the unique id (i.e., a `catid` value) of a superordinate category. As a result, subordination of categories needs to be deduced by means of combining the information from the identification columns – `t1` to `t7` and `subcat` – and comparing the identification values between categories. Storing the identification information separated over various columns hinders both retrieval of the identification string for a category and subsequent comparison of two such strings. Therefore, superordinate categories will be connected explicitly for the linguistic linked data form of TOE. Moreover, the identification string of each category will be stored and offered in a concatenated form rather than broken up in several segments⁶.

Thirdly, the TOE dataset conflates information on lexical senses and lexemes into a single structure: the lexeme table. The linked data terminology from OntoLex-Lemon disentangles these two notions, calling the former a `LexicalSense` and the latter a `LexicalEntry`. As the primary key of the lexeme table is unique per sense of a lexeme, each of these records is associated with a `LexicalSense` rather than a `LexicalEntry`. Although the existence and name of a `LexicalEntry` can be deduced from the TOE lexeme table, the TOE dataset contains insufficient information to determine which senses belong to the same lexical entry. According to the specification of OntoLex-Lemon, words “may be different lexical entries if they are distinct in part-of-speech, gender, inflected forms or etymology” (OntoLex). Although TOE indicates the part of speech per lexical sense (i.e., via the `pos` column in the category table), the thesaurus does not currently indicate their gender or inflected forms. As such, a `LexicalEntry` will be created for each `LexicalSense` until information is made available in the future on which of these deduced lexical entries are meant to be one and the same. Such information can be compiled and offered by parties other than the editors of TOE, owing to the new linked data form of the dataset⁷.

⁶ The reason as to why the TOE category table does not store its identification information in a concatenated string but spread over multiple columns is likely found in the development process of the thesaurus, which saw shifts in the technologies used and the identification for categories (TOE, ‘Creation of the *Thesaurus*’). One change in the identification system, for instance, is that subcategories have been provided with numbering since the first electronic edition.

⁷ Asserting an `owl:sameAs` relation between two `ontolex:LexicalEntry` instances will effectively indicate that the two are to be considered one and the same entry.

Lastly, some of the contents of TOE require linked data terminology that is more specific than that found in SKOS and OntoLex-Lemon alone. To illustrate, a label used to aid computers in determining the presentation order of senses may be a `hiddenLabel` according to SKOS. Such hidden labels are intended for machine processing rather than for people to read. However, the hidden label for TOE should convey that it is specifically meant for the purpose of ordering rather than, for instance, searching alternative spellings. For this label, a new linked data term has been coined for TOE that extends the standardized terminology from SKOS. This coined term can be found in Table 4, including the terminology from SKOS that it extends (indicated through the ‘>’ symbol). Next to this need specific to TOE, two other aspects of this thesaurus are in need of being captured in linked data – aspects shared by a great number of topical thesauri (Stolk, 2019).

The first aspect common in topical thesauri is a division of their topical systems into conceptual levels. As mentioned above, TOE distinguishes two such levels in its database: main categories (simply called categories) and subcategories. The distinction of such levels has been deemed important enough to be included by editors. Indeed, for some thesauri, including TOE, the presentation and navigation mechanisms rely on these distinctions.⁸ For a linked data form of TOE, then, this conversion follows the recommendations outlined by the compact *lemon-tree* model, which offers relevant terms such as `ConceptualLevel` and `conceptualDepth` – analogous to how tree levels can be represented using the XKOS (a well-known extension to SKOS used for statistics).

A second aspect, shared by all topical thesauri, is that they categorize lexical items. This is true both for thesauri that group lexical senses into sets of near-synonyms and those that do not. The *lemon-tree* model recognises the need to capture this loose form of categorization, for which it offers the `isSenseInConcept` property and indicates its relation to OntoLex terminology: the *lemon-tree* property is stated to be a more generic form (or super property) of OntoLex `isLexicalizedSenseOf`. This most basic form of categorization found in topical thesauri, then, can be automatically inferred by using the *lemon-tree* model alongside OntoLex for lexical senses in TOE that are asserted to lexicalize a given SKOS `Concept`. Figure 3 illustrates the resulting form for the sample content of TOE used throughout this paper. A combined presentation of this sample content is available in Figure 4. Prefixes are used to abbreviate the namespaces of data vocabularies, for which a mapping is provided in Table 5.

⁸ Levels more abstract in nature are typically meant to be navigated first and allow the user to make greater semantic strides, as it were, than conceptual levels more specific in nature.

Linked data property	Value obtained from legacy form TOE
ontolex:ConceptSet	
skos:prefLabel	The name of the lexicon as a whole (i.e., "Thesaurus of Old English")
tree:conceptualLevels	An ordered list of the category types distinguished in the lexicon
skos:Collection > tree:ConceptualLevel	
skos:prefLabel	The name of the category type (i.e., "Categories" or "Subcategories")
tree:conceptualDepth	The conceptual depth of the category type
skos:member	The URI for a category belonging to this category type
ontolex:LexicalConcept	
skos:prefLabel	The name of the category
skos:broader	The URI for the superordinate category
skos:notation	The identification of the category
skos:related	The URI for a cross-referenced category
skos:inScheme	The URI for the lexicon as a whole (see ontolex:ConceptSet)
skos:topConceptOf	The URI for the lexicon as a whole (property applicable only to the top-most categories in the lexicon)
ontolex:LexicalEntry	
skos:prefLabel	The name of the lexeme
skos:hiddenLabel > toe:orderLabel	The name of the lexeme, rewritten so as to enable computers to sort these variants alphabetically by conventional means
rdf:type	The URI for the class indicating the part of speech of the lexeme
rdf:type	The URI for the class indicating the distribution of the word forms of the lexeme
ontolex:LexicalSense	
skos:prefLabel	The name of the categorized lexeme
ontolex: isLexicalizedSenseOf	The URI for the category at which the categorized lexeme has been positioned (and is therefore known to lexicalize)
ontolex:isSenseOf	The URI for the ontolex:LexicalEntry associated with the lexeme

Table 4: Linked data terminology and corresponding TOE content
(grey rows across the width of the table state the type of resource that will be formed;
subsequent rows indicate which properties will be used to capture information for that
resource and what their value will be).

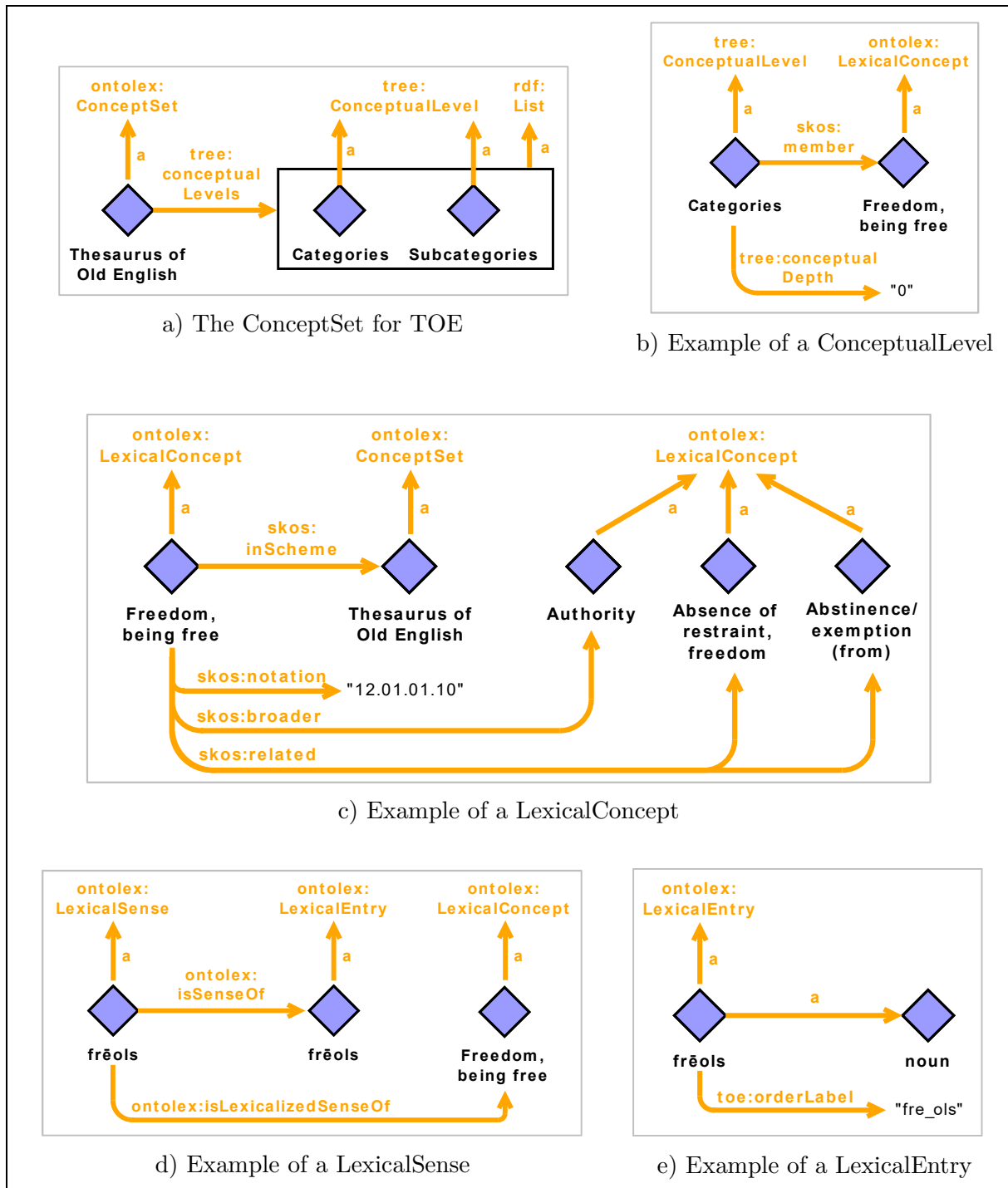


Figure 3: Linguistic linked data form of TOE
(diamonds represent linguistic linked data resources of TOE; arrows represent properties).

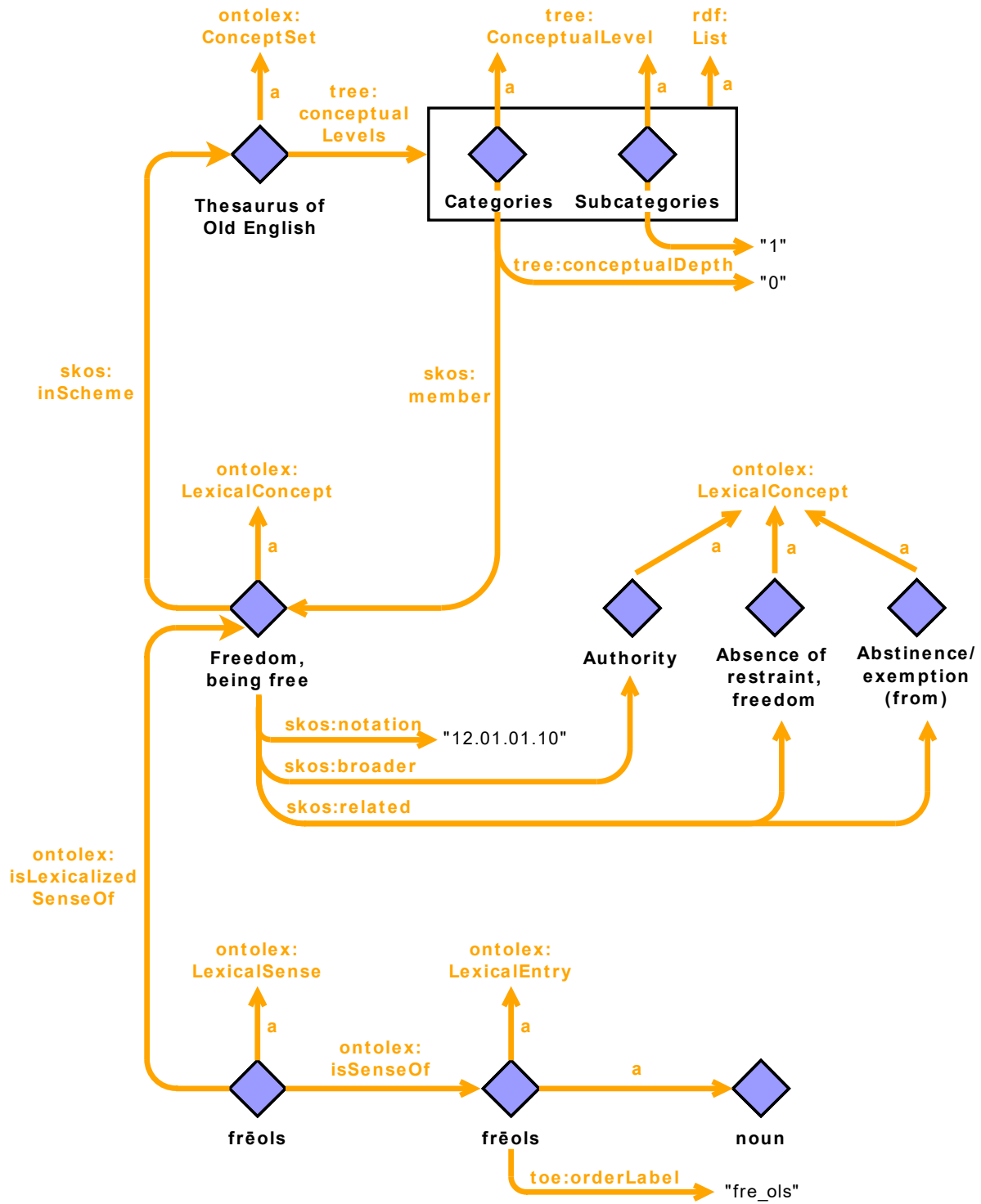


Figure 4: Linguistic linked data form of TOE (combining the examples provided in Figure 3).

Prefix	Namespace
ontolex:	http://www.w3.org/ns/lemon/ontolex#
owl:	http://www.w3.org/2002/07/owl#
rdf:	http://www.w3.org/1999/02/22-rdf-syntax-ns#
rdfs:	http://www.w3.org/2000/01/rdf-schema#
skos:	http://www.w3.org/2004/02/skos/core#
toe:	http://oldenglishthesaurus.arts.gla.ac.uk/
tree:	http://w3id.org/lemon-tree#

Table 5: Namespaces.

One further aspect needs to be discussed on bringing TOE content to the Semantic Web: the identification of each resource formed from TOE content. Bits of information on the Semantic Web are identified by a URI, typically in the form of an HTTP address. This holds for terminology from data vocabularies such as SKOS and OntoLex-Lemon, but also for instance data using such terminology. Best practices for coining URIs state that they should be simple, stable, and manageable (CoolURIs; CHIPS; SGOH). The first requirement entails that URIs need to be short and easy to remember; the second that they ought to be independent of the technology used to retrieve or visualize the content (as the software used may change); and the third that issuing new URIs should adhere to a straightforward strategy so as to be able to manage and maintain published content. With these requirements in mind, the following URI strategy has been adopted for the linguistic linked dataset of TOE. Each URI will be formed out of the following segments:

1. the Web domain of TOE (i.e., <http://oldenglishthesaurus.arts.gla.ac.uk/>),
2. the type of content the URI denotes (e.g., category, sense, entry), and
3. a unique number or string provided by the legacy form, if available.

The TOE category “Freedom, being free” (with `catid` 17189) thus gets the URI <http://oldenglishthesaurus.arts.gla.ac.uk/category/#id=17189> for its corresponding `LexicalConcept`. The lexical sense of the lexeme *frēols* (with `lid` 39488) gets <http://oldenglishthesaurus.arts.gla.ac.uk/sense/#id=39488>. This strategy has an additional advantage: it is aligned with the URI strategy in place for categories in the electronic edition of TOE hosted by the University of Glasgow. As a consequence, one can simply enter the URI of a category in a browser to view human-readable documentation on it. Adding linked data support to the electronic edition of TOE, as hosted by the University of Glasgow, is thus possible in the future without demanding a review or rework of the existing presentation. Having discussed both the original form of the TOE data and the desired linguistic linked data form, this paper will now turn to the conversion method employed to transform the former into the latter.

5. Conversion Process

Free digital tools already exist that facilitate a transformation from data in a tabular format to a linked data form. In selecting appropriate tools for the conversion of TOE from its legacy form to its desired linguistic linked data form, a number of requirements on the process need to be taken into account. These requirements, based on the premise that conversions ought to be reproducible by scholars with minimal effort, are listed in Table 6 and have been categorized according to priority⁹. Two requirements are mandatory, since these ensure an accurate conversion. The first is that the conversion process must accept tabular input either in an Excel spreadsheet or CSV format and provide transformed output in the RDF format (M1). The second requirement is that the process must be able to apply logic that relates the structure of the source to terminology from the desired linked data vocabularies (M2). The conversion logic for the TOE data has been described in Table 4. This logic also demands combining information from multiple tables, available in separate files. To illustrate, most of the information for lexical entries according to OntoLex-Lemon is found in the lexeme table of TOE. The part of speech of such an entry, however, is registered in another table of TOE: the category table.

Next to the requirements that are mandatory, three others have been formulated to which the process should adhere. Although not mandatory for an accurate outcome, these three requirements are geared towards increasing the maintainability and user-friendliness of the process. Firstly, the process should accept conversion logic in a form that has been standardized and is application-independent (S1). The alternative – relying on a format specific to a single tool – would limit the applicability, understandability, and reusability of the captured logic. Considering the availability of specific tooling and continued support from its creators are by no means guaranteed (as indeed seen for a number of conversion tools)¹⁰, great reliance on a single tool should be avoided. Secondly, the process should be executable by scholars without a background in software development (S2). To be more specific, it should be possible to obtain and install the necessary tools without first having to compile the source code. Moreover, the tools should provide a visual user interface rather than only a command-line execution mechanism. Lastly, the conversion process should be automatable so that it can be performed again with minimal effort after an update of the thesaurus data (S3).

The final requirement for the process, assigned a lower priority than the foregoing ones, is meant to facilitate deploying and utilizing the resulting linguistic linked data. Web-based platforms will be able to retrieve and query information from a thesaurus if its

⁹ The requirement prioritization follows the MoSCoW principles, developed by Dai Clegg et al. (1994).

¹⁰ Availability and support for the tools AnnoCultor, Aperture, and NOR2O have been discontinued.

conversion output has been stored in a database that facilitates access for linked data technology (C1). A database for linked data content is called a triplestore. Triplestores typically allow accessing their stored content via queries using the standard querying language SPARQL, which web applications can use to interact with the data.

Must have	
M1	Accept required input and output formats
M2	Apply required logic for conversion
Should have	
S1	Employ standardized form for logic
S2	Allow for scholars to perform each step
S3	Allow for automation of all steps involved
Could have	
C1	Store output in a triplestore with a query endpoint

Table 6: Requirements on the conversion process, categorized according to priority

The W3C provides a convenient overview of a number of tools that convert data into RDF (*ConverterToRdf*). Eighteen free tools listed there comply with requirement M1. These tools are listed in Table 7. Five of them appear to be discontinued, that is, they are no longer maintained or offered for download. Nine others do not comply with M2, either because they do not allow applying logic other than their default (Apache Any23) or because they cannot combine information from tables found in separate input files (RDF123; RDF Refine; csv2rdf4lod; Anzo for Excel; TabLinker; Excel2rdf; Sheet2RDF; Spread2RDF). The remaining four tools, then, conform to both mandatory requirements and should be able to convert the TOE legacy form into a linguistic linked data form. These tools are Datalift, Tarql, Virtuoso Sponger, and XLWrap.

One of the four remaining candidate tools for converting TOE data fails to meet requirement S1. This tool, XLWrap, defines its own form for capturing conversion logic, rather than using a standardized form (Langegger, 2017). A number of standardized forms for capturing conversion logic have been recommended by W3C. Two of these are specifically intended for logic converting tabular data into RDF: CSVW and R2RML. Unfortunately, these two forms are unsuitable for the conversion of TOE. The former cannot be used to combine information from multiple input files. The latter facilitates only relational databases as input and cannot be applied to Excel or CSV files. In fact, the three remaining tools – Datalift, Tarql, and Virtuoso Sponger – facilitate transformations utilizing another logic form: SPARQL. This query language, standardized by W3C, allows selecting patterns from an RDF source and constructing new RDF data that adheres to desired patterns.

Software	M1	M2	S1	S2	S3	C1
AnnoCultor	<i>(discontinued)</i>					
Anzo for Excel	+	-				
Apache Any23	+	-				
Aperture	<i>(discontinued)</i>					
Convert2Rdf	<i>(discontinued)</i>					
csv2rdf4lod	+	-				
Datalift	+	+	+	+	-	+
Excel2rdf	+	-				
NOR2O	<i>(discontinued)</i>					
RDBToOnto	<i>(discontinued)</i>					
RDF Refine	+	-				
RDF123	+	-				
Sheet2RDF	+	-				
Spread2RDF	+	-				
TabLinker	+	-				
Tarql	+	+	+	-	+	-
Virtuoso Sponger	+	+	+	-	+	+
XLWrap	+	+	-	-	+	-

Table 7: Software tools and the requirements they meet

The way in which SPARQL is used differs between Tarql on the one hand and Datalift and Virtuoso Sponger on the other. Tarql employs a unique approach by running SPARQL directly on CSV input rather than on RDF data. It does this by emulating patterns have been found based on the tabular input. Datalift and Virtuoso Sponger employ SPARQL in a two-step transformation. First, these tools apply a default, direct mapping to obtain RDF data that is “often more geared towards describing the structure of the data rather than the data itself” (Lefrancois et al, 2017)¹¹. This RDF data can subsequently be transformed to RDF data that uses the desired data vocabularies. In this second step, SPARQL (the standard query language for RDF data) is used to select patterns from the RDF source and construct new RDF data that adheres to the desired patterns. Indeed, this two-step approach is one that can be performed by end-users (using tools such as Datalink) but can also be automated (using a direct mapping application and any triplestore that supports SPARQL queries).

¹¹The alternative solution proposed by these authors, an extension to SPARQL, appears promising but has not been accepted yet as part of the SPARQL standard proper.

Moreover, this two-step approach is also applicable to formats other than CSV, which may well suit future conversions beyond TOE. The conversion process for TOE, then, will employ the following generic steps:

1. obtain an RDF graph that expresses the structure of the input data
2. store the RDF graph in a triplestore
3. obtain the RDF that adheres to the desired linguistic linked data form through SPARQL queries

Taking these steps will also ensure that the last of the requirements, C1, is met. In other words, the desired linguistic linked data form that has been obtained will be available for queries by platforms that intend to visualize or utilize the thesaurus information. In fact, these three generic steps, here applied to TOE data, should be applicable to the conversion of any topical thesaurus, including those with legacy formats other than tabular data.

For the tabular data of TOE, the first step of the conversion process can be performed by a number of tools. Apache Any23, CSVW implementations¹², Datalift, and Apache Jena all express the structure of such input data in a similar manner. The default logic that these tools share when processing a CSV file is as follows. Firstly, these tools create a node in RDF for each record from the input. Secondly, they add a relation to that node for each of the filled in cell values they encounter. The identification of this relation (i.e., its URI) ends in the column name¹³. An example snippet of such output can be found in Listing 1. To obtain such results using Jena, one simply has to install Apache Jena and run the following command (adjusted to the desired input filename and the output filename):

```
> riot "input.csv" > "output-graph.ttl"
```

¹² See the CSVW report for a list of implementations (*CSVW Reports*).

¹³ The initial letter of the column name is capitalized in the case of Apache Any23.

```
_:S39488 <file://C/lexemes.csv#lid> "39488" ;
<file://C/lexemes.csv#catid> "17189" ;
<file://C/lexemes.csv#word> "frēols" ;
<file://C/lexemes.csv#catorder> "1" ;
<file://C/lexemes.csv#oflag> "N" ;
<file://C/lexemes.csv#pflag> "N" ;
<file://C/lexemes.csv#gflag> "N" ;
<file://C/lexemes.csv#qflag> "N" ;
.
```

Listing 1: Snippet of RDF generated in the first step of the conversion process, based on the record for one of the senses of frēols (lid 39488) and expressed in the Turtle syntax.

The second and third steps of the conversion process require a triplestore. For this paper, the RDF4J triplestore is used to illustrate these steps. RDF4J offers a web-based interface, which allows users to set up a new repository for RDF content (see Figure 5) and therein store the intermediate RDF graphs obtained in step 1 (see Figure 6). Each of the graphs is assigned its own context in the repository, which will allow queries in the next step to select content accurately. Table 8 specifies the contexts used in the conversion process.

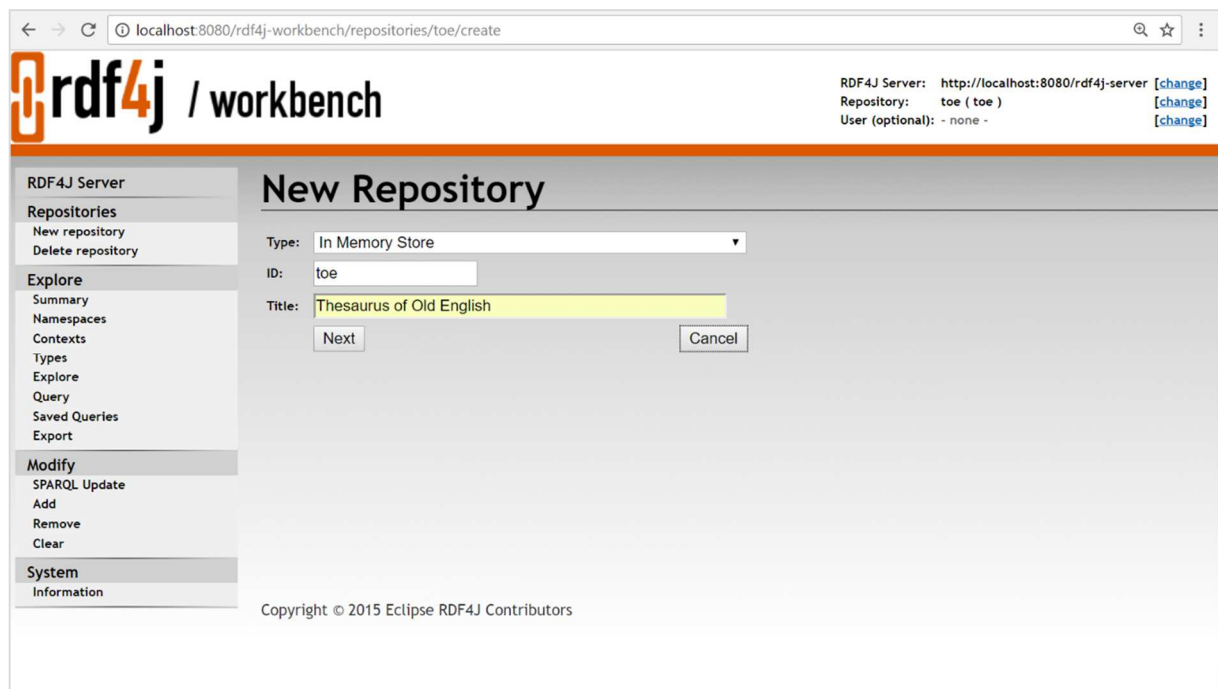


Figure 5: Creating a repository for TOE using the RDF4J user interface

Table of origin	Context
TOE category	<urn:toe:input:category>
TOE category-xref	<urn:toe:input:category-xref>
TOE lexeme	<urn:toe:input:lexeme>

Table 8: Contexts used upon adding RDF to the triplestore.

Figure 6: Adding RDF data to TOE categories using the RDF4J user interface.

In the third conversion step, queries are used to transform the available content in the repository to the desired linguistic linked data form. Such queries, written in SPARQL, can be executed via the RDF4J user interface (see Figure 7). Each query specifies a specific pattern that needs to be matched in the available content (in the WHERE clause of the query) and specifies another pattern that should be added as a result for each match (in the INSERT clause). Thus, patterns from the graph content of TOE can be transformed to patterns that conform to the desired outcome.

After the conversion, the resulting RDF will be available for querying and visualization. The intermediate RDF graphs that are uploaded in step 2 can be removed from the triplestore in order to ensure that only the final, desired form of the TOE dataset is indeed available in the repository. Automating the entire conversion process is also possible by means of a batch file. Both the batch file and queries that have been

employed in the conversion of TOE have been made available on GitHub¹⁴.

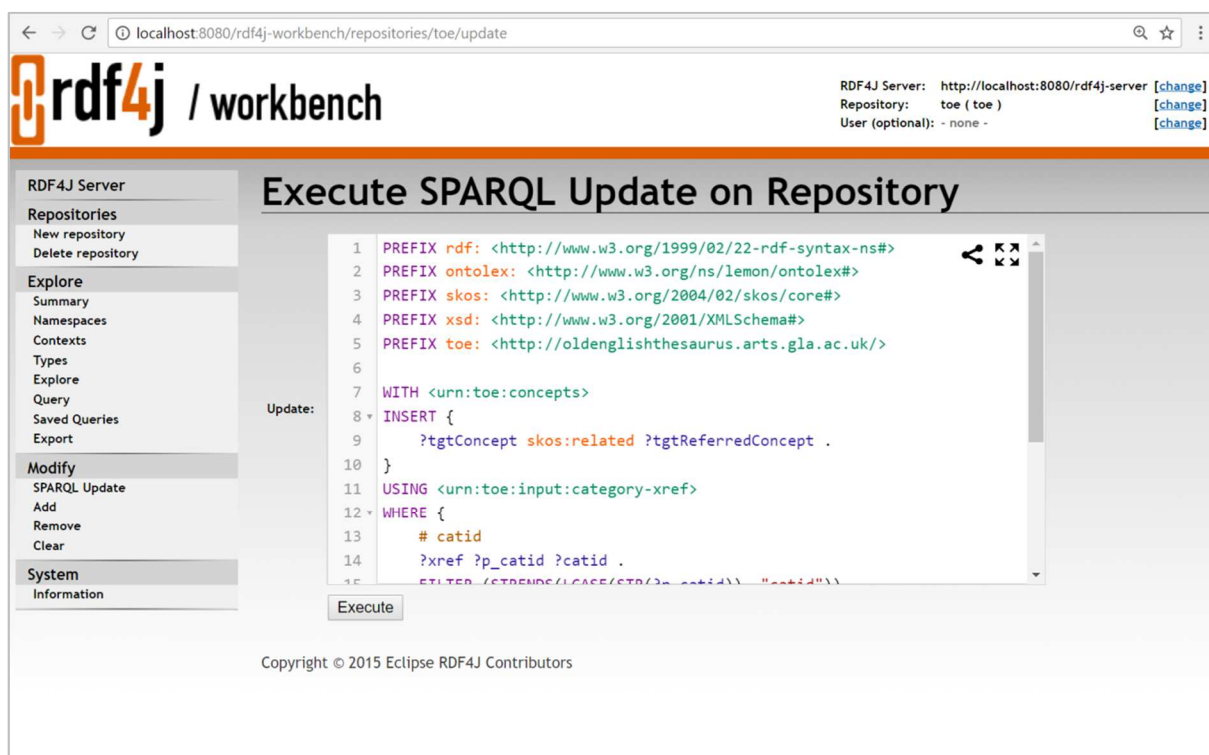


Figure 7: Executing a SPARQL update query using the RDF4J user interface

6. Conclusion

This paper has discussed the conversion of *A Thesaurus of Old English* from its legacy form to a linguistic linked data form utilizing OntoLex-Lemon, SKOS and *lemon-tree*. This conversion follows three steps: 1) obtaining an RDF graph that expresses the structure of the input data, 2) storing the graph in a triplestore, and 3) executing transformation logic using the standardized SPARQL language to produce the desired linguistic linked data form. Using SPARQL for capturing logic rather than a tooling-specific format ensures that the conversion process outlined does not rely on the existence of a single tool. Moreover, the three generic steps of the conversion process should be applicable to the conversion of any topical thesaurus – not just *A Thesaurus of Old English*. The results of the conversion discussed in this paper can be viewed in the online platform Evoke¹⁵.

The new digital form of the thesaurus is used in a number of projects in order to investigate whether linked data mechanisms can facilitate research into Old English language and culture. Some of these projects link lexical items with information to

¹⁴ <https://github.com/ssstolk/lld/toe/>

¹⁵ <http://evoke.ullet.net>

indicate their presence in a specific Old English text. Thus, subthesauri can be fashioned to look into specific contexts. Other projects establish links between existing lexicographic resources – connecting ones on Old Dutch and Old Frisian with the thesaurus. Doing so allows for reuse of the thesaurus macrostructure for other languages, but also for contrasting the degree of lexicalization present in these historical languages (e.g., the number of words that we know to have been available in Old Frisian to express a given concept compared to that for Old English). The findings of these and further projects will be presented at the Exploring Anglo-Saxon Eloquence pre-conference workshop at the 21st International Conference of English Historical Linguistics¹⁶.

7. Acknowledgements

The work described in this paper would not have been possible without the support of the Leiden University Centre for Digital Humanities for the Exploring Anglo-Saxon Eloquence project. Special thanks go out to the University of Glasgow, who have been kind enough to provide a license for working with the data of *A Thesaurus of Old English* and to give permission for distributing the resulting linked data form of the thesaurus on the Evoke platform.

8. References

- AnnoCultor*. Accessed at: <https://sourceforge.net/projects/annocultor/>. (4 June 2019)
- Anzo for Excel*. Accessed at: <https://supportcenter.cambridgesemantics.com/docs/glossary/Anzo-Excel>. (4 June 2019)
- Apache Any23*. Accessed at: <https://any23.apache.org/>. (4 June 2019)
- Apache Jena*. Accessed at: <https://jena.apache.org/documentation/io/>. (4 June 2019)
- Aperture*. Accessed at: <http://aperture.sourceforge.net/>. (4 June 2019)
- Bosque-Gil, J. et al. (2016). Modelling Multilingual Lexicographic Resources for the Web of Data: The K Dictionaries Case. In I. Kernerman et al. (eds.) *Proceedings of GLOBALEX'16 workshop at LREC'16*. Portorož, Slovenia, pp. 65–72. Available at: http://www.lrec-conf.org/proceedings/lrec2016/workshops/LREC2016Workshop-GLOBALEX_Proceedings-v2.pdf.
- Bremmer Jr, R. H. (2002). Treasure Digging in the Old English Lexicon, Review of *A Thesaurus of Old English*. *NOWELE*, 40, pp. 109–114.
- CASD: *A Concise Anglo-Saxon Dictionary for the Use of Students*. (1916). 2nd edition. New York: Macmillan.
- CHIPS: *Common HTTP Implementation Problems*. Accessed at: <https://www.w3.org/TR/chips/>. (9 June 2019)
- Clegg, D. & Barker, R. (1994). *Case Method Fast-Track: A RAD Approach*. Boston:

¹⁶ <https://icehl21.wordpress.com>

- Addison-Wesley.
- ConverterToRdf*. Accessed at: <https://www.w3.org/wiki/ConverterToRdf>. (20 December 2017)
- CoolURIs: *Cool URIs for the Semantic Web*. Accessed at: <https://www.w3.org/TR/cooluris/>. (9 June 2019)
- CSV2RDF: *Generating RDF from Tabular Data on the Web*. Accessed at: <http://www.w3.org/TR/csv2rdf/>. (9 June 2019)
- csv2rdf4lod*. Accessed at: <https://github.com/timrdf/csv2rdf4lod-automation/wiki>. 4 June 2019)
- CSVW Reports*. Accessed at: <https://w3c.github.io/csvw/tests/reports/>. (9 June 2019)
- Dance, R. (1997). Review of *A Thesaurus of Old English*. *Medium Ævum*, 66(2), pp. 312–313.
- Datalift*. Accessed at: <https://datalift.org/>. (4 June 2019)
- Declerck, T. et al. (2015). Towards a Pan European Lexicography by Means of Linked (Open) Data. In I. Kosem et al. (eds.) *Proceedings of eLex 2015*. Sussex, United Kingdom, pp. 342–355. Available at: http://www.dfki.de/web/forschung/iwi/publikationen/renameFileForDownload?filename=eLex_2015_22_Declerck+etal.pdf&file_id=uploads_2536.
- DOE: *Dictionary of Old English: A to I online*. Accessed at: <http://www.doe.utoronto.ca>. (4 June 2019)
- Evoke. Accessed at: <http://evoke.ullet.net>
- Excel2rdf*. Accessed at: <https://github.com/waqarini/excel2rdf>. (4 June 2019)
- Görlach, M. (1998). Review of *A Thesaurus of Old English*. *Anglia* 116(3), pp. 398–401.
- Kay, C. & Alexander, M. (2016). Diachronic and Synchronic Thesauruses. *The Oxford Handbook of Lexicography*. Oxford: Oxford University Press, pp. 367–380.
- Khan, F. (2016). Representing Polysemy and Diachronic Lexico-semantic Data on the Semantic Web. In *Proceedings of the Second International Workshop on Semantic Web for Scientific Heritage co-located with 13th Extended Semantic Web Conference (ESWC 2016)*. Heraklion, Greece, pp. 37–46.
- Klimek, B. & Brümmer, M. (2015). Enhancing Lexicography with Semantic Language Databases. In *Kernerman DICTIONARY News*, 23.
- Lefrancois, M. et al. (2017). A SPARQL Extension for Generating RDF from Heterogeneous Formats. In *Proceedings of the 14th International Conference of the European Semantic Web Conference*. Portorož, Slovenia, pp. 35–50.
- Lemon-tree*. Accessed at: <https://w3id.org/lemon-tree>. (4 June 2019)
- MySQL 5.7 Reference Manual*. <https://dev.mysql.com/doc/refman/5.7/en/>. (4 June 2019)
- NOR2O*. Accessed at: <https://github.com/boricles/nor2o>. (4 June 2019)
- OED: *Oxford English Dictionary Online*. Accessed at: <http://oed.com>. (4 June 2019)
- OntoLex-Lemon: *Lexicon Model for Ontologies*. Accessed at: <http://www.w3.org/2016/05/ontolex/>. (9 June 2019)
- R2RML: *RDB to RDF Mapping Language*. Accessed at:

- <http://www.w3.org/TR/r2rml/>. (9 June 2019)
- RDF123. Accessed at: <http://ebiquity.umbc.edu/project/html/id/82/RDF123>. (20 December 2017)
- RDF Refine. Accessed at: <http://refine.deri.ie/>. (20 December 2017)
- Roberts, J. (1978). Towards an Old English Thesaurus, *Poetica* 9, pp. 56–72.
- Roget: *Thesaurus of English Words and Phrases, Classified and Arranged so as to Facilitate the Expression of Ideas and Assist in Literary Composition*. (1852). London: Longman.
- SGOH: *Style Guide for Online Hypertext*. Accessed at: <https://www.w3.org/Provider/Style/URI>. (4 June 2019)
- Sheet2RDF. Accessed at: <http://art.uniroma2.it/sheet2rdf/>. (4 June 2019)
- SKOS: *SKOS Simple Knowledge Organization Reference*. Accessed at: <http://www.w3.org/TR/skos-reference/>. (9 June 2019)
- SPARQL: *SPARQL 1.1 Query Language*. Accessed at: <http://www.w3.org/TR/sparql11-query/>. (9 June 2019)
- Spread2RDF. Accessed at: <https://github.com/marcelotto/spread2rdf>. (4 June 2019)
- Spreadsheet-to-RDF Wrapper. Accessed at: <http://xlwrap.sourceforge.net/>. (4 June 2019)
- Stolk, S. (2019). Lemon-tree: Representing Topical Thesauri on the Semantic Web. In M. Eskevich et al. (eds.) *Proceedings of the 2nd Conference on Language, Data and Knowledge (LDK 2019)*. Leipzig, Germany. Accessible at: <https://doi.org/10.4230/OASIS.LDK.2019.16>.
- TabLinker. Accessed at: <https://github.com/Data2Semantics/TabLinker>. (4 June 2019)
- TOE: *A Thesaurus of Old English*. Accessed at: <http://oldenglishtesaurus.arts.gla.ac.uk>. (4 June 2019)
- Turtle: *RDF 1.1 Turtle*. Accessed at: <https://www.w3.org/TR/turtle/>. (9 June 2019)
- W3Schools.com. Accessed at: <https://www.w3schools.com>. (9 June 2019)
- XKOS: *An SKOS Extension for Representing Statistical Classifications*. Accessed at: <http://www.ddialliance.org/Specification/XKOS/1.0/OWL/xkos.html>. (9 June 2019)

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



SASA Dictionary as the Gold Standard for Good Dictionary Examples for Serbian

**Ranka Stanković¹, Branislava Šandrih¹, Rada Stijović²,
Cvetana Krstev¹, Duško Vitas¹, Aleksandra Marković²**

¹ University of Belgrade, Studentski trg 1, Belgrade, Serbia

² Institute for Serbian Language, SASA, Knez Mihailova 36, Belgrade, Serbia

E-mail: ranka@rgf.rs, branislava.sandrih@fil.bg.ac.rs, rada.stijovic@isj.sanu.ac.rs,
cvetana@matf.bg.ac.rs, vitas@matf.bg.ac.rs, aleksandra.markovic@isj.sanu.ac.r

Abstract

In this paper we present a model for selection of good dictionary examples for Serbian and the development of initial model components. The method used is based on a thorough analysis of various lexical and syntactic features in a corpus compiled of examples from the five digitized volumes of the Serbian Academy of Sciences and Arts (SASA) dictionary. The initial set of features was inspired by a similar approach for other languages. The feature distribution of examples from this corpus is compared with the feature distribution of sentence samples extracted from corpora comprising various texts. The analysis showed that there is a group of features which are strong indicators that a sentence should not be used as an example. The remaining features, including detection of non-standard and other marked lexis from the SASA dictionary, are used for ranking. The selected candidate examples, represented as feature-vectors, are used with the GDEX ranking tool for Serbian candidate examples and a supervised machine learning model for classification on standard and non-standard Serbian sentences, for further integration into a solution for present and future dictionary production projects.

Keywords: Serbian; good dictionary examples; automatization of dictionary-making; feature extraction; machine learning

1. Introduction

1.1 The aim of the paper

This paper outlines an approach to providing support for building different kinds of monolingual descriptive dictionaries of the Serbian language. The approach was motivated by the need for modernization of the dictionary-making process for the dictionary of the Serbian Academy of Sciences and Arts (SASA), a large monolingual thesaurus of Serbian, as well as for the production of new dictionaries of Serbian. The SASA dictionary is still developed traditionally, and its modernization could serve various different goals: speeding up the dictionary-making process, but also the development of a lexical database as the source for building new dictionaries of Serbian.

In the e-lexicography era, with the imperatives of faster dictionary-making and “smart lexicography”, special attention is devoted to semi-automatic selection of dictionary examples from corpora, and the presented approach supports the selection of dictionary examples making the process of dictionary development faster and more productive.

1.2 The role of dictionary examples

Dictionary examples play an important role in dictionary entries and they constitute, according to some authors, a “key microstructural element” of a dictionary (Kosem, 2017: 183). A good example is valuable from the aspects of both language reception and production. Examples have different roles, some of which are mentioned by S. Atkins and M. Rundell: they can complement the definition and help the user understand the meaning of the headword/lexical unit (their informative value); they should show the typical and natural way of behaviour of a word: syntactic patterns, collocations, as well as its colligational preferences – preferred form(s) of the paradigm, or the position(s) in the sentence; and since examples should help the understanding of the definition, they must be easy to understand – which means that their syntactic structure should be simple and their lexis not too difficult and uncommon. Informativeness, typicality with naturalness, and intelligibility are basic criteria for good dictionary examples (see more on these criteria in Atkins & Rundell 2008: 458–461).

However, many metalexicographers point out that it is not easy to find good dictionary examples in corpora. Kilgarrieff et al. (2008: 429) note that reading concordances is “an advanced linguistic skill”, and “the point of reading concordances – to pick up the common patterns that a word occurs in – is itself an abstract and high-level task”. This task is difficult even for trained lexicographers. In addition, finding good examples is time-consuming. The corpora are very big nowadays, the number of concordances one gets for a keyword is often too large, and it is impossible to read all of them. All this was the motivation for the development of GDEX, a tool designed for extraction of good dictionary examples (Kilgarrieff et al., 2008), now used not only by lexicographers, but also in language teaching and learning.

1.3 SASA-Dataset

The SASA dictionary is conceived as a thesaurus, meant primarily for native speakers. Its primary goal is to help understanding words from different kinds of texts (receptive use of dictionary). It covers a large portion of the vocabulary of the Serbian language, standard and vernacular, for the last 200 years. In Zgusta’s terms, it is a combination of the standard- and overall-descriptive dictionary (Zgusta, 1971: 212), which means that all marked lexis (dialectal, archaic or dated, jargon, etc.), as well as non-standard phonetic, morphological and syntactic forms and types of complements are labelled.

Each dictionary entry contains (or may contain) several subentries (one subentry for each lexical unit), and their descriptive definitions (sometimes definitions by synonyms). Every definition is followed by several (2 to 6) illustrative examples (examples are listed chronologically), with precise bibliographic references.

The first volume of this dictionary was published in 1959 (the project itself has been underway since the last decades of the 19th century), and the last, 20th volume was published in 2017 (the total number of volumes planned is 35). This is a long-term, time-consuming project.

Although the process of dictionary-making continues in the traditional way, there have been several initiatives for its modernization and acceleration. Digitization of the published volumes began in 2016, and the first exploitation of two digitized volumes was reported in Stijović and Stanković (2017). Dictionary entries from five volumes were automatically parsed and stored as a structured text in a lexical database, which offers the opportunity to use this data for extraction of different kinds of knowledge, as well as knowledge about examples.

This data-driven approach, combined with lexicographic expert knowledge, is the basis for the improvement of dictionary example selection which will be useful both for the production of different dictionaries of Serbian and the forthcoming volumes of the SASA dictionary.

Section 2 describes some steps towards modernization of the dictionary-making process and the development of the digital version of SASA dictionary, starting with retro-digitization process, followed by several ideas about modernization of dictionary-making and the description of the current, traditional practice of dictionary example selection. Section 3 presents a part of the feature distribution analysis of examples from five SASA dictionary volumes, while a comparison with feature distribution in sentence samples extracted from corpora is given in Section 4. The research focused on the development of the initial components of a model for example selection is presented in Section 5, followed by ideas for future work and some concluding remarks at the end of the paper.

2. SASA Dictionary

2.1 SASA Dictionary retro-digitization

The first ideas how to modernize the work on the SASA dictionary came many years ago (Sabo & Vitas, 1989). These ideas were later revitalized and various possibilities for updating the work on this dictionary were considered (Vitas & Krstev, 2015; Ivanović et al., 2016). The modernization of work finally began only in 2016 with digitization of printed volumes (Stijović & Stanković, 2017). Out of 20 volumes already

published, three were available as MS Word files, two as pdf files and others only in paper form. At the same time, a formal description of dictionary entry was produced, and a lexical database model was developed (Stanković et al., 2018).

The conversion of the SASA dictionary from unstructured text into a lexical database consisted of a thorough analysis of formatting conventions that were used for typesetting dictionary entries, as well as identification of triggers (such as special words, abbreviations or punctuation marks) used to introduce specific information. This analysis enabled the recognition of the entry structure: headword group, grammatical data, etymology, lexical units (senses), multiword expressions and proverbs (if any). Each lexical unit may contain linguistic labels (domain, style, time etc.), syntax patterns, definitions, related words, examples of usage, followed by bibliographic references.

2.2 Towards modernization of SASA dictionary-making

Transformation of the digitized text of the SASA dictionary into various standard structured formats and a lexical database was implemented using a custom software solution, with the primary goal to speed up the linear production process of the dictionary. This enabled the use of the lexical database for research purposes. After successful import of two volumes: the 1st and 19th into the database (Stanković et al., 2018), the process continued with another three volumes: 2nd, 18th and 20th.

Dictionary entries are represented by lexical entry elements in the database, with one or more lexical senses (units) that are further illustrated by examples. Each example is followed by information about the bibliographic source, the author, and optionally about the location, and indirectly related to information about the headword of dictionary entry, its part of speech and linguistic labels assigned to the headword and lexical unit. A classification of labels is also incorporated in the database to provide clustering of dictionary (sub)entries using several criteria: by domain (for terminology and specialized vocabulary), by region (dialect), register, style etc. Interlinking of related words is envisaged as more explicit, on the level of lexical units (senses), which will enable the reuse of dictionary content that already exists in the database.

The fine-grained structure of the database enabled the creation of a dataset of examples supported by a set of related information: headword/lexical unit the example is related to, part of speech, and linguistic labels. The dataset of examples derived from the SASA dictionary is a dataset of good dictionary examples that can serve various purposes: it can be used to procure examples for the SASA dictionary as well as for new dictionaries, but it can also be used for the development of a machine system for example selection.

From the analysis of samples of dictionary examples, metrics and example feature distribution can be derived, which can reduce the search space for relevant examples, for example, by setting the upper and lower limits for sentence length, based on the

most common length of example (in words, tokens and characters). Also, having in mind that this dictionary includes citations from a 200-year period, a time boundary can be set when extracting examples for some future dictionary of modern language.

About 12% of all examples in the digitized volumes of SASA dictionary contain lexis marked as obsolete (label *заст.*), 7% as dialect (*дијал.*), 4% as irregular (*некњ.*), 2% as vernacular (*нар.*), 2% as ephemeral (*необ.*) and the remaining 2% marked with labels for other types of non-standard lexis, in total 29%. These figures are approximative, since some examples contain lexis marked with several labels, and for this analysis only the first of them was taken into account.

2.3 The current practice of dictionary example selection

2.3.1 Criteria for example selection

Finding appropriate examples in a citation bank as big as the one for the SASA dictionary¹ (about five million paper slips, hand- or typewritten, only recently scanned and partially annotated with headwords) is a difficult and time-consuming job – a lexicographer has to read hundreds, sometimes even thousands of citations (for example, there are 2,830 citations for the preposition *po* ‘on’, ‘over’, ‘by’). When choosing illustrative examples for lexical units (LU) in the SASA dictionary, lexicographers are not guided by linguistic criteria alone. We will describe here briefly some of other criteria, primarily extralinguistic ones. The corpus of examples in paper form, used for this monolingual thesaurus, was made up of excerpts from resources written in Serbo-Croatian (SC), from the beginning of the 19th century to the present day, as well as about 300-word collections (for details see Stanković et al., 2018). Written texts, as well as word collections, come from what used to be the SC language territory. According to the Style Guide², lexicographers have to choose two to six examples for each LU, taking into account the following facts: a) each example should clearly show the meaning of the LU; b) they have to be from different parts of SC language territory; c) they should be from different periods, and listed chronologically, the oldest being the first, while the examples from word collections are given at the end, after all the examples from published sources; d) they should be written by renowned writers. What is not written in the Style Guide (and lexicographers learn it by word of mouth) is that

¹ It is important to emphasize that the citation bank for the SASA dictionary constantly gets up-dated and thus continues to grow – in the course of dictionary-building lexicographers continually consult reference literature (encyclopedias, different kinds of dictionaries, manuals etc.), and some of the recently published books, text-books etc. are also excerpted. The SASA dictionary contains only a small portion of these citations because of the described selection criteria.

² Упутство за обраду Речника, Београд: Институт за српск(охрватск)и језик САНУ (рукопис), 1959. и (допуњено) 2017 [A Style Guide for Dictionary-Making, Belgrade: SASA Institute for Serbo(-Croatian) (manuscript), 1959 and (supplemented) 2017].

it is not advisable to use more than one example of the same author. An exception to this rule can be made if there are not enough examples by other authors.

Only a few linguistic criteria are mentioned in the Style Guide. They can be paraphrased as follows: 1) each chosen example should show different relations of the headword with other words (the rection, for example); 2) it is recommended that every example represents a finished syntactic whole – with a subject and a predicate. It is even possible to add a missing sentence constituent, but it has to be in square brackets, as a mark of this kind of editorial intervention. (Though the excerpts are in the form of full sentences, the context they provide is sometimes insufficient, and it is necessary to provide a wider context.) As the first criterion is very important, it needs a more detailed explanation. Namely, the role of the examples is to convey the information about valency and rection of the headword in an implicit way (explicit syntactic information, if required by the Style Guide, is placed before the definition). Since the Style Guide for the SASA dictionary was written during the 1950s, there is no mention of collocations or of using examples to show the most frequent ones.

2.3.2 Editorial interventions and a control corpus

Sometimes it happens that additional examples are needed for a sense or lemma. There are two scenarios in such a case: 1) Experienced lexicographers may rely on their knowledge and invent an illustrative example. If such an example is typical for the standard language, the source is marked by the abbreviation Ed. ‘Editor’. The example for the noun *pivnica*, ‘pub’ is of this kind: *Najbolje je točeno pivo u češkim pivnicama (Ped.)*. ‘The best is draft beer in Czech pubs’ (Ed.). 2) An editor may also provide an example from the non-standard language, which usually means that he/she comes from a specific region; in such a case, the source is marked by the abbreviation of the editor’s name.

Editor’s intuition may and should be supported by the corpus data. It is common for lexicographers to look for examples in the corpus of contemporary Serbian (SrpKor, developed by D. Vitas and a group of collaborators from University of Belgrade, <http://www.korpus.matf.bg.ac.rs/korpus/>), which is being used as a control corpus, but they rarely refer to it, although all concordances are associated with data about the source (Vitas & Krstev, 2012; Utvić, 2014).

2.3.3 Allowed and recommended interventions on examples from the corpus

Examples from the corpus may be modified by lexicographers. It is advisable to shorten sentences that are too long, and this kind of intervention should be marked by an ellipsis (“...”). It is allowed to omit all irrelevant sentence constituents (different kinds of modifiers, words in enumerations etc.) or even a whole subordinate clause, if it is not important for illustrating the LU. Here is an example from which the beginning of

the sentence, as well as the relative clause were omitted, being irrelevant for the verb headword: *[omitted: U VII., VI. i V. razredu veliki broj slabih učenika došao je otuda, što su] mnogi učenici [omitted: , koji su iz matematike cele godine imali dobre ocene,] na ispitu [inserted: su] podobivali slabe ocene* '[omitted: In the seventh, sixth and fifth grade the number of bad students increased since] many of the students [omitted: , who had good grades in Mathematics during the school year,] got poor grades on the exam'. The same example shows an inserted part in square brackets, namely "su", the simple present tense form of the verb *to be*, 3rd person plural, which was removed with the first omission. This insertion enabled the editor to form a correct sentence shorter than the original one.

2.3.4 Summary of interventions

The dataset from five dictionary volumes comprises ~60,000 dictionary entries with ~105,000 lexical units (senses). Around 11,500 dictionary entries have headwords with several (numbered) lexical units. In the observed dataset, 70% of data entries have examples. According to the analysed dataset, approximately 71% of the examples were not shortened, 22% were shortened once, 6% twice, and 1% more than twice. Words were inserted (to clarify the meaning or to complement what is missing) in 7% of observed examples, while 93% were without any insertion. In total: 66% of the examples were not modified, 20% had one shortening and no insertions, 6% more than one shortening and no insertions, while 5% had an insertion but were not shortened and 2% had both insertions and shortenings. The number of editorial examples was relatively small, and we have not used these in our test set.

2.3.5 What should a good example contain?

As Atkins and Rundell (2008) point out, there is plenty of evidence when a lexicographer works with corpus data, trying to record how a word behaves, but not all of it is relevant for the description of a word's behaviour. The concept of lexicographic relevance is based on Fillmore's theory of frame semantics. The idea behind the concept is that a proper way to describe a word means that all the constructions it participates in should be identified as well as "all those through which its full semantic potential is to be expressed" (Atkins & Rundell, 2008: 252) should be recorded in the lexicographic database. The concept of lexicographic relevance was illustrated by the analysis of verbs, nouns and adjectives, since any word of this kind "cannot be used correctly if the constructions in which it participates are not known" (*ibid.*). Frame semantics links the meaning of a word with the syntactic contexts in which it occurs. To determine what is relevant for the semantic analysis implies identifying lexicographically relevant sentence constituents for verbs, nouns and adjectives.

An important conclusion by Atkins and Rundell (2008: 272) is that grammatical

contexts for discovering relevant information about keywords may differ depending on their part of speech. For example, if the keyword is a noun, lexicographically relevant co-constituents are its modifiers (the prototypical modifier of a noun in Serbian is an adjective phrase) and complements. If the keyword is an adjective, it is important, too, to consider its modifiers (for example, an adverb) and complements (noun phrases or prepositional phrases). For a verb keyword, it is important to note all its complements (objects, subject and object complements etc.).

The notion of lexicographic relevance may also be applied to the selection of good dictionary examples. The constituents important for proper analysis of an LU are also important for its illustrative examples. All relevant modifiers and complements, which affect the meaning of the LU, should be contained in the illustrative example. If a noun has a complement that affects its meaning, the complement should be represented in the example: *Tada se javila u njega velika ljubav i velika podobnost **za slikarstvo*** (paraphrase: ‘In that moment he felt a great affection and a great talent **for painting**’)³. If a keyword is a verb that in one of its senses takes a subject or object complement, then, of course, this complement has to be represented in the example: *On me smatraše **izgubljenom ovcom*** ‘He considered me **a lost sheep**’.

It is important to emphasize that “lexicographic relevance relates to what is relevant for an LU, and not to a lemma” (Atkins & Rundell, 2008: 150). We find similar considerations in Popović (2003). The author also believes that modernization of the description of both syntax and lexicography of Serbian standard language is needed. He points out that it is necessary to establish a relation between syntactic and lexicographic description. As for dictionaries, they should take into account the syntactic distribution of lexemes. Words from major word classes should be treated as central for certain types of syntactic units and syntactic information should be given systematically.

2.3.6 Is a context given in one sentence example enough for all word classes?

Some additional remarks are necessary. Conjunctions in Serbian are often at the initial position of the sentence, demarcating its beginning and delimiting it from the context that precedes it (Popović, 2004: 276–277). In such a case, semantic identification of the conjunction requires the context of the sentence that precedes the one beginning with the conjunction. For example, the conjunction *i* ‘and’, in one of its senses in the SASA dictionary, signals that an utterance comes as a conclusion, explanation, etc. of the sentence it follows. In this case it is necessary to adduce both the sentence beginning with a conjunction and the one before it: *Obeća, da će ovih dana otići. I održa reč* (‘He

³ A similar, bad example, missing this kind of noun complement, is mentioned in Atkins & Rundel (2008: 460): *One woman in every two hundred is **a sufferer*** (of what?).

promised he would leave one of these days. And he kept his promise’).

3. The features of dictionary examples

3.1 The role of example features

In order to facilitate example selection an extraction tool for representative sentences was developed – Good Dictionary EXamples, GDEX (Kilgarrieff et al., 2008), used today not only by lexicographers, but also in language teaching and learning. In this paper we present research aimed at the development of a GDEX method for Serbian that ranks corpus sentences and suggests the most appropriate ones.

As the gold standard for the development of our method, dictionary examples from five out of twenty volumes of the SASA dictionary (Stijović & Stanković, 2017), presented in Section 2, were used. The main reason for choosing examples from this dictionary as the gold standard was the fact that they were manually selected by experienced lexicographers⁴. In the first phase we automatically analysed various lexical and syntactic features of the gold standard examples, classified them and compared the results with the control corpus (both gold and control corpus are presented in Section 4). The initial set of features was inspired by Kilgarrieff et al. (2008) and Kosem (2017), and guided by recapitulation of features given in Kosem et al. (2019).

3.2 Feature extraction

Feature extraction is enabled by the development of a web service inspired by the work described in Kilgarrieff et al. (2008), Kosem (2017), and Kosem et al. (2019), which can presently extract 41 features. The developed service receives a text snippet as a string (in our case a sentence), which can have additional metadata attached (e.g. source, keyword/headword, labels), and returns a dictionary⁵ structure comprised of feature names and their values. The list of requested features can also be customized. The system is envisaged to process both the sentences from corpora and dictionary examples extracted from the lexical database. In the text that follows, the term sentence will refer to both dictionary examples and sentences from the control corpus (Section 4).

The implemented set of features is described by metadata, i.e. several attributes are assigned to each feature: code, description, processing level (char, word, and sentence), headword dependency (yes/no), weight (for weighted sum and use in our future model

⁴ They were chosen according to the principles described in previous sections (2.3.1 to 2.3.6) of this paper. Since these examples have been subject to multiple check-ups (the dictionary-making process goes through several phases), they can be considered a gold standard.

⁵ In Python terminology.

for ranking), type (categorical or quantitative), types of graphical representation and visualization parameters (range, bins). The feature list is not conclusive, and in the future, as a result of the present analysis, other features could be added, and additional metadata assigned to features, such as an eliminatory data range, preferred data range and the like.

For this research a subset of 14 features is taken into consideration:

- Character-based:
 - sentence_length: Number of all characters
 - no_digits: Number of digits
 - no_weird_chars: Number of characters ("#\$%&\'()*+,-/:;<=>?@[\\]^_`{|}~'„" ...)
 - no_commas: Number of commas
 - no_punctuation: Number of all punctuation marks

- Token-based:

no_all_tokens: Number of all tokens (contiguous sequences of characters e.g. words, numbers, punctuation marks; produced using NLTK's recommended tokenizer ⁶ *nltk.word_tokenize* (Bird et al., 2009))

- avg_token_len: Average token length
 - max_token_len: Max token length
 - no_all_words: Number of all words (contiguous sequence of letters)
 - avg_word_len: Average word length
 - no_capitalised_words: Number of words that begin with uppercase, which are not at the beginning of the sentence
 - no_rare_tokens: Number of tokens with frequency threshold in the referent corpus
 - avg_freq_in_corpus: Average word frequency in the referent corpus
- Syntactic features:
 - no_pronouns: Number of tokens tagged as pronouns

The set of features that were computed, but not taken into account in this paper, includes: count of blacklisted words, does the headword occur more than once, the number of lemmas that appear multiple times, does the sentence contain between 15 and 40 tokens, number of tokens that contain both alphabetic and numeric characters, number of tokens tagged as proper names, POS-tag of the first word in the sentence, the position of the headword in the sentence, does the sentence begin with a word from a stoplist, etc. The only features that are specific for this exact research are counts of

⁶ <http://www.nltk.org/>

ellipsis (deletions from the original sentences), inserted segments and lexicographic labels, but they are used for example classification, not for ranking (see Section 4).

The analysis showed that a group of features can be used as filter features, namely, as strong indicators that a sentence should not be used as an example. Sentences that have at least one non-zero value for any feature belonging to this group are categorised as negative samples (e.g. `blacklist_count`, `contains_web_or_email`). Features that were not taken into consideration in this analysis were mostly dependent on the headword, e.g. its position in the example. They were classified as headword dependent and will be part of future analysis.

3.3 API for feature extraction

The extraction of features is implemented as a web service⁷. This web service is also used for other tasks, such as text classification and corpus cleaning.

An example of the activation of this web service using *curl* in Unix is the following:

```
curl -d '{"data": "We are demonstrating the usage of our feature extractor!", "lang": "en", "kwic": "usage", "feature_names": ["sentence_length", "avg_word_len", "no_all_tokens"]}' -H "Content-Type: application/json" -X POST http://147.91.183.8:12347/features
```

and the fields are:

- `data` (string) – mandatory, contains text for which features are being extracted
- `lang` (string) – optional (the default value is “sr” for Serbian, but most of the features can be extracted for English, as well)
- `kwic` (string) – optional (only for headword-dependent features)
- `feature_names` (list of strings) – optional (if omitted, returns list of all feature values)

For the given example, the output would be:

```
{"sentence_length": 56, "avg_word_len": 5.222, "no_all_tokens": 10}
```

⁷ Extraction of GDEX features, <http://gdex.jerteh.rs/>.

4. Feature analysis

4.1 The gold and the control dataset

Each example extracted from the SASA dictionary for the gold dataset is supplied with a list of supporting information: volume, dictionary headword, headword's part of speech, linguistic labels (some of which are mentioned in Section 2.2), type of editorial intervention (if any) on the example (shortening or insertion) and a code for the bibliographical source. The size of the gold corpus is 133,904 examples, comprising 1,711,231 words or 10,577,723 characters. Within the gold dataset three types of partitioning were used: 1) by published volume (labelled D01, D02, D18, D19 and D20), 2) by type of lexis/language (labelled with DSS for standard Serbian and DNS for non-standard Serbian) and 3) by part of speech (POS) of the headword/keyword (N – nouns, V – verbs, A – adjectives, ADV – adverbs and X – other).

DSS partition contains sentences in contemporary language with examples that were not modified by editors. We presume that they would be good examples for some future dictionary of contemporary Serbian. DNS contains examples in languages other than standard Serbian (Church Slavonic, Čakavian, Kajkavian), and lexis marked with labels some of which are mentioned in subsection 2.2 (obsolete, dialect, non-standard, vernacular, ephemeral, loanwords, slang). A small number of examples with uncertain boundaries of dictionary entry elements, usually in phrases and proverbs, were excluded from the research, as well as examples from poetry that have the " | " delimiter between verses.

In addition to the corpus made of examples, we prepared a control dataset derived from various texts, which was used as a sample corpus for dictionary example extraction. The control dataset of example candidates was obtained from the digital library Biblisha⁸ (Stanković et al., 2017), SrpKor – the corpus of contemporary Serbian (Vitas & Krstev, 2012; Utvić, 2014) and Serbian ELTeC Collection⁹. It consists of several text collections of different types, which reflect text variability. For the first collection with contemporary novels (labelled CN), the sentences were extracted from seven novels written by contemporary Serbian writers and from seven novels written in German and translated to Serbian. In order to represent domain knowledge, two scientific journals (labelled SJ) were used: *The Journal for Digital Humanities Infotheca*¹⁰ and *Underground Mining Engineering*¹¹. The sample labelled DP, with 17 issues of the daily

⁸ <http://jerteh.rs/biblisha/>

⁹ *Distant Reading for European Literary History* (COST Action CA16204)
<https://distantreading.github.io/ELTeC/srp/index.html>

¹⁰ <http://infoteka.bg.ac.rs/index.php/en>

¹¹ <http://ume.rgf.bg.ac.rs/index.php/ume>

newspaper *Politika* published in 2001–2010, was retrieved from SrpKor. A part of the Serbian ELTeC was used, which contains 10 novels and excerpts from 15 novels that were all published 100 or more years ago (labelled ON for old novels). The system for Serbian text processing, based on comprehensive e-dictionaries and local grammar in the form of finite-state automata (Krstev, 2008) was used for sentence segmentation.

Concordances were extracted using appropriate regular expressions, to serve as candidate examples for corresponding headwords in volumes to come. They were bound by sentence delimiters and left/right context of up to 500 characters. The size of the control corpus was 30,104 sentences, comprising 908,980 words or 5,841,700 characters. A sample of 2,752 candidate examples (taken from all parts of the control corpus) was manually evaluated by two lexicographers: they evaluated 1,434 examples as inadequate (useless), 723 as inadequate but improvable with major changes, 441 as good examples in which only minor changes are required, and 154 as very good examples for which no changes are required.

4.2 Feature distribution in the gold dataset of good examples

The comparison by volumes did not show any significant deviations. All feature distributions were similar, as expected, given the same guidelines and methodology used for all published volumes in the last 70 years. Figure 1 presents frequency distribution by number of words in the examples. On the left side each volume of SASA dictionary is represented by a histogram with parts of speech in different colours. The right-hand side shows histograms of partitions of the control dataset.

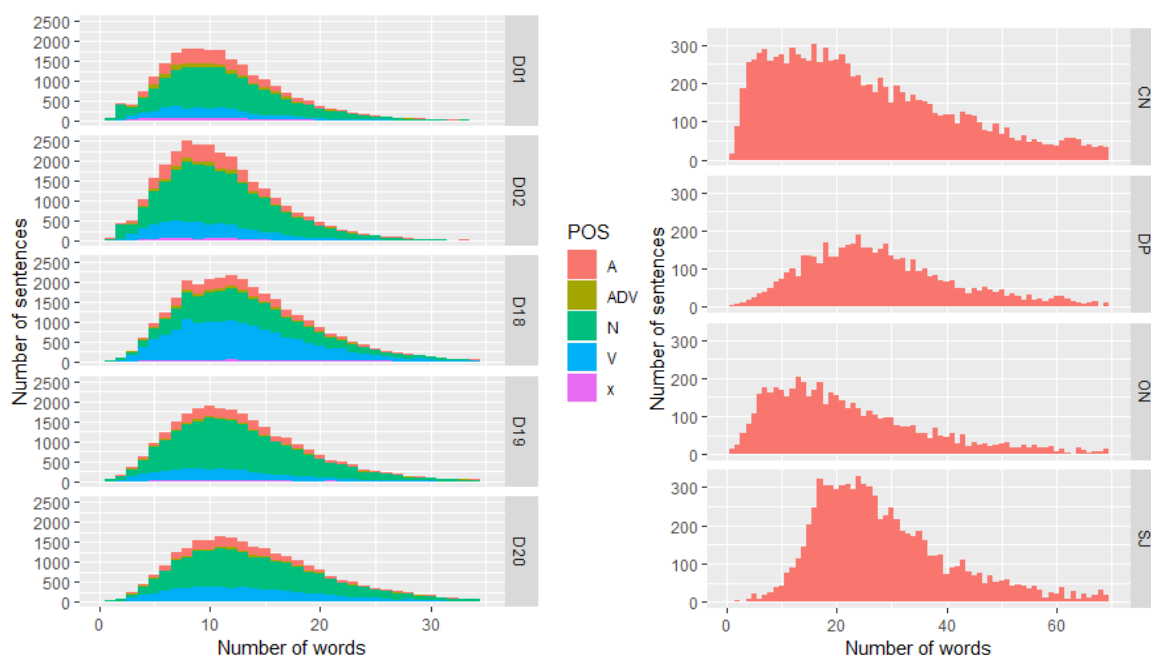


Figure 1: The histogram of the number of words in examples.

The histograms on the left show that sentences in the last three volumes tend to be slightly longer than in the first two, and that nouns (green) are the most numerous words. In volume D18 the number of verbs (blue) is considerably greater than in the other volumes, which can be explained by numerous verbs in this volume beginning with *o*, derived by the productive prefixes *od-* (allomorph *ot-*) and *o-*.

Comparison of lengths of examples for different parts of speech in the SASA dictionary shows that examples for adjectives and nouns tend to be longer than those for adverbs and verbs. Figure 2 presents corresponding boxplots, where the box represents the interquartile interval (IQR) with lower (Q1) and upper quartile (Q3), the middle bold line being the median (Q2), and the rhombus in the middle of the box presenting the average value, with POS on the x-axis and sentence/token length in characters on the y-axis. Dots present outlier examples longer than $Q3 + 1.5 \cdot IQR$.

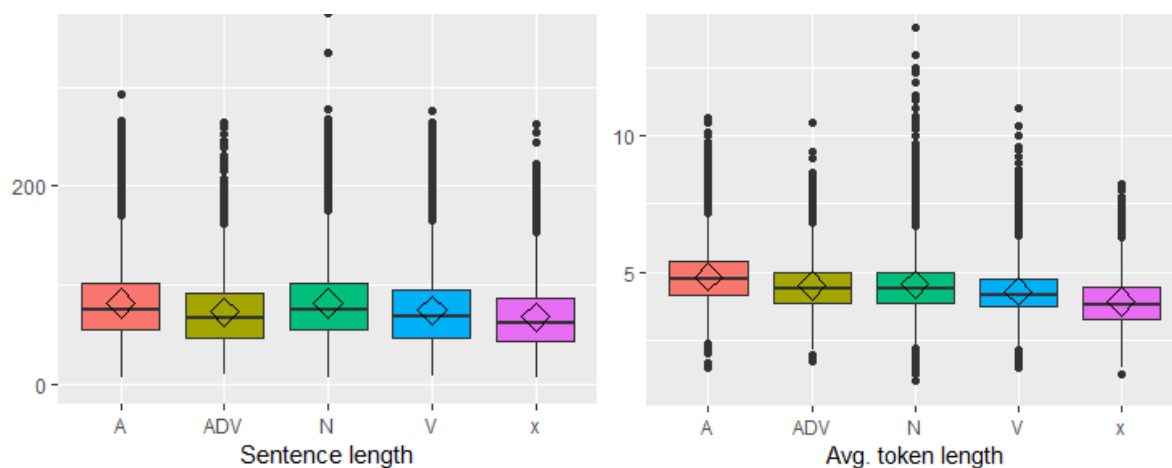


Figure 2: Boxplots showing sentence/token length per POS in the SASA dictionary.

4.3 Feature distribution on both corpora

Figure 3 presents a boxplot diagram of sentence length statistical values per partition (volume and text collection). It can be observed that the sentences in the control dataset partitions are longer than in any volume of the dictionary, that the dispersion for contemporary novels (CN) is the highest, that the average length of sentences in journals and daily papers is similar, and that old novels (ON) have shorter sentences than contemporary ones (CN).

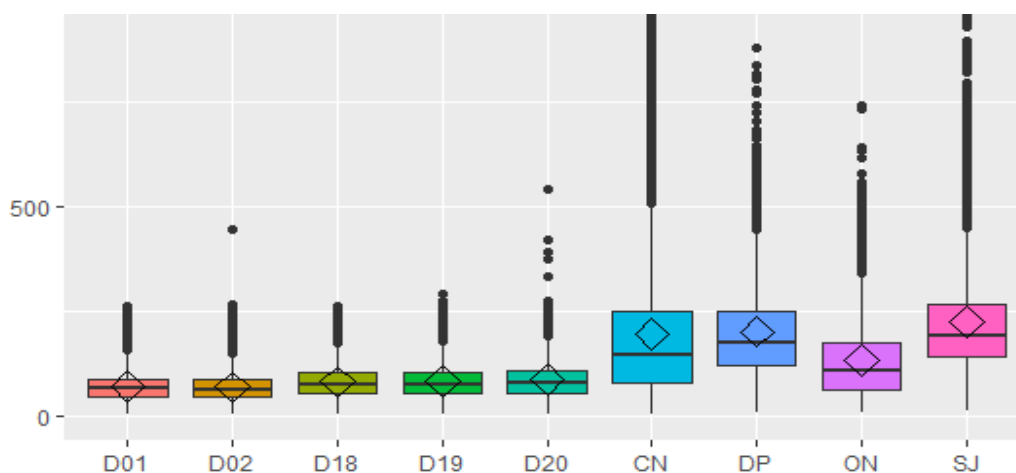


Figure 3: Boxplot of sentence (example) length (in number of characters) per partition.

The distribution of punctuation marks (normalized on sentence size) is presented in Figure 4 on the left: dictionary examples have less punctuation marks than the control corpus. The average word length is similar for all dictionary volumes, slightly shorter for novels and much longer for daily papers and even more for journals (Figure 4, right), probably due to the use of specific terminology, as expected.

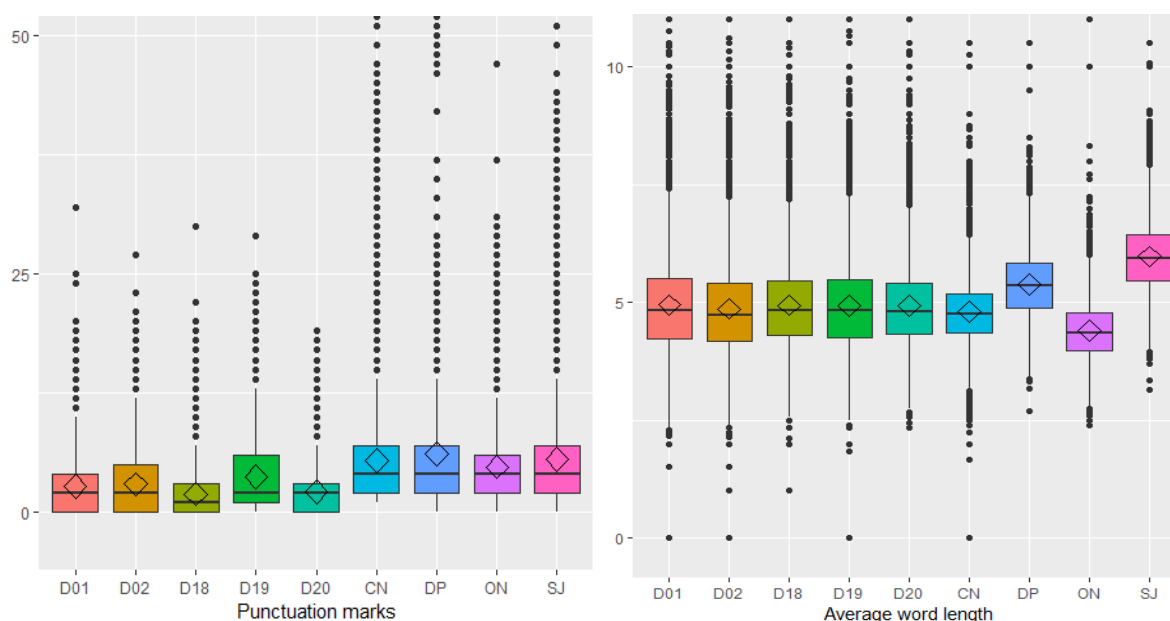


Figure 4: Boxplots for number of punctuation marks and average word length per partition.

According to the corpora, sentences in novels have more pronouns than examples in the SASA dictionary (Figure 5, left). The first two volumes have a very low median, which corresponds to the lexicographers' practice of choosing examples with nouns because they are easier to understand. Sentences extracted from daily papers and scientific journals also have very few pronouns, which can be explained by a greater need for precision in scientific and journalistic language.

In order to approximate and predict the ability of a user (with a specific profile) to understand a specific example, a “frequency indicator” was calculated for each example/sentence (Figure 5, right), as the average frequency of each word in it. The underlying assumption is that the more frequent the words in the example, the greater the possibility that the user will understand it. Word frequencies were obtained from SrpKorp2013 (Utvić, 2014). Examples from novels have higher frequency indicators, while these indicators are lower for examples from journals. The first two volumes of the SASA dictionary have a wider span of frequency indicators than other volumes (as expected, due to the type of the lexis contained in each volume; for example, the majority of the lexis beginning with a, contained in the first volume, is of foreign origin, while the second volume contains lexis mostly labelled as regional, obsolete, ephemeral, etc.).

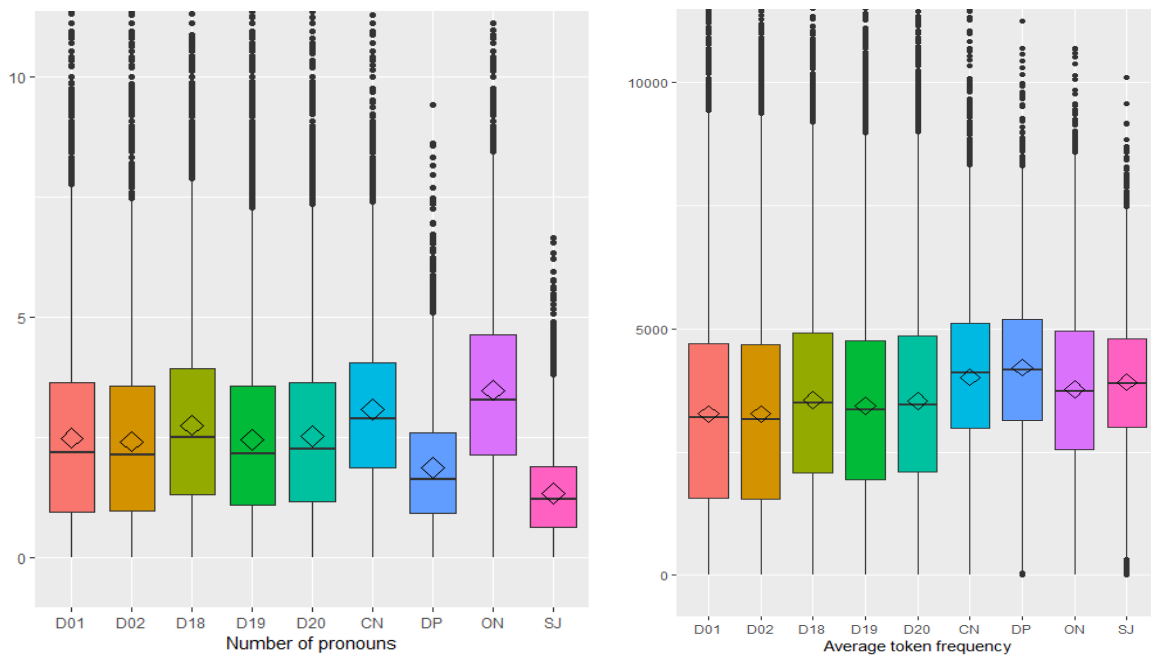


Figure 5: Boxplot of number of pronouns and token frequency per partition.

Figure 6 (left) shows that standard Serbian (DSS) and non-standard (DNS) in the dictionary have a similar distribution of the number of words in the examples, which means that there is no difference in this respect between good examples illustrating standard or non-standard lexis. On the other hand, the evaluated dataset has a wider range for inadequate examples (DNS (NO)), while a similar distribution with those in the dictionary. The results for other features also show that there are no significant differences between examples in DSS and DNS.

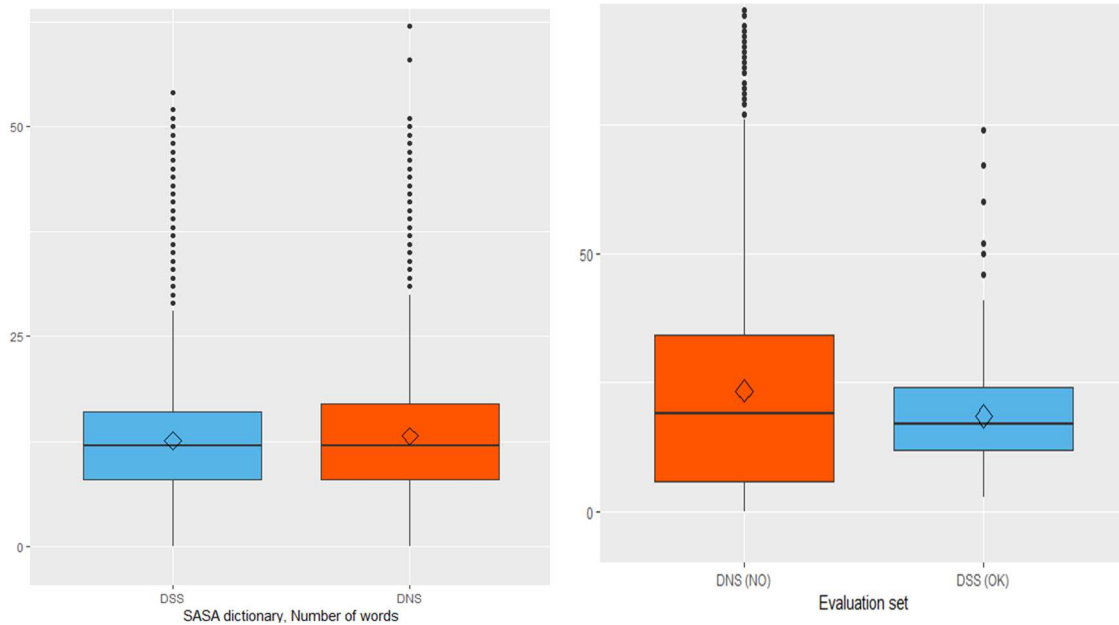


Figure 6: Boxplot of number of words per language type partitions.

Histograms and boxplots were supported by a data summary of calculated features, which offered the guidelines for data cleaning and control corpus preparation. We performed the preprocessing of both datasets we are using (SASA examples and control) and produced data summaries. These were analysed by lexicographers, on the basis of which parameters for potential example cleaning were deduced and threshold values for them were defined. Table 1 presents the data summary from SASA dictionary for five representative features.

Percentile	Sentence length	N° of digits	N° of words	Avg. word length	N° of stop words	N° UCase inside sent.
5th	28	0	5	3.6	0	0
40th	64	0	10	4	3	0
Median	73	0	12	4.8	4	0
65th	87	0	14	5.2	5	0
95th	150	0	25	6.6	10	2

Table 1: Data summary from SASA dictionary for selected features.

5. Preliminary model for identifying good dictionary examples

The future system for semi-automatic identification of good dictionary examples relies on the results of the outlined analysis and includes already developed modules for detection of good examples, as well as for detecting those that are not appropriate examples for standard language use. Filtering and ranking of examples can be

performed using rules obtained from analysed data (feature vectors) combined into a single score. The development of the GDEX function is inspired by the state of the art implementation¹² for which the following functions were developed: *blacklist()*, *greylist()* and *optimal_interval()*. For each feature the function *optimal_interval* uses four key percentiles from the gold SASA dataset (as shown in Table 1)¹³, where feature values lower than the first and higher than the last are assigned a score of 0.01, in the middle interval scores are 1, and between them a linear interpolation function is used. The four percentiles were computed for different key values, but final results will be deduced after a broader evaluation campaign, with parallel evaluation and adequate interrater agreement. For the *greylist* function only two key values are used (5th and 95th percentiles): values lower of the 5th are assigned a score of 1, higher than 95th a score of 0, and between them linear interpolation is used. Besides the solution with multiple assessments of features, we have also used the analytic hierarchy process (AHP), where each feature value is converted to a numerical value from 0 to 100 and a numerical weight (priority) is assigned to it (the sum of all weights being 1), which gave us better results. The precision calculated on the evaluation set for the first 100 ranked examples was 0.77, for the first 200 it was 0.70, for 400 it was 0.65, for 1,000 it was 0.6, etc. We believe that the results can be improved with additional rules, since the evaluators have noticed that some patterns and some types of sentences can indicate their inadequacy. For example, if the adverb of time or place is not the headword to be illustrated by the example, sentences beginning with these adverbs are not good examples, because they often need the preceding context (*Onda sam otputovao. ‘Then I left’*).

Sentences are ranked by a GDEX weighted sum of feature score values, which is then mapped to a user-friendly final score from 1 (poor, lowest 20%) to 5 (good, 20% highest), representing their suitability to serve as examples.

Sentences from the prepared dataset, represented as feature-vectors, were used as the dataset for a supervised Machine Learning (ML) model, which was then used in a GDEX classifier for contemporary Serbian sentences. Since the dataset of examples was unbalanced, with twice as many DSS examples as DNS examples, we have randomly extracted 44,808 (out of 89,096) examples with standard lexis from the DSS dataset and labelled them as ‘OK’ (positive class) and the same number of examples (44,808) from the DNS set with non-standard lexis (labelled as ‘NO’ – negative class). Since the manually evaluated sample was small it was replicated five times, yielding 7,165 ‘NO’ and 6,585 ‘OK’ examples.

We used the AdaBoost (Rätsch et al., 2001) algorithm’s implementation in Weka (Eibe et al., 2016), a suite of machine learning software. The trained model was evaluated in a 10-CV (cross-validation) setting, with the default Weka parameters for this algorithm.

¹² <https://www.sketchengine.eu/syntax-of-gdex-configuration-files/>

¹³ The 40th and 65th percentiles of the SASA dictionary for number of words are the same as the values in the example given to the Sketch Engine.

In the first decision step, the most distinctive feature, as expected, was *abbrev* (the indicator of the existence of a linguistic label). Namely, the corresponding rule is: “if the *abbrev* linguistic label is missing, there is a 92% chance that the sample is positive”. The confusion (error) matrix represents the features in predicted and actual classes: true positive 7,475 (0.68); false positive 3,581 (0.32); true negative 9,729 (0.87); false negative 1,451 (0.13). This result can be considered satisfactory; however, there is a serious issue – the existence of a linguistic label *abbrev* cannot be expected for corpora in general. Therefore, we wanted to build another classifier that uses other features.

The first step is feature analysis and feature selection. We first determined and visualised a Pearson correlation matrix that contains the correlation of features to manually assigned labels, where green represents a strong positive correlation, red a strong negative correlation, and yellow no correlation. After removing irrelevant features (those that have a very low correlation with *label*, like *avg_word_len*, or those that are highly correlated with each other, such as *max_word_len* and *max_token_len*), we represented each sample with the shorter feature vector (Figure 7).

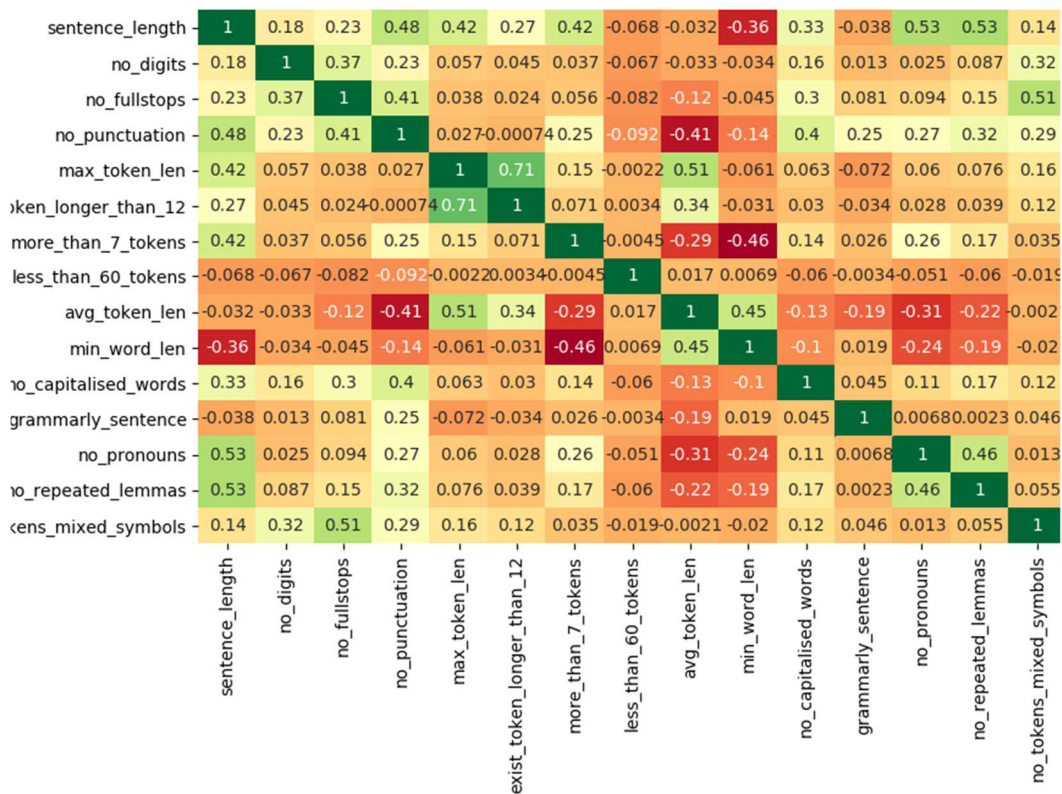


Figure 7: Pearson correlation matrix.

The gold dataset was split into a training and a validation set (20% of the dataset). The results of the Logistic Regression (Hosmer et al., 2013) classifier are given in Table 2, where NO stands for non-standard and OK for standard language.

		Precision	Recall	F1-score	Number of samples
NO	(NS)	0.84	0.68	0.75	11,056
OK	(SS)	0.73	0.87	0.79	11,180
ALL		0.78	0.77	0.77	22,236

Table 2: Results of the logistic regression binary classifier.

All metrics show better results for the negative class. Out of 11,056 negative samples in the validation set, 7,520 were classified as negative (68%, true negative), and the remaining ones as positive (23%, false positive). From 11,180 positive samples, 9,727 were classified as positive (87%, true positive), and the remaining ones as negative (13%, false negative).

The feature extractor is freely available, while the GDEX ranking and trained ML model are available for authorized users. The future system for semi-automatic identification of good dictionary examples implies the development of more modules, e.g. a user interface for feature extraction and for GDEX parameter fine tuning, but the evaluation of the first results of the developed core components is encouraging.

6. Future work and concluding remarks

The first results are encouraging, and they motivate further detailed analysis of other computed features and the introduction of new ones. Improvement of the weighted measure of features will follow, with a combination of expert knowledge and data training results.

Implementation of other features and criteria will be integrated into the web application and selections of parameters and features to be calculated will be enabled. Full system integration will combine the use of a lexical database with corpora exploitation via the developed web service and software. Since the work on digitization of other volumes of the SASA dictionary is continuing, more data is expected to bring more refined conclusions.

There is obviously a lot of room for improvement of the trained model, e.g. with the introduction of new features, by adding more samples, or using other state-of-the-art neural network architectures. Another future step is the model's evaluation on a control dataset – extraction and ranking performance is going to be tested by more lexicographers, with parallel evaluation and interrater agreement checking. Finally, we also plan to introduce flexible mapping of computing scores – from 1 (worst) to 5 (best) – and score our examples using them. This can be performed either by looking at the rules and constructing an equation, or by a trained classifier.

7. Acknowledgements

This research was partially supported by Serbian Ministry of Education and Science under the grants #178009, #III 47003 and #178003.

8. References

- Atkins, S. B. T. & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.
- Bird, S., Loper, E. & Klein E. (2009). *Natural Language Processing with Python*. O'Reilly Media Inc.
- Eibe, F., Hall, M. A. & Witten, I. (2016). *The Weka Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"*. Morgan Kaufmann, Fourth Edition.
- Hosmer Jr., D. W., Lemeshow, S. & Sturdivant, R. X. (2013). *Applied Logistic Regression*. Vol. 398. Hoboken, NJ: John Wiley & Sons.
- Ivanović, N., Jakić, M. & Ristić, S. (2016). Građa Rečnika SANU – potrebe i mogućnosti digitalizacije u svetlu savremenih pristupa. In S. Ristić et al. (eds.) *Leksikologija i leksikografija u svetlu savremenih pristupa*, Beograd: Institut za srpski jezik SANU, pp. 133–154. [The material of the Dictionary of the SANU - the needs and possibilities of digitization in the light of contemporary approaches (in Cyrillic)].
- Kilgariff, A., Husák, M., McAdam, K., Rundell, M. & Rychlý, P. (2008). GDEX: Automatically Finding Good Dictionary Examples in a Corpus. In E. Bernal & J. DeCesaris (eds.) *Proceedings of the XIII EURALEX International Congress, EURALEX 2008. Barcelona: Universitat Pompeu Fabra*, pp. 425–432.
- Kosem, I. (2017). Dictionary examples. In V. Gorjanc, P. Gantar, I. Kosem & S. Krek (eds.) *Dictionary of Modern Slovene: Problems and Solutions*. Ljubljana: University of Ljubljana, Faculty of Arts.
- Kosem, I., Koppel, K., Zingano Kuhn, T., Michelfeit, J. & Tiberius, C. (2019). Identification and Automatic Extraction of Good Dictionary Examples: the Case(s) of GDEX. *International Journal of Lexicography*, 32(2), pp. 119–137.
- Krstev, C. (2008). *Processing of Serbian – Automata, Texts and Electronic dictionaries*. Belgrade: Faculty of Philology, University of Belgrade.
- Popović, Lj. (2003). Integral sentence models and their importance for lexicographic description and corpus analysis [Integralni rečenični modeli i njihov značaj za lingvistički opis i analizu korpusa]. *Naučni sastanak slavista u Vukove dane*, 31(1), pp. 201–220. (In Serbian, cyrillic.)
- Popović, Lj. (2004). *Red reči u rečenici* [Word order in sentences]. Beograd: Društvo za srpski jezik i književnost Srbije. (In Serbian, Cyrillic.)
- Rätsch, G., Onoda, T., & Müller, K. R. (2001). Soft margins for AdaBoost. *Machine learning*, 42(3), pp. 287–320.
- Sabo, O. & Vitas, D. (1998). Mogućnost osavremenjivanja izrade rečnika na primeru

- Rečnika srpskohrvatskog književnog i narodnog jezika SANU i Instituta za srpskohrvatski jezik. In *IV međunarodni naučni skup „Računarska obrada jezičkih podataka”*, Portorož: Institut Jožef Stefan, pp. 375–384 [Possibility for modernizing the development of the dictionary on the example of the Dictionary of the Serbo-Croatian literary and vernacular language SASA and the Institute for Serbo-Croatian].
- SASA Dictionary: Речник српскохрватског књижевног и народног језика САНУ, I–XX (The Dictionary of the Serbo-Croatian Standard and Vernacular Language) (1959–2017). Београд: Институт за српски језик САНУ и САНУ.
- Stanković, R., Krstev, C., Vitas, D., Vulović, N. & Kitanović, O. (2017). Keyword-Based Search on Bilingual Digital Libraries. In A. Calì, D. Gorgan & M. Ugarte (eds.) *Semantic Keyword-Based Search on Structured Data Sources. COST Action IC1302 Second International KEYSTONE Conference, IKC 2016, Cluj-Napoca, Romania, September 8–9, 2016, Revised Selected Papers*, pp. 112–123. DOI:10.1007/978-3-319-53640-8_10.
- Stanković, R., Stijović, R., Vitas, D., Krstev, C. & Sabo, O. (2018). The Dictionary of the Serbian Academy: from the Text to the Lexical Database. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (eds.) *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*. Ljubljana, pp. 941–949. Available at: <https://euralex.org/category/publications/euralex-2018/>.
- Stijović, R. & Stanković, R. (2017). Digital edition of the SASA Dictionary: a formal description of the microstructure of the SASA Dictionary [Digitalno izdanje Rečnika SANU: formalni opis mikrostrukture Rečnika SANU]. *Naučni sastanak slavista u Vukove dane*, 47(1), pp. 427–440. (In Serbian, Cyrillic.)
- Utvić, M. (2014). The construction of reference corpus of contemporary Serbian [Izgradnja referentnog korpusa savremenog srpskog jezika] (Doctoral dissertation, University of Belgrade).
- Vitas D. & Krstev C. (2015). Blueprint for the computerized dictionary of the Serbian language [Nacrt za informatizovani rečnik srpskog jezika]. *Naučni sastanak slavista u Vukove dane*, 44(3), pp. 105–116. (In Serbian, Cyrillic.)
- Vitas, D. & Krstev, C. (2012). Processing of Corpora of Serbian Using Electronic Dictionaries. *Prace Filologiczne*, vol. LXIII (Warszawa), pp. 279–292.
- Zgusta, L. (1971). *Manual of Lexicography*. Praha: Academia.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



***eDictionary*: the Good, the Bad and the Ugly**

Marijana Janjić, Dario Poljak, Kristina Kocijan

Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences, University of Zagreb, Ivana Lučića 3, 10 000 Zagreb (Croatia)

E-mail: marijanajanjic@yahoo.com, poljak5@gmail.com, krkocijan@ffzg.hr

Abstract

On its own, learning a new language is an inherently daunting task. Combined with lacking or simply non-existent language resources, the task itself seems almost impossible. For some languages, this scarcity of available resources is even more obvious and further complicates the issue.

With an interdisciplinary approach, a team of linguists, language teachers, information scientists, and students themselves undertook a task of developing a learner's dictionary of Asian languages. With a great deal of care and discussion, an online e-dictionary was chosen as a platform for its ease of use, accessibility, and expandability, in lieu of a traditional printed dictionary.

Since *eDictionary* is built as a website, it is established as a platform, agnostic and available to everyone with Internet access. Furthermore, such a design allows a link to resources hosted on other web portals. To that end, cooperation was initiated with *Croatian Language Portal* and their Croatian dictionary with the aim of hyperlinking all of our Croatian lemmas to their word definitions. With the added benefits of giving users the ability to request new resources while keeping track of the request internally and allowing the updates of the whole language database seamlessly, the proposed solution to *eDictionary* provides user engagement and continuous integration that should benefit us all.

Keywords: e-dictionary; learner's dictionary; user engagement; Asian languages; Croatian

1. Introduction

History suggests that dictionaries in the form of word lists are a very old invention. From ancient Akkadian times to today, dictionaries represent important and valuable achievements in various cultures. Over the centuries, man has created different types of dictionaries with different purposes in mind. Among them is the dictionary aimed at learners of foreign languages, which is a version of a learner's dictionary - smaller in size than a general-purpose dictionary, with elements that enhance the learner's knowledge and skills in the target language. Modern times and the rapid advance of technology have made the existence of an online dictionary possible, not as a replacement for a printed dictionary, but as an addition to it.

Aside from the monolingual dictionaries, the new technology has also been used to develop student-oriented online dictionaries. A number of e-dictionaries have become popular for several reasons (Heuberger, 2016: 41): 1. the **size** of an e-dictionary is greater than that of a printed dictionary, and it is easier to alter; 2. e-dictionaries, unlike printed dictionaries, can **function** in more than one direction, i.e. any of the

languages included in it can be taken as a starting point (L1) and learning goal (L2); and 3. the possibility to include multimedia **features** (ex. recorded pronunciations, pictures).

Some good examples of online dictionaries include a dictionary for students of translation studies with a focus on Internet-related vocabulary (Alipour, Robichaud & L’Homme, 2015), a dictionary for language learners and other users (Deksne et al., 2013), and a dictionary for learners of Spanish (Renau & Battaner, 2011). Another interesting project is a multilingual lexicographic project for immigrants (Vacalopoulou & Efthimiou, 2015).

After giving a short overview of related work, the structure of the paper will take a closer look at the main idea behind this project, taking into account its upsides (the Good), downsides (the Bad), but also those aspects that could have been done better (the Ugly). Before the concluding remarks, a short analysis of the analytics will be provided.

2. Related work

Looking from the perspective of Croatian students, the search for web-based dictionary resources of Asian languages with Croatian as either the source or target language is like a scene from *Mission Impossible*. Most of the available resources have English as a link-language to the meaning of words from the Asian continent. If we were to operate under the assumption that all students know English well (and very well), and that this should not be considered an obstacle in using it to learn a third language, as different as any of the Asian languages, then things are all well and we can conclude our paper at this point.

However, this is not a valid assumption to make. Not all Croatian students have the same knowledge of English when starting university. Also, some language nuances are surely lost in translation, and even more so if they need to make their journey via multiple language groups (Slavic -> Germanic -> Asian and back). The lack of resources in one’s native language puts an additional burden on the student, as it forces them to become a learner of not just one foreign language, but two - the link language, as well as the target language. As our experience in learning and teaching Hindi and Sanskrit in Croatia shows, there are students who come into the classroom equipped with not just dictionaries that include entries in the target language and English, but also with English – Croatian and Croatian – English dictionaries. This means that they are familiar with English to some level, and that they are simultaneously tackling two foreign languages at different levels.

Our goal of building Croatian language resources for the benefit of students in Croatia stems from the question as to whether students would be more efficient and successful in mastering the target language if having to master the link language was removed from the equation. The overview of available literature on the use of dictionaries and

other linguistic resources in a foreign language classroom suggests that authors and teachers assume the presence and availability of foreign language learning resources in the students' native language. According to some, the importance of native language resources is particularly high at the beginning stages of learning a new language, as the role of context is negligible at that point (Pavičić Takač, 2008; Summers, 1988). This is precisely the situation that Croatian students face when they opt to learn one of the Asian languages. Thus, our decision to focus on the use of dictionaries in foreign language teaching and learning was supported by two facts: a) acquisition of new vocabulary presents an important part of language learning, especially at the beginner level; and b) some students have to overcome a considerable obstacle, which is mastering a link language.

Multiple experiments in a Croatian context (Dovedan et al., 2002; Družijanić Hajdarević et al., 2006; Lauc et al., 2006; Librenjak et al., 2012; Janjić et al., 2016a; Librenjak et al., 2016c) have reported on how well language learning and digital resources go together i.e. the learners were mostly positive about technology usage, which led to greater motivation and consequently to more frequent usage of resources, so resulting in better language acquisition and greater retention. Hence, it seemed reasonable to assume that an e-dictionary would be better received by students than its printed counterpart. The following section gives more information on the *eDictionary* project and its intended users.

3. The *eDictionary* Project

The idea of building the *eDictionary*¹ of Asian languages for the Croatian users emerged naturally during the work on the *MemAzija* project. The aim of the *MemAzija* project was to test the influence of technology in learning Asian languages (Librenjak et al., 2016c; Janjić et al., 2016a; 2016b; 2017b). In order to test its hypothesis, the research team developed a number of Croatian language resources for learning four Asian languages: Hindi, Korean, Japanese and Sanskrit. From that point on, it did not take long to see that a learner's dictionary aimed at learning Asian languages was long overdue.

The primary reason for building the dictionary was the lack of similar resources in Croatian. As a result, students used the available e-dictionaries that translated Asian languages to English, and vice versa. During the *MemAzija* project, the research team realized how useful e-tools were to new generations of students, as they used them frequently in order to learn new languages or further improve their language skills.

The dictionary was built mainly for Croatian students studying one of the included Asian languages. But as it turns out, students are a heterogeneous group, and therefore the focus was further narrowed down to those students just beginning their studies, as

¹ Dictionary is available at: <http://erjecnik.ffzg.hr/>.

it was deemed that the resources in Croatian would be most useful to them. This decision was beneficial for our project in two ways. Firstly, it allowed us to provide students with a tool that would let them study a new foreign language with more ease, as they would be able to focus fully on acquiring just the target language vocabulary. Secondly, it served as a clear starting point for what could have been a broad and aimless project. As there is no Croatian dictionary, or more specifically no e-dictionary, which targets students and combines different Asian languages, it seemed reasonable to start with a smaller project i.e. a dictionary for A1 – B1 learners and go on from there.

This focus has had an effect on the dictionary design and its information architecture in several different aspects:

- a) **dictionary mode**, i.e. choice between printed and online form,
- b) **the choice of lemmas** included in the dictionary, and
- c) **the structure of lemmas**.

For a more in-depth look at the choice and structure of lemmas included in the dictionary, please refer to Section 5.

Prior to building *eDictionary*, the research team had taken steps which were considered very important for both the end and front design of the dictionary, i.e. for the database and administrative dashboard design, as well as the user interface design. These steps included the **analysis** of existing e-dictionaries, **consultation** with lexicography experts, and a **survey**² of students' opinions regarding the preferable form and structure of an e-dictionary. All steps were equally useful to the research team. The last one, however, proved to be crucial as it showed what the primary user demographic considered important, relevant and beneficial for a learning tool to have.

Issues that students considered relevant when it came to the use of e-dictionaries for learning were helpful for determining what an e-dictionary should and/or should not include. The five most prominent of those issues were:

- a) different rules for typing in Asian alphabets often require downloading various additional programs;
- b) lack of off-line availability;
- c) lack of compatibility with non-desktop devices (tablet or mobile);
- d) no available Croatian translation;
- e) direct translations lack examples of usage.

² The survey was conducted in 2016 during the *MemAzija* project and it involved 82 learners of Asian languages, with 72 female and 11 male students. The larger group, 59 of them, spoke Croatian as L1, while other participants were native speakers of Serbian, Montenegrin, Bosnian, Slovenian, etc. In this paper, we will refer to this survey as the MAP Survey (*MemAzija Project Survey*).

Although there were some students who reported feeling content with the quality of the e-dictionaries they normally use, the majority reported feeling the opposite. Most students reported a lack of additional information (gender, part of speech, etc.) which they deem important for a dictionary aimed at students, an absence of particular context or even worse – the complete absence of some words. Another criticism that students had was that the dictionaries rarely function well in both directions (L1 – L2, L2 – L1). They also expressed a very clear dislike for advertisements on some websites that host e-dictionaries. Another criticism to note is that they pointed out that those dictionaries that translate one word into a number of L1 words make it difficult for them to discern the primary meaning of the word. And finally, according to students’ opinions, e-dictionaries should be closer to paper dictionaries in the sense that words with similar spelling and roots - words that would be found next to each other in a printed dictionary, should also appear together in an e-dictionary.

4. Introducing the Platform

When building a web platform there is much to take into consideration: from laying out the database and data flow to the sort of visual representation of elements on the page that would suit the user’s needs. As Heid et al. (2013:271) point out: “We have passed the stage of putting paper dictionaries on computer or simply designing electronic dictionaries in the same way as paper dictionaries”. Taking that into account, as well as the fact that this dictionary has a notably niche audience, we decided to adopt an architect’s approach to building it in order to satisfy our users’ needs. In other words, every feature was thoroughly planned out from the ground up. This approach has allowed us to future-proof the application by putting minimal constraints on adding new features or more languages.

Right at the start, it was obvious that one of *eDictionary*’s main features would have to be two-level expandability including a) **depth**, allowing for new words to be added to each existing language, but also b) **width**, allowing completely new language additions. To tackle that, we have designed the dictionary with Croatian words as meta words in a pivot table applicable to all languages. Another planned feature relying on that design element would allow users to compare their query in multiple languages at once, a feature specifically targeted towards students studying Sanskrit and Hindi.

In order to achieve our goals in a timely manner, we have designed a relational database model with Model-View-Controller (MVC) architecture in mind, and tried to delegate most of the heavy lifting to technology. According to Majeed and Rauf (2018) MVC provides three types of classes:

- A. **Model:** Model classes are used to implement the logic of data domains. These classes are used to retrieve, insert or update the data into the database associated with our application.
- B. **View:** Views are used to prepare the application interface through which users

interact with the application.

C. Controller: Controller classes are used to respond to and perform user-requested actions. These classes work with model classes and select the appropriate view that should be displayed to the user according to their requests.

Thanks to such a clear division between the MVC layers we were able to effectively break down the development requirements. That has in turn allowed us to focus completely on creating the first usable versions of *eDictionary* and getting it tested by students themselves. Because *eDictionary* uses a pivot language, the database must contain transfer tables with the pivot language and every other language. This kind of relational model requires smooth data manipulation using models and MVC architecture (Janjić et al., 2017a). Relying on open source technologies, we have opted for Laravel on the back-end and jQuery on the front-end, which has in turn allowed us to focus on the user experience. Additionally, *eDictionary* is entirely hosted on faculty servers, i.e. all the documentation, codebase, and the complete language database are securely backed up. All of that combined with the fact that technologies used in development of *eDictionary* are well-established and widespread, means that there are no technical obstacles for further development of the project.

When it came to user experience, the term “accessibility” came up most often. We wanted to focus on three device types for optimal accessibility: mobile phones, tablets and computers (Figure 1). Due to the sheer volume of data on display to our users, we had to ensure adequate accessibility for the smallest of devices from the very start. Using responsive design-driven methodology, we have managed to scale our design down to resolutions of 480x960 pixels, while still retaining all the features of the page.

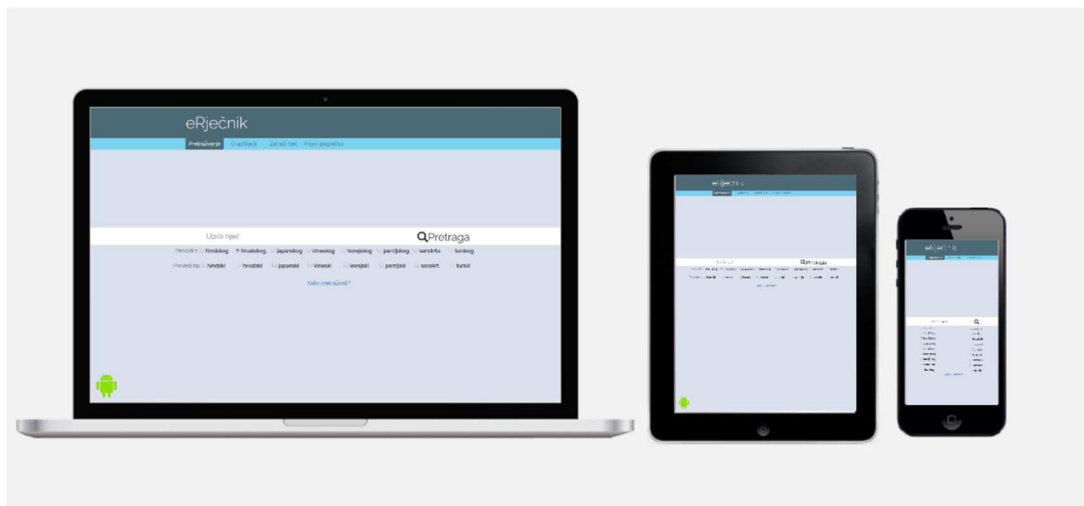


Figure 1: The *eDictionary* website as it appears on different resolution screens.

Further user experience improvements had to be made on the data querying front. Since we started working on the platform at the end of 2016, we had to take into consideration other non-dictionary web platforms which mainly use asynchronous calls

to the web server to deliver the data to the user without having to refresh the page. Thanks to our selection of development technologies, it was just a matter of agreeing on the data structure being sent to and from the server (Janjić et al., 2017a). In more technical terms, that means that we have opened up an API (Application Platform Interface) on our back-end Laravel server to receive data in JSON (JavaScript Object Notation) format via AJAX (Asynchronous JavaScript) calls and return the results in the same manner. That way we have achieved smooth and seamless data transmission between the user and the web server, therefore making *eDictionary* a full-fledged web application as defined by Paulson (2005).

Taking full advantage of the fact that the *eDictionary* is a website, all Croatian entries were made into hyperlinks. That means that each word in Croatian was connected to the well-established *Croatian Language Portal* (HJP) website, which already provides a single-language dictionary functionality for Croatian. This way, we made it possible for users to easily access any additional data they might find relevant without overwhelming them during the process (Heid et al., 2013).

Furthermore, due to the versatility of web 2.0 technologies *eDictionary's* expansion roadmap is not 'set in stone', but is rather expanding according to user needs. The way this works is through a *request system* that we have implemented, which makes it possible for users to request the needed resources for certain languages, as a team of language teachers or linguists develops requested resources for the next update of the internal database.

An internal monitoring tool was also prepared to provide administrators with analytics data on searched words, requested examples and reported errors, more or less covering the core functionality of *eDictionary*. Additionally, we are also using Google Analytics for data on user demographics, retention and bounce rate, as well as the type of platform users are accessing *eDictionary* on.

We have tried to cover as many fronts as possible while creating a web application that is both useful and not inherently limited in scope like a traditional dictionary. With prolonged use and administration, we have added new languages, new words and cooperated with domain experts to polish the core language learning functionalities.

4.1 Similar projects

Even though there are, at the time of writing, very few Croatian printed dictionaries targeting Asian languages³, we can still examine how other similar Croatian projects were developed. In this chapter, we will evaluate four Croatian web portals that serve as different types of e-dictionaries. The four web portals discussed are: the *Croatian*

³ Croatian-Japanese dictionary (2006) and Croatian-Turkish dictionary (2014).

Language Portal dictionary, which helped us with Croatian lemmas for *eDictionary*, *Croatian Encyclopedia*, *eGlava Online Valency Dictionary* and *Struna - Croatian National Termbank*. We are particularly interested in the technical implementations used for the four dictionaries, i.e. how they overcame some hurdles that we also faced, which features our dictionaries share and whether we are missing some features others consider crucial.

We chose these four projects in particular because of the shared similarity in niche target audiences, our affiliation with them (namely *Croatian Language Portal*), the fact that they came to be as a result of primarily academic efforts, and how well known they are.

4.1.1 *Croatian Language Portal*

Croatian Language Portal (cro. *Hrvatski jezični portal* or *HJP*) is a monolingual dictionary targeted at Croatian. It has emerged from collaboration between the publishing company *Novi Liber* and the *University Computing Centre – SRCE*. Armed with the prolific publication history of quality Croatian dictionaries by *Novi Liber*, the developers could easily construct a rich and detailed user interface for the dictionary.

Users of the *Portal* are given a plethora of information upon searching for a word, beginning with the word and its grammatical data, and followed by derived forms of the word in all cases/tenses/numbers, word definition(s) as found in printed dictionaries, in some cases even some example phrases, syntagmas, phraseology, onomastics, etymology and possibly even more. There is also a permalink feature for all searched words, which came in rather handy when we were connecting *eDictionary* Croatian entries with the existing definitions in *HJP*.

The *HJP* single language dictionary focuses on one thing and does it well, displaying all available language data for queried words. It is a synchronously loading site with a simple and straightforward design which translates well to mobile devices and computers alike.

The only downside to their design, one that we ran into during our own development process as well, is that it displays all of the language features even when there is no data associated with the searched word (for example, the title “Onomastics” appears to the user regardless of whether there is data to be shown or not). That, however, is almost a non-issue in contrast to the amount of presented and actually available information, since the user experience is not hampered in the least by this design “flaw”.

4.1.2 *Croatian Encyclopedia*

Croatian Encyclopedia is, as the name suggests, an encyclopaedia of the Croatian language. Developed by the Miroslav Krleža Department of Lexicography, it is presented as a single language dictionary web application that focuses on content presentation and professional explanations. The website offers a deep search functionality that goes not only through the lemma itself, but also through the explanations for all the occurrences of searched term.

Search results are colour-coded depending on where in the lemma or explanation the searched term was found. It can range from a direct hit, represented by dark red, when the searched term is present as a singular explained lemma. But it can also be found as part of the explanation and is then bright red. It should be noted that direct hits usually lead directly to the explanation, but other search results can also be accessed by clicking on “*Search further*” (cro. “*Traži dalje*”). We have found that colour-coding search results is a design element that may prove useful to our own application, especially for showing search results depending on target language(s). We will thus consider adding this upgrade in the future expansion of *eDictionary*.

The webpage is well designed for single-handed use on mobile devices, with a pop-up menu available on the bottom of the page and the action button positioned on its right side. This is another feature which is well thought out and will surely influence any future design revisions of *eDictionary*.

All explanations also serve as jumping-off points for further research, as some of the words are also hyperlinks to other lemmas and their explanations. This is an approach that, similar to our own, provides a natural bridge to even more relevant information on the subject, as proposed by Heid et al. (2013).

During our research for this paper, the only issue we encountered was slow page performance, with wait times for search results of up to 30 seconds. We did not, however, inspect this matter further since this can be attributed to many factors that are not directly controlled by the maintainers of the *Croatian Encyclopedia*. Still, we believe this to be an important issue, since wait times have been proven to be quite an important factor for user retention rate.

4.1.3 *eGlava Online Valency Dictionary*

eGlava Online Valency Dictionary is an online valency dictionary of Croatian verbs, developed within the project “*Valency Database of Croatian verbs*” at the Institute of Croatian Language and Linguistics. It contains valency descriptions for 900 verbs specifically built for linguists, teachers and students of Croatian language (Baza hrvatskih glagolskih valencija, 2019).

The website's sole purpose is listing all of the verb valences available in the database. There is no conventional search functionality through an input form, but rather an alphabetized list that can be filtered out by clicking on a specific letter. It is accompanied by an effective, albeit simple mobile design which keeps all website features accessible on all device types.

Some of the problems we encountered during research were mostly to do with misplaced links (i.e. some clickable elements throw the user back to the homepage), but otherwise the data it holds is presented exquisitely and in great detail. Furthermore, the site is also available in English, which is a feature still missing from *eDictionary*, but something that we strongly consider adding since we hope that our target group might also include non-native speakers of Croatian language in the future.

4.1.4 *Struna – Croatian National Termbank*

The last of the websites that we will discuss in this segment was also created under the leadership of the Institute for Croatian language. *Struna* is a website that focuses on standardized Croatian terminology for all professional domains.

Even though it is similar to the previously mentioned *eGlava Verb Valency Dictionary* in its narrow field of interest, the difference in website functionality is quite apparent. It offers both simple and advanced search options, mixes in attachments for certain defined terms similar to the *Croatian Encyclopaedia* website, and offers origin of the source for all defined terms.

Even though the content side of the website is meticulously crafted, there are some technical issues present that hamper the user experience. At the heart of said technical issues is the option to view the page in English – a feature that would be immensely useful, if only it were functional. Instead, what happens when the option is selected is that it breaks most of the hyperlinks on the website and instead returns *404 error pages* to user queries. There is also no responsiveness to speak of, so mobile use is strenuous at best.

But it is worth mentioning that both *eGlava* and *Struna* projects will be included in the ongoing *Mrežnik – Croatian Online Dictionary* (*Mrežnik – Hrvatski Mrežni Rječnik*) project (Hudeček & Mihaljević, 2017). That way, the content of both platforms will be unified and presented through a similar user interface.

After this analysis of projects similar to our own, we can conclude that we all had similar problems that were handled in similar ways – no matter the solution, the main focus was always on the content rather than the platform. This should not come as a surprise since the linguistic substance is the main reason users are visiting these sites, and in that regard all of them are very well executed. In the following section we will discuss the architecture of our own content.

5. The Soft Side of *eDictionary* – the Content

The target audience of *eDictionary* are both Croatian and Asian students learning Asian languages and Croatian, respectively. Because of this, Croatian was used as a source language, providing within *eDictionary* resources for six Asian languages (Hindi, Japanese, Chinese, Korean, Persian and Sanskrit) in varying degrees of fidelity. At the most, the dictionary is supposed to provide several key attributes for mastering a language. Among these attributes are translation, transliteration to Latin alphabet, grammatical notation and examples of usage. In cases where not all attributes are available at the time of the query, the base information always includes a translation and the link to the definition of the queried term in Croatian.

According to the MAP Survey of students' needs, a perfect e-dictionary for learning a new language would consist of lemmas that include a number of elements that we list here in the order of how many students selected them, starting with the most common one:

- a) translation
- b) grammatical information
- c) pronunciation
- d) examples of usage, phrases or sentences
- e) visual representation
- f) links to other resources, such as a lexicon or encyclopaedia that incorporates more elaborate definitions of particular cultural elements, products, ideas, etc.; links to other bilingual e-dictionaries with more elaborate lemma structures or to other monolingual dictionaries with more information
- g) orientation regarding the level at which a learner is supposed to master a particular entry (ex. A1 or B1).

We have tried to include as many of the listed elements as we could at the time of building *eDictionary*, while still maintaining the capability to include all of the suggested elements at some point in the future, depending on the availability of funds. The financial side of the project determined at an early stage that the pronunciation could not be included in *eDictionary* from the start due to high production costs. The same was concluded for visual representation.

Regarding the examples, however, the decision was made to provide them for a number of lemmas, with an open invitation extended to students and other learners and teachers of Asian languages in Croatia to send in their own examples. Their validity would then be evaluated by our project experts for each language and included in *eDictionary* if deemed valid. This decision was made with the intention of opening bidirectional communication between users and the authors. This would effectively result in expansion of the authors' roles, since they would now also serve as dictionary administrators as well, which could be seen as an opportunity for new classroom

activities where students are encouraged to look for new examples or new words that would be useful to them in their own studies.

The first version of *eDictionary* consisted of 5,953 Croatian entries. Currently, as a result of newly added lemmas, this number has increased to 6,172. However, this number is not evenly distributed among all included languages. The languages with the most entries at the moment are Hindi (2,232) and Japanese (1,330), while Persian has the least (156). In-between these two groups are Sanskrit (1,028), Chinese (762) and Korean (668). Some of the words are unique to one Asian language, while some exist in more than one. All listed languages can be compared among themselves, which learners of similar languages, like Hindi and Sanskrit, could potentially find useful.

The lemmas included in the first version of *eDictionary* are based on the learning/teaching programs used in Croatia, as well as the authors' experience as teachers and learners of the included Asian languages. With every new word request from the user, *eDictionary* becomes that much better of a learning tool that mirrors not only the teachers' perspective, but the students' as well. Hence, the decision about which lemmas to include has been greatly affected by practical experience and focused on the learner's perspective.

This, however, is not true for Sanskrit. Sanskrit is one language that stands out in *eDictionary* in terms of the methodology used for choosing the initial set of words. For all modern Asian languages, the vocabulary is similar to any other learner's dictionary in the world. But, since Sanskrit is not used for everyday communication, the same rules do not apply here. It is a classical language that Croatian students study at the Indology Department so that they could successfully read and analyse old Sanskrit documents (literature, philosophy, etc.). In other words, Sanskrit vocabulary does not contain those elements that would allow one to easily order from a restaurant, but it does contain elements relevant for the study of ancient Sanskrit texts. For that reason, the Sanskrit lemmas found in *eDictionary* are based on the frequency of their usage in texts that students often work with as they learn to master Sanskrit. The frequencies were based on the lexicographic work by Oliver Hellwig (2016) and can be found online as *Digital Corpus of Sanskrit*, hosted by the Cluster of Excellence "Asia and Europe in a Global Context" research facility.

eDictionary can be used in several directions, i.e. the Asian languages which were considered target languages (L2) in the project can also be used as L1 languages, i.e. source languages. In that sense, *eDictionary* can also be useful to students or other interested parties for a comparative search - for example from Hindi to other languages present in the database: Croatian, Chinese, Sanskrit, Japanese, etc. (Figure 2). With that in mind, *eDictionary* has the potential to become a multi-source online project for all the Asian languages concerned.

However, at this point, the *eDictionary* database is not suitable for learners of Croatian as a second language (L2), but only for native speakers where Croatian is the source language (L1). This is due to the fact that information that would be useful to L2 learners still needs to be added to our database, either as a direct entry or as a link to other existing projects that would serve this same purpose. One such project, for example, is *Mrežnik* - an online Croatian dictionary project that contains a separate module focused on learners of Croatian with 1,000 entries (Hudeček & Mihaljević, 2017). Although the *eDictionary* team had the intention of making the project accessible to L2 learners, the decision was made to use Asian languages only in the sense mentioned in Hannesdóttir (2015: 245-247), i.e. that lexical descriptions of languages in online dictionaries should be based on multiple accessibility rather than on the tradition of printed dictionaries.

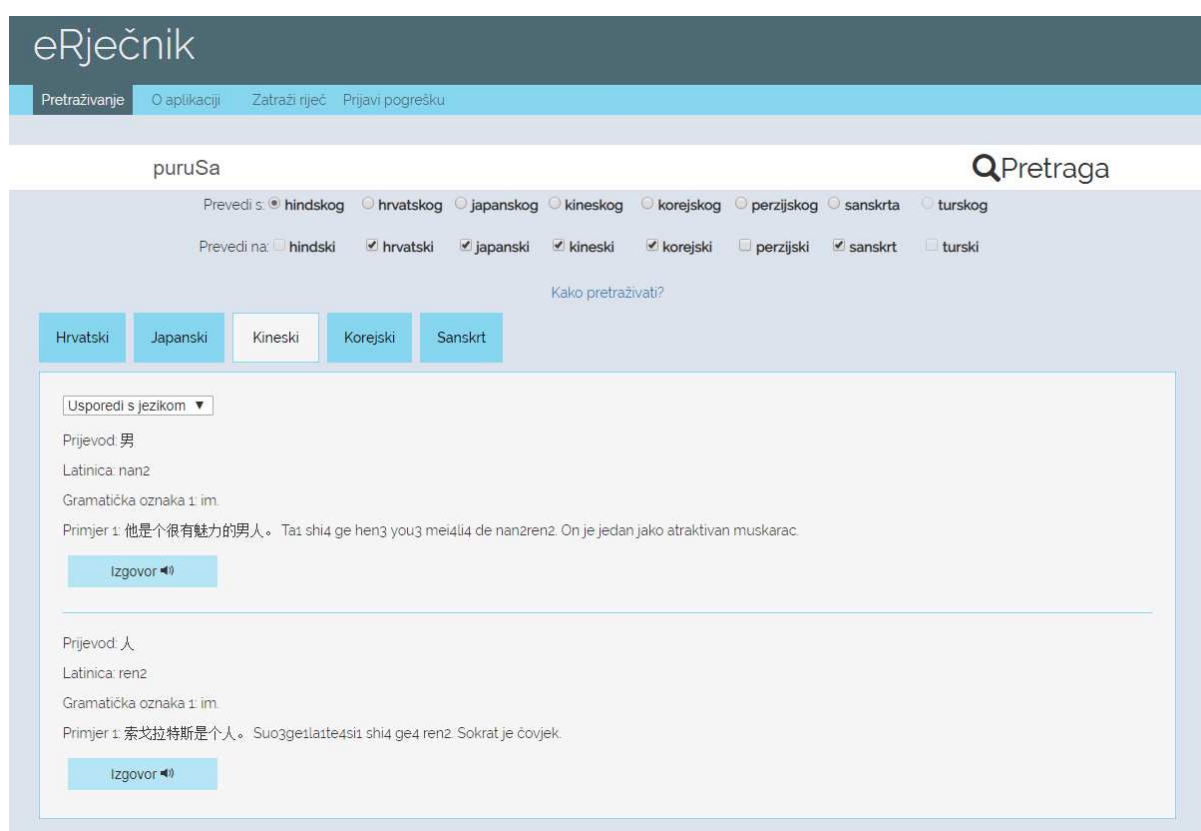


Figure 2: Example of a comparative search in *eDictionary* with Hindi as a source language.

6. Analysing the Analytics

Most of the *eDictionary* website functionality was designed in-house, including both the database model and request system. This approach lets us examine the data being recorded in the database at any point in time, namely the amount of word searches per language, which language pairs users are searching for and in which directions, as well as requested sound examples and words.

On a more global scope, we are relying on Google Analytics tracking to acquire broader data about our users. This data includes information of the country users are visiting from, number of visits, the average visit duration, and whether they are a returning or a new user.

This mixed approach to analytics allows for a more detailed overview of how the users are accessing and using our site. This method has already been described as effective by Lorentzen and Theilgaard (2012) in assessing user needs and planning on future improvements.

6.1 The Words

Looking into the global search history details, *eDictionary* users are most frequently searching with Croatian as the source language and Hindi as the target language, as presented in Tables 1 and 2 respectively.

SOURCE	COUNT	FOUND
CROATIAN	3052	1838
HINDI	521	252
SANSKRIT	200	81
JAPANESE	35	13
CHINESE	20	11
KOREAN	15	3
PERSIAN	10	5

Table 1: Count of searches, grouped by the source language

TARGET	COUNT	FOUND
HINDI	1135	929
NONE	711	710
JAPANESE	493	312
SANSKRIT	462	311
CHINESE	392	230
KOREAN	321	151
CROATIAN	257	257

Table 2: Count of searches grouped by the target language

Furthermore, we can also glean some more useful information from the search history of *eDictionary* (Table 3). Results are sorted alphabetically by the source language, i.e. the language of the queried word. The results per source language are sorted by the number of searches for each target language. Croatian is the most queried target language for all the other source languages. It is also the only source language that is paired with all the target language combinations, and is at the same time the most queried source language. These results are not surprising, since the expected dictionary users are dominantly Croatian students.

<i>Source Language</i>	<i>Target Language</i>	<i>Searches</i>	<i>Found</i>	<i>Total Searches</i>	<i>Total Found</i>
<i>Chinese</i>	Hindi	1	1	20	11
	Persian	1	0		
	Japanese	2	2		
	Korean	2	2		
	Croatian	6	6		
	<i>none</i>	8	0		
<i>Croatian</i>	Persian	77	13	3,052	1,838
	Korean	295	132		
	<i>none</i>	307	0		
	Chinese	362	205		
	Sanskrit	420	276		
	Japanese	469	294		
	Hindi	1,122	918		
<i>Hindi</i>	Persian	4	0	521	252
	Korean	17	14		
	Japanese	18	15		
	Chinese	24	20		
	Sanskrit	42	35		
	Croatian	168	168		
	<i>none</i>	248	0		
<i>Japanese</i>	Korean	3	2	35	13
	Croatian	11	11		
	<i>none</i>	21	0		
<i>Korean</i>	Japanese	2	0	15	3
	Croatian	3	3		
	<i>none</i>	10	0		
<i>Persian</i>	Japanese	1	1	10	5
	Chinese	1	1		
	Hindi	1	1		
	Croatian	2	2		
	<i>none</i>	5	0		
<i>Sanskrit</i>	Japanese	1	0	200	81
	Korean	4	1		
	Chinese	5	4		
	Hindi	11	9		
	Croatian	67	67		
	<i>none</i>	112	0		
Total		3,853	2,203		

Table 3: Detailed look into search history data.

The second most queried target language is **Hindi**. It is by far the most searched for target language with Croatian as a source language, and is also found as a target language for a small number of the source languages (Chinese, Persian, Sanskrit) and as a source language for most of the target languages (Chinese, Croatian, Japanese, Korean, Persian, Sanskrit). Similarly, despite the low number of total searches, **Chinese** as a source language is paired with five other target languages (Croatian, Hindi, Japanese, Korean and Persian) and is found as a target language for three (Hindi, Persian and Sanskrit). **Japanese** as a source language is only requested with Croatian and Korean as the target languages, but is found as a target language for all the other language combinations. Something similar is true for **Korean**, which is never requested as a target language except for Persian. **Persian**, with so far the lowest count of requests as a source language, was queried with Chinese, Croatian, Hindi and Japanese as target languages, but was also requested as a target language for Chinese, Croatian and Hindi source languages. Finally, **Sanskrit**, as a second most searched source language, was paired with Chinese, Croatian, Hindi, Japanese and Korean as target languages, and as a target language for Croatian and Hindi as source languages.

The detailed search history of source and target languages (Table 3) shows that there is quite a large number of searches with the target language “none”. This means that the users tried searching without any specified target language, and that could have happened in two use cases. The first is the possibility of unticking all the target language boxes and thus hitting search without choosing a target language, either on purpose or by mistake. The second entails users changing the target and request languages to the same value (i.e. Source: Hindi, Target: Hindi) where the Target language checkbox gets automatically unticked. The number of such requests (711) accounts for 18.45% of the total (3,853). This could indicate that the users want single language search functionality in conjunction with the existing multiple language translation functionalities. Since we as the designers of the platform are not its core users (Nielsen, 2008), that hypothesis will require further verification in the form of a user questionnaire before any further development.

However, a look at the *eDictionary* word request system (Table 4) shows some rather unfortunate results. The total of only 26 requested words, unevenly divided between Japanese, Hindi and Sanskrit may be a sign of either a job done well on the designers’ part when choosing those words to include in the dictionary, or a students’ lack of interest in actively helping the dictionary to develop.

LANGUAGE REQUESTS	
JAPANESE	19
HINDI	4
SANSKRIT	3

Table 4: Requested words per language.

Since the project was done in cooperation with students that will use it the most, we believe that the answer lies in our first assumption. This would mean that we have already covered the most common words that appear in curricular activities and the usual learning materials. And in spite of the fact that not many words were requested, they were all promptly added through the Admin panel of the website by language experts cooperating on the project.

6.2 The Users

Taking a look into the Google Analytics webpage, we can get some insight into probable explanations as to why there were not as many visits as we anticipated. The *eDictionary* website has had only a handful of active users during the past two years since it was published, the only exception being the first month of its publication (Figure 3).

The data on visitor nationality is in accordance with our expectations. The great majority of visitors, an overwhelming 80.42%, are users from Croatia (Figure 4). Surprisingly enough, the number of visits from the USA ranks it as second, with 3.11% or 28 unique users. Visits from around the region are expected due to the similarities in language and cooperation with colleagues from the region.

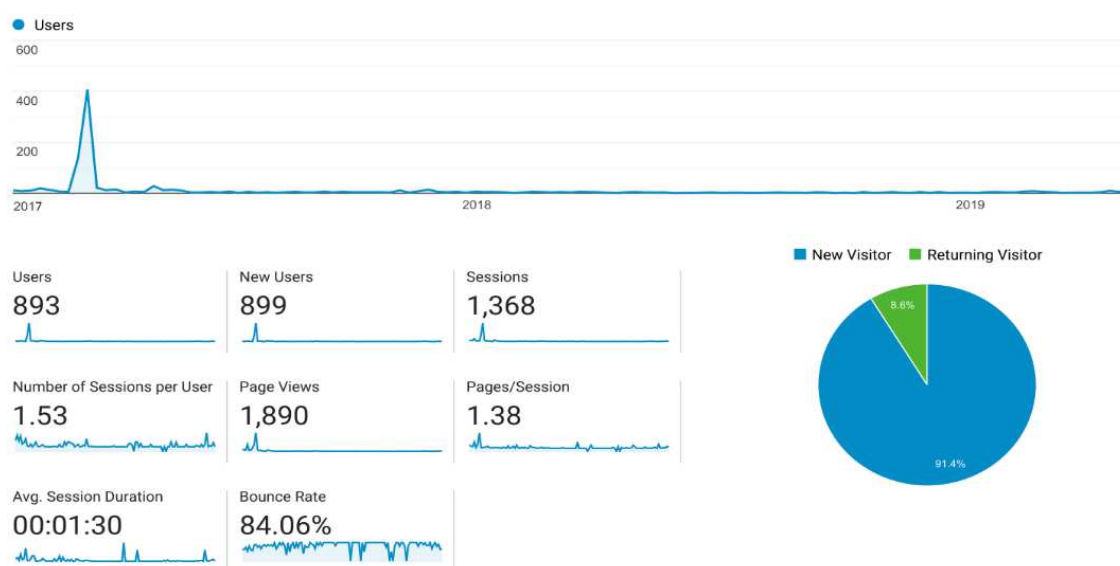


Figure 3: Number of users since 23.01.2017. to 25.05.2019.

Even though the technical side of *eDictionary* has been well thought out and technically polished, the analytics do not speak in favour of site usage. A thorough look through our e-mail system and integrated error reporting tool shows no indications of users having a buggy experience or requesting more materials. Still, somehow, the *eDictionary* project has not been able to collect more than a handful of returning users, which, considering the niche target audience of our project, though slightly discouraging, may not be that unexpected.

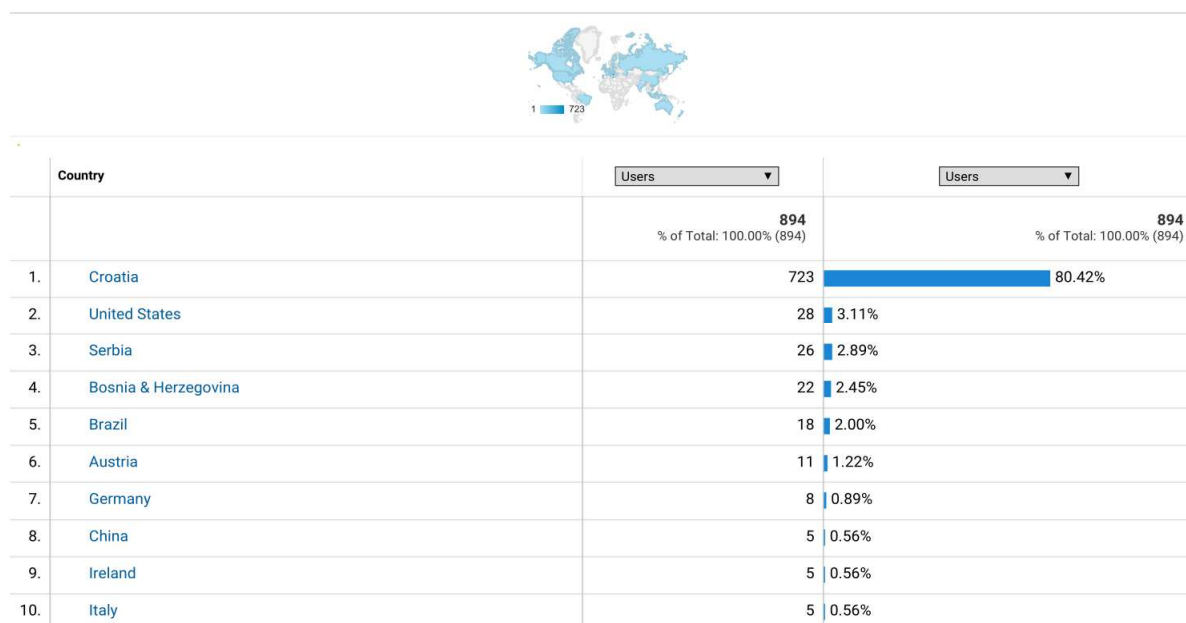


Figure 4: Country of users since 23.01.2017. to 25.05.2019.

7. Conclusion and Future Work

Considering everything that we have learned over the course of this project, the research team came to several conclusions that could help with future work. The creation of *eDictionary* was a rich learning experience for everyone involved in the process, from linguists and programmers to participating students. Students' input on their needs served as an important guideline for the project. However, the MAP students' survey covered just one small group of learners active at that particular moment, and their needs should be revisited and checked against the new generations of students.

Part of the job that was not covered well and should be altered in the future (the part that certainly falls into "the Ugly" category) has to do with the promotion of active use of *eDictionary* as a learning tool among new generations of students. The active role that was envisioned for students (word and pronunciation requests, sending in examples in target languages) turned out to be not so inviting for them. We believe that this could be changed through cooperation with teachers and active integration of *eDictionary* into curricula and lesson plans.

Future work would also entail further strengthening of the *eDictionary* database, including examples and grammatical information in coordination with users' observations. At this stage, it is only accessible as L1 material, understandable and manageable by native speakers. However, since the Croatian language is also used as a second language, one further step would be to make the Croatian database appropriate for such use as well.

8. Acknowledgements

This research was supported in part by the University of Zagreb Research Grant (43-917-1030). *eDictionary* would not be possible without the help of our students, both new and old, as well as their teachers. Their help is gratefully acknowledged. We thank Mr. Navaey for giving us guidance on the Persian resources.

9. References

- Alipour, M., Robichaud, B. & L'Homme M.-C. (2015). Towards an Electronic Specialized Dictionary for Learners. In I. Kosem, M. Jakubíček, J. Kallas & S. Krek (eds.) *Electronic lexicography in the 21st century: linking lexical data in the digital age*. Proceedings of the eLex 2015 conference, Ljubljana/Brighton: Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd., pp. 51–69.
- Deksne, D., Skadiņa, I. & Vasiljevs, A. (2013). The modern electronic dictionary that always provides an answer. In I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets & M. Tuulik (eds.) *Electronic lexicography in the 21st century: thinking outside the paper*. Proceedings of the eLex 2013 conference. Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut, pp. 421–434.
- Dovedan, Z., Seljan, S. & Vučković, K. (2002). Multimedia in Foreign Language Learning. In P. Biljanović & K. Skala (eds.) *Proceedings of the 25th International Convention MIPRO 2002: MEET + MHS*. Rijeka: Liniavera, pp. 72–75.
- Družijanić Hajdarević, E., Vučković, K. & Dovedan, Z. (2006). Računalo ili računalo uz pomoć računala. In *Proceedings of the 29th International Convention MIPRO, Rijeka*, pp. 283–287.
- Hannesdóttir, A. H. (2015). What is a target language in an Electronic Dictionary? In I. Kosem, M. Jakubíček, J. Kallas, & S. Krek (eds.) *Electronic lexicography in the 21st century: linking lexical data in the digital age*. Proceedings of the eLex 2015 conference. Ljubljana/Brighton: Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd., pp. 236–249.
- Heid, U., Prinsloo, D. & Bothma, T. (2013). Dictionary and corpus data in a common portal: state of the art and requirements for the future. *Lexicographica - International Annual for Lexicography / Internationales Jahrbuch für Lexikographie*. 28, 10.1515/lexi.2012–0014, pp. 269–291.
- Heuberger, R. (2016). Learners' Dictionaries: History and Development; Current Issues. In P. Durkin (ed.) *The Oxford Handbook of Lexicography*. Oxford/New York: Oxford University Press, pp. 25–43.
- Hudeček, L. & Mihaljević, M. (2017). The Croatian Web Dictionary Project – Mrežnik. In I. Kosem, C. Tiberius, M. Jakubíček, J. Kallas, S. Krek & V. Baisa (eds.) *Electronic lexicography in the 21st century*. Proceedings of the eLex 2017 conference. Brno: Lexical Computing CZ s.r.o., pp. 172–192.
- Janjić, M., Librenjak, S. & Kocijan, K. (2016a). Asian language teaching and learning - the influence of technology on students' skills in SL classroom. In T. Erjavec &

- D. Fišer (eds.) *Language Technologies & Digital Humanities* 2016, Ljubljana: Ljubljana University Press, Faculty of Arts, pp. 196-197.
- Janjić, M., Librenjak, S. & Kocijan, K. (2016b). Croatian Students' Attitudes Towards Technology Usage in Teaching Asian Languages - a Field Research, In *MIPRO 2016*, Rijeka, pp. 1051-1056.
- Janjić, M., Požega, M., Poljak, D., Librenjak, S. & Kocijan, K. (2017a). E-dictionary for Asian Languages. In I. Atanassova, W. Zaghouani, B. Kragić, K. Aas, H. Stančić, H. & S. Seljan (eds.), *INFuture2017 Proceedings: The Future of Information Sciences*, Zagreb: Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences, University of Zagreb, pp. 213-216.
- Janjić, M., Librenjak, S. & Kocijan, K. (2017b). Nastava stranih jezika: upotreba tehnologije. *Strani jezici: časopis za unapređenje nastave stranih jezika*, 44(4), pp. 232-243.
- Lauc, T., Matić, S. & Mikelić, N. (2006). Educational multimedia software for English language vocabulary. In *Proceedings of the 1st International Conference on Multidisciplinary Information Sciences and Technologies: InSciT2006, Vol. I: Current Research in Information Sciences and Technologies Multidisciplinary approaches to global information systems*, Merida, pp. 117-121.
- Librenjak, S., Janjić, M. & Kocijan, K. (2016a). Sustainable vocabulary acquisition in Japanese classroom with the help of Memrise. In M. Janesova, H. Kratochvilova, I.G. Rotaru, S. Pal, S. Anjali, R. Kratochvil, J. Grover & O. Beyhan (eds.) *Proceedings of International Academic Conference on Global Education, Teaching and Learning in Budapest*. Vestec: Czech Republic: Czech Institute of Academic Education z.s., pp. 54-61.
- Librenjak, S., Janjić, M. & Kocijan, K. (2016b). Computer assisted learning of Japanese verbs - Analysis of errors in usage by Croatian students. In J. Vopava, V. Douda, R. Kratochvil & M. Konecki (eds.) *Proceedings of MAC-ETL 2016*. Prague: Academic Conferences Association, pp. 262-273.
- Librenjak, S., Kocijan, K. & Janjić, M. (2016c). Improving Students' Language Performance Through Consistent Use of E-Learning: An Empirical Study in Japanese, Korean, Hindi and Sanskrit. *Acta Linguistica Asiatica*, 6(2), pp. 79-94.
- Librenjak, S., Vučković, K. & Dovedan Han, Z. (2012). Multimedia assisted learning of Japanese kanji characters. In P. Bijanović, Ž. Butković, K. Skala, S. Golubić, N. Bogunović, S. Ribarić, M. Čičin-Šain, D. Cisić, Ž. Hutinski, M. Baranović, M. Mauher & J. Ulemek (eds.) *Proceedings of 35. jubilee international convention on information and communication technology, electronics and microelectronics*, Rijeka, Croatian Society for Information and Communication Technology, Electronics and Microelectronics - MIPRO, pp. 1284-1289.
- Lorentzen, H. & Theilgaard, L. (2012). Online dictionaries – how do users find them and what do they do once they have. *Proceedings of the 15th EURALEX International Congress*, pp. 654-660.
- Majeed, A. & Rauf, I. (2018). MVC Architecture: A Detailed Insight to the Modern

- Web Applications Development, *Peer Review Journal of Solar & Photoenergy Systems*, vol 1, No 1, PRSP.000505.
- Nielsen, J. (17.3.2008). *Bridging The Designer – User Gap*. Accessed at: <https://www.nngroup.com/articles/bridging-the-designer-user-gap/> (June 1, 2019).
- Paulson, L. D. (2005). Building rich web applications with Ajax. In *Computer*, 38(10), pp. 14-17.
- Pavičić Takač, V. (2008). *Vocabulary Learning Strategies and Foreign Language Acquisition*. Clevedon/Buffalo/Toronto: Multilingual Matters LTD.
- Renau, I. & Battaner, P. (2011). The Spanish Learner's Dictionary DAELE on the Panorama of the Spanish E-lexicography. In I. Kosem & K. Kosem (eds.) *Electronic lexicography in the 21st century: new applications for new users*. Proceedings of the eLex 2011 conference, Ljubljana: Trojina, Institute for Applied Slovene Studies, pp. 221-226.
- Summers, D. (1988). The role of dictionaries in language learning. In R. Carter & M. McCarthy (eds.) *Vocabulary and Language Teaching*. London/New York: Routledge, pp. 111-125.
- Vacalopoulou, A. & Efthimiou, E. (2015). Multilingual lexicography for adult immigrant groups: bringing strange bedfellows together. In I. Kosem, M. Jakubíček, J. Kallas & S. Krek (eds.) *Electronic lexicography in the 21st century: linking lexical data in the digital age*. Proceedings of the eLex 2015 conference. Ljubljana/Brighton: Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd. pp. 315-326.

Websites:

- Baza hrvatskih glagolskih valencijsa. Accessed at: <http://valencije.ihjj.hr> (May 25, 2019)
- Hrvatska Enciklopedija. Accessed at: <http://www.enciklopedija.hr/> (May 20, 2019)
- Hrvatski Jezični Portal. Accessed at: <http://hjp.znanje.hr> (May 17, 2019)
- Struna | Hrvatsko strukovno nazivlje. Accessed at: <http://struna.ihjj.hr/> (May 12, 2019)
- Hellwig, O. (2010-2016). *DCS - The Digital Corpus of Sanskrit*. Berlin. Accessed at: <http://www.sanskrit-linguistics.org/dcs/> (July 15, 2019)

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



DiCoEnviro, a Multilingual Terminological Resource on the Environment: The Brazilian Portuguese Experience

Flávia Cristina Cruz Lamberti Arraes

Departamento de Línguas Estrangeiras e Tradução
Instituto de Letras, Universidade de Brasília, ICC – Ala Sul –
Sala B1 167/63 - Campus Universitário Darcy Ribeiro – Asa Norte –
Brasília/DF CEP: 70910-900
E-mail: flavialamberti@gmail.com

Abstract

DiCoEnviro is a multilingual terminological resource that contains terms in the field of the environment in different languages, i.e. French, English, Spanish, Portuguese, Italian and more recently Chinese. The present paper focuses on the Portuguese version of the resource in order to show how the terminological work has been developed particularly with the use of a Brazilian Portuguese corpus. More specifically the paper presents how DiCoEnviro i) represents the specialized meaning of the terms, ii) represents terminological structures within the environmental domain, and iii) uses lexical functions to establish connections between the terms within a lexical relation. The results show a selection of terms that belong to the environmental domain in Portuguese, particularly to deforestation, their analysis, linguistic description and representation of the most preferred lexical relations the terms establish among themselves. Terms and terminological relations for Portuguese in DiCoEnviro are under construction and our purpose is to increase the number of entries and relations that represent deforestation, as well as to expand the corpus to include other topics associated with the environment.

Keywords: environment; terminology; lexical-semantic approach

1. Introduction

DiCoEnviro is a multilingual terminological resource that contains terms from the subject field of the environment in different languages, i.e. French, English, Spanish, Portuguese, Italian and more recently Chinese. The research in Portuguese was initiated by Botta (2013) with the compilation of the Brazilian Portuguese corpus, selection and analysis of terms and preparation of entries.

The objectives sought by the development of a Portuguese version are: i) to investigate the field of the environment in Portuguese by means of the study of terms; ii) to identify the terms and their specialized meaning; iii) to reveal the terminological relations of the field and to represent them, iv) to establish interlinguistic relations among these languages; and v) to discover semantic frames by describing the linguistic property of terms.

The terminological work is based on the lexico-semantic approach to terminology (L’Homme, 2004a; 2004b; 2012; 2016; 2018). The approach is based on the following principles:

- i) the specialized domain is investigated based on the analysis of terms as lexical units;
- ii) terms are investigated based on the description of their specialized meaning;
- iii) terms are structured, i.e. they establish terminological structures that include two types of relations, the paradigmatic relations and syntagmatic relations.

This paper concentrates on explaining how the Portuguese version of DiCoEnviro describes the specialized meaning of the terms and their terminological structures within the environmental domain. This study refers to the first level of description provided by the resource¹. The paper is structured as follows. Section 2 provides information on the characteristics of the text corpus compiled in Portuguese by Botta (2013) and on the extraction of terms to develop the research. Section 3 focuses on the criteria used to identify the specialized meaning of lexical units. Section 4 concentrates on the linguistic description of terms, particularly the description of i) the lexical meaning and ii) the terminological structures. Section 5 provides details on the use of lexical functions as a model to describe the lexical relations. Finally, Section 6 draws some conclusions and mentions aspects that we wish to explore in the future.

2. Linguistic data

DiCoEnviro is a specialized dictionary which presents terms that belong to the environmental domain in different languages. The description of lexical units is heavily derived from a corpus, more specifically a specialized corpus containing environmental texts mainly from the subdomain of deforestation.

The Brazilian Portuguese corpus is composed of scientific and journalistic texts in the period between 1981 and 2012. The corpus was compiled by Botta (2013) and contains 136,910 words (types of words) in the scientific corpus and 139,943 in the journalistic one (Botta, 2013). The texts are stored in Intercorpus², an online concordancer, from which contexts are extracted. It produces KWIC (key word in contexts) concordance lines which are accessed in plain text by clicking on the keyword.

The Portuguese specialized lexical units on the subdomain of deforestation represent a distinct terminology. By applying automatic term extraction software called TermoStat, created by Drouin (2003), we can extract several lexical units that give us access to a

¹ The resource has three levels of description: i) a lexical resource, composed of lexical relations based on Melčuk et al. (1995); ii) contextual annotations and iii) semantic frames module.

² Chièze, E.; Polguère, A. (no date) available at <http://olst.ling.umontreal.ca/intercorpus/>.

selection of candidate terms in our specialized corpus³. Table 1 shows the list containing the first group of lexical units extracted by the software.

Candidate terms	Frequency	Score of specificity	Orthographic variants
floresta	1514	127.86	floresta____florestas
área	2880	118.91	área____áreas
solo	1406	115.16	solo____solos
atividade	1012	114.48	atividade____atividades
desmatamento	965	111.72	desmatamento____desmatamentos
manejo	865	105.34	manejo____manejos
espécie	1568	102.21	espécie____espécies
uso	1164	96.31	uso____usos
a	768	96.09	a____as
amazônia	695	94.84	amazônia
projeto	664	92.7	projeto____projetos
mata	688	83.34	mata____matas
plantio	539	82.72	plantio____plantios
ação	463	77.38	ação____ações
fator	454	76.62	fator____fatores
desenvolvimento	1304	72.79	desenvolvimento____desenvolvimentos
recurso	1032	70.2	recurso____recursos
pecuária	400	70.12	pecuária
vegetação	426	69.96	vegetação____vegetações
setor	369	69.06	setor____setores
carbono	435	68.72	carbono____carbonos
custo	926	66.62	custo____custos
sustentabilidade	341	65.29	sustentabilidade

Table 1: Automatic extraction of terms by TermoStat (Drouin, 2003) from the corpus compiled by Botta (2003).

Based on a reference corpus, which is a non-technical corpus, the software compares the behaviour of lexical units in both corpora and identifies the lexical items that are specific to the specialized corpus. The results are provided based on frequency and on the score of specificity (Drouin, 2003).

The list is further analysed manually by researchers in order to select true terms based on criteria to identify terms (L’Homme, 2004a: 64-66). The first criterion establishes

³ TermoStat may extract single-word and/or multi-word entries. However, the criteria applied to identify terms requires selection of single-word entries. The analysis may then identify compositional and non-compositional sequences as having a specialized meaning. A non-compositional sequence (a sequence whose meaning bears no relation to its parts or to some of its parts) is accepted as an entry; a compositional one is not for its components are regarded as entries themselves.

that we have a term when the lexical unit is closely related to the specialized domain. The list above presented by TermoStat offers us a list of lexical units that can be related specifically to the environment, such as *floresta*, *solo*, *desmatamento*, *espécie*, *amazônia*, *mata*, *vegetação*, *sustentabilidade*. Other criteria are applied when the link is not easily or clearly established. They are particularly applied to predicative units, such as verbs and activity nouns (e.g. *desmatar* and *desmatamento*) and adjectives, which are described in more detail in the next section.

3. Lexical units with a specialized meaning

In the lexico-semantic approach to terminology, terms are considered lexical units with a specialized meaning. This approach aims at investigating the terms with a specific focus on the description of their linguistic properties. Some lexical units are unanimously considered attached to a specialized domain (e.g. *floresta*, *solo* above). However other lexical units may not be directly associated with a specialized domain, particularly verbs, activity nouns, adjectives and adverbs (named predicative units).

In these cases, we may apply the second criterion proposed by L’Homme (2004a: 64), namely the analysis of the nature of the semantic arguments that interact linguistically with the lexical unit in focus. If the arguments are terms validated by the first criterion (i.e. they are related to a specialized domain), the lexical unit in focus is also a term. For example, the meaning of the verb *preservar* 1 requires two other arguments: 1. Someone (e.g. *homem*) or something (e.g. *sistema*) that preserves; 2. The thing that is preserved (e.g. *meio ambiente*, *floresta*). If the arguments are validated as terms by the first criterion, the predicative unit is also considered a term. *Preservar* is considered a term because *homem*, *sistema*, *meio ambiente*, and *floresta* are recognized as terms.

Other criteria were proposed by L’Homme (2004a: 64-66), namely i) a morphological relationship with a term, particularly those derived from word-formation processes. For example: the derivatives of *floresta*, such as verbs like *florestar*, *desflorestar*, *reflorestar*, and their nominal counterparts, *florestamento*, *desflorestamento*, *reflorestamento*, respectively; and ii) a paradigmatic relationship with the term. For example: a semantic relationship of quasi-synonym between *desmatamento* and *desflorestamento*, and an opposite relationship between, for example, *florestamento* and *desflorestamento*, a relationship of opposition in which both units represent a different perspective on a situation, ‘with trees’, and ‘without trees’ (Gagner; L’Homme, 2015).

Next, we show how DicoEnviro represents the linguistic description of the different kinds of terms mentioned above.

4. Linguistic description

DiCoEnviro includes different kinds of terms in contrast with typical terminological resources. It includes not only entities, usually denoted by concrete and physical things

(e.g. *biomassa*, *água*), but also verbs (*preservar*, *conservar*, *proteger*), nouns that denote activities (*preservação*, *conservação*, *proteção*) and adjectives (*degradado*, *desmatado*, *manejado*).

The linguistic descriptions of terms are placed in a terminological file which is divided in three main sections: i) a section that describes the specialized meaning of the term, ii) a section that presents the contexts; and iii) a section, named Lexical Relations, that describes the terminological structures established by the entry with other terms. The term to be described is extracted from contexts of occurrences; by default, three contexts are shown in the file. Next we present how the specialized meaning is represented and the types of terminological structures under attention in the resource.

4.1 Description of the specialized lexical meaning

Entities and predicative units are included in the DiCoEnviro. Entities, named in the literature semantic nouns (*noms sémantique* in Polguère, 2016: 164), are physical entities such as water, air, planet, plant, tree, etc. Their meaning is not a connecting one, and therefore no participants are expressed. We show below how DiCoEnviro represents this type of meaning taking as example the entry *água* 1 (Portuguese):

água 1 , n. f. a água	status: 2
Definição:	

Table 2: Meaning representation taken from the entry ÁGUA in the DicoEnviro.

The meaning of entities is to be expressed in a specific field for the definition (Definição in the terminological file). The specialized lexical meaning of a predicative unit is described based on the expression of its argument structure. A predicative unit is called a semantic predicate (*prédicats sémantiques*) in the literature (Mel'čuk et al., 1995: 76; Polguère, 2016: 162-163). Mel'čuk et al. (1995: 76) defines a semantic predicate as a:

“...connecting” meaning – it gathers other meanings in a semantic configuration arranged like a connecting tube that links the poles of a shelter in order to form the structure that supports the shelter. The semantic predicates designate actions, events, processes, states, properties, relations, etc in one word; this behaviour necessarily entails participants.⁴

Below we show how DiCoEnviro represents the meaning of a semantic predicate, the term *preservar* 1 selected from DiCoEnviro:

⁴ “...sens ‘liant’ - il réunit d’autres sens en des configurations sémantiques tout comme un tube de jonction réunit les pôles d’une tente pour former le squelette porteur de la tente. Les prédicats sémantiques désignent des actions, des événements, des processus, des états, des propriétés, des relations, etc, - en un mot, des faits qui impliquent nécessairement des participants” (Mel'čuk et al., 1995: 76).


preservar 1 v. tr.	status: 2
preservar: homem  ou sistema ~ meio ambiente, floresta 1	

Table 3: Argument structure extracted from the headword *preservar* 1 in the DicoEnviro.

Two typical participants are established in the argument structure of *preservar*: 1. The agent *homem* and the cause *sistema*; 2. The patient *meio ambiente* and *floresta*. Other contexts may reveal other arguments, other agentive participants such as *agricultor*, *fazendeiro*, *proprietário*; and other participants that are affected by the action of the verb, such as *bioma*, *espécie*, *fauna*, *flora*. However, the typical terms, i.e. the terms that seem to be more natural and frequent cooccurring with the term in focus, are the ones that are expressed first in the argument structure (L’Homme & Laneville, 2009).

4.2 Terminological structures

This section provides details on the types of lexical relations established between terms that are semantically related to the entry and how these relations are represented in the DiCoEnviro. This is based on the consideration that the lexical system of a language is not simply a list of lexical units, but a “vast lexical network: an extremely rich and complex system of lexical units connected to one another” (Polguère, 2016: 130). In this system each lexical unit has a value by means of which multiple types of relations are established. For this reason, the terms are thought to be structured within a system of relations established with other terms that belong to the specialized domain.

There are two major types of relations established by lexical units, as observed by Polguère (2016: 130):



1. Paradigmatic relations: they connect lexical units by means of semantic relations, which can eventually be accompanied by morphological ones. For example, the verbs *preservar* and *conservar* are quasi-synonyms; *preservar* and *proteger* are related meanings.
2. Syntagmatic relations: they link lexical units based on the most preferred combinations established in the syntactic axis of a language. For example, *preservar a área*, *~ a vegetação*, *~ a fauna*, *~ a flora*.

In the lexico-semantic approach the research focuses on the different types of paradigmatic and syntagmatic relations the terms establish among themselves. L’Homme (2004a: 83-118) names these *terminological structures* because they are identified within a specialized domain. Two types of terminological structures are envisaged: i) the classical lexico-semantic relations, composed of different types: taxonomic relations, synonymy and near synonymy, antonymy, meronymy; ii) other lexico-semantic relations particularly composed of combinations, such as collocations.

In DiCoEnviro the terminological structures are represented in the terminological file in a field named *Lexical Relations*. A list of terms that are semantically related to the entry is provided along with a short explanation of the relation. Terms that are available online are hyperlinked, allowing users to access their entries directly.




The *Lexical Relations* are composed of the following families: related meanings, opposites, types of, parts of speech and derivatives, combinations and others, as described below.

The family *Related Meanings* (*Voisins* in French; *Significados Relacionados* in Portuguese) includes the following relationships: near synonyms, related meaning and generic relation. For example, the entries *preservar* 1 and *conservar* 1 are analysed as ‘near synonyms’ because the data analysis shows that they may be interchanged in some contexts. On the other hand, *preservar* 1 and *proteger* 1 are analysed as related meaning (*sentido vizinho*) because they may not be interchanged and their argument structure displays a different configuration, as it is shown below:

preservar 1 , v. tr. status: 2
 preservar: homem  ou sistema ~ meio ambiente, floresta 1 
 Contexto(s)
 Relações lexicais

Explicação	Termos relacionados
Significados relacionados	
Quase sinônimo	conservar 1
Sentido vizinho	proteger 1

Table 4: Lexical relations extracted from the entry *preservar* 1 in DiCoEnviro.



proteger 1 , v. tr. status: 2
 homem  ~ recurso, espécie  contra degradação 1 
 Context(s)
 Lexical relations

Explanation	Termos relacionados
Significados relacionados	
Sentido vizinho	preservar 1 conservar 1

Table 5: Lexical relations extracted from the entry *proteger* 1 in DiCoEnviro.


The family *Opposites* (*Contraires* in French and *Opostos* in Portuguese) includes four main categories of opposite relationships: *antonym* (complementary and reversion),

opposite (near gradable, near reversible), *conversive* and *contrastive*⁵ (Gagné & L’Homme, 2016). DiCoEnviro considers, for example, that pairs such as *florestamento* and *desflorestamento* (English *afforestation* and *deforestation*) do not establish a canonical type of opposition (meaning the negation of one member of the pair necessarily entails the assertion of the other); they are considered, on the other hand, a type of reversible, a reversible 1.

florestamento 1 , n. m. status: 2
 florestamento: ~ da área  pelo homem para colocar árvore 
 Contexto(s)
 Relações lexicais

Explicação	Termos relacionados
Opostos	
Antônimo	desflorestamento 1 desmatamento 1

Table 6: Lexical relations extracted from the entry *florestamento* 1 in DiCoEnviro

desflorestamento 1 , n. m. status: 2
 desflorestamento: ~ de região  por homem para retirar árvore
 Contexto(s)
 Relações lexicais

Explicação	Termos relacionados
Significados relacionados	
	desmatamento 1
Opostos	
Antônimo	florestamento 1
Oposto	reflorestamento 1

Table 7: Lexical relations extracted from the entry *desflorestamento* 1 in DiCoEnviro

⁵ Gagné & L’Homme (2016) identified these different types of categories in a research based on data extracted from DiCoEnviro.

According to Gagné and L’Homme (2016: 16), “reversives 1 consist in a change of direction applied to an entity between two absolute states (...). Therefore, the initial state of the first member corresponds to the final state of the second member and vice versa, so both members represent a different perspective on a situation”.

Some lexical units establish an atypical type of opposition. In these cases, we add ‘near’ to the pairs. The terms *desflorestamento* 1 and *reflorestamento* 1 are considered ‘opposite’ (*oposto*) and not pure reversives (*antônimos*) because the change of direction, implied in a reversible case, is not an absolute state, i.e. *desflorestamento* 1 does not entail necessarily *reflorestamento*. The entries mentioned above are presented in Tables 6 and 7.

The family *Other Parts of Speech and Derivatives* (*Autres parties du discours et dérivés* in French and *Outras partes do discurso e derivados* in Portuguese) accounts for the morphological relations a term shares with the entry. For example: same meaning but different parts of speech: e.g. *desflorestar* (verb) → *desflorestamento* (noun); *desflorestar* (verb) → *desflorestado* (adjective). Table 8 shows the relationships represented in the DiCoEnviro.

desflorestar 1 , v. tr.

desflorestar: homem ~ mata ➕ para retirar árvore

ContextsLexical relations


Explanation	Related term
Other Parts of Speech and Derivatives	
Nome	desflorestamento 1
Uma mata que foi d.	desflorestado 1

Table 8: Lexical relations extracted from the entry *desflorestar* 1 in DiCoEnviro

The family *Types of* (*Sortes de* in French and *Tipos de* in Portuguese) accounts either for paradigmatic relations or syntagmatic relations (combinations). The paradigmatic relations contain single-word terms that represent, for example, a generic-specific relationship, i.e. the hyponyms related to the entry are represented (e.g. *floresta* is a ‘type of’ *ecossistema* – the generic). The syntagmatic relations involve properties and are expressed linguistically by the collocates of an entry. In the DiCoEnviro, the way the collocate combines with the entry is specified: e.g. *ecossistema* → ~ *aquático*; ~ *florestal*.

The family *Combinations* (*Combinatoire* in French and *Combinações* in Portuguese), on the other hand, accounts for syntagmatic relations that involve activities. The relations are also expressed linguistically by the collocates of an entry. The specification of the combination is represented as follows: *ecossistema* \rightarrow *ameaçar o* \sim ; or the nominalization: *ecossistema* \rightarrow *ameaça ao* \sim . Below we show the representation of these relationships in the entry *ecossistema*:

ecossistema 1 , n. m.

um ecossistema: \sim de floresta 1 

Contexts

Lexical relations

Explanation	Related term
Types of	
Que é relativo a uma área específica	\sim aquático \sim florestal 1 (...)
Combinations	
Alguém ou algo pode apresentar um risco ao e.	ameaçar 1 o \sim
Nome para alguém ou algo pode apresentar um risco ao e.	ameaça 1 ao \sim (...)

Table 9: Lexical relations (‘Types of’ and ‘Combinations’) extracted from the entry *Ecossistema* 1 in DiCoEnviro

5. Lexical functions

In the DiCoEnviro, the paradigmatic and syntagmatic relations are encoded in the database using lexical functions, LF, (Melčuk et al., 1995; Polguère, 2016). This system allows the encoding of the syntactic and semantic properties of paradigmatic relations and syntagmatic relations (i.e. collocations). For example: assuming that *desflorestar* 1 has the following argument structure:

DESFLORESTAR 1 : AGENTE {homem} \sim ORIGEM {mata} para retirar PACIENTE {árvore}

and that DESFLORESTAMENTO 1 and DESFLORESTADO 1 are related semantically, each relation will be defined based on lexical function, as follows:

S0 (DESFLORESTAR 1) = DESFLORESTAMENTO 1 (noun that conveys the same meaning)

A2 (DESFLORESTAR 1) = DESFLORESTADO 1 (the adjective that applies to the second argument of DESFLORESTAR)

If we were to encode the related term DESFLORESTADO to the entry DESFLORESTAR, the lexical relation would be assigned to “Other parts of speech and derivatives” due to the morphological and semantic relation between the terms: *desflorestado* is the adjective form of the verb *desflorestar*. The information that is inserted is shown below⁶:

```
<famille nom="Autres parties du discours et dérivés">
<lien-lexical>
<explication-ra>Uma <role-ref nom="Origem"/> que foi <lexie-ref/> </explication-ra>
<explication-tt>Uma <role-ref nom="Origem" lemme="mata"/> que foi <lexie-ref/>
</explication-tt>
<fonction-lexicale>A2Perf</fonction-lexicale>
<lien identificateur="desflorestado" numero-acceptation="1" xlink:type="simple"
xlink:show="replace" xlink:actuate="onRequest"
xlink:href="desflorestado.xml#_desflorestado1">desflorestado 1</lien>
</lien-lexical>
</famille>
```

Table 10: Encoding of the related term *desflorestado* 1 in the entry DESFLORESTADO 1

In the database of DiCoEnviro, three levels of explanation are provided for each relation: the first two are divided into two systems (L’Homme, 2012: 384-385): the first one (explication-ra) explains the relation in terms of semantic roles (e.g. Uma Origem que foi “entry”); the second one (explication-tt) refers to the typical term (e.g. Uma mata que foi d.). Then the lexical function (A2Perf) is indicated. Finally, a pointer to the related term is given (DESFLORESTADO1).

Each relation is encoded with the use of an LF based on the type of relation established with another term. Although the LFs are formally codified, the Web version of DiCoEnviro displays only explanations in natural language. Table 11 shows the relationships listed in Section 4 represented by means of LFs in English and Portuguese provided with a short explanation on the left.

⁶ An XML editor (Oxygen) is used to add entries to the database.

RELATION	EXAMPLE	LF
SIGNIFICADOS RELACIONADOS (RELATED MEANINGS)		
Quase-sinônimo (near synonym)	Preservar → conservar, maintain	QSyn
Sentido vizinho (related meaning) or	Preservar → proteger	Cf
OPOSTOS (OPPOSITES)		
Antônimo (Antonym)	Florestamento → desflorestamento, desmatamento	Rev1
Oposto (Opposite)	Desflorestamento → reflorestamento	QRev1
OUTRAS PARTES DO DISCURSO E DERIVADOS (OTHER PARTS OF SPEECH AND DERIVATIVES)		
Nome (Noun)	Desflorestar → desflorestamento	S0
Uma mata que foi d. (A forest that was deforested)	Desflorestar → desflorestado	A2Perf
TIPOS DE (TYPES OF)		
Tipo de (Type of)	Ecossistema → floresta	[Spec]
Que é relativo a uma área específica. (That concerns a specific location)	Ecossistema → ~ aquático, ~ florestal	Hypo - Lieu
COMBINAÇÕES (COMBINATIONS)		
Alguém ou algo pode apresentar um risco ao e. (Someone or something may cause the e. to be in a worse state)	Ecossistema → ameaçar o ~	Caus@AbleDegrad
Nome para alguém ou algo pode apresentar um risco ao e. (Someone or something may cause the e. to be in a worse state)	Ecossistema → ameaça ao ~	S0Caus@AbleDegrad

Table 11: Examples of terminological relations and lexical functions encoded in the DiCoEnviro

6. Concluding remarks

The paper presented the Portuguese version of DiCoEnviro referring particularly to its first level of description, the lexical resource. In this level, the resource concentrates on

the description of the specialized meaning of lexical units and on the description of the terminological structures established among the terms.

The lexical units include entities and predicative units. The meaning of predicative units is described in the argument structure section in which the core participants are stated. Subsequently, two types of terminological structures are described, one based on paradigmatic relations and the other on syntagmatic relations established among the terms. The lexical functions are the formal mechanism to encode the paradigmatic and syntagmatic relations in the database (Oxygen XML editor).

The coverage in Portuguese differs quite drastically from that in French and English. Data taken into account as of February 2018 (L’Homme, 2018) show French, English and Portuguese have the following number of entries and relations: English (982 entries, 11,942 relations), French (1,309 entries, 16,723 relations), and Portuguese (37 entries, 563 relations).

Terms and terminological relations for Portuguese in the DiCoEnviro are under construction and our purpose is to increase the number of entries and relations that represent deforestation, as well as to expand the corpus to include other topics (e.g. climate change, endangered species, recycling, sustainable development), associated with the environment.

7. Acknowledgements

The research is supported by funding from “Fundação de Apoio à Pesquisa do Distrito Federal” (FAP/DF) - Proposal call – Edital 3/2018).

I would like to thank professor Marie-Claude L’Homme from the Observatoire de Linguistique Sens Texte (OLST), University of Montreal, Canada, for the useful suggestions on the paper.

8. References

- Botta, M. Giacomini (2013). Comportamento dos termos do meio ambiente em textos de vulgarização. *TradTerm*, 22(1), pp. 185-210.
- Drouin, P. (2003). Term Extraction Using Non-Technical Corpora as a Point of Leverage. *Terminology*, 9(1), pp. 99–117.
- Gagné, A. M. & L’Homme, M.C. (2016). Opposite relationships in terminology. *Terminology* 22(1), 30-51.
- L’Homme, M. C. (2004a). *La terminologie: principes et techniques*. Montreal: Presses de l’Université de Montréal.
- L’Homme, M. C. (2004b). A Lexico-semantic Approach to the Structuring of Terminology. In *Computerm 2004, Coling 2004, Université de Genève, Geneva (Switzerland)*, pp. 7-14.

- L'Homme, M. C. (2012). Using ECL (Explanatory Combinatorial Lexicology) to discover the lexical structure of specialized subject fields. In J. Apresjan et al. (eds.) *Words, Meanings and other Interesting Things. A Festschrift in Honour of the 80th Anniversary of Professor Igor Alexandrovic Mel'čuk*. Moscow: RCK, pp. 378-390.
- L'Homme, M. C. (2016). Terminologie de l'environnement et sémantique des cadres. In *SHS Web of Conferences* 27 05010, *Congrès Mondial de Linguistique Française*.
- L'Homme, M. C & Laneville, M. È (2009). *Le dictionnaire fondamental de l'environnement*: Dictionnaire élaboré par l'équipe ÉCLECTIK. Accessed at: <http://olst.ling.umontreal.ca/dicoenviro/manuel-DiCoEnviro.pdf>. (June 2019)
- L'Homme, M. C., Robichaud, B. & Prével, N. (2018). Browsing the Terminological Structure of a Specialized Domain: A Method Based on Lexical Functions and their Classification. In *Language Resources and Evaluation*, LREC 2018. Myazaki, Japan.
- Mel'čuk, I., Clas, A. & Polguère, A. (1995). *Introduction à la lexicologie explicative: et combinatoire*. Louvain-la-Neuve (Belgium): Duculot.
- Polguère, A. (2016). *Lexicologie et sémantique lexicale*. Montreal: Les Presses de l'Université de Montréal.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



Modelling Specialized Knowledge With Conceptual Frames: The TermFrame Approach to a Structured Visual Domain Representation

Špela Vintar, Amanda Saksida, Katarina Vrtovec,

Uroš Stepišnik

Faculty of Arts, University of Ljubljana, Aškerčeva 2, SI – 1000 Ljubljana

E-mail: {spela.vintar, amanda.saksida, katarina.vrtovec, uros.stepisnik} @ff.uni-lj.si

Abstract

We describe an emerging knowledge base for karstology developed in line with the frame-based approach with data for three languages, English, Slovene and Croatian. An annotation framework was developed to identify the definition elements, semantic categories, relations and relation definitors in definitions of karst concepts extracted from specialized corpora. A multi-layered annotation was performed for sets of validated English and Slovene definitions. We present the distribution of semantic categories and typical definition frames for the most prominent semantic categories: surface and underground landforms, hydrological forms and geomes, for English and Slovene. The definition frames specify the typical properties of concepts we expect to be described, and in our case they were initialized by domain experts and then verified through corpus data. The structured domain representation resulting from the annotated corpus allows us to compare knowledge structures between languages, generate ideal definitions and experiment with domain visualisations, graphs and maps of geolocations.

Keywords: frame-based terminology; knowledge modelling; karstology; semantic annotation

1. Introduction

Domain terminologies are often thought of as structured and systematic networks of concepts which allow for efficient and unambiguous communication between experts. Traditional specialized dictionaries proved – through their alphabetic ordering alone – inadequate for representing concepts as abstract units of knowledge, but termbases in digital format can easily accommodate the concept-oriented approach and utilize the terminological entry as the tangible equivalent of the concept residing in the cognitive realm. Indeed, many online multilingual termbases such as IATE¹ or UMLS² embody this approach.

¹ <https://iate.europa.eu/>

² <https://www.nlm.nih.gov/research/umls/>

The frame-based approach to terminology (Faber et al., 2005; Faber, 2009; Faber et al., 2012) has provided a valuable new framework for representing specialized knowledge by combining linguistic information derived from specialized corpora with conceptual structures and by highlighting the fact that the cognitive frames underlying specialized communication are dynamic, context-, language- and culture-dependent (Leitchik & Shelov, 2007; Temmermann & Van Campenhoudt, 2014; Faber & Medina-Rull, 2017). Moreover, the concepts of a specialized domain should not be described in isolation but represented as nodes in an intricate knowledge network illustrating both generic and domain-specific relations between them. A widely known implementation of these principles is the EcoLexicon³, a multilingual knowledge base for the environmental domain.

The TermFrame project adopts the frame-based rationale, but adapts and extends existing methodologies with the following goals in mind:

- To build a comprehensive structured knowledge base for the domain of karstology in three languages – English, Slovene and Croatian;
- To develop modes of knowledge representation which can be used by linguists, terminologists, experts and data scientists alike, and which adequately show language- and context-dependent differences between knowledge frames;
- To explore new methods of knowledge extraction from specialized texts, so that our results can be generalized and applied to new languages and domains.

This paper focuses on the semantic annotation framework and the resulting resources which can serve both as input for knowledge visualization and as training data for future knowledge extraction tools. It is structured as follows: Section 2 gives a brief overview of related work on terminological definitions and their semantic structure from the Frame Semantics point of view. Section 3 describes the resources built and used in TermFrame, including the tools for term and definition extraction. In Section 4 we give a detailed explanation of our annotation framework and provide examples of annotated definitions, followed by some quantitative data from the annotated corpora and an illustration of the resulting domain representation in Section 5.

2. Definitions and frames

The terminological definition is the most concentrated means of communicating expert knowledge which helps users understand the meaning of a specialized lexical unit (Seppälä & Ruttenberg, 2013: 19). Although its structure was originally defined by Aristotle, the textual reality shows that authors use varying definition styles (Svensen, 1993: 117; Roche et al., 2009), while several attempts have been made to devise a

³ <http://ecolexicon.ugr.es/en/index.htm>

typology of definitions (Blanchon, 1997; Seppälä, 2007; Diki-Kidiri, 2000; Madsen/Thomsen, 2008; Pollak, 2010).

Here we refrain from delving deeper into the definition types and the factors which may influence the author to use a certain defining style over another, although some understanding of this variety is needed for automatic definition extraction, as we show in Section 3. It should be stressed, however, that the choice of semantic elements used to delineate specialized meaning is not arbitrary, and the frame-based approach helps us discern predominant definition templates or frames, or even guide definition formation, as shown for example in San Martin & L’Homme (2014) and Duran-Muñoz (2016). The definition template is usually related to the semantic category of the concept and reflects its role in the domain-specific event.

In our own previous work (Vintar & Grčić Simeunović, 2017), a cross-language analysis of definition frames in karstology revealed interesting differences between English and Croatian. Karst as a core concept is defined in Croatian mostly through its geomorphological features and settings, while in English we found several instances where karst or its subtypes were defined as the geomorphologic or hydrologic functioning of the karst processes. The underlying cognitive frame is in this case clearly language-dependent.

3. TermFrame resources

For the purposes of our research we built three corpora, Slovene, English and Croatian. The corpora contain relevant contemporary works on karstology and are comparable in terms of the domain and text types included. The corpora comprise scientific texts (scientific papers, books, articles, doctoral and master’s theses, glossaries and dictionaries) from the field of karstology, which in itself is an interdisciplinary domain partly overlapping with surface and subsurface geomorphology, geology, hydrology and other fields. Table 1 gives basic information about the corpus.

	English	Slovene	Croatian
Tokens	2,721,042	1,208,240	1,229,368
Words	2,195,982	987,801	969,735
Sentences	97,187	51,990	53,017
Documents	57	60	43

Table 1: The TermFrame corpora

Once the corpora were compiled we performed term and definition extraction and other knowledge mining steps described in Pollak et al. (2019). Definition candidates were extracted automatically with the pattern-based setup of ClowdFlows, which according

to previous research performs best (Pollak et al., 2012). At this time the tool yet has to be adapted to Croatian, hence the remainder of this paper reports results for English and Slovene only. Also, in the first stage definition extraction was performed on approximately half of the English corpus. The extracted sentences were manually validated to retain only contexts with valuable explanatory information about the karst concept. Given this relatively broad view many of our definitions do not necessarily comply with the traditional definition structure: in many cases the definiendum appears at the end of the sentence, the genus or hypernym may be missing, and several examples of extensional definitions were found. After validation the yield was 215 and 259 terms for English and Slovene, respectively.

The semantic annotation of definitions was performed in WebAnno, an open source server-based tool which allows users to specify the annotation layers, attributes and tagsets, and perform annotation, curation and monitoring (De Castilho et al., 2014). In our workflow, each definition was annotated by two persons (linguists), then curation was performed by a domain expert. Regular meetings of all annotators and curators were organized to discuss ambiguities and consolidate the annotation procedure.

4. Annotating definitions in TermFrame

4.1 The annotation framework

The development of the annotation framework is an essential step in domain modelling as it attempts to produce a mapping between the cognitive level representing expert knowledge, the textual reality describing this knowledge, and a formal level with structures, categories and relations. The primary purpose of such a mapping is to allow for an accurate and functional representation of the domain. At the same time, a secondary purpose is to provide insight into linguistic features which may be used for automatic knowledge extraction not just in the domain of choice, but potentially also in other domains. Our project team consists of linguists, a cognitive scientist, a karstologist and several experts in NLP, and has developed a framework able to accommodate both these purposes.

The annotation consists of five layers:

1. Definition element. This layer identifies the following elements of the definition: DEFINIENDUM (the term which is being defined), DEFINITOR (the defining phrase of the definition, usually a verbal phrase), GENUS (the hypernym or superordinate term), and SPECIES (the hyponym or subordinate term; relevant in extensional definitions). Though not annotated, the IS_A relation is implicit between DEFINIENDUM and GENUS (sandstone IS_A rock), and SPECIES and DEFINIENDUM (doline IS_A karst depression).

2. **Semantic category.** This is a hierarchical framework which used the EcoLexicon conceptual hierarchy as a starting point, but was adapted to karstology in collaboration with domain experts. It uses five top-level categories (for details see Figure 1). The concepts represented by the categories were modelled according to the basic karstologic approach (Ford & Williams, 2007; Jennings, 1985) corresponding to surface and subsurface karst landforms (Landform) and a number of related processes (Process). Other categories included typical karst environments (Geome), materials, processes and landforms closely connected to karst environments (Entity/Element/Property) and typical methods and tools in karstology (Instrument/Method).

3. **Relation.** We use a set of 16 relations, each of which marks a specific property or feature of the definiendum. Relations may span over several words or phrases and do not necessarily overlap with the two previous layers. Thus, in the example sentence in Figure 2 the relation COMPOSED_OF is expressed in the text

by of freshly formed gypsum.

 The following relations were defined by domain experts according to the geomorphologic analytical approach (Pavlopoulos et al., 2009) considering spatial distribution (HAS_LOCATION; HAS_POSITION), morphography (HAS_FORM; CONTAINS), morphometry (HAS_SIZE), morphostructure (OCCURS_IN_MEDIUM; COMPOSED_OF), morphogenesis (HAS_CAUSE), morphodynamics (AFFECTS; HAS_RESULT; HAS_FUNCTION), and morphochronology (OCCURS_IN_TIME). Additional relations were applied for general properties (HAS_ATTRIBUTE; DEFINED_AS), and for research methods (STUDIES; MEASURES).

4. **Relation_definitior.** This layer was introduced to facilitate potential knowledge extraction experiments, but also for easier access to the concept features expressed by the relations. In the example below, the composition of the definiendum *sandstone* is expressed by the phrase *made of cemented quartz sand*, where *made of* is the relation definitior.

5. **Term_canonical.** This layer was added primarily for term normalization purposes in elliptic constructions, for example in *water discharge and velocity* we may add water velocity as the canonical or full version of the term.

Typically, the definition has one definiendum, although in our corpus and domain it is not uncommon to list term variants for certain karst phenomena; in such cases (see below) all synonymous term variants were marked as definienda. We may find definitions without a genus, for example extensional or functional definitions. In the case of extensional definitions listing members of a class we mark hyponyms as SPECIES.

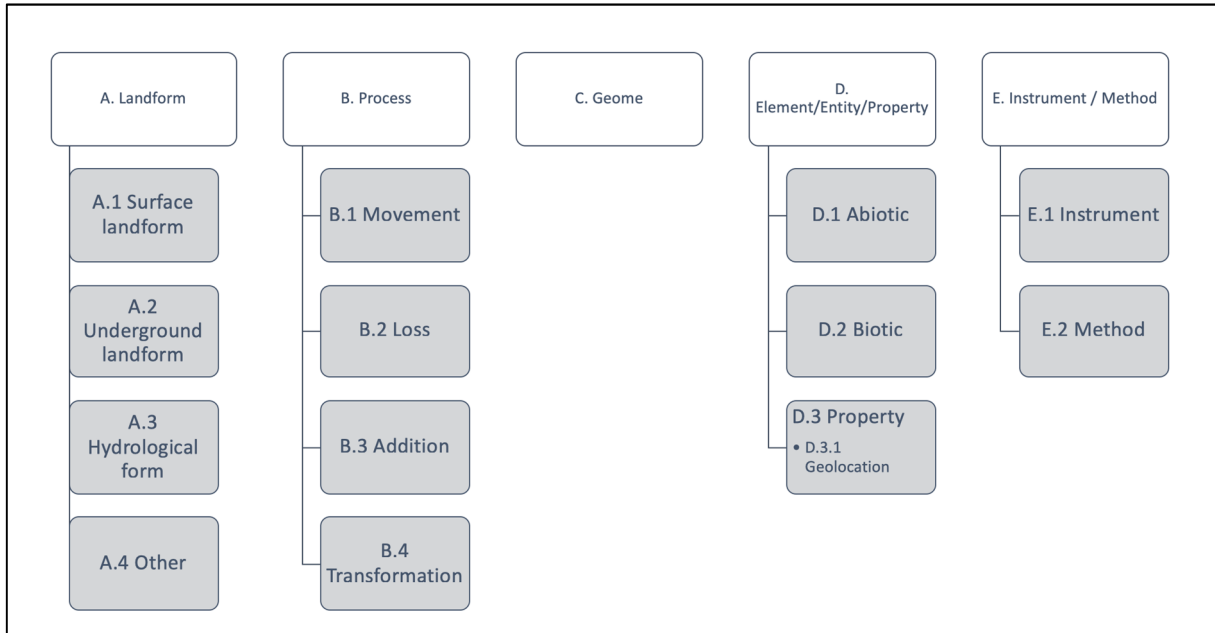
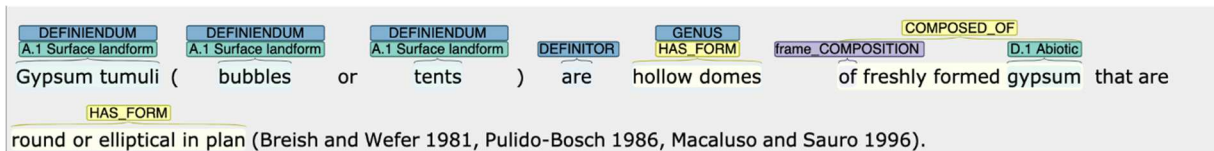


Figure 1: Semantic categories in TermFrame.

Semantic categories are assigned to terms or term-like expressions pertaining to karstology, whereby some categories (e.g. D.1 Abiotic) include terms from the broader domains of geography, geology and chemistry. The definiendum must always be assigned a semantic category, and it is expected that the genus – if present – will share the same category as the definiendum.



The choice of relations in a definition is not arbitrary, rather there are certain logical connections between the semantic category and the relations which are used to define it. Such connections can help us predict the relations to be found in a definition. Thus, a surface landform is typically defined using one or several of the following relations: HAS_FORM, HAS_CAUSE, HAS_SIZE and HAS_LOCATION; whereas processes will typically be defined through the HAS_CAUSE, HAS_RESULT, HAS_ATTRIBUTE, OCCURS_IN_TIME and AFFECTS relations. These initial assumptions about definition templates in karstology were formulated by the domain expert prior to the annotation stage. One of the goals of the TermFrame project is to verify such assumptions and compare corpus evidence from three languages with the “ideal” definition template. On the other hand, the ideal template may serve as an aid for generating complete definitions from annotated corpus data.

4.2 Distribution of categories and relations in the English and Slovene

TermFrame corpora

In total, 1,061 English and 1,332 Slovene terms were assigned categories, of which 215 English and 286 Slovene terms were definienda. Figure 4 shows the distribution of categories for all annotated terms; we see that in both languages the most frequent category is D.1 Abiotic, followed by surface and underground landforms and geomes. Abiotic elements are frequent categories in definitions because they comprise all kinds of natural entities not specific to karst, such as *bedrock*, *calcite*, *deposit*, *limestone*, *ridge*, *sediment*, etc. Amongst the definienda, the most frequent category for both English and Slovene is surface landform (73/119) followed by geomes in Slovene and underground landforms in English.

A geome is a geographical environment or landscape. We find numerous definitions for geomes denoting either types of karst (*cryptokarst*, *fluviokarst*, *glacier pseudokarst*) or subsurface environments, usually defined by their hydrologic function (*epikarst*, aquifer, conduit system, subcutaneous zone). Geomes seem more frequent in Slovene, but in fact this is due to numerous definitions for the same concept (e.g. 14 definitions for *kontaktni kras*, six for *kras*).

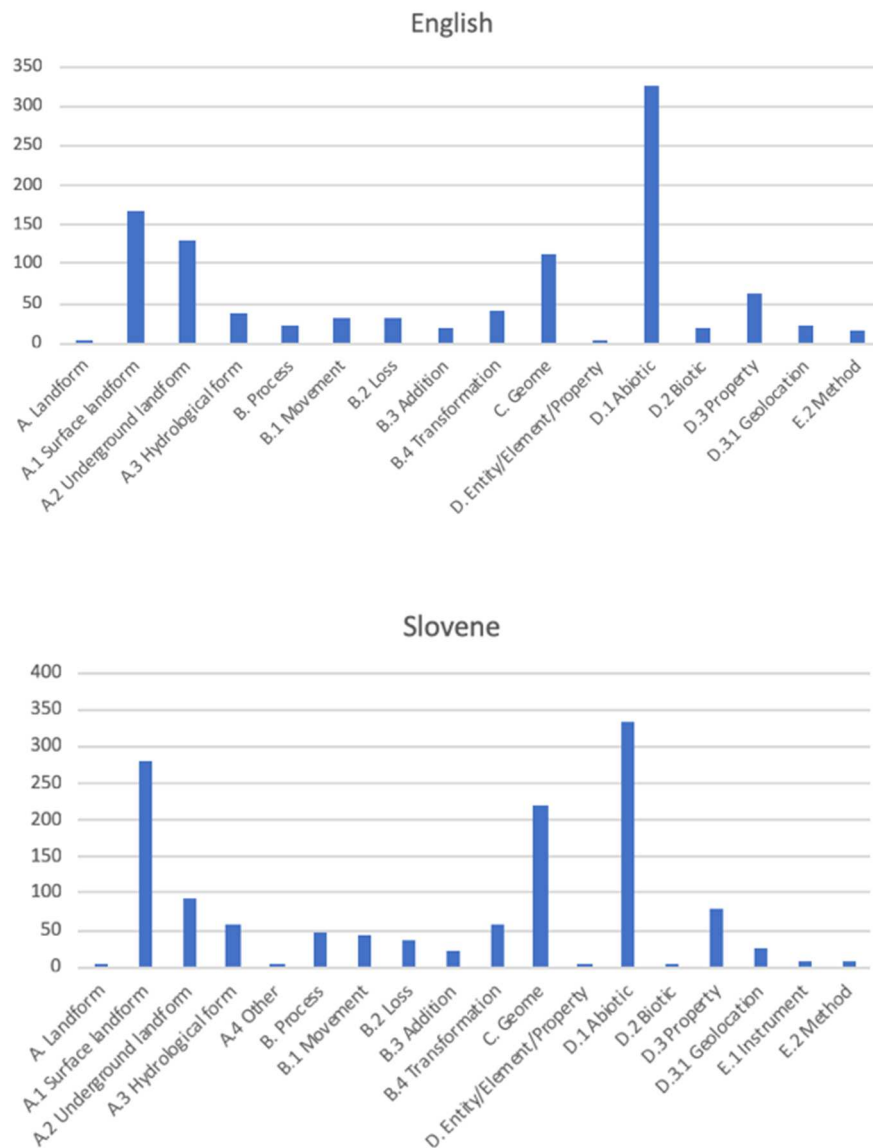


Figure 4: Frequency of categories assigned to English and Slovene terms.

A total of 382 relations were marked in English and 482 in Slovene. In both languages karst concepts are most frequently described through their spatial distribution (HAS_LOCATION), followed by morphography and morphogenesis (HAS_FORM, HAS_CAUSE). This is in accordance with the basic concept of geomorphology (as well as karstology) as a science (Jennings, 1985; White, 1988) that focuses primarily on the shape of landscape features (morphography) and the processes forming them (morphogenesis). Other relations have a similar distribution, apart from the rather general HAS_ATTRIBUTE relation, which appears more frequently in Slovene than in English. Clearly though the frequencies alone do not tell us much about how concepts

are defined in karstology. Looking at the relations occurring with specific semantic categories enables us to discern definition templates.

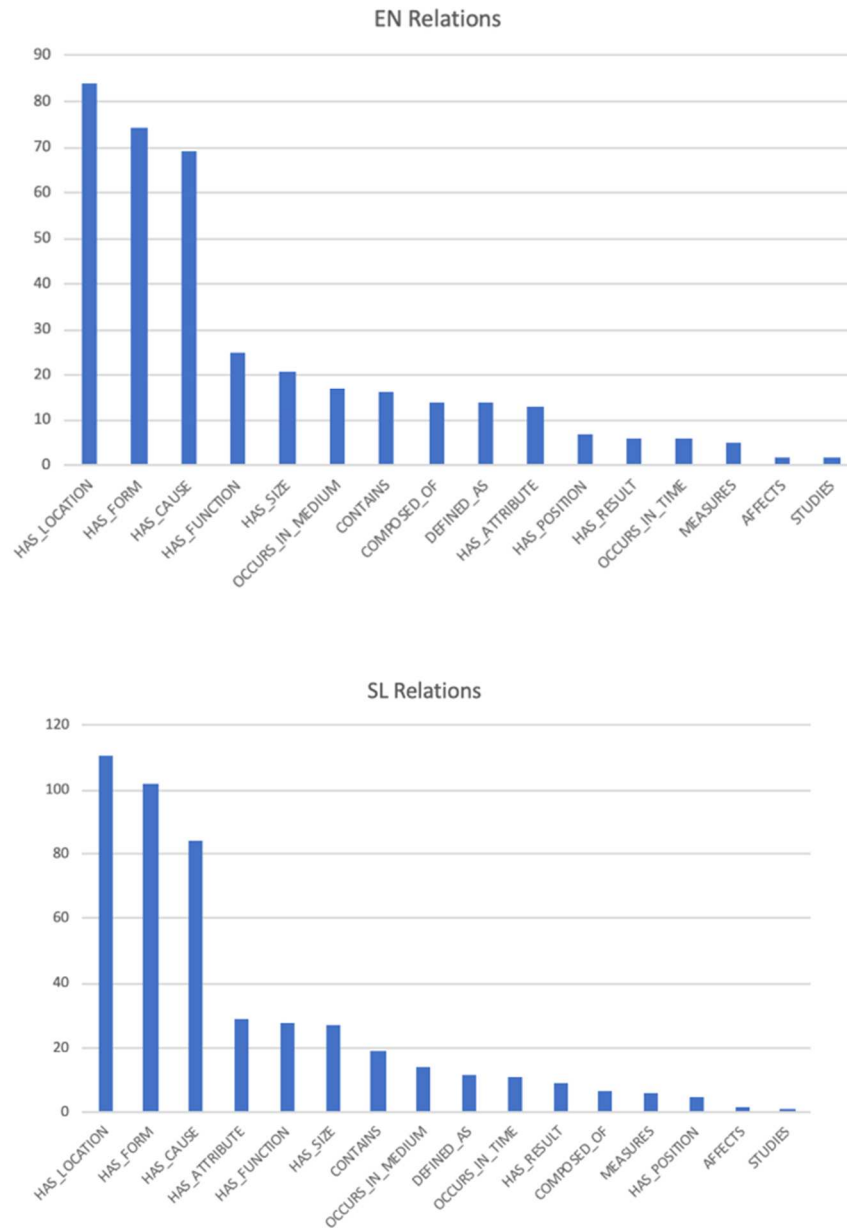


Figure 5: Frequencies of relations found in English and Slovene definitions.

An average definition for a surface landform contains only two relations out of the four typical ones for this category: form, size, location and cause. Sometimes the relation coincides with the genus, as the example in Figure 6 shows. The CONTAINS relation is more frequent with the underground landforms than with other landforms. Thus, *blue holes* contain *tidally influenced waters*, *marginal caves* contain *troglobiotic species*, *vertical shafts* contain *shattered rock and sediment* etc. It is not surprising that

speleothems as subsurface voids have a more pronounced tendency to *contain* something than surface landforms.

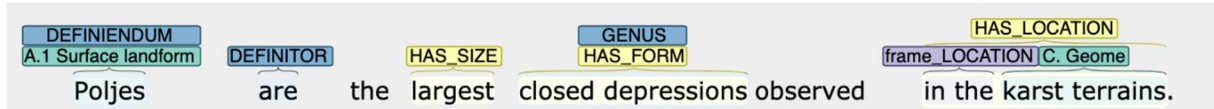


Figure 6: Definition for *polje*.

In contrast to surface and underground forms, hydrological forms are more frequently defined through their function and time pattern. As the examples below illustrate, water is an important agent in karst and hydrological forms are the points in the karst system which may function as storage or transmitters.

Geomes are the second most frequent definiendum category in the Slovene corpus and the third in the English one. We find definitions for environments such as karst and its subtypes (denuded karst, open karst, contact karst, doline karst, epikarst, fluviokarst, hypogene karst, paleokarst, fengcong karst, shilin etc.), but also other large entities and their subparts (aquifer, aquiclude, phreatic zone, zone of vertical circulation etc.). The higher number of geomes in Slovene may be due to the high variability of karst landscapes in Slovenia, which are very actively studied and described by local karstologists.

The most frequent relations used to define geomes in both languages are HAS_CAUSE, HAS_LOCATION, CONTAINS, HAS_ATTRIBUTE, HAS_FORM, HAS_FUNCTION. Interestingly, in English we find three instances where the relation HAS_RESULT is used to define a geome, while no such cases were found for Slovene. The HAS_RESULT relation conceptually requires an agent as subject, in other words a geographical entity would need to instigate some natural activity in order to produce results. In previous work (Vintar & Grčić Simeunović, 2016) we have shown that the cognitive frames underlying definition templates may be language- or culture-dependent, and here we find further evidence for this by defining a geome as a process (see Figure 7). It would appear that English definitions emphasize the morphogenetic aspect, while the Slovene ones prefer the morphodynamic properties of the karst environment as part of the karst system.

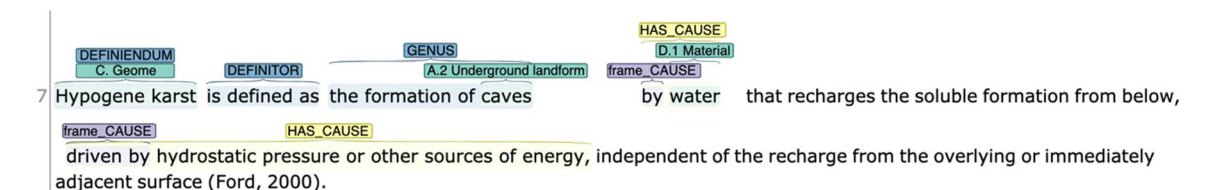


Figure 7: Definition for *hypogene karst*.

5. Towards domain modelling

The TermFrame corpus annotation imposed a rich multi-layered structure onto the previously unstructured content of a large set of documents. The annotation has so far been limited to definitions, although the present annotations can be used for machine learning to extract additional bits of knowledge and the relations among them. The development of a domain representation suited to the needs of experts, researchers, terminologists and lay users remains the primary future task of the project, but several possible directions have already been identified.

For many key concepts in karst we have found several definitions, whereby different authors emphasize different aspects of the definiendum depending on the context, text type and other factors. The identification of the prototypical or ideal definition frame allows us to generate a complete definition from the relations found in different definitions.

blue hole	
<i>Category: underground landform</i>	
IS_A	subsurface void
HAS_FORM	open to the Earth's surface, extending below sea level
CONTAINS	tidally influenced waters of fresh, marine or mixed chemistry
HAS_CAUSE	carbonate deposition and dissolution cycles controlled by glacial sea-level fluctuations
OCCURS_IN_MEDIUM	carbonate banks and islands

Figure 8: Generating a complete definition frame from several definitions.

Representing the structure of the domain in a graph allows us to see the size of individual concept category hubs, explore nodes and their neighbours, view nodes belonging to several categories and much more (Figure 9). Visualization experiments are underway also for unsupervised detection of communities, see Miljković et al. (2019: 12).

Karstology is essentially a subdomain of geography, and most of the features we explore and represent occur as tangible objects, often sites of interest, in various karst landscapes of the world. Since our corpus contains numerous references to geographical entities, one possibly useful representation is displaying instances of a particular karst feature on a map. Figure 10 presents a map depicting the geolocations of caves extracted from our English corpus using GeoNames.org for co-ordinates.

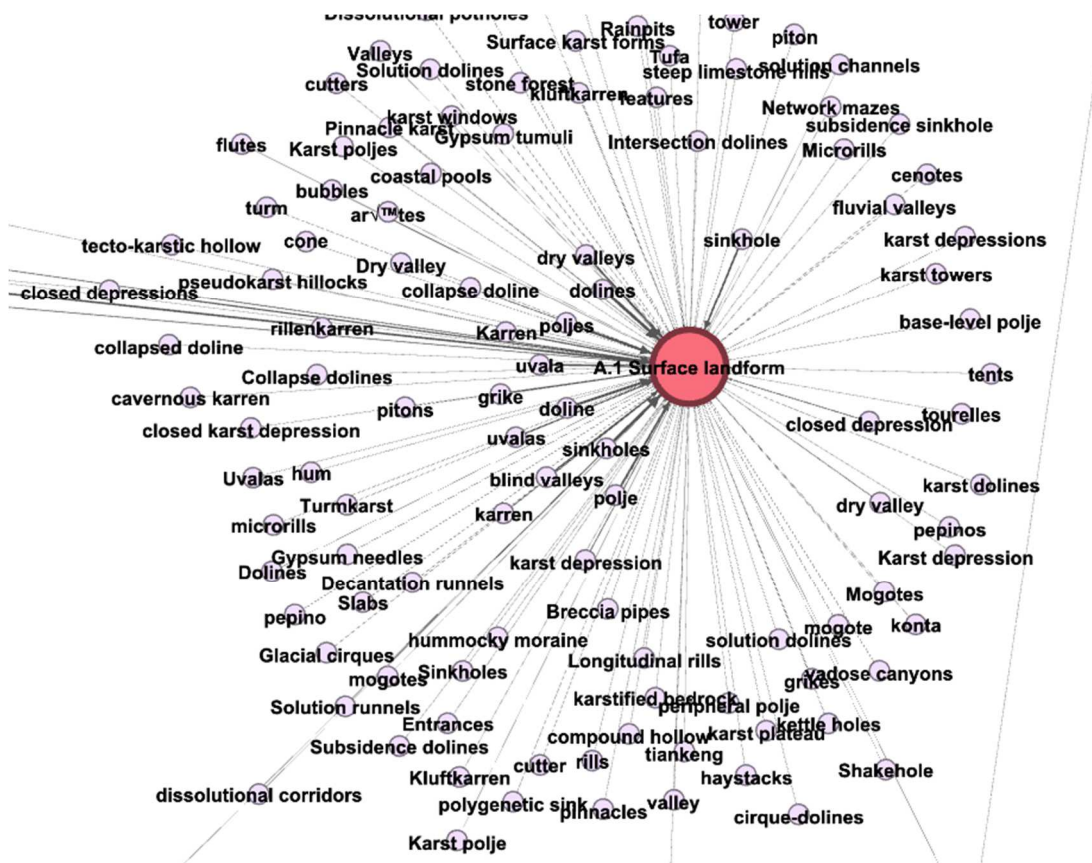


Figure 9: A section of the domain graph representing *surface landforms*.

6. Conclusions

We describe the first stages of the TermFrame project with the construction of a trilingual comparable corpus of karstology and the development of a multi-layer framework for semantic annotation. Analyses of the annotated definitions in English and Slovene allow us to draw conclusions about the cognitive frames underlying knowledge structures in the selected domain, in particular the definition templates for each semantic category. So far these seem similar for both languages, with some differences in frequency distribution and the occurrence of the HAS_RESULT relation to define geomes in English but not Slovene.

Our future plans are to explore the potential of relation definitors in combination with semantic categories to automatically extract or predict relations. Several experiments are underway to extract meaningful knowledge through graph modelling.

7. Acknowledgements

This work is funded by the Slovenian Research Agency grant J6-9372 TermFrame: Terminology and knowledge frames across languages, 2018-2021.



Figure 10: Map of caves mentioned in the TermFrameEN corpus.

8. References

- Blanchon, E. (1997). Point de vue en terminologie. *Meta*, 42(1), pp. 168-173.
- De Castilho, R. E., Biemann, C., Gurevych, I. & Yimam, S. M. (2014). WebAnno: a flexible, web-based annotation tool for CLARIN. In *Proceedings of the CLARIN Annual Conference (CAC) 2014*, Soesterberg, Netherlands.
- Diki-Kidiri, M. (2000). Une approche culturelle de la terminologie. *Terminologie Nouvelles* 21, Rifal, pp. 58-64.
- Duran-Muñoz, I. (2016). Producing frame-based definitions. *Terminology*, 22/2, pp. 223-249.
- Faber Benítez, P., Márquez Linares, C., & Vega Expósito, M. (2005). Framing Terminology: A process-oriented approach. *Meta: Journal des traducteurs/Meta: Translators' Journal*, 50(4).
- Faber, P. (2009). The Cognitive Shift in Terminology and Specialized Translation. *MonTI. Monografías de Traducción e Interpretación*. 1: pp. 107-134.
- Faber, P. (ed.) (2012). *A Cognitive Linguistics View of Terminology and Specialized Language*. Berlin, Boston: De Gruyter Mouton.
- Faber, P., León-Araúz, P., & Reimerink, A. (2016). EcoLexicon: new features and challenges. *GLOBALEX workshop, Portorož*, pp. 73-80.
- Faber, P. & Medina-Rull, L. (2017) Written in the Wind: Cultural Variation in Terminology. In M. Gryviel (ed.) *Cognitive Approaches to Specialist Languages*. Newcastle-upon-Tyne: Cambridge Scholars, pp. 419-442.
- Ford, D. & Williams, P. D. (2007). *Karst Hydrogeology and Geomorphology*. Wiley, Chichester.

- Jennings, J. N. (1985). *Karst Geomorphology*. Basil Blackwell, Oxford, pp. 293ff.
- Leitchik, V. M. & Shelov, S. D. (2007). Commensurability of scientific theories and indeterminacy of terminological concepts. In B.E. Antia (ed.) *Indeterminacy in terminology and LSP: Studies in honour of Heribert Picht*. Amsterdam: John Benjamins, pp. 93-106.
- Madsen, B. N., & Thomsen, H. E. (2008). Terminological Principles Used for Ontologies. *Managing Ontologies and Lexical Resources*, pp. 107-122.
- Miljković, D., Kralj, J., Stepišnik, U. & Pollak, S. (2019). Communities of related terms in Karst terminology co-occurrence network. *Proceedings of eLex 2019*.
- Pavlopoulos, K., Evelpidou, N. & Vassilopoulos, A. (2009). *Mapping Geomorphological Environments*. Springer, Berlin Heidelberg.
- Pollak, S., Vavpetic, A., Kranjc, J., Lavrac, N. & Vintar, Š. (2012). NLP workflow for on-line definition extraction from English and Slovene text corpora. *Proceedings of KONVENS*, pp. 53-60.
- Pollak, S., Repar, A., Martinc, M. & Podpečan, V. (2019). Karst exploration: extracting terms and definitions from karst domain corpus. *Proceedings of eLex 2019*.
- Roche, C., Calberg-Challot, M., Damas, L., & Rouard, P. (2009). Ontoterminology: A new paradigm for terminology. In *International Conference on Knowledge Engineering and Ontology Development*, pp. 321-326.
- San Martin, A. & L'Homme, M.-C. (2014). Definition Patterns for Predicative Terms in Specialized Lexical Resources. In *Proceedings of LREC14*, pp. 3748-3755.
- Seppälä, S. (2007). La définition en terminologie: typologies et critères définitoires. In *Terminologie & Ontologies: Théories et Applications* (TOTh 2007), pp. 23-43.
- Seppälä, S. & Ruttenberg, A. (2013). *Survey on Defining Practices in Ontologies*. Concordia University: Montreal. (<http://www.webcitation.org/6O2zethfp>)
- Svensén, B. (1993). *Practical Lexicography: Principles and Methods of Dictionary-Making*. Oxford University Press.
- Temmerman, R., & Van Campenhoudt, M. (Eds.). (2014). *Dynamics and Terminology: An interdisciplinary perspective on monolingual and multilingual culture-bound communication* (Vol. 16). Amsterdam: John Benjamins.
- Vintar, Š. (2010). Bilingual term recognition revisited: The bag-of-equivalents term alignment approach and its evaluation. *Terminology*, 16(2), pp. 141-158.
- Vintar, Š. & Grčić Simeunović, L. (2017). Definition frames as language-dependent models of knowledge transfer. *Fachsprache* 1-2/2017, pp. 43-58.
- White, W. B., (1988). *Geomorphology and hydrology of karst terrains*. Oxford university press, Oxford.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



The Lexicographer's Voice: Word Classes in the Digital Era

Geda Paulsen, Ene Vainik, Maria Tuulik and Ahti Lohk

Institute of the Estonian Language, Estonia

E-mail: geda.paulsen@eki.ee, ene.vainik@eki.ee, maria.tuulik@eki.ee, ahti.lohk@eki.ee

Abstract

The present study examines the role of word classes in contemporary lexicography using examples from Estonian. Since Estonian is a morphologically rich language, the results may be extendable to other languages with abundant morphology. Two research questions are examined: i) What are the problems and practices of lexicographers when determining word classes? and ii) What are the needs and expectations of lexicographers for a possible digital tool that would facilitate word class identification? The results of a metalexicographic survey carried out among 23 Estonian lexicographers show the relevance of word classes as a categorial frame in their lexicographic work. There is a need to improve or reconsider the (theoretical and technical) factors influencing the process of PoS tagging. A reliable software application (provisionally a PoS evaluator) easing the decision making process would be welcome. According to the ideas suggested by the respondents, the solution would be an improved morphological and syntactic parsing system with respect to the present solutions, and a corpus-driven application presenting statistics with regard to the morphosyntactic distribution of an ambiguous word with access to the data source.

Keywords: lexicography; word classes; metalexicographic survey; Estonian

1. Introduction

The challenge of modern lexicography is to create digital tools which would be able to present meaningful and reliable generalizations over a large amount of raw data in a way that meets the specific needs of lexicographers, including inter alia the procedures of word class categorization¹. Although several studies indicate that lexical categories are far from clear-cut or self-apparent (see, e.g., Mark, 2015; Croft, 2001; Culicover 1999), grouping lexemes is vital for the information that dictionaries provide. The task of PoS markup has not disappeared in the digital era of lexicography, when stand-alone dictionaries are increasingly replaced by unified and standardized databases. On the contrary – integrated all-purpose root databases comprise all kinds of information for as many lexemes as possible. The data models of such databases typically include a PoS unit (e.g. Tavast et al., 2018).

The aim of this study is to clarify the role of word class categorization in contemporary lexicographic work in Estonian. It comprises the first step of a project that aims to develop a corpus-driven solution, tailored to the needs of lexicographers. We believe that the results can be extended to other languages with abundant morphology when

¹ Throughout the study, we use the notions word class and part of speech (PoS) synonymously.

planning e-lexicographic projects with similar goals.

In order to establish the extent of the open class problem for lexicographic work, the present practices, and the needs and expectations for language technology, we have conducted and carried out a metalexicographic survey with questions such as: Are word classes even a necessary and useful concept in modern lexicography? How challenging is word-class categorization for Estonian lexicographers? What are their actual expectations for the possibilities of the digital era in that respect?

First, we provide a background to the study in Section 2 presenting a short summary of the general traits of Estonian along with its most recent word class systematization and the treatment of word classes in some Estonian dictionaries (subsection 2.1). In subsection 2.2 we explain the setup of the semi-structured interviews with lexicographers. Section 3 focuses on the delineation and analysis of the data concerning our first research question, i.e. the problems lexicographers experience in connection with word classes. The second main question of this study, the solutions for aiding the lexicographer in categorization and presentation of word classes, is addressed in Section 4. The results are then summarized in Section 5.

2. Background and details of the study

2.1 Estonian and its word class system

Estonian is a Finno-Ugric language spoken by about 1 million people in the Estonian Republic and abroad. Although Uralic languages are considered agglutinating, Estonian morphosyntax is generally more fusional and analytic than that of the northern branch of Finnic languages (Finnish, Karelian, Veps etc.), which are characterized by a high degree of allomorphy and grammatical syncretism (see Viitso 2007, Remes 2009). Estonian can be described as a morphologically rich language: words inflect (nouns and adjectives for number and case; (finite) verbs for mood, tense, person and number; adjectives and adverbs for degrees of comparison) and are subject to agreement. There are approximately 100 native derivational suffixes, and new words are productively formed by compounding (Kerge, 2016: 3228). Nouns and adjectives decline for 14 morphological cases; nominative, genitive and partitive are traditionally considered grammatical cases and the remaining 11 cases are held to be semantic. For instance, spatial relationships are expressed by inner (illative, inessive, elative) and outer (allative, adessive, ablative) locative cases, besides adpositions. Verbs have, in addition to the abovementioned finite conjugational forms, infinitival, converbal and participial forms. A typical Estonian adjective normally agrees with its head noun in case and number (see (1)); a verb agrees with its subject in person and number (2):

- (1) *kirju-de-st* *koer-te-st*
 piebald-PL-ELA dog-PL-ELA

- (2) *te jook-si-te*
 you run-PST-2PL

The common categorization of word classes involves two main types: content words (typically nouns, verbs, adjectives, and adverbs) and function words (adpositions, pronouns, conjunctions, etc.). The criteria of categorization are generally based on morphosyntactic properties, but cross-linguistically these can only be identified semantically (Haspelmath, 2001). There are diverse approaches to the Estonian word classes, related to different language varieties and different methodological perspectives: see Kaalep et al. (2000) for contemporary written language and automatic morphological tagging, Habicht et al. (2011) for old written language, Lindström et al. (2006) for dialectal language and Hennoste (2002) for (contemporary) spoken language. The latest general word class system for Estonian proposed by Erelt (2017: 58–61) divides Estonian words into four main classes based on syntactic and semantic criteria²:

1. autonomous content words (verbs, nouns, adjectives, numerals and adverbs) that occur independently in a phrase and convey their denotative meaning obvious without context,
2. autonomous functional or substitution words (pronouns, proadverbs)
3. non-autonomous functional words or auxiliaries (auxiliary verbs, affixal adverbs, adpositions, conjunctions),
4. syntactically independent pragmatic words or particles (modal adverbs, interjections).

Word class can be seen as a link between grammar and lexis, providing a hint to a word's general meaning, its paradigmatic (morphological) behaviour and sentential function. It also gives the non-native user an idea about the uses of a particular word in the language and what patterns of grammar it should follow. In the Estonian lexicographic tradition, word class information is part of the description of a word's lexical behaviour, but PoS tagging is somewhat sporadic and this task is (implicitly) assigned to certain dictionaries, not all. Tagging all words with a PoS label is complicated, as since it is being a morphologically rich language Estonian is characterized by a tendency where inflected word forms may shift their lexical categorial status in respect to the base word.

There is a tradition of presenting PoS information in some of the general monolingual dictionaries (e.g. *The Explanatory Dictionary of Estonian* (EKSS, 2009), *The Dictionary of Estonian* (DicEst)³ and *The Dictionary of Estonian Word Families* aim at systematic PoS markup) and in particular, in learner's dictionaries (e.g. *Collocations*

² Morphologically, the Estonian words fall into inflected (verbs, nouns and adjectives) and uninflected words (particles).

³ EKSS and DicEst cover word class information over the Estonian lexis most comprehensively, however, even these dictionaries do not tag all headwords with word class information.

Dictionary (ECD)⁴, and *The Basic Estonian Dictionary* (BED)) as well as bilingual dictionaries. As a rule, PoS is not marked in orthographic⁵, onomastic, terminological, dialect, or etymological dictionaries. For instance, *The Dictionary of Standard Estonian* (DSE) marks word class traditionally only in certain exceptional cases relevant for advisory purposes. The question of word classes has become more topical along with the development of the Ekilex database and dictionary writing system, an integrated lexical resource, where PoS belongs to the structure of every lexical entry (Tavast et al., 2018). In the most recent output of the lexicographic resources, the language portal Sõnaveeb⁶ ('Wordweb, 2019'), the explicit marking of word classes is an ultimate goal.

Regarding the technical side of word class marking, there are two parallel sets of labels for word classes in the Estonian lexicographic tradition – one using loanwords of international origin (e.g. *substantiiv* 'noun') and the other using coined Estonian terms (e.g. *nimisõna* 'lit. name word'), which are more transparent. The two sets of terms basically address different users: the international terms are for experts and the transparent ones for learners. Most dictionaries use abbreviations of international terms (*v* for verbs, *adv* for adverb etc.), but the language portal Sõnaveeb and BED use non-abbreviated native Estonian terms.

2.2 Details of the study

2.2.1 Methods

To clarify the opinions and experience of professional dictionary-makers about the role of word class in their everyday work, we conducted a metalexicographic survey in the form of semi-structured oral interviews containing both open-ended questions and opinion ratings on a Likert-type scale. The target group were lexicographers working with the Estonian language in monolingual or multilingual dictionaries; the participants were informed beforehand about the general topic of the survey (the challenges of parts of speech categorization in their lexicographic practices). The interviews were carried out by three interviewers in February and March 2019 and lasted approximately 30 minutes each. All the conversations took place privately at the lexicographers' work environment. The conversations were taped, transcribed and analysed content-wise (searching for qualitatively different opinions). The numerical data were subjected to simple scoring.

⁴ In the ECD, displaying the collocational behaviour of the 10 000 most frequent words in Estonian, PoS tagging has crucial relevance as the grouping of collocates is based on their word class affiliation.

⁵ The paradigmatic affiliation of the entries is traditionally indicated by inflectional types (*muuttüübid*) marked by numerical indices.

⁶ <https://sonaveeb.ee/> (25.5.2019).

2.2.2 Respondents

Altogether 23 lexicographers (F=21, M=2) participated in the survey. For comparison: in the cross-European survey of lexicographic practices only 8 of the Estonian lexicographers participated, which – in the context of that study – was a rather high rate (Kallas et al., 2019: 7). We suppose that the collegial one-to-one setting and oral form of the survey facilitated participating in our study. The majority of our respondents were current employees of the Institute of the Estonian Language, the institution producing and publishing most of the academic dictionaries in Estonia. Only a few respondents were from Tartu University or some other institution.

The lexicographers' work experience varied from 0.5 to 48 years, averaging at 18 years. More than 10 (48%) of them had worked in this field for at least 20 years. The European lexicographer's average work experience is approximately the same, with a slightly smaller proportion of professionals (35.6%) having more than 20 years' experience. It was pointed out in the cross-European survey that the profession of a lexicographer tends to be a lifelong one (Kallas et al., 2019: 8).

The experience of our respondents was also impressive in terms of content; and altogether 38 different dictionaries were mentioned as their past or current projects. The variety of dictionaries included both monolingual and bilingual dictionaries, both standard Estonian and dialects, both descriptive and prescriptive ones, among others. Altogether 22% of the respondents had some experience with general monolingual dictionaries (such as EKSS, DicEst, the *Dictionaries of Standard Estonian*, the *Basic Estonian Dictionary*, the *Dictionary of Word Families*, the ECD etc.); 26% had worked with specific monolingual dictionaries, such as the *Dictionary of Estonian Dialects*, *Estonian Etymological Dictionary*, *Low German Loanwords in Estonian*, etc). The smallest proportion (4%) had only worked on bilingual dictionaries, such as Estonian-French, Estonian-Finnish and Finnish-Estonian, Russian-Estonian, German-Estonian, etc. Almost half (47%) of our respondents had experience with multiple types of dictionaries.

Considering the length and range of the working experience, it is quite remarkable that most of the interviewees (83%) had worked or were currently working with an electronic dictionary writing system (mostly the institute's own in-house software EELex). The percentage of lexicographers using corpora (and specific tools for corpus search, such as Sketch Engine) was 74%. While many lexicographers mentioned using the Estonian National Corpus, some other more specific corpora were also mentioned, such as the Corpus of Old Literary Estonian, the Educational Corpus of Estonian etc. Some of the lexicographers reported using a corpus that had been specifically designed for compiling the dictionary at hand. In comparison with European lexicographers, on average, our respondents' use of IT resources and tools was 10% higher (see Kallas et al., 2019).

In conclusion, the interviewed lexicographers have been quite flexible to adjust to the rapid changes in the field of the lexicographers' workflow while the institutions have

provided good technological support. This is the background for the lexicographers' expectations for IT resources and tools: a long experience as a lexicographer and a certain degree of familiarity with the affordances that the electronic era, in principle, could provide make the lexicographers wish for even better tools and technological support.

3. The challenge(s) of word classes in lexicographic work

3.1 Can we manage without word classes?

As a point of departure, we focus on the general role of word classes in our respondents' everyday tasks. The interviewees were encouraged to reflect on the necessity of PoS markup and on whom it benefits. The respondents generally considered the task of PoS markup rather important and beneficial. All the interviewed lexicographers agreed that the PoS information is presented for "the user". Some lexicographers emphasised the role of PoS for a regular user (pupils, language learners, teachers) and took into account their restricted ability to cope with the overwhelming lexicographic information, while the rest took the perspective of expert users (linguists, lexicographers) and provide information as detailed as possible, "because it is in the interests of the researchers". Lexicographers saw themselves among the potentially beneficiary parties of the PoS information. If a dictionary has already been PoS-tagged, a professional has analysed the material, and the earlier work of the colleagues needs to be (re)valued.

The respondents ranked the necessity of the word class information in the dictionary they currently worked with on a 5-point scale. As the response rates in Figure 1 show, the results varied from "very necessary" to "somewhat unnecessary", while none of the lexicographers rated it "completely unnecessary":

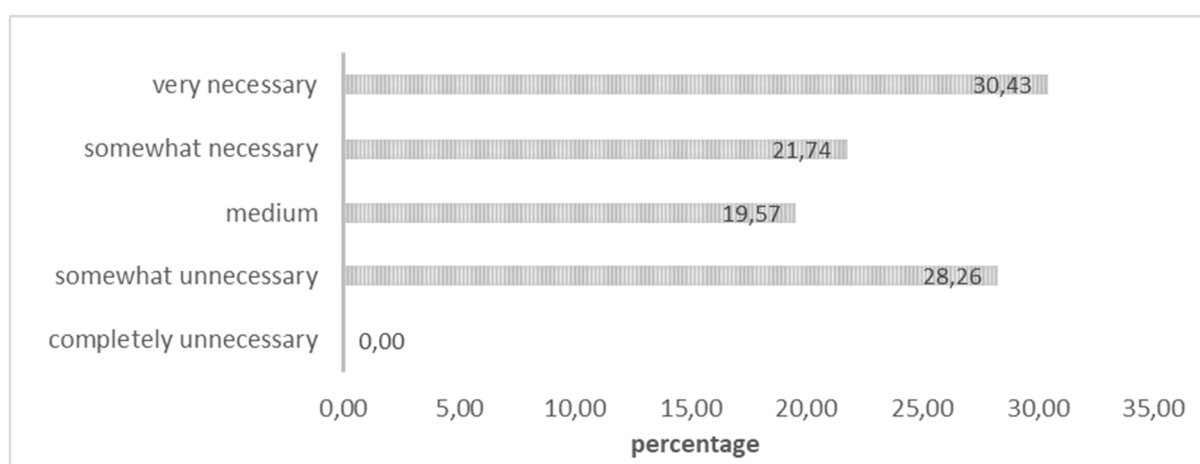


Figure 1: The necessity scale.

There are two peaks in the diagram: 30% of the respondents claimed that PoS marking is very necessary and 26% considered it somewhat unnecessary. This polarization of opinions can be explained by variation in the representation of word classes (mandatory or occasional) in different dictionaries, motivated by the assumed needs of the target groups. As mentioned in Section 2.1, the PoS tagging is also determined by tradition and a certain division of labour. Although the necessity rating of PoS information was primarily based on the respondents' ongoing projects, the previous experience seemed to influence the assessments. In addition, word classes seem to provide a general logical structure for organisation of the material (they “help to categorize the material in one’s mind”).

The lexicographers’ reflections reveal the relevance of word classes in the systematization of the material and a need for as clear criteria of classification as possible. Since word class categorisation involves different linguistic levels, the balance may be swaying towards one or another of them. Some of the respondents emphasised the relevancy of semantics in the specification of a word’s categorial affiliation: “I definitely take meaning into consideration, because semantic features define the essence of the word”. Other respondents base the judgment on syntactic properties, considering a word’s typical function in a sentence. Problems arise from the classical description of, for example, the noun as the argument of a clause – whenever a noun tends to occur in another function, for instance as an adverbial, its syntactic properties (and thus the word class attributes) will change. The respondents consider morphology to be the main source of the word forms departing from a paradigm, seen as a special characteristic of morphologically rich languages. Although the resulting ambiguous cases complicate information retrieval in databases, they also reveal ongoing lexical changes. Yet the respondents engaged in ascertainment of word class boundaries on a daily basis would still prefer an “ideal” situation where every word has a definite word class label.

3.2 Different dictionaries – different challenges

To a great extent, the problems faced depend on the properties related to the dictionary type the lexicographer is working on – its object of description, purpose, target group and other factors.

The lexicographers working with bilingual dictionaries generally use a database of Estonian that contains the most frequent words with word classes already defined (the Estonian-X dictionary⁷). The respondents belonging to this group generally assessed word class categorisation as not a too complicated task. The interviewees working with general and specific monolingual dictionaries (see the distribution of dictionaries the respondents work with in Section 2.2) are not that unanimous. For the most part, the dictionaries in the general monolingual group require explicit formulation of the word

⁷ The database is available at <http://exsa.eki.ee/exsalogin.cgi> (25.5.2019).

class, except for the DSE. The lexicographers experiencing particularly challenging problems with PoS tagging work either with the DicEst, ECD (general dictionaries), or the dictionary of old written Estonian (a specific dictionary). The impression of interviewers is that the lexicographers compiling the dictionaries that provide a systematic markup of word classes manifest a particularly deep sense of responsibility, as their work results will be source material for other lexicographers. In other dictionaries labelled as specific in our study, the word class category is generally not a prominent issue, even though it may be a topic puzzling the lexicographer in the background.

For instance, in compiling an etymological dictionary the word class category is not the primary concern, as the main focus is on the origin of a word stem and words with the same stem are gathered in a same entry. Hence, the derivative relations appear as most important; in case there are doubts regarding a word's root form, the most plausible variant is preferred and a word class suggestion often appears in the definition of the word. The lexicographers compiling the dictionary of old written Estonian handle specific problems such as different stages of lexicalization-grammaticalization compared to contemporary language and a rather limited availability of linguistic sources, which complicates the determination of the developmental stages of a word and hence also the determination of the word class. The compilation of dialect dictionaries involves analogical tasks compared to the etymological and old written language ones, but the work has its own logic, since the object of description basically originates in colloquial spoken language and data collections based on fieldwork.

The lexicographers were invited to assess the challenge of the task of word class categorisation among the other tasks they perform in their everyday occupation with dictionary compilation on a 5-point scale: “easy”, “pretty easy”, “medium”, “challenging”, and “very challenging”. The assessments followed the normal distribution with the peak of 56% at the point “medium” and 22% on both “pretty easy” and “challenging”. None of the respondents used the extremes of the scale. The results reflect the factors related to the somewhat different challenges of the compilers of different types of dictionaries and those related to the ambiguity of certain forms, which will be discussed in more detail below.

3.3 The natural flux of word classes in Estonian and its implications for lexicographic work

A characteristic feature of Estonian is that the inflected word forms tend to move from their basic lexical categorial status to another. For instance, the boundaries between adpositions, nouns and adverbs in Estonian are considered to be rather fuzzy (see, for example, Grünthal, 2003), and there are always words and word forms in a transition stage, appearing both as standard nouns and as part of more or less fixed expressions with more abstract meanings (see, for example, Paulsen, 2018, 2019). The natural flux of words from one word class to another – detectable in changes in their

syntactic/pragmatic function and, occasionally, in a shift of meaning – was reflected in the reasoning of our interviewees, too.

Lexicographers are trained to recognize words by their word class membership and in most instances it is not a critical concern. In less self-evident cases, for instance when the actual usage of the word (or its form) in the corpus shows idiosyncratic tendencies, the lexicographer may be in a difficult position. The respondents were encouraged to bring up examples of some particularly striking cases. Almost all of them could think of such examples, the total number of tokens being 145. The number of different examples (types) was 127. The average number of critical examples per person was six, which falls well within the limits of one’s short-term memory. In reality, some of the respondents had prepared for the interview and brought up more examples; four respondents declared that they either did not have any considerable problems with word classes or they could not remember the exact problems.

Figure 2 presents the distribution of the examples across the “classical” word classes. The thicker the line in the figure, the higher the proportion of the examples falling between the two categories located at the ends of the line. The noun sits in the centre of the diagram because it has the highest proportion of “overlapping” cases: altogether 35% of the total number of examples enjoyed “dubious” membership with this category (and with adverbs, adpositions and adjectives, respectively). Adjectives appeared as the second most “slippery” word class, with 26% of the total number of critical examples.

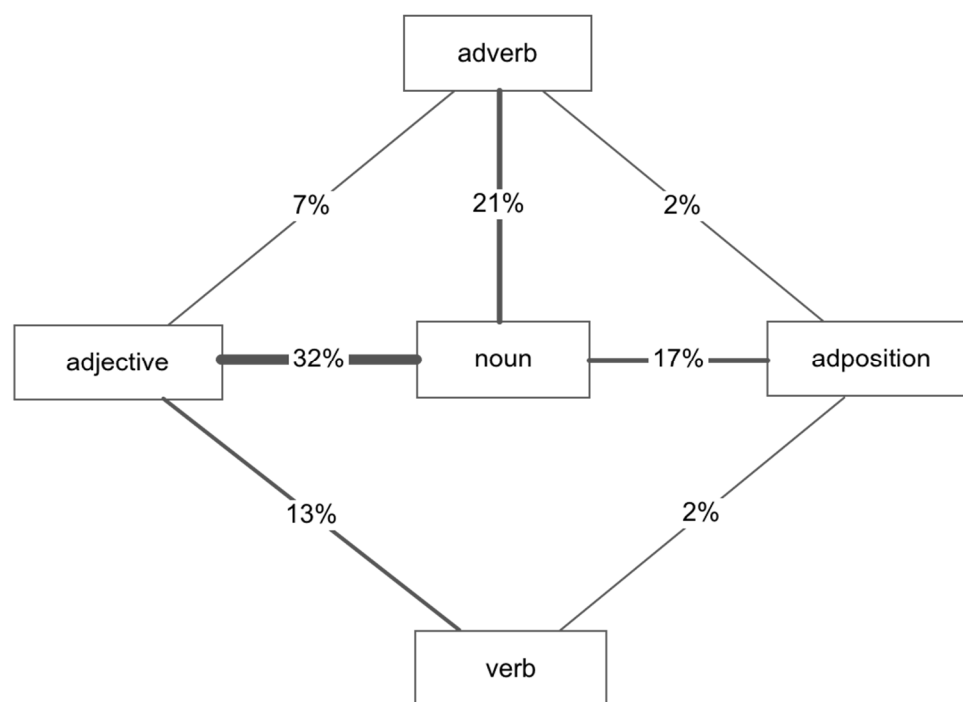


Figure 2: Distribution of the critical examples along their word class membership.

As a generalisation, we can say that the lexemes the most complicated to define either i) occur in two or more lexical classes keeping the same base form (as examples (3) and (7) below), ii) are in a transition phase from one word class to another (e.g. the inflected forms developing new functions, see examples 4–6) or iii) only appear in certain inflected forms (8). From a nominal perspective, the main source of the word class shift lies in the semantic cases, especially the locative case forms of nouns that develop autonomous uses both semantically and syntactically (see example (4)). The main source of categorial shift for verbs are infinitives/converbs (5), participles (6), and nominalisations (*ela-mine* [live-NOM] ‘habitation’).

(3) S → ADJ

Koer poiss roni-b puu otsa
 dog boy climb-3SG tree.GEN tip.ILL
 ‘The naughty boy is climbing the tree’

(4) S → ADV

Mu-l on tema-st kõri-ni
 I-ADE be.3SG he-ELA throat-TERM
 ‘I have had enough of him / I’m fed up with him’

(5) S → ADP

Pole riigi huvi-des makse alandada
 do.not state.GEN interest-CONV tax.PL.PART reduce
 ‘it is not in the interest of the state to lower the taxes’

(6) V → ADJ

Ta and-i-s töö-le hävita-v-a hinnangu
 he give-PAST-3SG work-ALL destroy-PTCP.GEN judgement.GEN
 ‘He gave the work a devastating assessment’

(7) ADV → ADJ

- a. *Õpetaja on alati abivalmis.*
 teacher be.3SG always help.ready
 ‘The teacher is always helpful’
- b. *Mõned on abivalmi-m-ad kui teise-d*
 some be-3PL help-ready-COMP-PL than other-PL
 ‘Some are more helpful than others’

(8) ?

Üritus toimus noor-te eestvõtte-l
 event happen-PAST-3SG youth-PL.GEN front.grasp-ADE
 ‘The event took place on youth initiative’

How do lexicographers solve the puzzle of classifying ambiguous words? The reported strategies were lookup in other dictionaries, checking the grammars, consulting relevant research, using syntactic tests, and looking at the distribution of the given word in a corpus. If these measures do not suffice, they turn to colleagues – after discussions and consideration of different possible perspectives, the team of lexicographers may decide the PoS markup collectively, by voting.

As for IT-solutions, the lexicographers mainly use the corpus query software Sketch Engine (Kilgarrieff et al., 2014), particularly word sketches and concordance queries. Only a few of them were aware of the Sketch Engine function “Lempos” showing the distribution of a lemma in certain positions. Some respondents use automatic morphological analysis⁸ to get an idea about possible alternative interpretations of the word. In short, only a fraction of the respondents use an IT solution in the decision-making process of PoS markup. The lexicographers confirmed repeatedly that they search for information, evidence and opinions, but the final decision about the PoS markup is up to them. There is no automatic PoS markup in the current practice of lexicographers, and they thought it would be almost impossible to use one, mostly because of the questionable reliability – “I will trust only myself as a researcher”.

3.4 Multiple or zero tagging? Practical implications

The PoS categorisation and markup is not only of theoretical interest, as it has numerous practical implications on lexicographic reasoning. The lexicographer has to take a quick stand on the forms undergoing grammaticalisation or lexicalisation, fix the base word class in case a word belongs to different word classes within one morphological paradigm (like adjectives and nouns in Estonian) and consider the diverging opinions expressed in the linguistic literature. The approach to the word class affects the whole structure of a dictionary starting with the number and the organisation of entries. The PoS categorisation problem can, for instance, be solved by presenting the questionable form as a subheadword instead of a separate independent headword; the PoS tag of the subheadword can then be omitted (this is the solution used in the EKSS). This is a way to present the items that are (yet) not fully lexicalised or grammaticalized, indicating an ongoing change. However, as mentioned in Section 2.1, the development of the database and dictionary writing system Ekilex and the integrated language portal Sõnaveeb (‘Wordweb’) set completely new demands for the lexicographer, as the goal is to provide the PoS information for every lexical entry.

Most lexicographers agreed that it is acceptable and even inevitable that some headwords have two PoS tags, e.g. *haige* ‘ill’ (ADJ) and ‘patient’ (S). An argument for this was that the dictionary should reflect the actual usage: If the words tend to be used in different kinds of constructions typical of different PoS, then the dictionary must display it. It is also expected to facilitate comprehension for language learners by

⁸ artur.eki.ee/morf (25.5.2019).

explicitly tagging the two possible ways of usage instead of having the users study the examples and make their own inferences.

Some respondents were more dubious about multiple tagging of the same headword and stressed that it can be accepted only in cases when the meaning of different parts of speech is “exactly the same”. The typical example of such a case is *all* ‘under, below’, used either independently (as an adverb) or as part of a phrase (as a postposition). It was argued instead that they should be presented as separate headwords except when the cases are semantically strictly identical. Again, the motivation behind the one-to-one relationship in the description was “user needs”.

The lexicographers agreed that the degree of specification of the PoS markup depends on the type and purpose of the dictionary. There was an opinion that everyone would benefit from at least one reliable source (a kind of “master dictionary”) assembling the word class information. Ideas differed on how to deal with ambiguous words. Some lexicographers trust that every word can be classified, even if it seems difficult at the beginning. Others are less idealistic, and propose that sometimes it would be practical to present the questionable form not as a fully independent headword but as a subheadword without a special PoS tag (like the solution in EKSS discussed above). There are also lexical items other than words (idioms, multiword expressions, phrasal verbs) that could hardly be tagged for PoS. It was pointed out that there are other possibilities to demonstrate the usage of the word, such as by presenting examples. There was an agreement that in the vague cases a word cannot be classified properly without context. Another practical question concerning the corpus data was “What is the sufficient degree of frequency?”.

4. Expectations for solutions

The second main research question of this study concerns the possibilities for facilitating the PoS-categorization task in lexicographic workflow. The lexicographers were asked about solutions they could think of when dealing with complicated cases of word class identification.

4.1 Could we just change the classification?

The system of PoS marking in a language holds as a part of the general agreement about the linguistic categories and no single lexicographer nor group of lexicographers can easily change it. The lexicographer must adopt the existing system and find reasonable practical solutions. Would the word class system need an adjustment into a more suitable one? This question has two possible answers: The classification can either be generalised and schematised into more heterogeneous groups, thus increasing the average number of class members, or the system can be elaborated by increasing the number of classes and creating specific labels for the classes of “ambiguous” cases with a more homogeneous class membership as a result.

The respondents presumed that the system could in principle be changed if it would match the actual usage and become more comprehensible for the user. They pointed out that dictionary-wise the word class labels vary anyway: Some dictionaries distinguish between prepositions and postpositions, while others use the more comprising term adposition; some dictionaries mark just adjectives, while others tag also its subclass of indeclinable adjectives etc. The EKSS uses 17 different tags for word classes because in addition to the traditional labels (noun, adjective, adverb, etc.) some specific ones have been created, such as abstract noun, diminutive, proper noun, (adjective-like) participle, actor noun, and action noun.

The attitudes towards potential changes differ notably. The lexicographers oriented to the needs of regular users (particularly learners) prefer a simple and elegant PoS markup: just a few word classes with transparent native terms. The respondents focusing on the needs of expert users are against losing the attained level of granularity and seek for continuous enhancement. They prefer the present system of PoS labels and are ready to welcome a more precise and detailed system, if justified. Some respondents are aware of the heightened need of precision for natural language processing applications and are therefore in favour of finer granularity. They admit, however, that not every detail known to the lexicographer or to the “system” needs to be presented to the regular user. In the case of an e-dictionary, an adjustable interface conforming to the needs of different users could be a solution.

Some of the lexicographers mentioned that a good system of PoS markup could be a hierarchical one containing both more general classes and the more specific ones (subclasses as well as subclasses of subclasses). Such a system would remind one of the general prototype model of human categorisation with its basic, superordinate and subordinate levels of knowledge (see Rosch et al., 1976). The present system of PoS in Estonian follows, in some respects, such a hierarchical model: The words are divided into inflected vs uninflected words, content vs function words, and further into specific classes with their specific combinations of meaning, form and function (see Section 2.1).

4.2 Visions of a PoS evaluator

The lexicographers were encouraged to share their conception of an ideal IT tool that would help them solve the ambiguous cases of word class affiliation. They expressed certain scepticism and even reluctance towards this idea – mostly because the respondents got the impression that they were expected to present a fully conceived technical solution in detail. However, some of them were disappointed with lexicographical IT tools in general. They shared a suspicion that no perfect tool would be possible, and envisaged themselves correcting the mistakes made by an automatic system. The respondents who were more aware of the technical nuances pointed out that no system can work better than the underlying automatic tagging (both morphological and syntactic) of corpora, and thus any result relying on the same tagged corpus would present results similar to those of Sketch Engine.

The ideas the lexicographers came up with can be divided into (structurally) simple and complicated ones. The relatively simple, not particularly corpus-driven solutions make helpful information easily available and facilitate the exchange of information among lexicographers:

- 1) **A database of ambiguous cases**, collecting the earlier (also divergent) lexicographic judgments with eventual reference to the corpus data the definitions rely on. The result would be like a “master dictionary”, where the lexicographers can test their intuition or find analogical cases to base their judgments on. Such a solution requires a group of experts charting all ambiguous cases and making justified decisions about their PoS. Although the data can be updated and changed by the lexicographers themselves, it would be an off-line solution by nature – it would not refresh automatically when the corpus data are updated (illustrative examples can be added by lexicographers). The examples gathered in this study (see Section 3.3) can serve as a starting point for such a database, and these cases can also be used for extraction of similar cases from large corpora.
- 2) **A lexicalisation-grammaticalisation scale**. A word (form) should match a set of explicit criteria in order to get a certain PoS tag. The (grammatical, distributional) criteria would be included as a module in the lexicographers’ workbench (EELex or now Ekilex). The problem with this solution is that it differs only a little from the lexicographers’ current task, saving their time and energy only by making the criteria easily accessible. The solution relies on the “classical” understanding of category membership (the necessary and sufficient conditions), and it is unclear whether it would produce sufficient solutions for the ambiguous cases that share the criteria of many classes or lack some necessary condition of the main class.
- 3) **A set of smart syntactic tests to “try out” the PoS membership**. Lexicographers use “testing it mentally” in their everyday practice. For example, if an ambiguous participial form is agrammatical in the comparative form or in a phrase with the intensifier *väga* ‘very’, there is a question of a verb form rather than of an adjective. This kind of test could help the lexicographer to make a proper decision about the PoS. Such a solution requires a group of experts to refine the system of adequate tests. The task of PoS evaluation would be facilitated, but the decision relies on the lexicographer’s grammaticality judgment of composite phrases.

The main idea of a more advanced PoS evaluator is a corpus-driven tool that searches the corpus and presents the (up-to-date) statistics of the morphosyntactic distribution of an ambiguous form on the lexicographer’s desktop. The behaviour of a questionable word would be compared to the corresponding profiles of the typical members of different PoS and the percentage of overlap and discrepancy would be revealed. The

prototypical PoS profiles (in terms of syntax, morphology, semantics) should first be established in the corpus data. The respondents prefer a visualised output with an indication of the dominant PoS profile and the degree of predominance. The tool should generalise over the results, suggest qualitative distinctions, and provide access to the original data the statistics is based on (concordances with a gateway to the context). The raw material should be presented according to its relevance, showing explicitly which criteria of the particular PoS are satisfied and which are not. Statistics about the presence of a semantic shift would also be welcome. Basically, the tool should be similar to Sketch Engine but even more advanced and reliable.

There were different ideas about the scope of the task that the PoS evaluator should perform. There is no need for such an application for the typical “well-behaving” word forms. The respondents imagine an application providing a desktop window where one can insert the search term and receive its statistics and tendencies related to a PoS. Presuming such a rather narrow task, the other steps of the lexicographic workflow would remain the same. Some of the lexicographers came up with a broader view of the task: The tool would analyse the corpus for “suspicious” word forms (e.g. nouns that appear mostly or only in locative case forms), create a list of the potential new headwords and then analyse them in detail, according to the lexicographer’s choice. Such an automated procedure would draw the lexicographers’ attention to certain changes in usage that would otherwise remain unnoticed.

The respondents would prefer to have the tool as a module in their habitual work environment, either as part of their workbench (EElex, Ekilex) or as part of the corpus searching tool (e.g. Sketch Engine). Some of the lexicographers envisaged that the PoS evaluator would be useful not only for lexicographers, but also for the general public. In that case an application with a simplified interface is needed – the information served to a language learner should not be too abundant or complicated.

How to arrive at such a system is a task for the future. The aspects of knowledge, mentioned in relation with the “simple solutions” (a database of critical cases, a scale of explicit criteria, a set of discriminative tests), will be useful sources of information also when striving for an automated PoS markup.

5. Conclusion

This study aims to grasp the lexicographers’ experiences and visions regarding word class categorisation and to relate these ideas to the changed paradigm of modern lexicographic work. PoS categorisation is a topical issue in Estonian lexicography, as the current trend is to avoid omitting tags as well as the multiplicity of PoS markup. The ultimate aim is to provide a word class tag for every dictionary entry in the main database (regardless of whether the end-product contains or displays the PoS tags). This trend is dictated by the data model of the Ekilex database and dictionary writing system and the design of its main output, the language portal Sõnaveeb.

The results of the survey indicate that in lexicography word classes provide a categorial frame that is in the background, even if PoS tagging is not an explicit task in the dictionary a lexicographer works with. Changing the word class label of a word is a long-term process and the changes are not made easily; the lexicographer has to take into account the fact that every decision may add new boundaries and ambiguous spots. Is it necessary to take a more flexible approach to lexical category membership (see also Smith 2015)? What if all words cannot be PoS-tagged?

The first research question our study focuses on is the problems and practices of lexicographers dealing with PoS categorisation problems. Three issues were pointed out as the linguistically most problematic: the lexemes that i) occur in two or more lexical classes keeping the same base form, ii) are in a transition phase from one word class to another, or iii) only appear in certain inflected forms. Morphology was considered the main reason for the word forms departing from a paradigm. Regarding the possible reformation of the current Estonian word class system, opinions diverged: Considering the needs of regular users, a more general system was seen as preferable, but for the expert users a more fine-grained system was preferred. As an “applied approach” to word classes, the idea of a flexible display (applicable to a lexicographic root-database and dictionary writing system like Ekilex) emerged, taking into account both the needs of dictionary users and those of the experts.

The main concern is how to make well-grounded decisions based on the deluge of linguistic material. All in all, the lexicographers consider numerous aspects of their work but are also open to innovative solutions if they see the advantages. The respondents actually working with word class identification expressed a need to improve the factors influencing the process of PoS tagging, but also a certain scepticism towards an “ideal machine” that would be able to solve the categorisation issues characteristic of natural languages.

This leads us to the second focus of this study: the expectations lexicographers have with regard to modern technology-related solutions. We can conclude that despite a grain of scepticism, the lexicographers would welcome a reliable software solution to ease the decision-making process. In general, there is indeed a need for an improved morphological and syntactic parsing system, as well as for detection of changes in words’ semantic behaviour, and the latter is perhaps the most difficult to achieve. The solution would be a corpus-driven application presenting the statistics over the morphosyntactic distribution of an ambiguous word with access to the data source.

All in all, the lexicographers share an acute sense of responsibility related to the PoS judgment. They show remarkably high levels of empathy by having in mind both the regular user (or its conception), when making their proposals, e.g., an application of a possible technological tool with a simplified interface, and colleagues, when conceptualising the applied database assembling the judgments on difficult phenomena. Moreover, the potential PoS evaluator was considered useful not only for lexicographers

but also for the general public.

6. Acknowledgements

This work was supported by the Estonian Research Council grant PSG227.

7. Abbreviations

3 = third person; ALL = allative case; COMP = comparative; CONV = converbal; GEN = genitive case; ILL = illative case; PART = partitive case; PTCP = participle; PL = plural; SG = singular; TERM = terminative; TRA = translativ.

8. References

- Croft, W. (2001). *Radical Construction Grammar: Syntactic Theory in Typological Perspective*. Oxford: Oxford University Press.
- Culicover, P. W. (1999). *Syntactic Nuts*. Oxford University Press: Oxford.
- Erelt, M. (2017). Sissejuhatus süntaksisse [Introduction to syntax]. In M. Erelt & H. Metslang (eds.) *Eesti keele süntaks*. Tartu: Tartu Ülikooli Kirjastus, pp. 537–564.
- Habicht, K., Penjam, P. & Prillop, K. (2011). Sõnaliik kui rakenduslik probleem: sõnaliikide märgendamise vana kirjakeele korpus [‘Parts of speech as a functional and linguistic problem: annotation of parts of speech in the corpus of Old Written Estonian’]. *Estonian Papers in Applied Linguistics*, 7, pp. 19–41. <https://doi.org/10.5128/ERYa7.02>
- Haspelmath, M. (2001). Word classes and parts of speech. In P. B. Baltes & N. J. Smelser (eds.) *International encyclopedia of the social and behavioral sciences*, pp. 16538–16545. Amsterdam: Pergamon.
- Hennoste, T. (2002). Suulise kõne uurimine ja sõnaliigi probleemid. In R. Pajusalu, I. Tragel, T. Hennoste & H. Õim (eds.) *Teoreetiline keeleteadus Eestis*, pp. 56–73. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 4. Tartu: Tartu Ülikool.
- Kaalep, H.-J., Muischnek, K., Rääbis, A. & Habicht, K. (2000). Kas tegelik tekst allub eesti keele morfoloogilistele kirjeldustele? [‘Do the available morphological descriptions of Estonian work on a real text?’]. *Keel ja Kirjandus*, 9, pp. 623–633.
- Kallas, J., Koeva, S., Kosem, I., Langemets, M. & Tiberius, C. (2019). *Lexicographic practices in Europe: a survey of user needs*. Accessed at: https://elex.is/wp-content/uploads/2019/02/ELEXIS_D1_1_Lexicographic_Practices_in_Europe_A_Survey_of_User_Needs.pdf (08 June 2019)
- Kerge, K. (2016). Word-formation in the individual European languages: Estonian. In P. O. Müller, I. Ohnheiser, S. Olsen & F. Rainer (eds.) *Word-Formation. An International Handbook of the Languages of Europe*. Vol. 5. Handbooks of Linguistics and Communication Science 40. Berlin, New York: De Gruyter, pp. 3228–3259. <https://doi.org/10.1515/9783110424942-009>
- Lindström, L., Bakhoff, L., Kalvik, M.-L., Klaus, A., Läänemets, R., Mets, M., Niit,

- E., Pajusalu, K., Teras, P., Uibo, K., Veismann, A. & Velsker, E. (2006). Sõnaliigituse küsimusi eesti murrete korpuse põhjal. In E. Niit (ed.) *Keele ehe. Tartu Ülikooli eesti keele õppetooli toimetised* 30. Tartu: Tartu Ülikool, pp. 154–167.
- Paulsen, G. (2018). Manner and adverb: Fuzzy categorial boundaries in collocations. *Estonian Papers in Applied Linguistics*, 14, pp. 117–135. doi:10.5128/ERYa14.07
- Paulsen, G. (2019). Sõnaliigipiiridest kollokatsioonide vaatenurgast: erikäändelised noomenadverbid [Word class boundaries and collocations: The Estonian nominal adverbs in special cases]. *Estonian Papers in Applied Linguistics*, 15, pp. 121–137. doi.org/10.5128/ERYa15.07.
- Remes, H. (2009). *Muodot kontrastissa. Suomen ja Viron vertailevaa taivutusmorfologiaa. Acta Universitatis Ouluensis Humaniora B 90*. Oulu, Oulun yliopisto.
- Rosch, E., Mervis, C. B., Gray, W., Jason, D. & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, 8, pp. 382–439.
- Smith, M. C. (2015). Word categories. In J. R. Taylor (ed.) *The Oxford Handbook of the Word*. OUP Oxford: Kindle Edition.
- Tavast, A., Langemets, M., Kallas, J. & Koppel, K. (2018). Unified Data Modelling for Presenting Lexical Data: The Case of EKILEX. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (eds.) *Proceedings of the XVIII EURALEX International Congress: EURALEX: Lexicography in Global Contexts, Ljubljana, 17–21 July 2018*. Ljubljana University Press, Faculty of Arts, pp. 749–761.
- Viitso, T.-R. (2003). Rise and Development of the Estonian Language. In M. Ereht (ed.) *Estonian Language* (Linguistica Uralica Supplementary Series 1.). Tallinn: Estonian Academy Publishers, 130–230.

Dictionaries:

- BED: *Eesti keele põhisõnavara sõnastik* [The Basic Estonian Dictionary]. (2014). Kallas, J., Tiits, M., Tuulik, M. (eds.); Jürviste, M., Koppel, K., Tuulik, M. (compilers). Tallinn: Eesti Keele Sihtasutus. <https://sonaveeb.ee>
- DicEst: *Eesti keele sõnaraamat* [The Dictionary of Estonian]. (2019). Langemets, M., Tiits, M., Uibo, U., Valdre, T. & Voll, P. (eds.); Kuusik, K., Kuusk, K., Langemets, M., Tiits, M., Uibo, U., Valdre, T. & Voll, P. (compilers). Institute of the Estonian Language. <http://www.sonaveeb.ee>
- DSE: *Eesti õigekeelsussõnaraamat* [The Dictionary of Standard Estonian]. (2018). Raadik, M., Ereht, T., Leemets, T. & Mäearu, S. Tallinn: Eesti Keele Sihtasutus. <http://www.eki.ee/dict/qs/>
- ECD: *Eesti keele naabersõnad* [The Estonian Collocations Dictionary]. (2019). Kallas, J., Koppel, K., Paulsen, G. & Tuulik, M. Institute of the Estonian Language. <http://www.sonaveeb.ee>
- Eesti keele sõnapäevad. Tänapäeva eesti keele sõnavara struktuurianalüüs* [The

- Dictionary of Estonian Word Families. A structural analysis of the contemporary Estonian lexis*. Vol. I-II. Vare, S. (2012). Tallinn: Eesti Keele Sihtasutus.
- EKSS: *Eesti keele seletav sõnaraamat I–VI* [*The Explanatory Dictionary of Estonian*]. “Eesti kirjakeele seletussõnaraamatu” 2., täiendatud ja parandatud trükk. Margit Langemets, Mai Tiits, Tiia Valdre, Leidi Veskis, Ülle Viks, Piret Voll (Toim.). Eesti Keele Instituut. Tallinn: Eesti Keele Sihtasutus, 2009.
<http://www.eki.ee/dict/ekss/>
- Sõnaveeb* [Wordweb]. (2019). The Language Portal of the Institute of the Estonian Language. Hein, I., Kallas, J., Koppel, K., Langemets, M., Männiko, K., Nurk, T., Viks, Ü., Laubre, M., Ukkivi, R., Tavast, A., Lastovets, S. & Rautam, S. (eds.).

Corpora:

- Kallas, J., & Koppel, K. (2018, March 26). *Eesti keele ühendkorpus 2017* [*The Estonian National Corpus*]. Center of Estonian Language Resources.
<https://doi.org/10.15155/3-00-0000-0000-0000-071E7L>
- Corpus of Old Written Estonian. Prillop, Külli. (2013, January 9). *Vana kirjakeele korpus*. [*Corpus of Old Written Estonian*]. Center of Estonian Language Resources. <https://doi.org/10.15155/1-00-0000-0000-0000-000751>
- Kallas, J. & Koppel, K. (2018, April 23). *Estonian Corpus for Learners 2018 (etSkELL)*. Center of Estonian Language Resources. <https://doi.org/10.15155/3-00-0000-0000-0000-073351>.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



Repel the Syntruders! A Crowdsourcing Cleanup of the Thesaurus of Modern Slovene

Jaka Čibej, Špela Arhar Holdt

Centre for Language Resources and Technologies (Faculty of Arts, Faculty of Computer and Information Science), University of Ljubljana, Večna pot 113, 1000 Ljubljana, Slovenia
E-mail: jaka.cibej@cjvt.si, spela.arhar@cjvt.si

Abstract

The Thesaurus of Modern Slovene is the largest open-source digital collection of Slovene synonyms, published in March 2018 by the Centre of Language Resources and Technologies of the University of Ljubljana. The Thesaurus was initially compiled entirely automatically and allows users to contribute toward improving the resource by adding suggestions for missing synonyms and/or by evaluating both the synonym candidates from the initial database as well as the suggestions added by other users. As an automatically generated language resource, however, the initial database of the Thesaurus includes a certain degree of noise. In the paper, we present two crowdsourcing activities aimed at cleaning up the database. The first is a targeted annotation campaign aimed at evaluating multi-word synonym candidates in the Thesaurus, and the second is an analysis of user votes provided directly in the Thesaurus interface. Both scenarios are examples of an effective postprocessing method for an automatically generated language resource and demonstrate that crowdsourcing can play an important role in smart lexicography, especially in the case of less-resourced languages.

Keywords: crowdsourcing; synonyms; Slovene; thesaurus; digital lexicography

1. Introduction

Crowdsourcing has demonstrated its value in numerous scientific endeavours, as demonstrated by a number of successful initiatives that have channelled the power of the crowd to great effect: in the field of linguistics, natural language processing has embraced crowdsourcing as a method to clean noisy datasets (Fišer et al., 2014), annotate language data (Fort et al., 2014), or collect user estimations and judgments (Snow et al., 2008). In the field of lexicography, a number of important steps towards the implementation of crowdsourcing in lexicographic workflows have also been made – see, for example, Čibej et al. (2015) for a proposed modular crowdsourcing workflow model for lexicography, or Abel and Meyer (2013) for an overview of different types of user contributions to online dictionaries – and although user involvement in lexicographic projects for digital dictionaries is not entirely new (as shown by Lew (2014) numerous collaborative lexicographic projects are available online, the most

noted among them being the Urban Dictionary¹ and Wiktionary²), crowdsourcing differs from collaborative lexicography in the fact that the former is usually more restricted in how users can contribute to the compilation of a digital dictionary (i.e. they solve a relatively narrow, predefined task and require a platform on which to solve it, as opposed to the free-for-all approach often employed by collaborative dictionaries). In addition, crowdsourcing can take place at any stage of dictionary compilation, both pre- and post-publication. Although only a handful of good practice examples showcase the implementation of crowdsourcing in lexicographic workflows (Kosem et al., 2018), the rise and proliferation of digital-born dictionaries is paving the way to a more crowd-oriented form of dictionary compilation. The goal of this paper is to present one such dictionary project, the Thesaurus of Modern Slovene (Krek et al., 2018a), and the way crowdsourcing is being used to clean it. We present the results of two instances of crowdsourcing activities aimed at cleaning up the noise in the Thesaurus: (a) the votes provided by the dictionary users and collected directly through the dictionary interface in the first year since its publication, and (b) a more targeted crowdsourcing campaign for students of linguistics in order to evaluate a set of multi-word synonym candidates.

The paper is structured as follows: in Section 2, we present the Thesaurus of Modern Slovene, its compilation and overall design. In Section 3, we describe the results of the targeted crowdsourcing campaign focusing on multi-word synonym candidates. In Section 4, we present an analysis of the upvotes and downvotes on synonym candidates collected directly through the Thesaurus interface. We conclude with a discussion and some directions for future work in Section 5.

2. The Thesaurus of Modern Slovene

The Thesaurus of Modern Slovene is the largest open-source digital collection of Slovene synonyms. It was published in March 2018 by the Centre of Language Resources and Technologies of the University of Ljubljana as the first example of a *responsive dictionary* (Arhar Holdt et al., 2018), a new type of language resource that is defined by the following characteristics: first, it is a born-digital and digital-only dictionary, designed with the needs, requirements, and advantages of the digital medium in mind. Second, its database was initially compiled entirely through automatic methods that were tested and evaluated beforehand. Third, both the database and the language resource were made openly accessible to the language community immediately after the automatic compilation to provide a large amount of automatically extracted language data which is relevant, but this contains a certain degree of noise. Fourth, because of its digital nature, the dictionary is frequently updated and all changes are tracked through versions and with timestamps at the level of entries. Finally, the responsive dictionary features one or more ways to allow users to contribute to its development.

¹ <https://www.urbandictionary.com/>

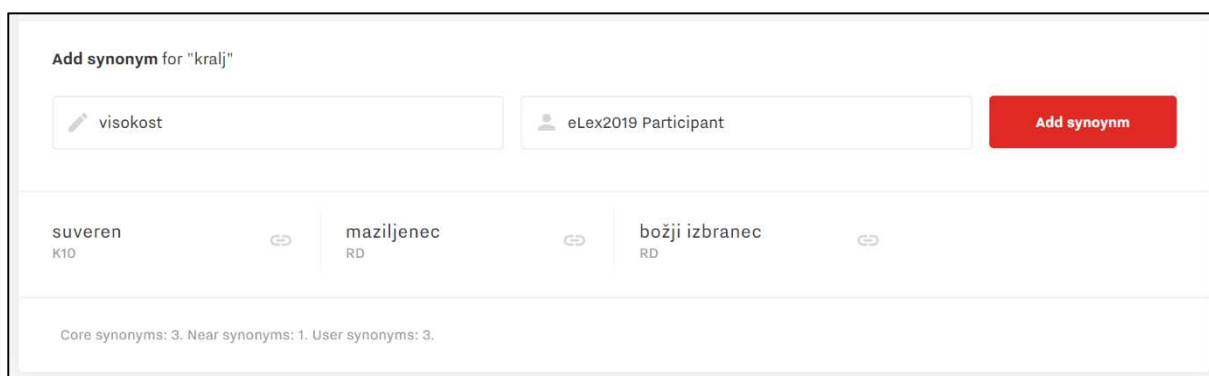
² <https://en.wiktionary.org/>

Through this, it responds to changes in language on the one hand and the knowledge/consensus of its users on the other.

As the first example of a responsive dictionary, the Thesaurus of Modern Slovene was initially compiled automatically with co-occurrence graphs (for a more detailed description of the methodology, see Krek et al., 2017) using existing language resources, namely The Oxford®-DZS Comprehensive English-Slovenian Dictionary and the Gigafida Reference Corpus of Written Slovene. The Thesaurus database was made available in the CLARIN.SI repository (Krek et al., 2018b) under the Creative Commons Attribution-ShareAlike 4.0 International licence (CC BY-SA 4.0).

A custom interface was developed to enable the language community to contribute toward improving and further developing the resource in two ways: (1) by adding their own suggestions of missing synonym candidates to a particular entry; and/or (2) by evaluating both the synonym candidates from the initial database as well as the suggestions added by other users by upvoting or downvoting them.

Users can add synonyms through a special form integrated in the interface (Figure 1). No registration is required – the user can enter a username and the suggested synonym in the designated fields and then click the Add Synonym button. The suggestion is instantly displayed in the user synonym section.



Add synonym for "kralj"

visokost

eLex2019 Participant

Add synonym

suveren K10

maziljenec RD

božji izbranec RD

Core synonyms: 3. Near synonyms: 1. User synonyms: 3.

Figure 1: Adding user synonyms to the Thesaurus.

Users can upvote or downvote existing synonym candidates by hovering over the candidate and clicking the upvote (green) or downvote (red) button (Figure 2). They can also cancel the vote if they misclicked. During dictionary updates, votes are taken into account so that downvoted synonyms can be excluded from the dictionary, while upvoted synonyms can be ranked higher.

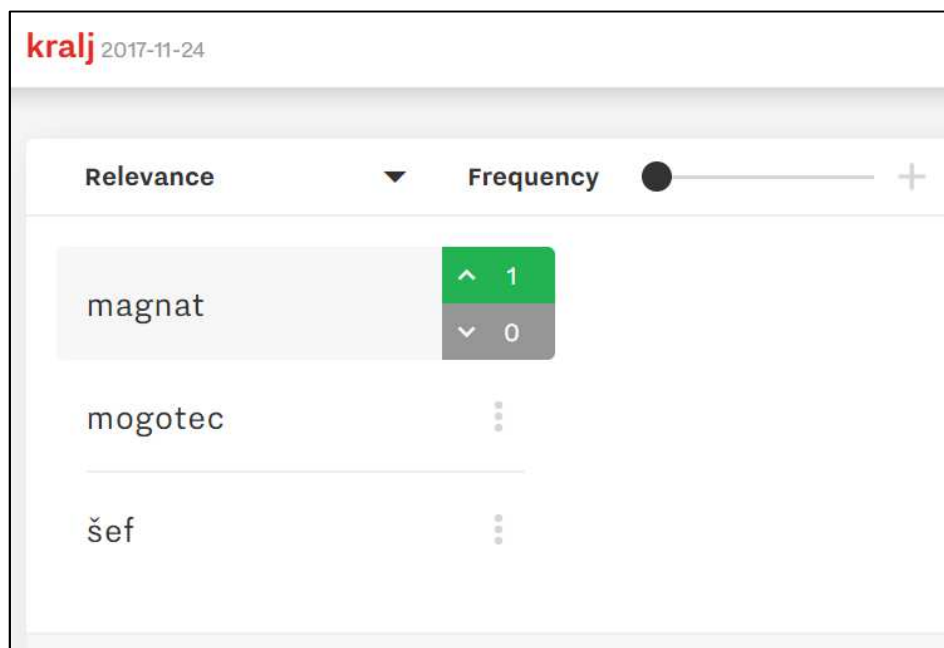


Figure 2: Voting for synonym candidates in the Thesaurus.

The users' reactions to the Thesaurus were predominantly positive. The number of user added synonyms and votes (more on this in Section 4) indicates that users are positively inclined toward user involvement in the Thesaurus. This demonstrates that the automatic compilation of language resources can be efficient both in terms of time and financial investment, particularly when development is continued in the post-publication phase and complemented by user involvement.

2.1 Noise from automatic synonym extraction

As an automatically generated language resource, the initial database of the Thesaurus of Modern Slovene includes a certain degree of noise. These methodology-related problems were clear in the beginning, so we decided to focus on them. Because the synonyms in the Thesaurus were extracted from Slovene translation equivalents of English headwords, multi-word synonym candidates are sometimes only descriptive approximates of concepts that are lexicalized in English but not in Slovene, e.g. *rooming-in* as an English loanword and *24-urno sobivanje novorojenčka in matere* 'a 24-hour cohabitation of a newborn and their mother'. Problematic categories of automatically extracted synonyms include feminine-masculine word pairs that ended up under the same entry (e.g. the word *učitelj* 'teacher [masculine]' is listed as a synonym under the headword *učiteljica* 'teacher [feminine]'), inadequate candidates arising from structural irregularities or inconsistencies in The Oxford®-DZS Comprehensive English-Slovenian Dictionary, and multi-word synonym candidates that border on paraphrases, definitions, partial repetitions, or descriptions.

Because multi-word synonym candidates were easy to identify and were the most obviously problematic category (as well as manageable in size), we decided to organize a targeted crowdsourcing campaign to identify the synonym candidates to be removed in the next Thesaurus update. We describe the campaign in Section 3.

3. Crowdsourcing multi-word synonym candidates

The goal of the crowdsourcing experiment was to exclude inadequate multi-word synonym candidates from the dataset. In this section, we describe the infrastructure utilized in the data and the platform used, the task design process, and the results of the experiment.

3.1 Data preprocessing

Version 1.0 of the Thesaurus of Modern Slovene contains 368,117 headword-synonym pairs (not counting the synonym candidates added by users); 162,719 of these pairs (44%) contain a multi-word synonym candidate or headword. The Thesaurus is structured in such a way that each synonym can also appear as a headword, so the number of unique pairs (in terms of their components) containing a multi-word string is 84,128.

Not all pairs were included in the crowdsourcing task. The data was preprocessed to make sure the workload was manageable and to remove inadequate pairs that were easy to identify automatically (through a set of rules).

The excluded categories were the following:³ (a) pairs containing two two-word synonym candidates, both containing the reflexive pronoun *se* (e.g. *prelomiti se* and *zlomiti se* ‘to break’; 4,510 pairs); (b) pairs containing a number of problematic words often used in descriptive synonym candidates, e.g. the verbs *biti* ‘to be’, *začeti* ‘to begin’, *končati* ‘to finish’, the preposition *brez* ‘without’, and the nouns *prebivalec* ‘inhabitant [male]’ and *prebivalka* ‘inhabitant [female]’ (6,141 pairs); (c) pairs that overlapped to a great extent (e.g. *hoditi z dolgimi koraki* ‘to walk with long steps’ and *začeti hoditi z dolgimi koraki* ‘to begin walking with long steps’, 5,517 pairs); (d) pairs that contained a synonym candidate with a terminological/field label (e.g. *zoologija* ‘zoology’, 14,581 pairs); and (e) pairs that contained masculine and feminine synonym candidates (e.g. *industrijski psiholog* ‘industrial psychologist [male]’ and *industrijska psihologinja* ‘industrial psychologist [female]’, 5,334 pairs). The final set of pairs after the automatic preprocessing contained 18,635 headword-synonym pairs. The pairs not included in the experiment will be further analyzed and most likely removed from the Thesaurus.

³ Some synonym pairs were assigned to multiple categories.

3.2 Crowdsourcing platform

The platform we used in the experiment was PyBossa⁴, an open-access Python-based crowdsourcing platform that features a great deal of flexibility, especially with regard to task design and interface optimization with the aim of greater user-friendliness. It also features an API, and allows data upload/download in .JSON format. The crowdsourced data can be downloaded at any stage of the crowdsourcing process. However, PyBossa does not include any quality control functions (e.g. inter-annotator agreement measures, automatic gold-standard comparison), so these were handled outside the platform using external custom-made Python scripts. PyBossa has already been used with great success in previous work, e.g. for annotating collocations for the Collocations Dictionary of Modern Slovene (Kosem et al., 2018).

3.3 Task Design and Crowdsourcer Recruitment

We designed a custom interface for the task (Figure 3). In each task, the crowdsourcer was presented with 10 headword-synonym pairs and three options to choose from: *Da* ‘Yes’ if the units were synonymous, *Ne* ‘No’ if they were not, and *Ne vem* ‘I don’t know’ if they were uncertain. The crowdsourcer had to tick all ten pairs to be able to finally click *Shrani* ‘Save’ and proceed to the next task.

Several measures were taken to reduce the number of mistakes during crowdsourcing. Radio buttons were used in order to reduce the number of misclicks (and the crowdsourcer could revise their annotation within each batch of 10 pairs as many times as they wanted). In addition, the user got a pop-up alert if they forgot to tick one of the buttons.

Six crowdsourcers⁵ were recruited for the task, all of them students of linguistics at the University of Ljubljana. They were familiarized with the Thesaurus and the goal of the task during an introductory briefing session. The guidelines were not overly specific, as the task is very similar to the voting system already enabled in the dictionary interface. The crowdsourcer’s main task was to provide their subjective judgment on whether the given headword-synonym pair would be useful in the Thesaurus.

⁴ <https://pybossa.com/> – DOI: 10.5281/zenodo.3239980

⁵ The targeted campaign is not a typical example of crowdsourcing (as it relies on a limited group of preselected crowdsourcers) and is actually a mock-up of a full-scale crowdsourcing campaign because it uses the same methodology which is independent of this specific project. The crowd used could be significantly larger and more diverse, and may be such in our future work. For the sake of simplicity, we refer to all these activities as crowdsourcing in the paper.

The image shows a web interface for a crowdsourcing task. It contains two identical-looking task blocks. Each block has a header with a synonym pair in a box, followed by a label 'Sopomenki:' and three radio button options: 'Da', 'Ne', and 'Ne vem'. The first task's header is 'agitator || dekllica za vse' and the second's is 'agitator || deček za vse'. At the bottom right of the interface is a blue button labeled 'Shrani'.

Figure 3. PyBossa task interface for annotating multi-word synonym pairs.

Based on the testing phase in which each synonym pair took approximately 7-8 seconds on average to evaluate (rounded up to 10 seconds), the crowdsourcing campaign was foreseen to take approximately 153 hours to complete. At the standard rate of the University of Ljubljana for student work (€7), it would cost approximately €1,071.

3.4 Results

In total, 56,745 responses were collected during the crowdsourcing task, three (in some rare cases four⁶) for each headword-synonym pair. The mean response time per synonym pair (after removing outliers above 15 seconds) was approximately 8.4 seconds, with 7.9 as the median value. The total time spent on tasks was approximately 92 hours (with a cost of cca. €650 in student work). In general, approximately 57% of the responses were positive and approximately 41% were negative, while only 1.6% were undecided. We tracked inter-annotator agreement for each crowdsourcer pair with two measures: the percentage of identical answers, and Cohen's kappa coefficient. The average percentage of identical answers between annotators was 71% (ranging from 63% to 79%), while Cohen's kappa coefficient ranged from 0.33 to 0.55 with an average of 0.42, which indicates fair to moderate agreement.

We also measured to what extent the crowdsourcers agreed on whether a given pair is adequate or inadequate by calculating information entropy for the responses of each

⁶ The number of responses was limited to three per synonym pair, but if a crowdsourcer started solving a particular task and another crowdsourcer was presented with it before the first crowdsourcer was done with it, both responses were registered, which sometimes resulted in four responses per synonym pair.

pair. An information entropy of 0 indicates perfect agreement (e.g. all responses are positive), while higher values indicate various degrees of disagreement (e.g. a value of 1.58 indicates a response combination akin to Yes-No-I Don't Know). The results are shown in Table 1.

Information Entropy	Frequency	Percentage
0	10,136	54.39
0.81	203	1.09
0.92	7,943	42.62
1	89	0.48
1.5	7	0.04
1.58	257	1.38

Table 1: Evaluated synonym pairs by information entropy.

More than half (54%) of the pairs featured complete agreement between crowdsourcers, while the majority of the rest featured a slightly mixed response (42%). A more detailed distribution is shown in Table 2, where response combinations are also grouped into categories by agreement.

A thorough manual categorization of the annotated data is beyond the limits of this paper, but we nevertheless show a brief overview of the results which indicate that several groups or categories can be formed based on different response combinations. We list illustrative examples for both the pairs with complete agreement and pairs with mixed responses in order to provide some insight into the results.

A large number of inconclusive or mixed responses was likely caused by unfamiliarity with the annotated synonym candidates (e.g. infrequent words, terminological units, or loanwords). On the other hand, disagreement occurs with pairs in which their semantic similarity is clear, but they are interchangeable only in specific language contexts.

Response Combination	Frequency	Percentage
Complete Agreement		
Yes, Yes, Yes	6,245	33.51
Yes, Yes, Yes, Yes	156	0.84
No, No, No	3,616	19.40
No, No, No, No	115	0.62
Mixed Response		
Yes, Yes, Yes, No	117	0.63
Yes, Yes, No	4,267	22.90
Yes, Yes, I don't know	228	1.22
Yes, No, No	3,163	16.97
Yes, No, No, No	81	0.43
No, No, No, I don't know	5	0.03
No, No, I don't know	200	1.07
Inconclusive Response		
Yes, Yes, No, I don't know	2	0.01
Yes, No, No, I don't know	4	0.02
Yes, No, I don't know	257	1.38
Yes, No, I don't know, I don't know	1	0.01
No, I don't know, I don't know	45	0.24
Yes, I don't know, I don't know	40	0.21
I don't know, I don't know, I don't know	4	0.02
Yes, Yes, No, No	89	0.48

Table 2. Evaluated synonym pairs by response combinations.

The examples with agreement that the multi-word unit is relevant for the Thesaurus include e.g. pairs in which **the multi-word unit is an explanation of a (frequently)**

single-word headword. This is often an explanation of loanwords or neologisms (*sendvičarna* ‘sandwich store’ – *trgovina s sendviči* ‘store with sandwiches’; *absentizem* ‘absenteeism’ – *odsotnost z dela* ‘absence from work’; *glosirati* ‘to gloss’ – *pojasniti v opombi* ‘to explain in a notation’), but explanations of more frequent vocabulary also occur (*zmagati* ‘to win’ – *priti na prvo mesto* ‘to get first place’; *zmleti* ‘to crush’ – *zdrobiti v prah* ‘to grind into dust’; *pravljichen* ‘fairytale-like’ – *kot iz pravljice* ‘like in a fairytale’). What is interesting to note is that the crowdsourcers did not see these examples as redundant and irrelevant, but want to keep them in the Thesaurus. The same is true for **multi-word pairs that differ only in a single word** and in which the two differing words are synonymous (e.g. *razdeliti na pokrajine* ‘to divide into regions’ – *razdeliti na province* ‘to divide into provinces’; *sprejem s koktejli* ‘cocktail reception’ – *zabava s koktejli* ‘cocktail party’; *obsoditi na pogubo* ‘to condemn to oblivion’ – *obsoditi na propad* ‘to condemn to downfall’). The third group includes examples which **differ in part-of-speech structure**, e.g. a pair of nominal phrases containing an adjectival or prepositional attribute (*cestni davek* ‘road tax’ – *davek za uporabo cest* ‘tax for road use’; *brivski pribor* ‘shaving kit’ – *pribor za britje* ‘kit for shaving’; *bralna očala* ‘reading glasses’ – *očala za branje* ‘glasses for reading’), or pairs containing instrumental-case and genitive-case phrases (*s spretnimi prsti* ‘with nimble-fingers’ – *spretnih prstov* ‘lit. [of] nimble fingers, nimble-fingered’). In some pairs, one of the units features a **semantically light verb** (*dati nižjo oceno* ‘to give a lower grade’ – *znižati oceno* ‘to lower the grade’; *dati novo ime* ‘to give a new name’ – *preimenovati v* ‘to rename [to]’). The most interesting examples (although rare) are the ones in which one or even both components of the pair are **idiomatic expressions** (e.g. *dati zeleno luč* ‘to greenlight’ – *uradno odobriti* ‘to officially approve’; *pogled resnici v oči* ‘lit. staring truth in the eye’ – *spust na realna tla* ‘lit. a descent to solid ground’).

On the other hand, examples that the crowdsourcers unanimously want removed from the databases include pairs that were semantically linked in the original bilingual dictionary but **are not themselves synonymous**, e.g. *trikrat tedensko* ‘three times per week’ – *vsake tri tedne* ‘once every three weeks’; *sod za olje* ‘oil barrel’ – *vinski sod* ‘wine barrel’; *speti v čop* ‘to put up [hair] in a ponytail’ – *splesti v kito* ‘to braid [hair]’. Similarly, in some examples **synonymy is limited to a specific context** (*ne sprejeti* ‘to not accept’ – *vreči na izpitu* ‘to fail [someone] on an exam’; *nizek* ‘low’ – *s plosko peto* ‘with a flat heel’). Other negatively evaluated examples include pairs in which one of the components contains a **semantically specific complement** (*zvižati se* ‘to squirm’ – *zvižati se kot črv* ‘to squirm like a worm’, *za kuhanje* ‘for cooking’ – *za kuhanje na visoki temperaturi* ‘for cooking on high heat’) or a **prepositional verb** (*iti do* ‘to go to’ – *spustiti se v* ‘to descend into’; *iti k* ‘to go to’ – *videvati se z* ‘to see [someone]’), as well as **inadequately paired masculine and feminine variants** occurring in more complex structures, which prevented them from being filtered out automatically (e.g. *študent medicine* ‘[male] student of medicine’ – *študentka medicinske fakultete* ‘[female] student of a Faculty of Medicine’). In rare cases, inadequate pairs stem from **errors during the automatic export** of synonym candidates (e.g. *official* – *na visokem položaju* ‘in high places’, *people* – *na visokem položaju* ‘in high places’).

Examples with mixed responses contain less clearly defined groups. Predominantly positively evaluated examples include pairs from the above-defined groups that are characterized by a certain degree of **semantic similarity**, but do not necessarily overlap in terms of synonymy in different contexts, e.g. *zaljubljen v gledališče* ‘in love with the theater’ – *zaljubljen v oder* ‘in love with the stage’; *zakopati v jamo* ‘to bury in a cave’ – *zakopati v luknjo* ‘to bury in a hole’). In some examples, the vocabulary is **infrequent or specialized**, and as such presumably not familiar to the crowdsourcers (e.g. *primogenitura* ‘primogeniture’ – *pravica prvorojenca do nasledstva* ‘the right of the firstborn to inheritance’; *kiras* ‘cuirass’ – *prsni del oklepa* ‘the chest part of armor’). Similar categories can be identified in predominantly negatively evaluated pairs: semantically similar pairs (but not similar enough), e.g. *zapreti v kletko* ‘to put in a cage’ – *zapreti v kurnik* ‘to put in a chicken coop’; *upodobiti na fotografiji* ‘to portray in a photograph’ – *upodobiti na sliki* ‘to portray in a portrait’) and pairs with infrequent or specialized vocabulary (*blindirati* – *obložiti s ploščicami* ‘to insulate with panels’).

The most interesting examples for further analyses are the ones found in the inconclusive responses. These predominantly consist of **infrequent, specialized vocabulary** likely unfamiliar to the crowdsourcers (*barg* ‘barge’ – *kanalski tovorni čoln* ‘channel cargo boat’; *alkova* ‘alcove’ – *okno v tinu* ‘window in a corner’), but also of **stylistically marked vocabulary** (*crkniti* ‘to drop dead’ – *zrušiti se od utrujenosti* ‘to collapse of exhaustion’), **loanwords** (*digest* – *zbirka izvlečkov iz člankov* ‘a collection of article excerpts’), or more complex examples of **masculine-feminine pairs** (*liftboy* – *uslužbenka pri dvigalu* ‘[female] employee at the elevator’). A separate category consists of examples in which the responses were certain, but divided (e.g. Yes, Yes, No, No). Besides the already mentioned pairs with **limited synonymy** (e.g. *torba za cunje* ‘bag for clothes’ – *vreča za cunje* ‘sack for clothes’; *medsebojno povezovanje* ‘interconnection’ – *navezovanje poslovnih stikov* ‘forming business contacts’), an interesting category are the examples with **phraseological components** that certain crowdsourcers may not have recognized (*iti zraven* ‘to come with’ – *priti v paketu* ‘to come with the package’; *nasloniti se nazaj* ‘to lean back’ – *sprostiti se* ‘to relax’). Disagreement can also be observed with phrases that express the **(im)perfectiveness of the action**, e.g. *razdražiti* ‘to irritate [perfective]’ – *iti na živce* ‘to go on [smn’s] nerves [imperfective]’; *nadeti si tančico* ‘to put on a veil’ – *nositi tančico* ‘to wear a veil’, *izbruhniti v smeh* ‘to burst into laughter’ – *pokati od smeha* ‘to be bursting of laughter’.

Among the already mentioned 54% of pairs with complete agreement, approximately 34% were evaluated as adequate and 20% as inadequate. While only 2.37% of all evaluated synonym pairs resulted in a response that was completely inconclusive, and the remaining 43% could be resolved through a majority vote at this stage, a seventh annotator was recruited to provide additional votes for pairs with mixed or inconclusive responses. Table 3 shows the results for the ambiguous pairs when taking into account the responses made by the seventh annotator.

Response Combination	Frequency	Percentage
Predominantly Positive		
Yes, Yes, Yes, Yes, No	63	0.77
Yes, Yes, Yes, No	2,373	28.97
Yes, Yes, Yes, No, No	87	1.06
Yes, Yes, Yes, No, I don't know	2	0.02
Yes, Yes, Yes, I don't know	135	1.65
Predominantly Negative		
Yes, No, No, No	2,103	25.68
Yes, No, No, No, No	56	0.68
Yes, No, No, No, I don't know	4	0.05
No, No, No, No, I don't know	4	0.05
No, No, No, I don't know	145	1.77
Yes, Yes, No, No, No	61	0.74
Inconclusive Response		
Yes, Yes, No, No	2,700	32.97
Yes, Yes, I don't know, I don't know	23	0.28
Yes, Yes, No, I don't know	204	2.49
Yes, No, No, I don't know	168	2.05
Yes, No, No, I don't know, I don't know	1	0.01
Yes, No, I don't know, I don't know	28	0.34
Yes, I don't know, I don't know, I don't know	1	0.01
No, No, I don't know, I don't know	30	0.37
No, I don't know, I don't know, I don't know	2	0.02

Table 3. Evaluated ambiguous synonym pairs with additional annotations.

With the addition of another vote, 2,660 synonym pairs start to converge more toward Yes and 2,373 toward No (with 60-80% of votes in favour of one or the other), while 3,157 still remain inconclusive (approximately 17% of all the multi-word synonym pairs included in the crowdsourcing task). Out of these, 2,700 examples keep conflicting responses (Yes, Yes, No, No). Many of these are pairs in which synonymy is limited to a specific context, while one component is semantically wider or inclusive, e.g. *redčiti se* ‘to thin’ – *začenjati dobivati plešo* ‘to begin to go bald’, *prečkanje puščave* ‘the crossing of the desert’ – *vožnja čez puščavo* ‘the drive across the desert’. In some examples, the multi-word units differ in a single word, but the substitution reduces the degree to which the units are interchangeable in use, e.g. *znova se sestati* ‘to have a reunion’ – *znova se zbrati* ‘to regroup’; *zelo smešna zgodba* ‘a very funny story’ – *zelo smešna šala* ‘a very funny joke’; *zbirati se v bazen* ‘to gather in a pool’ – *zbirati se v tolmun* ‘to gather in a pond’. Another category difficult to evaluate consists of examples with a semantically specific complement that is part of the original phrase, but is commonly left out in language use, e.g. *človek z dna* ‘a man from the bottom’ – *človek z dna družbene lestvice* ‘a man from the bottom of the social scale’; *dijak tretjega letnika* ‘third-year student’ – *dijak tretjega letnika srednje šole* ‘third-year high-school student’. The same is true of examples in which the synonym candidate is explanatory, but semantically too narrow or too vague (*čarodej* ‘magician’ – *praktikant črne magije* ‘practitioner of black magic’; *nož za sir* ‘cheese-cutting knife’ – *naprava za rezanje sira* ‘cheese cutter’). Some examples are problematic because of the (im)perfectiveness of the expressed action (*pihati od jeze* ‘to be seething with anger’ – *ujeziti se* ‘to become angry’), prepositional verbs (*vrniti se na* ‘to return to’ – *znova stopiti v* ‘to reenter into’), and (potential) phraseological units (*prinesti na krožniku* ‘to bring on a plate’ – *servirati na pladnju* ‘to serve on a platter’). On the other hand, at this stage there seem to be no more problematic examples with rare or specialized vocabulary, masculine-feminine pairs, paraphrases of part-of-speech structures, methodological extraction errors, or results that clearly are (or are not) synonymous.

This outcome is in line with our expectations based on previous findings: after the automatic compilation of the Thesaurus, an evaluation of the dataset was conducted by experts (linguists and lexicographers) on a random subset of headword-synonym pairs (not limited to multi-word synonym candidates). The goal of the task was to evaluate synonyms as either good, acceptable, or poor. The results of the expert evaluation are shown in Figure 4 (see Arhar Holdt et al., 2018 for more on the evaluation).

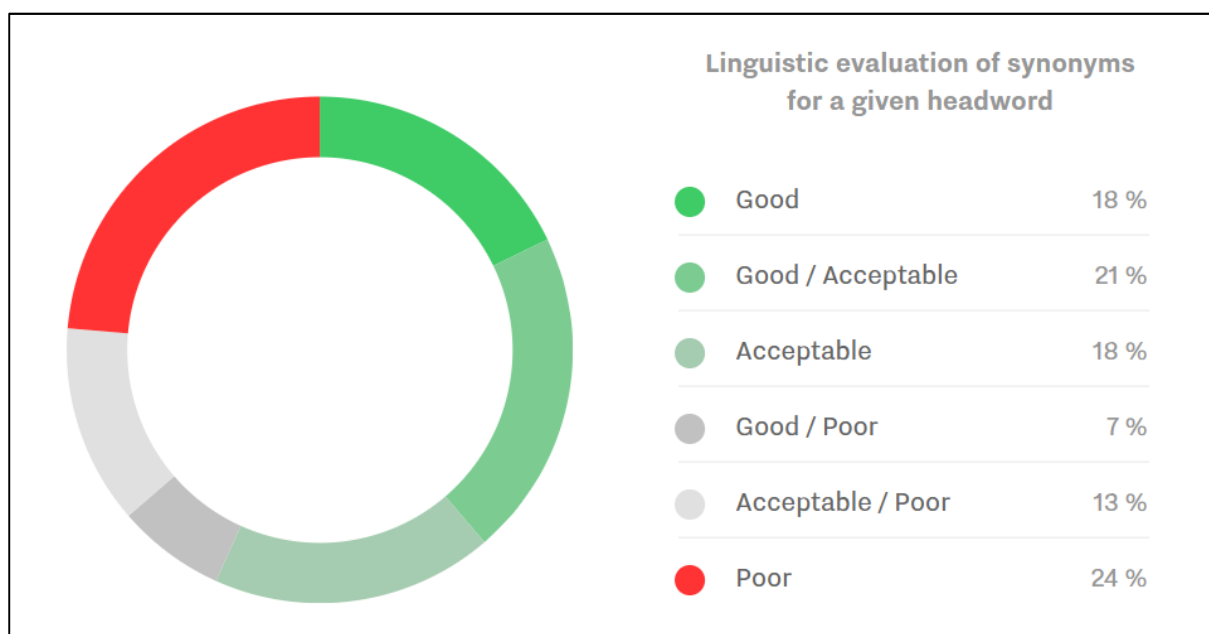


Figure 4: Expert evaluation of the synonyms in the Thesaurus of Modern Slovene.

Even with expert annotations, 20% of the evaluated synonym pairs showed considerable disagreement: 7% were simultaneously evaluated as good and poor, while 13% were rated as both acceptable and poor. This demonstrates that evaluating synonymy is not a trivial or one-dimensional task, and that some examples pose a challenge for both experts and non-experts. Synonymy is highly context-dependent, and to some extent subjective. With the Thesaurus of Modern Slovene, however, we took a greedier approach and opted to treat ambiguous synonyms as potentially adequate rather than exclude them, as the evaluation has shown that at least some users might find them useful.

4. Collecting user votes through the interface

Parallel to the targeted crowdsourcing campaign described in Section 3, user votes were also collected directly from the interface of the Thesaurus (for all synonym candidates, not just multi-word ones). In the period since the release of the Thesaurus (March 2018 – 3 June 2019), a total of 26,253 user votes was collected, 24,214 (92%) of which were upvotes and 2,039 (8%) were downvotes. The majority of votes (21,886, or 83%) was collected for the original Thesaurus synonyms, while a smaller portion (4,367, or 17%) was collected for the user-added synonyms. In this paper, we concentrate on the original Thesaurus synonyms.

A total of 17,904 headword-synonym pairs in the Thesaurus (5% of the entire dictionary) received at least one vote, the majority one vote (15,307 pairs) or two (2,035 pairs). 16,938 pairs received at least one upvote and 1,340 pairs received at least one downvote.

The results indicate that the users are positively inclined to the automatically compiled data (taking into account the headwords they have queried so far), with only 7.5% of the voted pairs having been downvoted, and even fewer (968, or 5.4%) having received only downvotes and no upvotes.

We list here several examples of user-voted synonym candidates, with the number of upvotes and downvotes in brackets. As can be expected, the most votes were collected for the example entries *zelen* ‘green’, *ideja* ‘idea’, and *spati* ‘to sleep’, which are often used during Thesaurus presentations as demonstrative examples (for voting as well); *zelen* – *mlad* ‘young’ (17+, 9-), *zelen* – *bled* ‘pale’ (11+, 8-). Disagreement in votes also occurs with terminological words or words of foreign origin (*splentitis* ‘spleentitis’ – *vnetje vranice* ‘spleen infection’, 6+, 6-; *hiša* ‘house’ – *polje* ‘field’, astrology, 1+, 4-; *izdajalec domovine* ‘traitor of the country’ – *kvisling* ‘quisling’, 2+, 3-) or stylistically marked words (vulgar: *drek* ‘shit’ – *en kurec* ‘piece of shit’, 2+, 3-). Pairs that have been more consistently downvoted include explicitly pejorative components (*teta* ‘auntie’ – *poženščen peder* ‘effeminate faggot’, 0+, 3-) or masculine-feminine pairs (*babica* ‘grandmother’ – *dedek* ‘grandfather’, 0+, 12-), as well as examples in which the vocabulary is general and stylistically neutral, but not synonymous enough (*domišljav* ‘pretentious’ – *zadovoljen* ‘satisfied’, 0+, 1-; *domišljav* ‘pretentious’ – *samozavesten* ‘confident’, 0+, 1-).

With positively voted candidates it is more difficult to pinpoint the reasons for the upvotes. It appears that users upvote very prototypical, unambiguous, and widely interchangeable synonyms (*lakomnost* ‘avarice’ – *pohlepnost* ‘greed’, 12+, 0-; *lep* ‘beautiful’ – *čeden* ‘handsome’, 12+, 2-; *groziti* ‘to threaten’ – *pretiti* ‘to menace’ 4+, 0-; *prebrisan* ‘ingenious’ – *premeten* ‘cunning’, 4+, 0-). Examples with single upvotes indicate that users vote for them systematically: presumably the same user votes for most of the synonym candidates in a headword (*abolirati* ‘to abolish’ – *ukiniti* ‘to cancel’, 1+, 0-; *abolirati* – *razveljaviti* ‘to cancel’, 1+, 0-).

An overview of the multi-word synonym candidates with votes (which this paper focuses on) shows that 868 examples with three or more words received user votes in the interface. Sixty of these include negative votes (e.g. pairs with redundantly repeated parts and/or masculine-feminine pairs (*paravojak* ‘[male] para-soldier’ – *pripadnica paravojaške organizacije* ‘[female] member of a paramilitary organization’, 0+, 2-), non-synonymous pairs (*človek* ‘human’ – *nabit z energijo* ‘full of energy’; 0+, 2-), and similar. As already mentioned, because of the methodology implemented in the compilation of the Thesaurus, multi-word units as headwords are often unusual (and are as such less queried), while the votes they receive are rather sporadic.

In the ideal scenario, the entire Thesaurus would be evaluated through user votes. A quick estimate reveals that this is not impossible: with cca. 184,000 headword-synonym pairs to be evaluated in the entire dictionary (not counting inverted pairs and user-added synonyms) and presupposing that four votes per pair would be enough to

distinguish the worst candidates from the best, a total of 736,000 thousand votes would have to be collected. So far, the Thesaurus has been accessed from approximately 38,000 different IP-addresses. If a quarter of these would vote on synonyms (taking into account that on average, a vote on a synonym pair takes 8 seconds), each individual would have to spend approximately 13 minutes (not necessarily all at once) voting on synonyms.

However, for this to be successful, user motivation is key. The Thesaurus already features Tasks of the Day (shown in Figure 5), a special subsection on the homepage that invites users to evaluate up to four headwords in which the synonyms have not yet received any votes. Additional features in a similar vein are planned.

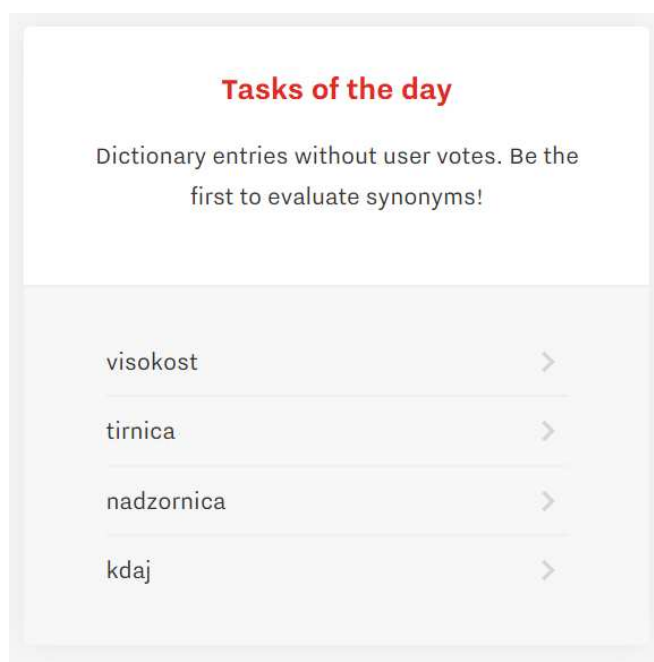


Figure 5: Tasks of the Day in the Thesaurus.

Furthermore, the Thesaurus logs show that only 48% of the Thesaurus has been queried so far (cca. 51,000 headwords out of 105,473), which means that half of the content has not even been seen by users yet. This suggests that new manners of presenting the content of the dictionary to the users are required, such as targeted crowdsourcing campaigns or gamification. This is part of our future work and we discuss it in more detail in the conclusion.

5. Conclusion and future work

In the paper, we presented two crowdsourcing activities aimed at cleaning up the first version of the Thesaurus of Modern Slovene. The targeted crowdsourcing campaign has processed 18,365 multi-word synonym candidates and resulted in a dataset of 5,882 negatively evaluated candidates (3,509 with complete agreement), 9,027 positively

evaluated candidates (6,367 with complete agreement), and 3,456 ambiguous candidates, which means the experiment cleaned up 81% of the multi-word synonym candidates included in the task. The votes collected through the Thesaurus interface resulted in a dataset of 17,904 synonym candidates with votes (5% of all synonyms candidates in the Thesaurus). Both datasets will be taken into account during the next Thesaurus upgrade in order to reduce the amount of irrelevant noise.

The results of both crowdsourcing activities have also produced a number of interesting findings. The results of crowdsourcing are similar to the results of the expert evaluation of the Thesaurus, which indicates that the method is indeed applicable to language resource compilation. The evaluation has also shed light on measures to be taken in the future: because of its automatic origin, the Thesaurus also consists of some unusual headwords that would otherwise not have been included (e.g. *zbirati se v bazen* ‘to gather in a pool’ – *zbirati se v tolmun* ‘to gather in a pond’). This also raises the question of the degree to which the Thesaurus is limited in terms of data, as it contains only the phrases used as translation equivalents for English headwords in The Oxford®-DZS Comprehensive English-Slovenian Dictionary. This calls for a more thorough analysis of the relevance of not only the synonym candidates, but the headwords as well, especially the ones that have not yet been queried by users (i.e. the ones not in the 48% of the Thesaurus queried so far). In the interface this issue might be addressed by providing users with the option to downvote headwords as well.

On the other hand, the voting system has provided votes for only 5% of the Thesaurus so far, which indicates that lexicographers would require more features to involve users in the cleanup. Even the most motivated users currently have no way of systematically contributing toward the improvement of the Thesaurus (other than by solving tasks of the day or by searching for random headwords). The evaluation presented in this paper has shown that while an evaluation of the entire Thesaurus is possible, user motivation is key. The inclusion of the Thesaurus data in a gamified environment would be an even more efficient and expedient manner of crowdsourcing user votes. To tackle this issue a mobile game is already in development as part of our future work. More targeted and short-term crowdsourcing campaigns extended to a larger crowd and aimed at solving specific problems would also be beneficial, particularly in combination with future updates that will add new, automatically extracted synonyms and evaluate user synonyms added through the interface. And last but not least, the annotated data we only briefly analysed in the paper offers great potential for linguistic studies on synonymy and the development of a concept of synonymy based on usefulness, as defined by the collective intuition of the language community.

6. Acknowledgements

The research presented in this paper was conducted within the project titled *The Thesaurus of Modern Slovene: By the Community for the Community* (2018–2019), which is financially supported by the Ministry of Culture of the Republic of Slovenia.

The authors acknowledge the financial support from the Slovenian Research Agency (research core funding No. P6-0411, Language Resources and Technologies for Slovene). The research was conducted within the framework of the CA160105 eNetCollect COST Action. The authors would also like to thank all the users of the Thesaurus of Modern Slovene and the crowdsourcers who participated in the synonym evaluation campaign: Haris Agović, Ajda Diaci, Zoran Fijavž, Barbara Gorišek, Tajda Liplin Šerbetar, Angelika Markič, and Jana Šter.

7. References

- Abel, A. & Meyer, C. (2013). The dynamics outside the paper: user contributions to online dictionaries. In I. Kosem et al. (eds.) *Proceedings of eLex 2013*, pp. 179–194.
- Arhar Holdt, Š., Čibej, J., Dobrovoljc, K., Gantar, A., Gorjanc, V., Klemenc, B., Kosem, I., Krek, S., Laskowski, C. & Robnik Šikonja, M. (2018). Thesaurus of Modern Slovene: By the Community for the Community. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (eds.) *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*. ISBN 978-961-06-0097-8). Ljubljana: Znanstvena založba Filozofske fakultete. 2018, pp. 401-410.
<https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/118/211/3000-1>
- Čibej, J., Fišer, D. & Kosem, I. (2015). The role of crowdsourcing in lexicography. In I. Kosem et al. (eds.) *Electronic lexicography in the 21st century: linking lexical data in the digital age: proceedings of eLex 2015 Conference, 11-13 August 2015, Herstmonceux Castle, United Kingdom*. Ljubljana: Trojina, Institute for Applied Slovene Studies; Brighton: Lexical Computing. 2015, pp. 70-83.
https://elex.link/elex2015/proceedings/eLex_2015_05_Cibej+Fiser+Kosem.pdf
- Fišer, D., Tavčar, D. & Erjavec, T. (2014). sloWCrowd: A crowdsourcing tool for lexicographic tasks. *Proceedings of LREC 2014*.
- Fort, K., Guillaume, B. & Chastant, H. (2014). *Creating ZombiLingo, a Game With A Purpose for dependency syntax annotation*. Proceedings of the Gamification for Information Retrieval (GamifIR'14) Workshop, Amsterdam, The Netherlands, April 2014.
- Kosem, I., Krek, S., Gantar, P., Arhar Holdt, Š., Čibej, J. & Laskowski, C. (2018). Collocations Dictionary of Modern Slovene. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (eds.) *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*. Ljubljana, Ljubljana University Press, Faculty of Arts, pp. 989-997. <https://e-knjige.ff.unilj.si/znanstvena-zalozba/catalog/view/118/211/2939-1.pdf> (25 August 2018).
- Krek, S, Laskowski, C., Robnik-Šikonja, M., Kosem, I., Arhar Holdt, Š., Gantar, P., Čibej, J., Gorjanc, V., Klemenc, B. & Dobrovoljc, K. (2018b). *Thesaurus of Modern Slovene 1.0*. <http://hdl.handle.net/11356/1166> (3 June 2019)

- Krek, S., Laskowski, C., Robnik-Šikonja, M., Kosem, I., Arhar Holdt, Š., Gantar, P., Čibej, J., Gorjanc, V., Klemenc, B. & Dobrovoljc, K. (2018a). *Sopomenke 1.0: Thesaurus of Modern Slovene*, <https://viri.cjvt.si/sopomenke> (17 June 2019).
- Krek, S., Laskowski, C. & Robnik-Šikonja, M. (2017). From translation equivalents to synonyms: creation of a Slovene thesaurus using word co-occurrence network analysis. In I. Kosem et al. (eds.) *Proceedings of eLex 2017: Lexicography from Scratch, 19-21 September 2017, Leiden, Netherlands*, pp. 93-107. <https://elex.link/elex2017/wp-content/uploads/2017/09/paper05.pdf>.
- Lew, R. (2014). *User-generated content (UGC) in online English dictionaries*. OPAL - Online publizierte Arbeiten zur Linguistik 2014.4: 8-26. https://repozytorium.amu.edu.pl/bitstream/10593/5011/1/Lew_GAL.pdf
- Snow, R., O'Connor, B., Jurafsky, D. & Ng., A. Y. (2008). Cheap and Fast – But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, Honolulu, October 2008, pp. 254–263. <https://www.aclweb.org/anthology/D08-1027>.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



Communities of Related Terms in a Karst Terminology Co-occurrence Network

Dragana Miljkovic¹, Jan Kralj¹, Uroš Stepišnik², Senja Pollak^{1,3}

¹ Jožef Stefan Institute, Ljubljana, Slovenia

² University of Ljubljana, Ljubljana, Slovenia

³ University of Edinburgh, UK

E-mail: dragana.miljkovic@ijs.si, jan.kralj@ijs.si, uros.stepisnik@gmail.com, senja.pollak@ijs.si

Abstract

Karst science is an attractive field of interdisciplinary research with rich terminology. This study was performed as part of a project aiming at developing novel approaches to terminology extraction and visualization, in line with the understanding of knowledge, as represented in texts, as conceptually dynamic and linguistically varied. The aim of this paper is to investigate how powerful graph-based methods can be used for visualizing and analysing domain terminology. In order to detect communities in karst terminology, we analyse the frequently co-occurring karst terms in a scientific corpus of karstologic literature. The most frequent co-occurrence pairs, which included ten or more co-occurrences within the whole corpus, are delivered as input to the Louvain community detection algorithm and visualized as a domain graph. The resulting data was evaluated by domain experts who found that the detected term groups are meaningful and correspond to different types of karst phenomena. The results are further discussed in relation to more standard topic modelling approaches, using Latent Dirichlet Allocation and Non-negative Matrix Factorization algorithms.

Keywords: karstology; co-occurrence network; community detection algorithm; network visualization; topic modelling

1. Introduction

Karst science, or karstology, is a well-researched discipline with rich terminology, consisting of many expressions referring to regionally specific phenomena. Contemporary research of the topography that is referred to as a ‘karst geomorphologic system’ or simply ‘karst’ includes numerous scientific disciplines that study the karst environments worldwide; however, the earliest research on karst primarily regards Classical Karst, which is located in western Slovenia. Consequently, karstologists use many local Slovenian scientific terms and toponyms for typical geomorphological karst structures not only when writing in Slovene, but also in English and other languages. In this paper, we focus on karst texts in English.

This study was undertaken as part of the TermFrame project¹, which is based on contemporary findings in the field of terminology and cognitive linguistics, and aims to

¹ TermFrame project web site: <http://termframe.ff.uni-lj.si/>

develop novel methods that can be utilized in the field of terminology research. The focus of these novel methods is on corpus-based approaches to extraction and visualization of terminological knowledge, including text and graph mining and advanced data representation techniques.

Recent attempts in terminological science understand knowledge, as represented in texts, as conceptually dynamic and linguistically varied (Cabr , 1999; Temmerman, 2000; Kageura, 2002). Research advances in cognition have contributed to the Frame-Based Terminology (Faber, 2012; Faber, et al., 2006), which focuses on representing dynamic knowledge and investigating cultural elements in cognitive structures (Rodr guez Redondo, 2004; Grygiel, 2017), while projects such as EcoLexicon² attempt to visually represent concept networks. While a limited number of studies have used graph-based approaches in the fields of terminology and lexicography (Meyer & Eppinger, 2018; Krek et al., 2017) and for language comparison ( skrlj & Pollak, 2019), we believe that these methods are still to be fully explored, as they present the potential for novel research of specialized knowledge, as well as for new possibilities of knowledge representation that can be inspiring to contemporary lexicography. We believe that the graph-based method for exploring term co-occurrences can contribute to the needs of frame-based terminology, aiming at facilitating user knowledge acquisition through different types of multimodal and contextualized information (Gil-Berrozpe et al., 2017). This type of graph-based tool also has potential for future data representation in the field of e-lexicography (Granger, 2012), where multimodal data and hybridization between different types of language resources (e.g., dictionaries, encyclopaedias, term banks, lexical databases, translation tools) are commonly observed.

The focus of the present work in the scope of the above-mentioned project is to apply graph-based methods to the terminology of karst research. This has motivated us to explore co-occurrences of the specific karstology terms and visualize the results. Another motivation for the visualization of results is that domain experts are often able to interpret information faster when viewing graphs as opposed to tables (Brewer et al., 2012). More generally, as evident by the rising field of digital humanities, digital content, tools, and methods are transforming the entire field of humanities, changing the paradigms of understanding, asking new research questions and creating new knowledge (Hughes et al., 2015; Hughes, 2012). The work complements the results in karst terminology research presented in Vintar et al. (2019), where frame-based annotation of karst definitions is introduced, and in Pollak et al. (2019), where the authors present the results of term, definition, and triplet extraction from karst literature.

This paper is structured as follows: after presenting the background technologies and related work in Section 2, Section 3 introduces our method, which is based on

² <http://ecolexicon.ugr.es/en/index.htm>

community detection of terms extracted from a karstology corpus and their visualization in the form of a network; along with Section 4, the two sections represent the main contribution of the paper. In Section 5, we discuss the results in relation to a more standard topic modelling methods approach, and we conclude this paper in Section 6.

2. Background technologies and related work

This section presents a brief overview of the state-of-the-art of the fields related to our study methods, including co-occurrence and visualization, community detection algorithms and topic modelling.

2.1 Co-occurrence approach and visualization

Scientific literature in different fields can be explored through a search for the co-occurrences of domain-specific terms and their frequencies. A co-occurrence of two terms means that the terms coexist in the text within a certain window. The idea behind detecting co-occurrences of terms is that closely related terms will appear together more frequently. Moreover, co-occurrences can reveal hidden patterns and interesting features in the texts that are being analysed. For example, the co-occurrence analysis might detect spam messages (Krestel & Chen, 2008) or find meaningful knowledge from biological literature in a systematic and automated way (Al-Aamri et al., 2017). Co-occurrence is also used widely in text classification (Figueiredo et al., 2011) and categorization (Luo & Zincir-Heywood, 2004).

There is a difference between first-order and second-order co-occurrence approaches. For the first-order co-occurrence, one would simply count how many occurrences of one token there are within a specified distance of the particular occurrence of another token and build a vector presentation of the results. A second-order co-occurrence vector would represent some aggregation over the token representations, and in the simplest case this is a sum (Maldonado & Emms, 2012).

Representation of co-occurrence pairs in the form of a network is a common way to aid the domain experts with exploration of research results. Such representations can be used for various purposes, such as word sense disambiguation, which represents a challenge in natural language processing field (Duque et al., 2018). Li et al. (2018) report the discovery of new information in the biomedical domain based on the analysis of the structural characteristics of the co-occurrence network. Additionally, co-occurrence networks are increasingly used when analysing users' behaviour on social media (Correia et al., 2016).

In the field of lexicography, co-occurrence networks have been used with the aim of building a new Slovene thesaurus from data available in a comprehensive English–Slovene dictionary (Krek et al., 2017).

2.2 Community detection algorithms

When co-occurrence networks become too large and complex, their visual inspection becomes difficult. One way to explore complex networks more easily is to use community detection algorithms.

Community detection algorithms can be split into several classes based on the underlying idea that guides the algorithms. It must be noted that a strict split between the different methods is impossible, as these methods are not developed in isolation. For example, many methods that are not strictly classified as modularity-based algorithms still use the concept of modularity in one of their steps.

Divisive algorithms are algorithms that find the community structure of a network by iteratively removing edges from the network. The most widely used algorithm among divisive algorithms is the Girvan Newman algorithm (Girvan & Newman, 2002), which removes the network edges with the largest centrality measure. The reasoning behind this is that edges which are more central to a graph are the edges most likely to cross communities. An alternative algorithm is the Radicchi algorithm, which calculates the edge-clustering coefficient of edges in order to determine which edges must be removed. Here, the reasoning is that edges between communities belong to fewer cycles than edges within communities.

Modularity-based algorithms form the majority of community detection algorithms. While, as mentioned above, the concept of modularity (Newman & Girvan, 2004) is used in almost all algorithms to an extent (especially when attempting to determine the best clustering from a hierarchical clustering of nodes), the algorithms in this class use modularity more centrally than other algorithms. The most prominent modularity-based methods are the Louvain algorithm (Blondel et al., 2008) and the Newman greedy algorithm (Newman & Girvan, 2004). Other methods include variations of the greedy algorithm (Wakita & Tsurumi, 2007), simulated annealing (Guimerà & Amaral, 2005), spectral optimization of modularity via a modularity matrix (Newman, 2006a; Newman, 2006b) or via the graph adjacency matrix (White & Smyth, 2005), and deterministic optimization approaches (Duch & Arenas, 2005).

Spectral algorithms find communities in networks by analysing the eigenvectors of matrices derived from the network. The community structure is extracted either from the eigenvectors of the Laplacian matrix of the network (Donetti & Muñoz, 2004) or from the stochastic matrix of the network (Capocci et al., 2005). In both cases, the idea behind the algorithms is that eigenvectors extracted from the network will have similar values on indices that belong to network vertices in the same community. First, a computation of several eigenvectors belonging to the largest eigenvalues is performed. The resulting eigenvectors form a set of coordinates of points, each belonging to one network vertex, with clustering of these points corresponding to community detection of network vertices.

Another important community detection algorithm is the InfoMap algorithm (Rosvall et al., 2009). This is based on the idea of minimal description length of the walks performed by a random walker traversing the network. The communities in InfoMap are determined by constructing so-called codebooks, which are used to describe walks on the network – corresponding to communities in the network, codebooks yield on average shorter average descriptions of walks. Finally, in the most recent rapid development of network embedding algorithms, some researchers have begun using embedding-based methods for network community detection (Li et al., 2018).

2.3 Topic modelling

In this section, we cover topic modelling, i.e. methods used for discovering various topics that appear in a collection of documents. Topic modelling methods are well-established in the field of text modelling, and can be considered as alternative approaches to co-occurrence community detection. Methods for topic modelling can rely on linear algebra, such as Vector Space Model (VSM) (Becker & Kuropka, 2003) or Matrix Factorization (NMF) (Paatero & Tapper, 1994), while others are based upon statistical distributions, for example Latent Dirichlet Allocation (LDA) (Blei et al., 2003). When using both NMF and LDA for topic modelling, two matrices are constructed from the document-term matrix: the document-topic and topic-term matrices. The topics are derived from the contents of the documents, and the topic-document matrix describes data clusters of related documents. LDA usually performs well when it comes to identifying coherent topics, whereas NMF provides incoherent ones (Stevens et al., 2012). While VSM is based on a similar principle as NMF, it has significant limitations when processing long documents as they have poor similarity values. Because the corpus analysed for the purposes of this paper includes both short and long documents (doctoral dissertations, dictionaries, etc.), this specific method was excluded from consideration.

The aim of this paper is to analyse the communities in karst terminology by analysing the co-occurrence network of frequently co-occurring karst terms in the scientific corpus of karst literature. We defined a co-occurrence of terms as their coexistence in the same sentence, while in order to qualify as frequently co-occurring, a term pair had to occur at least ten times over the span of the entire corpus. We decided to start inspecting karst corpus gathered for the purpose of the TermFrame project with basic first-order co-occurrence vectors and present the results of co-occurrence terms in the form of community network, as it is easily comprehended by domain experts. For our research, we used three leading algorithms in the community detection field: Label propagation, Louvain, and InfoMap. The InfoMap and Label propagation algorithms did not yield meaningful results: both identified one large community and several singletons. For this reason, the Methodology, Results, and Discussion sections all focus exclusively on the results obtained using the Louvain algorithm. We also discuss the results from the community detection experiment in relation to two topic modelling approaches, LDA

and NMF, while the exploration of second-order co-occurrence approaches will be explored in future work.

3. Methodology

First, we tokenized and lemmatized our collection of scientific literature and the corresponding term list. Next, first order co-occurrences of pre-specified terms were identified within the corpus. After this, the Louvain community detection algorithm was used to find the communities of co-occurrence pairs. The schematic of the methodology used in this study is shown in Figure 1, with each step further explained below.

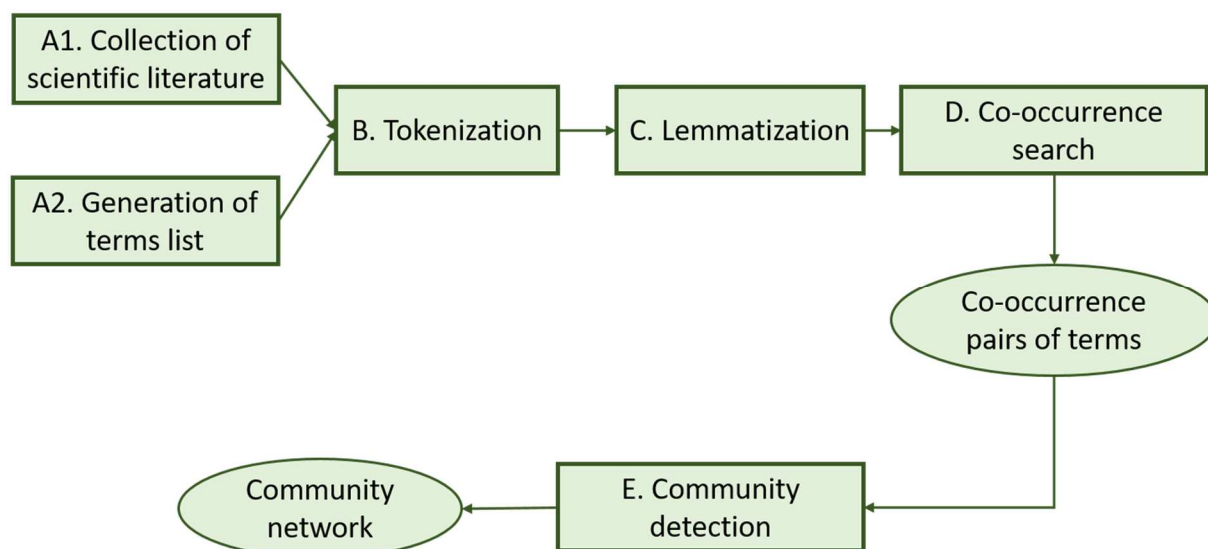


Figure 1: The schematic of the methodology.

A1. Collection of scientific literature represents the compilation of 25 scientific karstology texts, including papers, doctoral dissertations, and the glossary of cave and karst terminology. This corpus was compiled as part of the TermFrame project and is an extended version of earlier work (Vintar & Grčić Simeunović, 2016).³

A2. Generation of terms list was performed as a two-phase process. First, relevant terms were automatically extracted from the TermFrame corpus using the LUIZ-CF term extractor (Pollak et al., 2012), which is a variant of LUIZ (Vintar, 2010) refined with scoring and ranking functions. The terms were validated by the domain expert and were used to compile a term list along with the previously acquired terms from the QUIKK termbase⁴. This process of term extraction and evaluation is presented in more detail in Pollak et al. (2019).

³ We used the corpus version v1.0.

⁴ <http://islovar.ff.uni-lj.si/karst>

B. Tokenization was performed using the NLTK Tokenizer for Python.

C. Lemmatization was performed using the Lemmagen tool (Juršič et al., 2010).

D. Co-occurrence search was performed automatically by the Python script, which stores in a separate file the co-occurring term pairs and the number of their co-occurrences in the whole TermFrame corpus.

E. Community detection was performed using the Louvain algorithm (Blondel et al., 2008), which works by decreasing the modularity of the network, a function that measures the density of links inside communities compared to links between communities. The modularity of a network is defined as:

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j)$$

where A_{ij} denotes the weight of the edge between nodes i and j (in our case, the number of co-occurrences), k_i denotes the degree (sum of all adjacent edge weights) of node i , and m denotes the total sum of weights in the network. The term c_i denotes the community to which node i is assigned, meaning the sum above runs over all pairs of i, j where i and j belong to the same community.

4. Results and discussion

For the purposes of this research, we compiled a list of 452 karst terms drawing from a corpus of karstology texts which contained 108,769 sentences in total. Both the list and the literature were tokenized and lemmatized prior to the co-occurrence search, which yielded a list of 10,990 unique co-occurrence pairs using 426 unique lemmatized terms, as well as the data regarding co-occurrence frequency.

The initially obtained co-occurrence pairs would result in a complex network that would be difficult to represent in a comprehensible manner. To simplify the visualization, co-occurrence pairs with frequencies of ten or less were removed from the subsequent analysis. This left us with 1,247 co-occurrence pairs (see Table 1).

	Initial co-occurrence list	Filtered co-occurrence list
Number of co-occurrence pairs	10,990	1,247
Number of unique terms	426	309

Table 1: The summary of the initially obtained co-occurrence list and the filtered version, which contains only the co-occurrence pairs with frequencies of 10 or more.

The 20 most frequent co-occurrence pairs extracted from the karst corpus are listed in Table 2.

ID	Term 1	Term 2	Frequency of appearing	ID	Term 1	Term 2	Frequency of appearing
1	cave	karst	1688	11	limestone	dolomite	368
2	cave	passage	1482	12	cave	karren	349
3	cave	limestone	739	13	solution	karren	319
4	cave	spring	735	14	karren	limestone	311
5	cave	speleothem	664	15	cave	pit	288
6	cave system	cave	597	16	limestone	marble	282
7	cave	gypsum	512	17	karst	spring	270
8	cave	calcite	468	18	karst	term	261
9	karst	limestone	464	19	cave	canyon	261
10	calcite crust	cave	381	20	karst	doline	259

Table 2: The list of common co-occurrence pairs extracted from the karst corpus sorted from most to least frequent.

The filtered co-occurrence pairs served as input for the Louvain algorithm for community detection. Starting with each node in its own community, the algorithm iteratively works in two stages. In the first stage, it searches for the optimum pairs or groups of communities to merge into a larger community and thus increase the modularity of the partition. In the second stage, the algorithm reduces the network to a coarser network based on the discovered communities. The two-stage procedure is then repeated until no increases in modularity can be made. This results in a hierarchy of network node clusters, which can then be cut at any level to produce a clustering of the network nodes. In our case, the algorithm resulted in a three-layer hierarchy. The top level consisted of only two communities and the bottom level of single-node communities. The middle layer was the only layer containing non-trivial information about the structure of the co-occurrence network, and it was therefore subject to further analysis.

The middle layer of the hierarchy, discovered by the Louvain algorithm, consisted of eight communities. Next, we visualized the network using the Barnes-Hut approximation of the force-directed layout to calculate optimal node positions (Jacomy et al., 2014). The discovered communities were then displayed on the network visualization by colouring nodes corresponding to the communities they belong to (see Figure 2).

The karst domain experts analysed the resulting network and found the network visualization particularly interesting, as the communities (listed below) were found to correspond to different types of karst phenomena.

- Community 0: Exokarst landforms ('kamenitza', 'grike', 'stone forest'), which are the result of direct effects of dissolution of bedrock exposed on the surface;
- Community 1: Subsurface landforms, speleogenetic features, and cave environments (e.g. 'passage', 'flowstone deposit', 'cave system'). This community comprises all types of underground voids typical for karst environments regardless of their morphogenesis, including characteristic mechanical and chemical fills within.
- Community 2: Surface karst landforms and environments (e.g. 'uvala', 'doline', 'karst terrain') which are a product of surface and subsurface karst processes, materialising as relief forms or terrain types.
- Community 3: Karst hydrologic processes, environments, and methods (e.g. 'karst recharge', 'groundwater basin', 'tracer test') incorporate all karst aquifer types, the processes within them, and methods concerning their research.
- Community 4: Karst geology representing terms related to karst lithology (e.g. dolomite), minerals ('calcite') and processes affecting them (e.g. 'dissolution')
- Community 5: Includes only two terms (karrenfield, phreatic-cave), which is not enough to define the topic field.
- Community 6 includes only two terms ('turbulent flow', 'laminar flow'), which is not enough to define the topic field.
- Community 7 includes only two terms ('vadose zone', 'phreatic zone'), which is not enough to define the topic field.

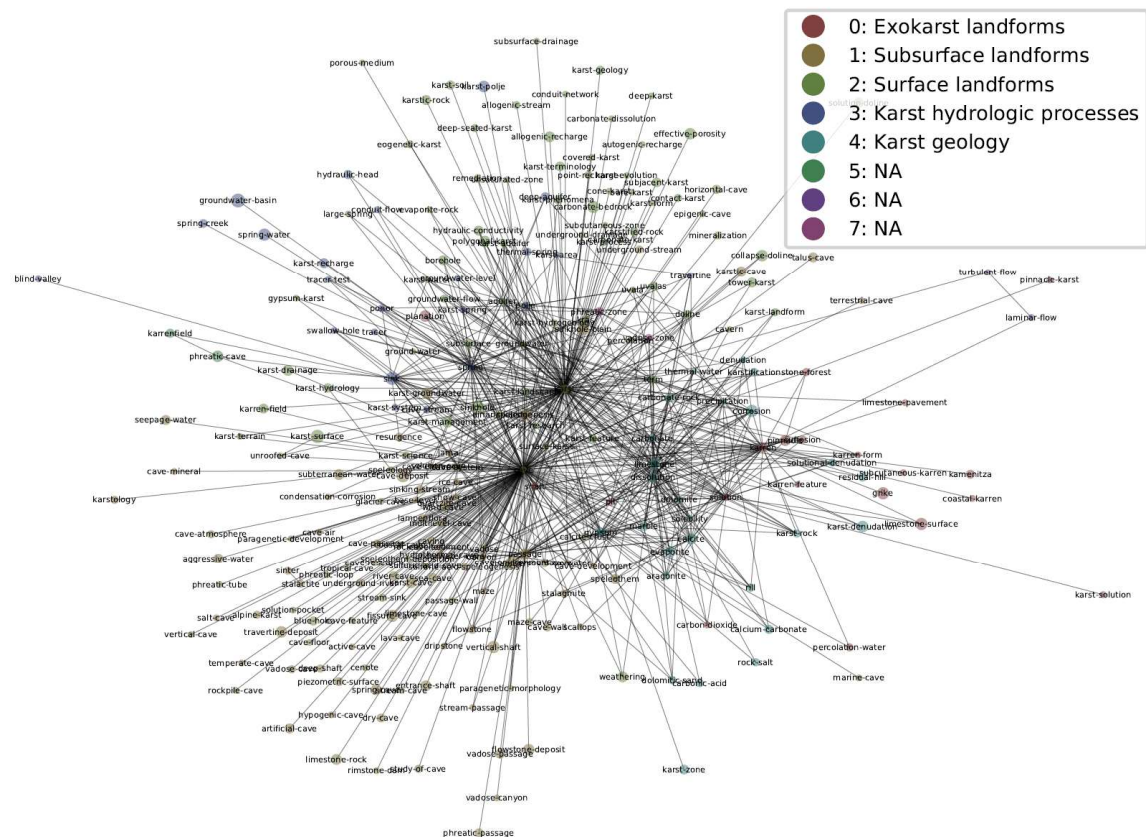


Figure 2: The co-occurrence network, visualized using a force-directed layout, showing the communities discovered within the network. The colours of the nodes correspond to the communities the nodes belong to.

5. Topic modelling experiments

As graph-based modelling is a relatively novel field for harvesting knowledge from specialized corpora, this section discusses our results with respect to more standard topic modelling approaches. For the purpose of this research, we used LDA and NMF algorithms, implemented within a Scikit-learn Python module. The algorithms searched through the complete corpus of 25 documents (described above) containing 108 769 lemmatized sentences, presenting the domain expert with the 25 most important words for each topic. The domain expert subsequently evaluated whether the derived topic words adequately represent specific subfields of karstology. In Table 3, we list the topics and the topic words identified by the NMF and LDA algorithms, which were estimated as meaningful groups by the domain expert. To enable further comparison of results with the community detection experiment, the number of topics was set to eight for both algorithms.

NMF	Topic 0: SPELEOLOGY	cave passage entrance long know study world km large deep map exploration bat sediment mammoth example explore stream important contain river site animal speleothem state
	Topic 1: KARST HYDROLOGY	water flow spring table level aquifer zone high groundwater discharge surface underground stream sea conduit phreatic supply resource fresh mix air rise sink temperature time
	Topic 5: KARST GEOMORPHOLOGY	rock form figure surface limestone large develop small carbonate karren passage process area 10 soil solution high occur dissolution doline lower feature cover sediment deposit
	Topic 6: SPELEOBIOLOGY	species family subterranean know troglobitic habitat include genera number genus group population troglomorphic bat fauna large occur troglobite terrestrial aquatic marine represent small order environment
	Topic 7: GENERAL METHODOLOGY (KARST)	use method data term model technique land date tracer place study time site widely approach human dye analysis test trace map measure determine source work
LDA	Topic 0: SPELEOLOGY	cave sediment passage type channel wall 20 place contain small like 12 width speleothem vertical significant 100 2001 possible figure direction rillenkarren floor stream scale
	Topic 2: KARST GEOLOGY	rock large limestone carbonate cover deposit upper surface gypsum forest dissolution area stone protect calcite earth line layer bed joint various material analysis salt fracture
	Topic 5: KARST HYDROLOGY	water flow spring zone soil deep high aquifer karst surface occur groundwater slope natural condition table value depression low erosion increase result point temperature climate

Table 3: Topic modelling results with Non-negative Matrix Factorization (NMF) and Latent Dirichlet distribution (LDA) applied to karst literature

From a karstologic point of view, the following topics extracted by means of the NMF method describe various aspect of karstology, i.e. different scientific fields regarding karst research:

- Topic 0: Speleology incorporates topic words that are directly referring to cave processes, cave-related landforms, or toponyms regarding to research of caves (i.e. speleology).
- Topic 1: Karst hydrology topic words comprise a variety of terms describing karst aquifers and their study.
- Topic 5: Karst geomorphology topic words correspond to a variety of surface landforms and processes, as well as words labelling their properties.
- Topic 6: Speleobiology topic words are related to cave biota and habitats.
- Topic 7: General karst methodology topic words incorporate a combination of various terms describing research methods from different karst research fields.

LDA identified only three topic groups meaningful to the domain expert, compared to the five identified by NMF:

- Topic 0: Speleology (see NMF Topic 0).
- Topic 2: Karst geology words regarding karst rocks, minerals, and processes concerning them.
- Topic 5: Karst hydrology (see NMF Topic 1).

NMF and community detection experiments have some overlaps in results, such as karst hydrologic processes and karst surface landforms and environments, as well as a partial topic overlap with terms related to speleology.

The results of our proposed community detection methodology have identified several specific topics as evaluated by the expert; however, it can be hard to determine to which extent this is to be attributed to term pre-selection, the community detection algorithm, or to the visualization of results. A detailed study of the role of each component is beyond the scope of this paper, but we believe that graph-based methods coupled with visualization offer great opportunities for investigating terminology as dynamic systems.

An overview of the number of meaningful communities identified by the proposed community detection approach and topic modelling methods (NMF and LDA) is presented in Table 4. All of the topics listed in this paper were manually evaluated by a domain expert. Community detection differs from the topic modelling approaches in that it takes pre-specified terms as input, while topic modelling approaches take as

input all words in the corpus documents. For this reason, a deeper quantitative comparison between these approaches is not feasible.

Number of meaningful topics		
Community detection algorithm	Topic modelling (LDA)	Topic modelling (NMF)
5	3	5

Table 4: Quantitative overview of the discovered topics with topic modelling and graph-based methods.

6. Conclusions and future work

In this work, we used a list of terms extracted from karst scientific literature and then performed a network analysis of karst terminology, wherein the network was constructed from co-occurring karst terms. The community detection algorithms described in this paper grouped specialized terms into semantically related topics, which were also visually presented as coloured nodes in the graphs. In addition, we approached the same corpus from the viewpoint of more standard topic modelling techniques, using LDA and NMF as our main tools.

In future work we plan to include the exploration of second-order co-occurrences, embedding-based topic modelling, and combining graph-based term and community detection methods. In addition, we consider performing a systematic comparison of graph-based community detection and topic modelling approaches, as well as evaluating if term extraction can contribute to these approaches.

Furthermore, we plan to use network representation in the form of triplets {subject, predicate, object}, which can also be a source of identifying novel semantic relations. Within the scope of the TermFrame project, a multi-layer semantic annotation has been performed and the most frequent conceptual frames for specific semantic categories explored. By combining information from manual annotations and the proposed network-based techniques, new knowledge about conceptual frames, semantic relations, and topics could be observed. The potential of graph-based topological analysis lies also in its power to explore structural information, which could reveal potential language and culture-driven differences if, for example, applied to larger comparable corpora of karst texts in different languages.

7. Acknowledgements

This work was financed by Slovenian Research Agency grants J6-9372 (Terminology and Knowledge Frames across Languages - TermFrame) and P2-0103 (Knowledge Technologies). This paper is supported by European Union’s Horizon 2020 research

and innovation programme under grant agreement No 825153, project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media). The results of this work reflects only the authors' views, and the European Commission is not responsible for any use that may be made of the information it contains.

8. References

- Al-Aamri, A., Taha, K., Al-Hammadi, Y., Maalouf, M., & Homouz, D. (2017). Constructing Genetic Networks using Biomedical Literature and Rare Event Classification. *Scientific Reports*, 7(1), pp. 2045-2322.
- Becker, J., & Kuropka, D. (2003). Topic-based vector space model. *Proceedings of the 6th International Conference on Business Information Systems*. Colorado Springs, USA.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, pp. 993-1022.
- Blondel, V., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, P10008.
- Brewer, N. T., Gilkey, M. B., Lillie, S. E., Hesse, B. W., & Sheridan, S. L. (2012). Tables or Bar Graphs? Presenting Test Results in Electronic Medical Records. *Medical Decision Making*, 32(4), pp. 545-553.
- Cabré, M. T. (1999). *Terminology: Theory, methods, applications*. Amsterdam; Philadelphia: J. Benjamins Publishing Company.
- Capocci, A., Servedio, V. D., Caldarelli, G., & Colaioni, F. (2005). Detecting communities in large networks. *Physica A: Statistical Mechanics and its Applications*, 352(2), pp. 669-676.
- Correia, R. B., Li, L., & Rocha, L. M. (2016). Monitoring potential drug interactions and reactions via network analysis of Instagram user timeliness. *Pacific Symposium on Biocomputing*. 21. Kohala Coast, Hawaii, USA: World Scientific, pp. 492-503.
- Donetti, L., & Muñoz, M. A. (2004). Detecting network communities: a new systematic and efficient algorithm. *Journal of Statistical Mechanics: Theory and Experiment*, 2004(10), P10012.
- Duch, J., & Arenas, A. (2005). Community detection in complex networks using extremal optimization. *Physical review E*, 72(2), pp. 027104.
- Duque, A., Stevenson, M., Martinez-Romo, J., & Araujo, L. (2018). Co-occurrence graphs for word sense disambiguation in the biomedical domain. *Artificial Intelligence in Medicine*, 87, pp. 9-19.
- Eronen, L., & Toivonen, H. (2012). Biomine: predicting links between biological entities using network models of heterogeneous databases. *BMC Bioinformatics*, 13(1), pp. 119.
- Faber, P. (ed.). (2012). *A Cognitive Linguistics View of Terminology and Specialized Language*. Berlin, Boston: De Gruyter Mouton.

- Faber, P., Montero Martínez, S., Castro Prieto, M. R., Senso Ruiz, J., Prieto Velasco, J. A., León Arauz, P. & Vega Expósito, M. (2006). Process Oriented Terminology Management in the Domain of Coastal Engineering. *Terminology*, 12(2), pp. 136.
- Figueiredo, F., Rocha, L., Couto, T., Salles, T., André Gonçalves, M., & Meira Jr, W. (2011). Word co-occurrence features for text classification. *Information Systems*, 36(5), pp. 843-858.
- Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12), pp. 7821-7826.
- Grygiel, M. (2017). *Cognitive Approaches To Specialist Languages*. (M. Grygiel, Ed.) Cambridge Scholars Publishing.
- Guimerà, R. & Amaral, L. A. (2005). Functional cartography of complex metabolic networks. *Nature*, 433(7028), pp. 895-900.
- Hughes, L. M. (2012). Using ICT methods and tools in arts and humanities research. In L. M. Hughes (ed.) *Digital Collections: Use, Value and Impact*. London, UK: Facet Publishing, pp. 123-134.
- Hughes, L., Constantopoulos, P., & Dallas, C. (2015). Digital Methods in the Humanities: Understanding and Describing their Use across the Disciplines. In S. Schreibman, R. Siemens, & J. Unsworth (eds.) *A New Companion to Digital Humanities*. John Wiley & Sons, pp. 150-170.
- Jacomy, M., Venturini, T., Heymann, S., & Bastian, M. (2014). ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLoS One*, 9(6), e98679.
- Juršič, M., Mozetič, I., Erjavec, T., & Lavrač, N. (2010). LemmaGen: multilingual lemmatisation with induced Ripple-Down rules. *Journal of Universal Computer Science*, 16(9), pp. 1190-1214.
- Kageura, K. (2002). *The Dynamics of Terminology: A Descriptive Theory of Term Formation and Terminological Growth*. Amsterdam, Philadelphia: John Benjamins Publishing Company.
- Krek, S., Laskowski, C., & Robnik-Šikonja, M. (2017). From Translation Equivalents to Synonyms: Creation of a Slovene Thesaurus Using word co-occurrence Network Analysis. In I. Kosem et al. (eds.) *Electronic lexicography in the 21st century: Proceedings of eLex 2017 conference*. Leiden, the Netherlands: Lexical Computing CZ s.r.o., Brno, Czech Republic, pp. 93-109.
- Krestel, R., & Chen, L. (2008). Using co-occurrence of tags and resources to identify spammers. *ECML PKDD*. Antwerp: Springer, pp. 38-46.
- Li, T., Bai, J., Yang, X., Liu, Q., & Chen, Y. (2018). Co-Occurrence Network of High-Frequency Words in the Bioinformatics Literature: Structural Characteristics and Evolution. *Applied Sciences*, 8, 1994.
- Li, Y., Sha, C., Huang, X., & Zhang, Y. (2018). Community detection in attributed graphs: an embedding approach. *Thirty-Second AAAI Conference on Artificial Intelligence*. New Orleans, Louisiana, USA: AAAI .

- Liu, Y., McInnes, B. T., Pedersen, T., Melton-Meaux, G., & Pakhomov, S. V. (2012). Semantic relatedness study using second order co-occurrence vectors computed from biomedical corpora, UMLS and WordNet. *2nd ACM SIGHIT International Health Informatics Symposium*. Miami, Florida, USA: ACM, pp. 363-372.
- Luo, X., & Zincir-Heywood, A. N. (2004). Combining word based and word co-occurrence based sequence analysis for text categorization. *Machine Learning and Cybernetics*. Shanghai: IEEE, pp. 1580-1585.
- Maldonado, A., & Emms, M. (2012). First-order and second-order context representations: geometrical considerations and performance in word-sense disambiguation and discrimination. *JADT 11th International Conference on the Statistical Analysis of Textual Data*. Liege, France, pp. 676-686.
- Meyer, P., & Eppinger, M. (2018). fLexiCoGraph: Creating and Managing Curated Graph-Based Lexicographical Data. In J. Čibej, V. Gorjanc, I. Kosem, & S. Krek (eds.) *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*. Ljubljana, Slovenia: Ljubljana University Press, Faculty of Arts, pp. 1017-1022.
- Newman, M. E. (2006a). Finding community structure in networks using the eigenvectors of matrices. *Physical review E*, 74(3), pp. 036104.
- Newman, M. E. (2006b). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23), pp. 8577-8582.
- Newman, M. E., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical review E*, 69(2), pp. 026113.
- Paatero, P., & Tapper, U. (1994). Positive Matrix Factorization: A Non-Negative Factor Model with Optimal Utilization of Error Estimates of Data Values. *Environmetrics*, 5(2), pp. 111-126.
- Pollak, S., Repar, A., Martinc, M., & Podpečan, V. (2019). Karst exploration: extracting terms and definitions from karst domain corpus. In I. Kosem et al. (eds.) *Proceedings of eLex 2019*. Sintra, Portugal, pp. 934-956.
- Pollak, S., Vavpetič, A., Kranjc, J., Lavrač, N., & Vintar, Š. (2012). NLP workflow for on-line definition extraction from English and Slovene text corpora. *Proceedings of KONVENS 2012*. Vienna, Austria: ÖGAI, pp. 53-60.
- Rodríguez Redondo, A. L. (2004). Aspects of cognitive linguistics and neurolinguistics: conceptual structure and category-specific semantic deficits. *Estudios ingleses de la Universidad Complutense*, 12, pp. 43-62.
- Rosvall, M., Axelsson, D., & Bergstrom, C. T. (2009). The map equation 178.1. *The European Physical Journal Special Topics*, 178(1), pp. 13-23.
- Stevens, K., Kegelmeyer, P., Andrzejewski, D., & Buttler, D. (2012). Exploring Topic Coherence over many models and many topics. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Jeju Island, Korea: Association for Computational Linguistics, pp. 952-961.

- Škrlj, B., & Pollak, S. (2019). Language comparison via network topology. *Proceedings of the 7th International Conference on Statistical Language and Speech Processing*. LNCS 11816. Springer.
- Temmerman, R. (2000). *Towards New Ways of Terminology Description: The Sociocognitive-approach*. Amsterdam, Philadelphia: John Benjamins Publishing Company.
- Vintar, Š. (2010). Bilingual Term Recognition Revisited. The Bag-of-Equivalents Term Alignment Approach and Its Evaluation. *Terminology*, 16(2), pp. 141-158.
- Vintar, Š., & Grčić Simeunović, L. (2016). Definition frames as language-dependent models of knowledge transfer. *Fachsprache: internationale Zeitschrift für Fachsprachenforschung, - didaktik und Terminologie*, 39(1-2), pp. 43-58.
- Vintar, Š., Saksida, A., Stepišnik, U., & Vrtovec, K. (2019). Knowledge frames in karstology: the TermFrame approach to extract knowledge structures from definitions. In *Proceedings of eLex 2019*, Sintra, Portugal.
- Wakita, K., & Tsurumi, T. (2007). Finding community structure in mega-scale social networks:[extended abstract]. *16th international conference on World Wide Web*. Banff, Alberta, Canada: ACM, pp. 1275-1276.
- White, S., & Smyth, P. (2005). A spectral clustering approach to finding communities in graph. *SIAM International Conference on Data Mining*. 5. Newport Beach, California, USA: SIAM, pp. 76-84.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



How Can App Design Improve Lexicographic Outcomes? Examples from an Italian Idiom Dictionary

Valeria Caruso¹, Barbara Balbi², Johanna Monti¹,

Roberta Presta²

¹ Università degli Studi di Napoli 'L'Orientale', Via Duomo 219 - 80138, Napoli (Italia)

² Università degli Studi Suor Orsola Benincasa, Via Suor Orsola 10 – 80135, Napoli (Italia)

E-mail: vcaruso@unior.it, barbara.balbi@centroscienzanuova.it,

roberta.presta@centroscienzanuova.it, jmonti@unior.it

Abstract

Despite the growing number of smartphone apps used in everyday tasks, lexicographic applications are still rarely discussed. Studies focus mainly on the usability of available tools, but contributions concerning the development of dictionary apps are almost non-existent.

In this paper, three different design solutions are presented to implement a dictionary app for Italian idioms, having foreign learners as prospective users. Prototypes were sketched according to *Human-centred design* principles and by applying a participatory approach in which users contribute to the design process.

To offer a trustworthy tool, special attention was also paid to the lexicographic data provided. To this end, the *OWID Sprichwörterbuch* model was enriched with specific information to support foreign speakers, whose communicative needs had been tested in a production task with Italian idioms.

The presentation of three prototypes is specifically addressed to highlight design solutions which can guarantee descriptive richness.

Keywords: dictionary Apps; electronic lexicography; *Human-centred design*; lexicographical functions; interactive systems

1. Introduction

This paper reports on the main features of a dictionary app prototype of Italian idioms for learners. The report will focus on the design concept and app features highlighting the interdisciplinarity of the project and the hybrid methodology used to investigate the best solutions to the challenges of the new media, i.e. smartphones.

Theoretical issues will be discussed throughout the paper while presenting the different stages of the app design: i) a post-consultation study (Fuertes-Olivera & Tarp, 2014) on the ability of target users to extract information from existing dictionaries; ii) a co-design protocol to merge experts' point of view with users' expectations and needs; iii) a final discussion on the best dictionary prototype to be tested with real users in the next research step.

It is worth noticing that the contribution deals with electronic dictionaries released as

smartphone applications. For this reason, we call them dictionary apps in place of “pocket electronic dictionaries” (or PED)¹, which was in common use before the smartphone revolution occurred to refer to “a small hand-held calculator-type reference work containing basic vocabulary in one or more languages” (Hartmann & James, 1998).

Nowadays, however, smartphones have evolved dramatically from the calculator format, and exploration of lexicographic applications for these devices is still in its infancy. In this paper possible contributions from the field of ergonomic design will be shown, with the hope that they could stimulate further debates and experiments.

1.1 The future of dictionaries and the dictionary apps of the future

In his vision about the future of dictionaries, Rundell (2012: 29) emphasizes that these tools will morph into services integrated into other software and stand-alone-products will decrease dramatically in number. However, if we focus on specific tasks that specific users might be interested to perform, one could also foresee different scenarios for the future implementations of dictionaries. For example, learners might profit from tools designed to increase specific skills, and the more we focus on single abilities, the more mobile apps can provide valuable assistance. Two options could guarantee, in fact, a future for dictionaries, as Amsler (cited in Lew & de Schryver, 2014) notes: “It’s a matter of either having lexical knowledge that nobody else has or displaying lexical knowledge in ways that are so convenient that other means of access are less attractive”. Essentially, this paper deals with the second option, focusing on key features of mobile applications. As IT experts and coders generally maintain, screen constraints and hardware limitations demand simple software in mobile devices, but simplicity is a more general concern for smartphones that deserves special attention.

In the years 2004-2008, smartphone apps contributed significantly to the process of the ‘eversion’ of cyberspace, as novelist William Gibson (2010) calls it, “a shift from virtual reality to mixed reality” (Jones, 2014). Today’s media tend in fact “to move out of the box and overlay virtual information and functionalities onto physical locations [thus creating] environments in which physical and virtual realms merge in fluid and seamless ways” (Hayles cit. in Jones, 2014). Focusing on smartphone users, Simonsen (2014: 260) notes that they navigate “in both the physical world and in the user interface of the mobile device at the same time”. This overlapping works as long as virtual data fit real-world issues, and the way data are provided is paramount. Different electronic devices – e.g. PCs, tablets, smartphones – can assist with different types of situations, as well as for different tasks. In particular, task complexity affects the type of device adopted by users, as reported by Simonsen (2014: 253): websites on PC screens to acquire extensive knowledge, smartphones to get a piece of missing information.

¹ The term was introduced by Taylor & Chan (1994).

1.1.1 Narrowing the scope

Restricting the scope of activities is a key feature which can make for valuable mobile apps. Tailoring information is important not only with respect to the type of task to be performed, but also for the amount of data to be managed by the electronic tool, as is also underlined by Simonsen (2015: 88): “The empirical data [...] show that different tasks call for different data sets and different access methods are required when using a dictionary app”.

Dictionaries with restricted macrostructure, e.g. collocation or idiom dictionaries, can be compiled more easily for mobile apps, since the scope of consultation is restricted from the beginning to a specific type of linguistic data. This reduces information overload and helps lexicographers accomplish some requirements of lexicographic description more easily, such as the need for a microstructural organization to comply with the lexicological properties of words.

Different word types – such as phrasal verbs and fixed phrases, or pragmatic markers and conjunctions – require different descriptions (Wiegand & Smit, 2013), which can be provided using specific data types within the dictionary articles. As an example, in the next sections (§ 3) we briefly report on some information needs related to idiomatic expressions that general language dictionaries are not able to fulfil when the user is an L2 speaker. We collected evidence by administering a test on the use of Italian idioms by foreign learners. The dictionary app described in this paper is instead particularly consistent at the “presentation level” (Müller-Spitzer, 2013), because all articles have the same microstructure which, however, can be split in different views, accessible by several actions.

2. Dictionary apps in the literature

The current debate on electronic lexicography is focused on complex tools developed as PC software, but research on dictionaries for handheld devices is still rare. However, the concept of an electronic dictionary is extremely broad and wide-ranging: “collections of structured electronic data that can be accessed with multiple tools, enhanced with a wide range of functionalities, and used in various environments” (de Schryver, 2003a: 146). Under this respect, dictionary apps should figure among the key concerns of this field, and debates should cover usability issues as well as technological solutions to fill information voids.

Existing research on dictionary apps has instead explored i) common features of available resources (e.g. Gao, 2013; Vitayapirak, 2013), ii) business models in the publishing market (Winestock & Jeong, 2014), iii) users’ interactions with these tools (Curcio, 2014; Marelllo, 2014; Simonsen, 2014, 2015; Vitayapirak, 2013). Marelllo and Simonsen, for example, adopt interesting methodologies and protocols to study the way users interact with mobile dictionaries, but the apps they have tested are rather

conventional, offering just a couple of smart features such as all-text-searches and a bar code reader in the medical tool used by Simonsen (2014, 2015). Some of the apps' shortcomings are also underlined in the papers. For example, Marelllo suggests microstructural implementations, while Simonsen complains about the interactional constraints of mobile devices which "drastically" reduce "information access success", thus urging that "mobile lexicography [...] reinvent itself" (Simonsen, 2014: 259).

Unfortunately, the revolution will not take place unless editors change their business model, which consists of developing one app "for one print dictionary", as Winestock and Jeong (2014) note, describing the app market. For the future implementations of dictionary apps, these authors suggest app aggregators, in which one initial dictionary can be implemented with special 'adds-on': different component parts addressing specific skills or features. A simplified version of this model is already available in the Chinese-English dictionary app released by Pleco. It is possible to suggest that similar tools are implemented in the future with search masks to access the different component parts of the app where each type of lexical unit is described according to its features.

This vision goes not very far from the segmentation of knowledge that *Lexicographical Function Theory* (Tarp, 2008) has claimed for electronic dictionaries, thus creating monofunctional tools (Tarp, 2012) in which users find different dictionaries addressing a specific lexicographic topic (e.g. general language, specialized language, collocations or idioms) from the perspective of different tasks to be performed with the dictionaries. Following this theory, Kwary (2013) outlines two different app concepts for the target users of Indonesian business people, who need to acquire news from the international market very quickly. The first software has the same functions that ebook readers implemented around the time of Kwary's paper: text-integrated dictionaries offering word meanings or translations as tooltips. The other tool goes in the opposite direction, listing the latest business headlines and giving access to a dictionary through a search bar where words can be typed or drag-and-dropped directly from the headlines.

In the current research, the same assumptions on lexicographic functions have been followed to define an app concept suited to the target users of advanced foreign speakers of Italian who wish to improve their language proficiency. For this reason, the app deals exclusively with idiomatic expressions, which are among the target skills of advanced levels (from B2 onwards) of linguistic certifications in CEFR (Common European Framework of Reference for Languages).

3. Monitoring users' needs in language tasks with idioms

To better support users' needs, a preliminary study of available dictionaries was carried out (Caruso, 2016). Idioms are in fact demanding for their semantics as well as for their morphosyntactic properties, since they are "fixed in their lexical structure (however, this does not exclude a certain limited variation), and they must be, at the same time, semantically reinterpreted units (i.e. they do not point to the target concept directly

but via a source concept) and/or semantically opaque” (Dobrovol’skij & Piirainen, 2005: 40). In Italian, for example, *dare la mano* (‘shake hands’) and *dare una mano* (‘help someone’) have different semantic and pragmatic meanings, despite the single variation in the noun determiner (a definite, *la*, or an indefinite article, *una*). *Darsela a gambe* (en. ‘to escape, running fast, from a complicate situation’) is instead extremely difficult to inflect (e.g. *Maria se l’è data a gambe*) and even to be searched for in the dictionary, because the lemma form is given in the infinite tense with agglutinated placeholder pronouns.

Ten Chinese and eight Vietnamese university students learning Italian in Naples were administered a test to assess their ability to extract information on idioms when using an authoritative general language dictionary, such as the *Vocabolario Treccani* (VT) online. The participants had been living in Italy for six months when the test was administered, and eight of them had a B2 certificate of proficiency, the others a C1 certification. The majority (55%) had been studying Italian for three years, others (28%) for two, and a smaller group for four years.

In a pre-test homework activity, students were asked to search for all the idioms listed in the *VT* articles for the words *testa* (‘head’) and *mano* (‘hand’) after having attended a lesson on the concept and features of idiomatic expressions, illustrated through Italian examples. After three days they were given a gap-filling exercise with missing idioms, having *testa* or *mano* as their “key constituents” and presenting an “image component”: “a specific conceptual structure mediating between the lexical structure and the actual meaning of figurative units” (Dobrovol’skij & Piirainen, 2005: 14). During the test, students had to choose the right idiom from a list which provided the explanations contained in the *VT* dictionary.

The results prove the inability of this type of users to extract information from the general language dictionary (Caruso, 2016). Only 56% of their answers were correct, since they either failed to select a semantically suitable expression (56%) or a correct inflectional form (43%). In correlating the type of explanation to students’ scores, the analysis showed that positive scores correlate with full-sentence explanations, written in a natural language style, as well as with those illustrating shifts from literal to abstract meaning. Concerning mistakes related to the inflectional form, they are caused by a lack of awareness about how idiom constituents inflect or do not change. Students’ proficiency level and years of study of the language do not correlate with better performance (Caruso, 2016).

4. Data types and lexicographic organization

In line with other studies on the role of imagery in idiom learning (see Szczepaniak & Lew, 2011), our data demonstrate the relevance of etymology in understanding figurative idioms, since it explains the shift from the literal to the metaphorical meaning and helps speakers build the “mental image” of the expression (see Dobrovol’skij, 2016:

23; Dobrovolskij & Piirainen, 2005). Another key concern for foreign speakers is the morphosyntactic explanation, thus inflexion tables should display paradigmatic declension exhaustively and remark unadmitted forms. For example, *Mettersi le mani nei capelli* (lit. ‘to put one’s hand in the hair’) conveys the idea of ‘despair’ by depicting² the conventional gesture of putting one’s hand in the hair (*capelli*) and is not used at the imperative form, nor can it convey all type of speech acts, such as giving advice or reproach someone.

Therefore, having as reference the lexicographic data types contained in the *OWID Sprichwörterbuch* (Steyer & Ďurčo, 2013), we added some features to support foreign learners more effectively. In particular, semantics is illustrated along with the etymology and literal meaning, whilst participants and valency structure are specifically addressed for verbal idioms, to explain the event the idiom describes thoroughly. This type of annotation is inspired by Frame Semantics (Fillmore, 1985), although Frames or Frame Elements listed within FrameNet (Fillmore et al., 2003) are not maintained within the app. Intuitive descriptors are used in their place to help users understand idiom syntax and semantics more accurately. The participants and valency structure, labelled “struttura linguistica” (en. ‘linguistic structure’), is annotated as follows:

	Mettere le corna	[a qualcuno]	[con qualcuno]
Maria	ha messo le corna	al suo fidanzato	con Fabrizio
[il traditore]		[la persona tradita]	[l’amante] ³

Additionally, in order to improve app effectiveness we highlighted unattested uses and word forms. For example, in *Mettere le corna* (en. ‘to cheat on someone’): “Parte non modificabile: le corna, non si può cambiare il genere, il numero e l’articolo. SBAGLIATO: mettere ~~il corno~~, mettere ~~la corna~~, mettere ~~un corno~~, mettere ~~una corna~~, mettere ~~i corni~~”⁴.

4.1 Lexicographical functions to create tripartite access to data: one

dictionary for writing, one for understanding and one for learning

To reduce information overload, we sketched a provisional microstructural organization for three different monofunctional app dictionaries of idioms addressing the

² According to Burger (2010: 63-64) it is a *Kinegramm*.

³ En.

	Cheat	[on someone]	[with somebody]
Maria	cheated	on his boyfriend	with Fabrizio
[the betrayer]		[the one who is betrayed]	[the lover]

⁴ English: “Unmodifiable word constituents: *le corna* (lit. ‘the horns’), article, gender, and number variation are not allowed. WRONG FORMS: mettere ~~il corno~~, mettere ~~la corna~~, mettere ~~un corno~~, mettere ~~una corna~~, mettere ~~i corni~~”

corresponding functions:

- dictionary for idiom understanding, type of data included:
 - Meaning (describing the idiom meaning and emphasizing the ‘image component’)
 - Literal meaning
 - [Participants and valency structure]⁵
 - Affective meaning
 - Stylistic meaning
 - Pragmatic and social meaning
- dictionary for using idioms:
 - Meaning
 - Unadmitted lexical variations
 - Affective meaning
 - Stylistic meaning
 - Pragmatic and social meaning
 - Contexts of use
 - Texts genera
 - [Connectors]
 - Typical modifiers
 - [Negative transformations]
 - [Syntactic transformations]
- dictionary for leaning idioms:
 - Meaning
 - Literal meaning
 - etymology
 - [Inflectional forms (active, passive, pronominal/impersonal/reciprocal voice)]
 - Lexical variations

Being useful for different functions, some data are displayed in more than one dictionary, as this is one of the main concerns in building monofunctional dictionaries: avoiding data redundancies whilst preserving descriptive adequacy. In the next section, the focus on usability required by the design protocols will prove its effectiveness in solving similar issues.

5. Design protocols to enhance dictionary usability

The idiom dictionary prototype developed so far has been released following the *Design thinking* (Plattner et al., 2014) protocol introduced by Hasso Plattner at the Stanford Institute of Design. This approach guides design processes to meet the standards of

⁵ Square brackets include data types used only for verbal idioms licensing a syntactic structure.

Ergonomics of human-system interaction, classified by the International Organization for Standardization (ISO) as 9242-210 in the *Standards catalogue*, which is specifically addressed to *Human-centred design for interactive systems*.

5.1 General principles of *Human-centred design*

The guidelines provided for *Human-centred design* aim at making computer-based interactive systems more usable “by focusing on the users, their needs and requirements, and by applying human factors/ergonomics, and usability knowledge and techniques” (ISO 9241-210: vi). The paradigm seems to be particularly promising in the field of electronic lexicography, especially when the dictionary moves into the handy format of a smartphone app. This approach lays down four key principles for design:

- encourage users’ active involvement in the design process to better understand their needs and task requirements;
- evaluate the distribution of functions to be performed by the user and by the technology he/she uses;
- iterate design solutions;
- adopt a multi-disciplinary approach to systems design.

The involvement of Human Factors (hence, HF) in the development of interactive systems is paramount. They work side by side with project stakeholders and technical implementers to guarantee that ergonomics principles concerning people’s capabilities, user experience and usability are covered from the beginning to the end of the project: from the concept outline to its prototyping and testing sessions with real users, followed by a re-design process of the tool. It is worth noting that in the field of electronic lexicography, the iteration of development phases has also been applied by de Schryver (2013) in his *Simultaneous feedback* protocol for dictionary compilation, where “user behaviour influence the presentation of lexicographic data through a direct feedback loop” (Lew & de Schryver, 2014).

Another key component of *Human-centered design* is the “context of use” (ISO 9241: 210), which is defined by the users, tasks and environments in which the system works. HF specialists employ social science techniques to define specific features of each “context of use” of the interactive system, i.e. contextual inquiries, interviews, focus groups, brainstorming, questionnaires, and co-design workshops with adequate stakeholders are common tools for this type of investigation.

5.2 Design methodology

The development framework used for the dictionary app is a well-known design protocol in five stages known as *Design thinking* (Platter et al., 2014). It is based on the active collaboration and involvement of stakeholders, i.e. the dictionary users and the

lexicographers (or “subject matter experts”), in a design process guided by HF specialists during co-design sessions, where participants work in pairs through the following time-constrained phases:

- Empathize: the participant acting as a designer (afterwards “designer”) poses questions to the participant acting as a user (afterwards “user”) to understand his needs and expectations about the system to be designed;
- Define: the user’s characteristics and his needs are the focus when outlining core features of the interactive system;
- Ideate: a range of possible solutions are sketched by the designer and, afterwards, are evaluated with the user to assess if they meet his/her needs;
- Prototype: after having selected the best ideas from the sketched solutions, the designer outlines a single proposal;
- Test: the prototype is evaluated together with the user to identify strengths and weaknesses.

For the current research, the *Design thinking* model was implemented by a co-design workshop (Halloran et al., 2009) which allows the relevant stakeholders to take part in the design process: thus dictionary users and lexicographers (or “subject matter experts”) have been working side by side with the designers, sketching dictionary prototypes on paper. Designers (or HF specialists in charge of the system design) were researchers from the University of Naples Suor Orsola Benincasa, and stakeholders were lexicographers from the University of Naples ‘L’Orientale’ and 14 Chinese learners of the Italian language. The ideational process started with a preliminary interview of lexicographers, by which HF specialists could gather insights from the experts’ point of view regarding dictionary features, shortcomings during consultation, state-of-the-art electronic tools and lexicographic theory.

5.3 The co-design workshop

Twenty-four people participated in a co-design session: four Italian lexicographers, six Italian designers and 14 Chinese students learning Italian at the University of Naples ‘L’Orientale’, having a B2 or C1 certification of proficiency in this language.

To make users more aware about the tasks to be performed with the app, an introductory presentation was made, and users were assigned reading, writing, and learning tasks in which idioms were involved. In the same session, idiom features were briefly explained together with the lexicographic data (those listed in § 4.1) that dictionaries can provide to assist users with these demanding lexical units.

The co-design session aims to collect information about users’ ideas and expectations as well as their needs when consulting a monolingual idiomatic dictionary, i.e. the way they approach lexicographic tasks and the type of expectations they have about a dictionary compiled as a mobile application. This is done by letting users “empathize”

with designers in sessions of role-playing activities, during which the user acts as a designer and is in charge of prototyping the interactive system with the designers' tools, following the steps of a design framework. Participants annotate their findings, needs, ideas and even draw prototype sketches of their solution proposals on paper sheets that are collected at the end of the co-design session. These materials contain meaningful insights, inspirations and well-focused needs coming from the community of project stakeholders (i.e. learners, lexicographers and designers as well) and are used to design the first prototype.

However, this is only the first stage of the iterative cycle of *Human-centred design*, consisting of a proposal of a first set of prototypes to be used as test materials with prospective users. After a first testing session, an improved solution is re-designed and new design cycles, typically two or three, will take place before the final tool is released.

5.4 Output of the co-design workshop

From the analysis of the co-design session materials it emerged that users' needs were focused both on the content and functional requirements of the tool. Fig. 1 presents the results of this, showing the percentage of participants who responded with each need.

Afterwards, users' needs were arranged and classified into three types:

- (i) goals: the aims for which the user wants to use the app;
- (ii) generic features: what the user expects to find in the app, because of the standard features of many other apps he/she uses;
- (iii) specific features: functionalities and content that are specific for the idiom dictionary app. Content is related to social and motivational aspects, while functionalities are linked to cognitive and epistemic aspects (Buccini & Padovani, 2007).

5.4.1 Goals

What mostly motivates students in using the app is the desire to be able to master idiomatic expression in conversations (**Communication** in Fig.1) and in real-life situations. The other goal-related needs are:

- **Learning**, an important objective for language certifications;
- **Culture**, a type of knowledge which can be improved through a deeper understanding of the origins of idiomatic expressions;
- **Teaching**, a key concern of lexicographers who wish to rely on apps for teaching purposes during class hours;
- Finally, **Entertainment**, because participants acknowledged that the stories behind idiomatic expressions are often surprising and entertaining.

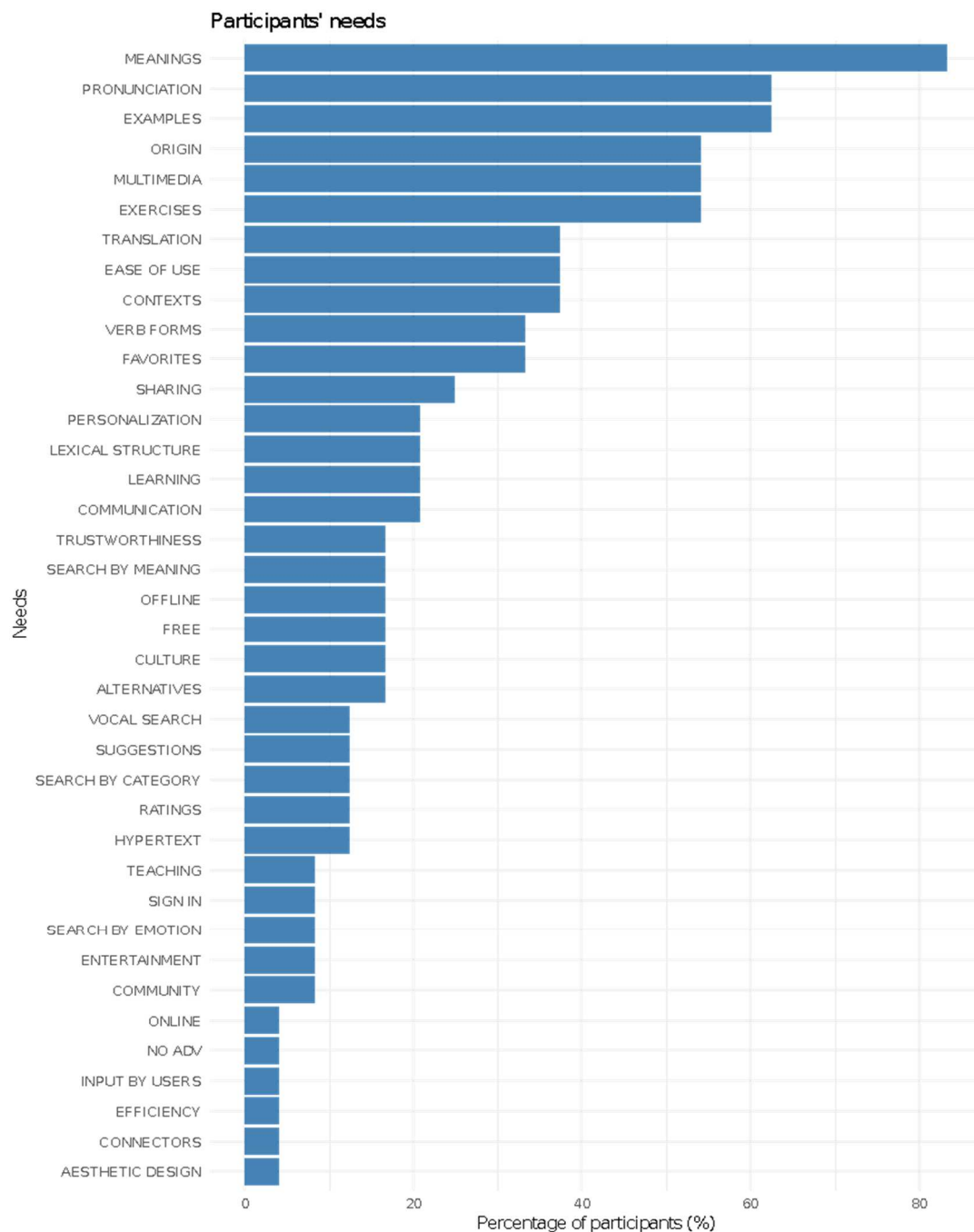


Figure 1: The complete listing of users' needs extracted from the co-design session

5.4.2 Generic features

Generic features are apps' standard features not specifically relevant for dictionaries. Some of them are related to social and motivational needs, and in particular to the possibility that the users will rate and produce content, thus inheriting interaction models typical of social networks:

- **Ratings:** users can rate the quality of the app content for each idiomatic expression;

- **Community:** the app has a forum for discussion with other users;
- **Production:** users can add content, for instance new examples.

Users also declared that they were interested in selecting and sharing (**Sharing**) favourite content (**Favorites**) and creating a community via the **Sign up** features.

5.4.3 Specific features

It is worth noting that all participants expressed their wish to understand the various different shades of idiomatic meaning (**Meanings**), with a particular interest in the main stylistic, affective and literal components. Also in line with the primary goal of improving communication and conversational skills, **Pronunciation** and examples of use (**Examples**) were considered as must-have features in the app, together with **Exercises** to fulfil learning goals.

Further requirements concerned **Multimedia** content and notes on idiom **Origin**, which fulfils cultural, learning and entertainment needs, because the story behind the idiom is typically easily memorable and nice to know. **Translation** is one of the top needs, too, even if it was presented as off the topic, because we wanted to focus on a monolingual dictionary. A related need is having an integrated term dictionary (**Hypertext**) that enables users to search for the meaning of single words appearing in an idiom. Other features addressed search options, i. e. **Vocal search**, and the **Search by meanings, emotions, and context/categories**, which are alternative ways of retrieving idiomatic expressions by selecting a group of tags.

The list of requirements also included a clear explanation about the **Context of use**, **Inflection**, the **Lexical structure**, or the idiom invariant constituents, the **Connectors** which typically introduce idioms in the discourse, and **Alternatives**, or other idioms to be used in place of the one under consideration. Finally, usability features were included as these are essential for a valuable design (Blythe & Monk, 2018): i.e. **Ease of use**, addressing readability and understandability, **Effectiveness**, **Trustworthiness** of data, and the **Aesthetic design**, meaning that the interface is expected to look modern and not overwhelm users with information.

6. Designing prototypes

Based on the priorities that emerged from the co-design session, a first app prototype was sketched to summarize the needs and priorities related to a monolingual idiom dictionary to develop an artefact that could be used in testing sessions with real users.

The prototype is developed for iOS devices, following the Apple *Human Interface Guidelines* and using the software *Sketch*, which allows for dynamic linking of the user interface views by tapping on the envisioned interactive components, and can be easily used in testing sessions with real users.

In this first stage of prototyping, we focus mainly on the app structure and its content, leaving aesthetic details for a second round of prototyping after having collected users' feedback. The challenge to meet is to combine information access efficiency with content completeness, thus merging expert knowledge with users' desires, prioritized with the co-design experiment (§ 5.4).

The Home interface (fig. 2) shows a tab bar (at the bottom of the app screen) giving access to the main app sections related to the main goals identified by users:

- **Search view**⁶: (corresponding to the book icon in the tab bar) is a rather traditional search interface;
- **Idiom categories view**: (multiple squares icon) allows users to search idioms by tags, thus making search options more advanced than in traditional electronic dictionaries;
- **Practice view**: (graduate cap icon) gives access to an exercise section;
- **Favourites view**: (star icon) collects the user's preferred content;
- **Settings view**: (human figure icon) gives access to setting options.

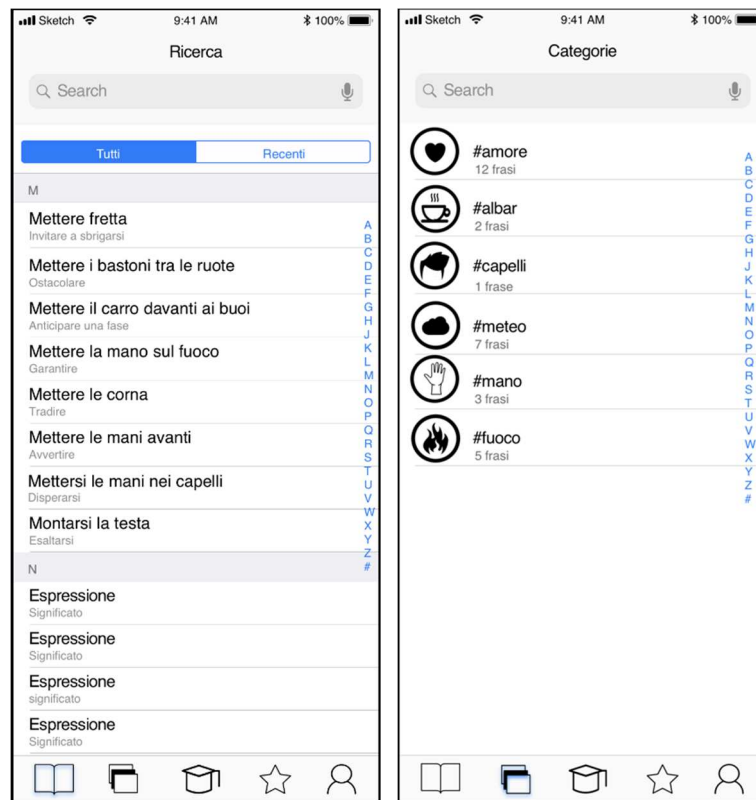


Figure 2: Home/Search interface and Home/Search by categories

⁶ Speaking about apps, we adopt the current terminology in use to address their component parts. The 'view' is what the user sees displayed on the screen. For these terms, please refer to Human Interface Guidelines provided by operating systems developers, such as iOS or Android.

The first tab bar buttons address search sections specifically developed for fast comprehension and production tasks. Indeed, the Home Search interface ('Ricerca', first button, book icon) allows for searching by idiom forms and meanings as well: by typing a word in the search bar, or pronouncing the desired expression, the app returns the idioms related to the word, which is included in the idiom or represents its meaning. The Idiom Categories View (Categorie), on the other hand, enables the search of the idiom by “tags” (e.g., #meteo, #love, #school, #fear, ...): a tag can express a topic, a situation, a place, an emotion, and other. Each idiom can have multiple tags, thus allowing the user to make searches by emotions and contexts of use, as they were desiderata coming from the co-design session. Practice view (Pratica) answers to the need for mastering idiomatic expressions and addresses the ‘learning function’ which, therefore, is not devised in the form of a separate dictionary, as proposed before the design process began (see § 4.1). **Favourites** allows for rapid access to those idioms that users have already gone through, and **Settings** allows for the personalization of the app from the aesthetics and content perspectives, e.g. interface colour modes, configuration of exercises, and so on.

The core of the design effort is the **Home view**, because it should give access to lexicographic data in a way that can be successfully used by a mobile application user: synthetically and in a recognizable form, because of the limited display space and interaction time constraints of mobile apps. Users, in fact, expect more rapid interactions with these devices than with paper books and other electronic devices (see also Simonsen, 2014). With this in mind, the priority list from the co-design session was rescheduled by the designers and lexicographers to develop a more consistent arrangement of data, and the priorities were set as follows:

- **Meanings:** main, literal, affective and stylistic meaning. For each meaning type, explanations and examples are provided.
- **Origin:** etymology has a storytelling power which is useful to understand and memorize idiomatic expressions, whilst enhancing the app entertainment dimension.
- **Contexts of use:** provides attested uses in different situations, places, text typologies or registers.
- **Inflexion, Lexical Structure, Connectors, and Alternatives** are described in § 5.4.3.

Directives from the experts provide meaningful hints along two dimensions: the provision of different data types according to the tasks the user is about to perform (e.g. inflectional information for production tasks); and the ordering of these data (e.g. pronunciation and morphological transformations near the lemma sign).

Given this overview, access to lexicographic data is discussed using three different possible approaches to prototyping the dictionary.

6.1 Prototyping approach A

The first approach (Fig. 3) is the most straightforward, displaying lexicographic data in different *search zones* (Wiegand et al., 2013), labelled with the name of the data types therein contained, which are accessible in a scrollable way. This solution allows the user to access the main information immediately, without having to tap other interactive components, like buttons. Moreover, interested users can read lexicographic data by scrolling the view, which is a quite natural gesture, and this could facilitate in understanding the labels (or *data-identifying entries*, Gouws, 2014), such as “emotive meaning” (it.: ‘significato emotivo’), thanks to the data provided under each heading.

On the other hand, users have to scroll the view in order to find information that might be useful for their tasks, and more experienced users might get frustrated in navigating the entire view and its contents each time.

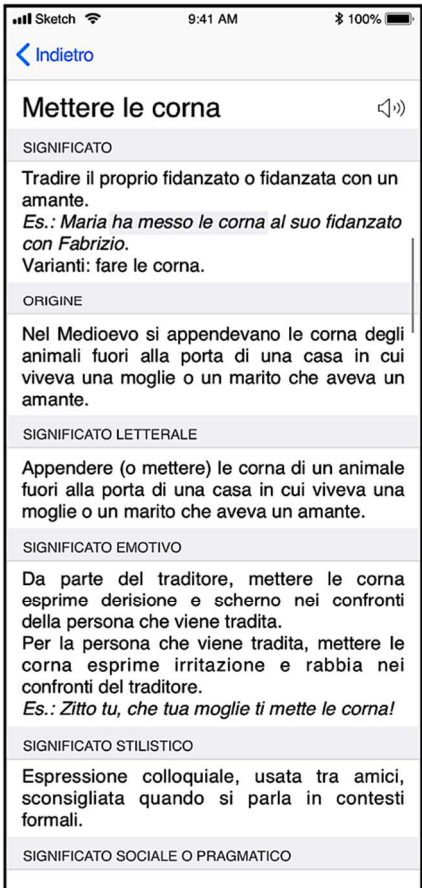
	<p>Cheat on someone</p> <p>MEANING To betray one's boyfriend or girlfriend with a lover. <i>Ex.: Maria has cheated on her boyfriend with Fabrizio.</i> Admitted lexical substitutions: fare le corna (instead of 'mettere le corna')</p> <p>ORIGIN In Medieval age, people used to hang horns at the front door of houses where unfaithful wives or husbands lived</p> <p>LITERAL MEANING To hang animals' horns at the front door of houses where unfaithful wives or husbands lived</p> <p>EMOTIVE MEANING Used by the traitor, 'mettere le corna' (en. 'cheat on someone') expresses scorn and derision for the betrayed person. Used by the betrayed person, 'mettere le corna' expresses anger towards the betrayer. <i>Ex. You shouldn't say a word, as long as your wife cheats on you!</i></p> <p>STYLISTIC MEANING Colloquial expression, used with friends, not recommended in formal contexts.</p> <p>SOCIAL OR PRAGMATIC MEANING</p>
--	---

Figure 3: Prototyping approach A on the left, English translation on the right

6.2 Prototyping approach B

A second possibility is a scenario in which each type of lexicographic data is provided by accessing a dedicated row of a table view, like the one in Fig. 4. This solution has the advantage of simplifying the view by showing exclusively data-identifying entries, or lexicographic labels. This minimalist approach has the obvious disadvantage of increasing the time needed to access data, and the number of actions needed for task completion. Besides, less experienced users might be not familiar with lexicographic labels and could get confused to the point of quitting the app.

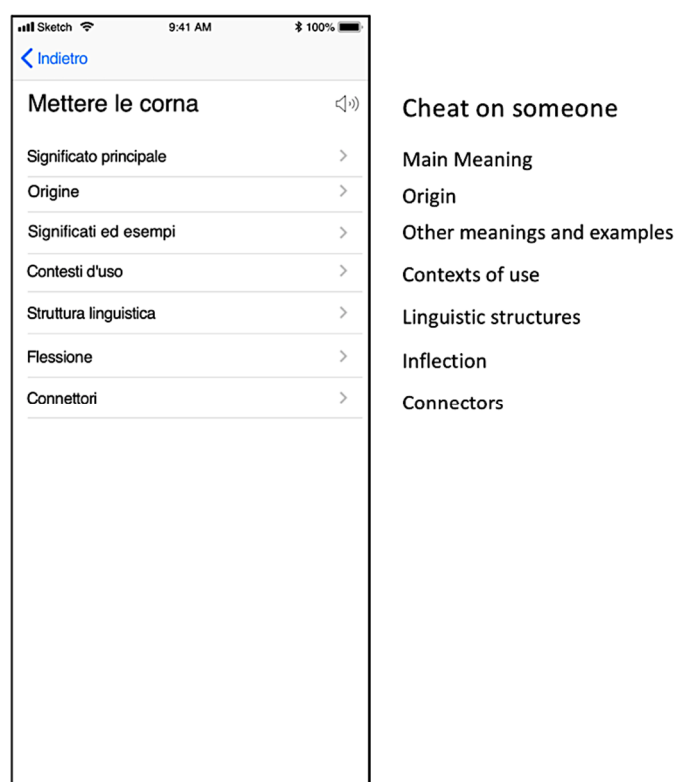


Figure 4: Prototyping approach B, English translation on the right

6.3 Prototyping approach C

A third prototyping solution is the result of a hybrid design, providing in a single view the idiom's general meaning and different ways to access other data. To achieve the aim of consulting more explanations about the idiom, two possible design solutions have been sketched, prototypes C1 and C2, which guarantee that users find general information quickly, and can then decide to acquire more data using specific lexicographic assistance if needed.

6.3.1 Prototype C1

In prototype C1, a segmentation bar⁷ allows the user to choose the situation of use, thus accessing different data types, as happens in the monofunctional dictionaries discussed before. Choosing between one of the available options in the segmentation bar, i.e. **Comprensione** (en. comprehension) and **Produzione** (en. production), the app filters data suited for comprehension (meanings, origin, context of use) from that suited for production (lexical structure, verb forms, connectors, see Fig. 5). In this way, for example, users who need to perform a comprehension task are provided only with the necessary lexicographic content in a scrollable way.

The advantage of such a solution is that users are fluently guided through the data types better suited to the different tasks, as advocated by the *Lexicographical Function Theory*, while inheriting the strengths and weaknesses of approach A.

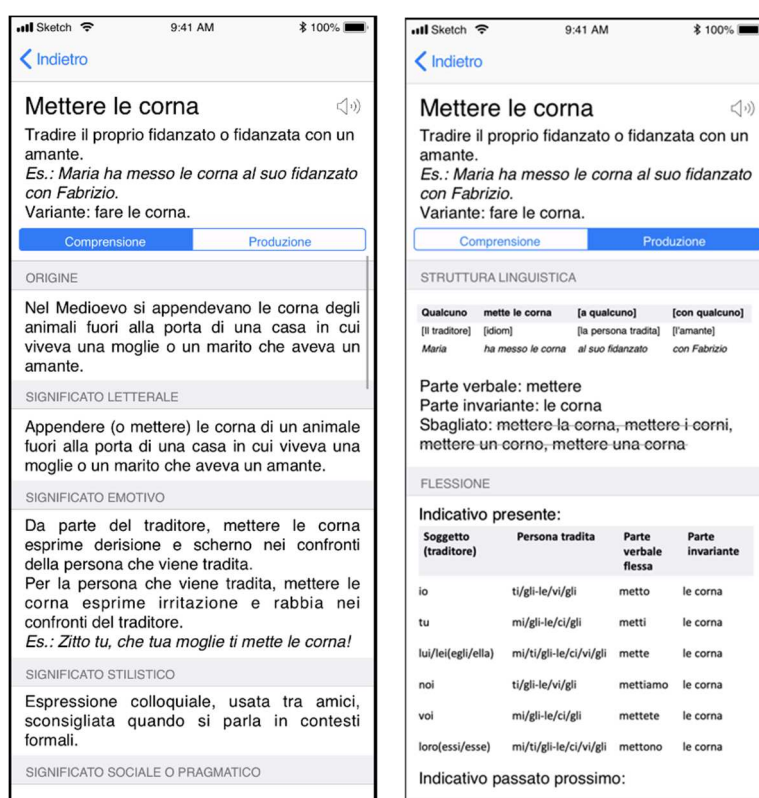


Figure 5: Prototype C1 – On the left: comprehension-oriented task view; on the right: production-oriented task view

⁷ See “Segmented controls”: <https://developer.apple.com/design/human-interface-guidelines/ios/controls/segmented-controls/>.

6.3.2 Prototype C2

The second prototype inherits the structure of approach C with a clearer indication of what is recommended for comprehension and production tasks (Fig. 6).

Users are provided at a glance with the list of the app contents divided per task ('Informazioni per la comprensione', eng: Information for understanding; 'Informazioni per la produzione', eng.: Information for production), thus helping them in constructing a mental model of what is needed for comprehension and production activities. Lexicographic data can be accessed instead by tapping on the labels and opening a new view, thus the space available for lexicographic descriptions is larger than in prototype C1, which is particularly valuable to manage inflexion tables (compare the corresponding views in prototypes C1 and C2, in Figs. 5 and 7).

To sum up, while the access to lexicographic content in C2 is pushed one tap forward in comparison with prototype C1, such a structure conveys more information to the user in an easier way. The table view structure inherits the advantages highlighted in the B approach, while it reduces the disadvantages by employing *structural indicators* that suggest the type of data better suited for specific task completion, as happens in monofunctional dictionaries.

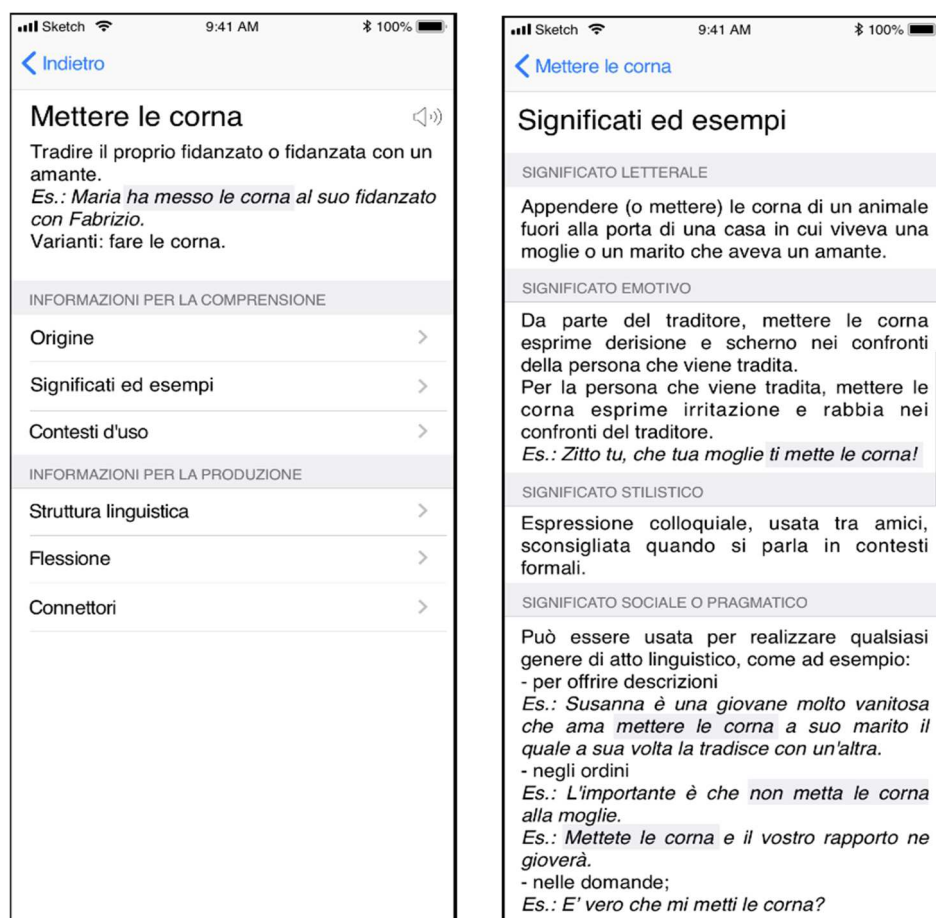


Figure 6: Prototype C2 – Examples of meanings details

Flessione

FORME IN USO

- Negativa
Es.: *Va bene tutto, purchè tu non mi metta le corna!*
- Attiva
Es.: *Lui è un tipo che mette le corna.*
- Passiva
Es.: *A me sono state messe le corna parecchie volte.*
- Impersonale
Es.: *Se si mettono le corna, per me il rapporto è naufragato.*
- Reciproca
Es.: *Le coppie italiane si mettono le corna.*

CONIUGAZIONI

Attiva >

Passiva >

Coniugazione attiva

INDICATIVO PRESENTE

Soggetto (traditore)	Persona tradita	Parte verbale flessa	Parte invariante
io	ti/gli-le/vi/gli	metto	le corna
tu	mi/gli-le/ci/gli	metti	le corna
lui/lei(egli/ella)	mi/ti/gli-le/ci/vi/gli	mette	le corna
noi	ti/gli-le/vi/gli	mettiamo	le corna
voi	mi/gli-le/ci/gli	mettete	le corna
loro(essi/esse)	mi/ti/gli-le/ci/vi/gli	mettono	le corna

INDICATIVO PASSATO PROSSIMO

Soggetto (traditore)	Persona tradita	Parte verbale flessa	Parte invariante
io	ti/gli-le/vi/gli	ho messo	le corna
tu	mi/gli-le/ci/gli	hai messo	le corna
lui/lei(egli/ella)	mi/ti/gli-le/ci/vi/gli	ha messo	le corna
noi	ti/gli-le/vi/gli	abbiamo messo	le corna
voi	mi/gli-le/ci/gli	avete messo	le corna
loro(essi/esse)	mi/ti/gli-le/ci/vi/gli	hanno messo	le corna

INDICATIVO IMPERFETTO

Soggetto (traditore)	Persona tradita	Parte verbale flessa	Parte invariante
io	ti/gli-le/vi/gli	mettevo	le corna
tu	mi/gli-le/ci/gli	mette	le corna

Figure 7: Prototype C2 – Examples of verb forms details

7. Conclusions and future work

This paper has pointed out key features of the recent digital revolution to introduce basic principles of app design for smartphones. With the “eversion of cyberspace” (Gibson, 2010) information has become ubiquitous, but the way users access data – whether through PCs, tablets or smartphones – makes for completely different knowledge experiences. With regard to smartphones, the focus should be on how data can fit real-life situations at a glance, displayed on small screen views, and reachable by a few, fluid actions.

The discussion on possible design solutions in Section 6 has shown how *Lexicographical Function Theory* can contribute to dictionary app design, offering valuable criteria for data arrangement. For example, using *structural indicators* (i.e. labels) to suggest data for the tasks to be performed, prototype C2 guides users through data consultation whilst preserving a minimalist interface, because the information is displayed in separate views. At the same time, recommending data for specific tasks, instead of building separate monofunctional dictionaries, gives users the option of selecting data autonomously, thus customizing their consultations. This solution also avoids repetitions that may occur in compiling separate, monofunctional dictionaries (see §4.1),

since the same type of information may be beneficial for different actions performed with the dictionary support.

On the other hand, *Human-centred design* offers new protocols, which put users and usability issues on the centre stage. To increase data accessibility, for example, the ‘learning dictionary’ has been transformed into a training section provided only with exercises. The learning component is, in fact, a more general function that is fulfilled by all the dictionary component parts: from the advanced search functionalities (e.g. searches by tags), to the rich semantic descriptions (literal, stylistic, pragmatic meaning) and morpho-syntactic explanations (inflexion tables, linguistic structure, connectors). Assuming this point of view, data selection becomes easier, because dictionary functions are reduced to production and reception tasks, while the co-design workshop offers other valuable insights for compiling the dictionary. In contrast to what one might expect, for example, etymology proved to be among the users’ top-rated features, therefore this data type should be displayed not only to improve idiom comprehension and learning, but also to fulfil an entertainment function.

In the next research step, the prototype solutions presented so far will be assessed with real users to implement a re-design cycle based on users’ feedback. Evaluation criteria will deal with prototype usability for different lexicographic tasks (comprehension, production, learning) and according to objective and subjective measurements (e.g., the time for task completion, user satisfaction, and so on).

8. Acknowledgements

The authorship contribution is as follows: Valeria Caruso is author of Sections from 1 to 4 and 7 (except the last paragraph written by Roberta Presta); Johanna Monti of 4.1 and 5; Barbara Balbi of 5.2; Roberta Presta of 5.1; and from 5.3 to 6.3.2, with the exception of lexicographic statements and concepts, added by Valeria Caruso.

9. References

- Blythe, M. & Monk, A. (eds.) (2018). *Funology 2: From usability to enjoyment*. Cham: Springer.
- Buccini, M. & Padovani, S. (2007). Typology of the experiences. In I. Koskinen & T. Keinonen (eds.) *DPPI 2007 - Proceedings of the 2007 International Conference on Designing Pleasurable Products and Interfaces*. New York: ACM Press, pp. 495-504.
- Burger, H. (2010). *Phraseologie. Eine Einführung am Beispiel des Deutschen*. Berlin: Erich Schmidt.
- Caruso, V. (2016). Dizionari elettronici e apprendimento delle espressioni idiomatiche: monitoraggio dei bisogni e prospettive future. In F. Bianchi & P. Leone (eds.) *Linguaggio e apprendimento linguistico. Metodi e strumenti tecnologici*. Milano: Studi AItLA, pp. 173-189.

- Curcio, M.-N. (2014). Die Benutzung von Smartphones im Fremdsprachenerwerb und -unterricht. In A. Abel, C. Vettori & N. Ralli (eds.) *Proceedings of the XVI EURALEX International Congress: The User in Focus*. Bolzano/Bozen: Institute for Specialised Communication and Multilingualism, pp. 15-19.
- de Schryver, G.-M. (2003). Lexicographers' Dreams in the Electronic-Dictionary Age. *International Journal of Lexicography*, 16(2), pp. 143-199.
- de Schryver, G.-M. (2013). The Concept of Simultaneous Feedback. In R.-H. Gouws, U. Heid, W. Schweickard & H. E. Wiegand (eds.), pp. 548-556.
- Dobrovol'skij, D. (2016). The notion of "inner form" and idiom semantics. In S. Viellard (ed.) *Proverbes et stéréotypes : formes, forme et contextes*, Paris: Université Paris-Sorbonne - CNRS. Accessed at: <http://eurorbem.paris-sorbonne.fr/spip.php?article696&lang=fr> (14 June 2019)
- Dobrovol'skij, D. & Piirainen, E. (2005). *Figurative language: cross-cultural and cross-linguistic perspectives*. London: Emerald.
- Fillmore, Ch. (1985). Frames and the semantics of understanding. In *Quaderni di semantica*, pp. 222-254.
- Fillmore, Ch., Johnson, Ch. & Petruck, M. (2003). Background To Framenet. *International Journal of Lexicography*, 16 (3), pp. 235-250.
- Fuertes-Olivera, P.-A. & Tarp, S. (2014). *Theory and Practice of Specialised Online Dictionaries*. Berlin & New York: de Gruyter.
- Gao, Y. (2013). On the application of dictionaries: from a Chinese perspective. In I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets & M. Tuulik (eds.) *Proceedings of the eLex 2013 conference. Electronic lexicography in the 21st century: thinking outside the paper*. Ljubljana & Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut, pp. 213-224.
- Gibson, W. (2010). Google's Earth. In New York Times, August 31, 2010, <http://nytimes.com/2010/09/01/opinion/01gibson.html>.
- Gouws, R. H., Heid, U., Schweickard, W. & Wiegand, H.-E. (eds.) (2013). *Dictionaries. An International Encyclopedia of Lexicography: Supplementary volume. Recent Developments with Special Focus on Computational Lexicography*, Berlin, Mouton: de Gruyter.
- Gouws, R.-H. (2014). Article structures: Moving from Printed to e-Dictionaries. *Lexikos*, 24, pp. 155-157.
- Halloran, J., Hornecker, E., Stringer, M., Harris, E., & Fitzpatrick, G. (2009). The value of values: Resourcing co-design of ubiquitous computing. *CoDesign*, 5(4), pp. 245-273.
- Hartmann, R.-R.-K. & James, G. (1998). *Dictionary of Lexicography*. London and New York: Routledge.
- ISO 9241: 210. *Ergonomics of human-system interaction: Part 210. Human-centred design for interactive systems*. Geneva: ISO.
- Jones, S. E. (2014). *The Emergence of the Digital Humanities*. New York & London: Routledge.
- Kwary, D.-A. (2013). Principles for the design of business dictionaries on mobile

- applications. *HERMES-Journal of Language and Communication in Business*, 50, pp. 69-81.
- Lew, R. & de Schryver, G.-M. (2014). Dictionary Users in the Digital Revolution. *International Journal of Lexicography*, 27(4), pp. 341–359.
- Marello, C. (2014). Using Mobile Bilingual Dictionaries in an EFL Class. In A. Abel, C. Vettori & N. Ralli (eds.) *Proceedings of the XVI EURALEX International Congress: The User in Focus*, Bolzano/Bozen: Institute for Specialised Communication and Multilingualism, pp. 63-83.
- Müller-Spitzer, C. (2013). Textual structures in electronic dictionaries compared with printed dictionaries: A short general survey. In R.-H. Gouws, U. Heid, W. Schweickard & H. E. Wiegand (eds.), pp. 367–381.
- Plattner, H., Meinel, C. & Leifer, L. (eds) (2014). *Design thinking Research. Understanding Innovation*. Berlin & Heidelberg: Springer.
- Rundell, M. (2012). The Road to Automated Lexicography: An Editor's Viewpoint. In S. Granger, & M. Paquot (eds.) *Electronic Lexicography*. Oxford: Oxford University Press, pp. 15-30.
- Simonsen, H.-K. (2014). Mobile Lexicography: A survey of the mobile user situation. In A. Abel, C. Vettori & N. Ralli (eds.) *Proceedings of the XVI EURALEX International Congress: The User in Focus*, Bolzano/Bozen: Institute for Specialised Communication and Multilingualism, pp. 15-19.
- Simonsen, H.-K. (2015). Mobile Lexicography: Let's Do it Right This Time. In I. Kosem, M. Jakubíček, J. Kallas & S. Krek. (eds.) *Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference*, Ljubljana/Brighton: Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd.
- Steyer, K. & Ďurčo, P. (2013). Ein korpusbasiertes Beschreibungsmodell für die elektronische Sprichwortlexikografie. In J. M. Banayoun, N. Kübler & J. P. Zouogbo (eds.) *Parémiologie, Proverbes et formes voisines*. Sainte Gemme, Band 3, pp. 219-250.
- Szczepaniak, R. & Lew, R. (2011). The Role of Imagery in Dictionaries of Idioms. *Applied Linguistics*, 32(3), pp. 323–347.
- Tarp, S. (2008). *Lexicography in the Borderland between Knowledge and Non-knowledge*. Berlin & New York: de Gruyter.
- Tarp, S. (2012). Online dictionaries: today and tomorrow. *Lexicographica*, 28 (1), pp. 253–268.
- Taylor, A. & Chan, A. (1994). Pocket Electronic Dictionaries and Their Use. In W. Martin et al. (eds.) *Proceedings of the 6th Euralex International Congress*. Amsterdam: Vrije Universiteit, pp. 598–605.
- Vitayapirak, J. (2013). Mobile Dictionary Use and the Design of an Electronic Bilingual Dictionary on Mobile Phones for Thai Users. In *ASIALEX Proceedings*, pp. 115-118.
- Wiegand, H.-E. & Smit, M. (2013). Microstructures in printed dictionaries. In R.-H. Gouws, U. Heid, W. Schweickard & H. E. Wiegand (eds.), pp. 149-214.

- Wiegand, H.-E., Beer, S. & Gouws, R.-H. (2013). Textual Structures in Printed Dictionaries: An Overview. In R.-H. Gouws, U. Heid, W. Schweickard & H. E. Wiegand (eds.), pp. 31-73.
- Winestock, C. & Jeong, Y.-K. (2014). An analysis of the smartphone dictionary app market. *Lexicography*, 1(1), pp. 109-119.

Dictionaries:

- OWID-Sprichwörterbuch: *Online-Wortschatz-Informationssystem Deutsch-Sprichwörterbuch*. Accessed at <http://www.owid.de/wb/sprw/start.html>. (10 June 2019)
- Pleco: *Pleco Chinese Dictionaries*. Accessed at: <https://www.pleco.com>. (10 June 2019)
- VT: *Vocabolario Treccani*, Istituto della Enciclopedia Italiana. Accessed at <http://www.treccani.it/vocabolario/>. (10 June 2019)

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



TEI Encoding of a Classical Mixtec Dictionary Using GROBID-Dictionaries

Jack Bowers^{1,2,3}, Mohamed Khemakhem^{1,4,5,6}, Laurent Romary¹

¹ Inria-ALMAAnaCH - Automatic Language Modelling and ANALysis & Computational Humanities, Paris, France

² EPHE - École Pratique des Hautes Études, Paris, France

³ ACDH - Austrian Center for Digital Humanities, Vienna, Austria

⁴ UPD7 - Université Paris Diderot - Paris 7, Paris, France

⁵ CMB – Centre Marc Bloch, Berlin, Germany

⁶ BBAW – Berlin-Brandenburg Academy of Sciences and Humanities, Berlin, Germany

E-mail: iljackb@gmail.com, mohamed.khemakhem@inria.fr, laurent.romary@inria.fr

Abstract

This paper presents the application of GROBID-Dictionaries (Khemakhem et al., 2017; Khemakhem et al., 2018a; Khemakhem et al., 2018b; Khemakhem et al., 2018c), an open source machine learning system for automatically structuring print dictionaries in digital format into TEI (Text Encoding Initiative) to a historical lexical resource of Colonial Mixtec ‘Voces del Dzaha Dzahui’ published by the Dominican Fray Francisco Alvarado in the year 1593. The GROBID-Dictionaries application was applied to a re-organized and modernized version of the historical resource published by Jansen and Perez Jiménez (2009). The TEI dictionary thus produced will be integrated into a language documentation project dealing with Mixtepec-Mixtec (ISO 639-3: mix) (Bowers & Romary, 2017, 2018a, 2018b), an under-resourced indigenous language native to the Juxtlahuaca district of Oaxaca Mexico.

Keywords: Mixtec; TEI; GROBID-Dictionaries

1. Introduction to the resource

This paper presents the creation of a TEI dictionary of the earliest lexical resource¹ of a Mixtec language: the Vocabulario published by the Dominican Fray Francisco Alvarado in the year 1593². This resource was automatically converted from PDF format to a structured TEI dictionary using the application GROBID-Dictionaries (Khemakhem et al., 2017; Khemakhem et al., 2018a; Khemakhem et al., 2018b; Khemakhem et al. 2018c), an open source machine learning system for automatically

¹ Not including the codices which were pictographic and not specific to any local variety of Mixtec. Though the author did not represent all the features of the language such as tone, nasalization among other features is resource is thus the first with any representation of the phonetic characteristics of a Mixtecan Language (Jansen & Perez Jiménez, 2009).

² This document was likely compiled from the few existing resources at the time, namely the Doctrina en Lengua Mixteca by Fray Benito Hernández published in 1567 and 1568 respectively from sources compiled in Teposcolula Mexico (Jansen & Perez Jiménez, 2009).

structuring print dictionaries in digital format into TEI (Text Encoding Initiative). The PDF source used in the transformation is from a re-organized and modernized version of the historical dictionary published by Jansen and Perez Jiménez (2009). The TEI dictionary produced contains roughly 26,600 entries and related entries.

The Mixtec variety sampled by Alvarado to create this vocabulary was that of Yucu Ndaa (Teposcolula) *dzaha dzavui*³, which according to the sources is thought to have been used as a *lingua franca* of the Mixteca region at the time and the language is presently in the field of Mixtecan commonly referred to as “Classical Mixtec” or “Colonial Mixtec” (Jansen & Perez Jiménez, 2009).

The vocabulary was produced by the Orden de los Predicadores (O.P.) aka. the Dominican Order, who wanted to learn the language as part of the evangelization efforts in order to be able to communicate with Mixtecs in their own language for the purposes of conversion. In this same year a grammar was published by Fray Antonio de los Reyes (also of the Teposcolula - Yucu Ndaa variety)⁴.

There are several inter-related potential uses of the output of this endeavour⁵ for philological, linguistic, anthropological purposes including: 1) the creation of a machine searchable data set for the study of the Yucu Ndaa variety itself, and/or the historiographical and philological issues related to the collection and specifics of the vocabulary collected; 2) creating an open, highly structured resource for other Mixtecan lexical projects; 3) combining the first two to potentially create a more cohesive body of pan-Mixtecan resources and a set of vocabulary for cross Mixtecan comparison; 4) the TEI format can easily be exported into other formats (e.g. tab separated plain text, etc.) for non-TEI users, i.e. the format is fully extensible.

And in line with the above, this endeavour was undertaken in order to integrate the contents of this historical resource into a TEI-based language documentation project dealing with Mixtepec-Mixtec (ISO 639-3: mix) (Bowers & Romary, 2017, 2018a, 2018b), an under-resourced indigenous language native to the Juxtlahuaca district of

³ In the present day, there are dozens of Mixtec varieties with different levels of mutual intelligibility, estimates range from 52 (Simons & Fennig, 2018) to 85 distinct varieties (Instituto Nacional de Lenguas Indígenas, 2015).

⁴ Both the source of the document (Jansen & Perez Jiménez, 2009) and Mesolore provide excellent overviews of issues relevant to the study and understanding of the contents of the vocabulary and thus those seeking a more extensive description thereof, should consult these studies.

⁵ Note these benefits discussed are on top of the essential work done by Jansen and Perez Jiménez (2009) who made the resource much more user-friendly in implementing a number of normalizations, altering the entries to Mixtec -> Spanish, provided an indepth discussion of the source and its context, and provided a vision of the resource as a potential basis for pan-Mixtecan etymological and philological comparison. We share this vision and assert that the application of TEI enables the use of the resource as a machine and human readable database.

Oaxaca Mexico⁶. Mixtepec-Mixtec (spoken in the Juxtlahuaca district of Oaxaca) like Teposcolula is in the “Mixteco Alto” region, and the linguistic relation between modern Mixtepec-Mixtec and the historical variety Yucu Ndaa is quite clear in a significant portion of the vocabulary.

2. OCR technology and indigenous language dictionaries

In recent years there have been growing efforts to apply OCR to digitize indigenous language resources, which is increasingly necessary as language communities are seeking to make the limited materials they have more widely available and to avoid situations where paper copies of content are not only inaccessible but at risk of complete loss if physical copies fall victim to any number of potential man-made or natural disasters.

Maxwell and Bills (2017) discuss the application of OCR methods in creating a structured, machine readable XML lexicon for indigenous language resources, including Tzeltal-English, Muinane-Spanish and Cubeo-Spanish dictionaries. Additionally, Ranaivo-Malançon et al. (2017) discuss the conversion of Melanau-Mukah-Malay and Iban-Malay indigenous language dictionaries from PDF sources into HTML files, which were then parsed using a Python HTMLParser to extract the dictionary content to be saved as comma-separated plain text files.

More advanced approaches using machine learning techniques have been seen since in recent years. The most successful one that showed enough potential for scalability and adaptation is the cascading parsing of print dictionaries implemented in GROBID-Dictionaries (Khemakhem et al., 2017; Khemakhem et al., 2018a; Khemakhem et al., 2018b; Khemakhem et al., 2018c). The technique is based on Conditional Random Fields (CRF) (Lavergne et al., 2010) which allow, along with dedicated libraries for manipulating PDF documents, the end-to-end extraction of lexical structures into TEI compliant resources. The extensibility of GROBID-Dictionaries, along with being language agnostic, have motivated our present work to speed up the process of building a structured resource for the historical Mixtec language resource.

2.1 Different versions of the resource

In both the automatic structuring process used to create the TEI dictionary and in the specifics of the content, the history of the organization of the lexical resource plays a significant role. While the original dictionary created by Alvarado was Castilian - Mixtec, the version by Jansen and Perez Jiménez (2009) was transformed to be Mixtec - Castilian. Below is an example of the original Castilian - Mixtec entry structure taken from the PDF version with the original structure created by Mesolore. Not only is this

⁶ Mixtepec-Mixtec is an Otomonguean language spoken by roughly 9,000 – 10,000 people, and in addition to the native communities in Mexico, it is also spoken by communities of several people living in California, Oregon, Washington, Florida and Arkansas in the United States.

lexicon Castilian based, but it is organized in such a way that an entry often contains multiple Mixtec forms, has unclear indicators of grammatical information, the components of the Mixtec items are not appropriately delimited, in some cases they were not consistently spelled, and finally in many cases the Mixtec forms had other senses that were placed in separate entries. The original content was thus not a user-friendly resource.

Aceptar persona. Yodzacainuundi
yositoninondita, f. coto, yotniño
nuundita, yonaquai nuundita,
yonaquaicahandisita.

Figure 1: Dictionary structure prior to the restructuring of Jansen and Perez Jiménez (2009).

Jansen and Perez Jiménez (2009) split up the contents of this into five separate entries and applied several normalizations to the orthographic representation to produce a more uniform convention. These changes both improve the organization of the Mixtec content and more clearly reflect the linguistic structure. The results of which are shown below⁷:

yodza cay noondi: deshollejar; abajar la cabeza para mirar algo profundo;
acceptar persona; anillo poner en el dedo; echar los ojos en algo; inclinarse
bajando la cabeza para mirar hacia abajo; poner los ojos en algo para hurtarlo;
poner los ojos en algo que parece bien
yosito ninondita, futuro coto: aceptar persona
yotniño nuundita: aceptar persona
yona quay nuundita: aceptar persona
yona quay cahandi sita: aceptar persona

Figure 2: Revised version of the entry shown above, separated into four separate entries in Jansen and Perez Jiménez (2009)

The changes made in the aforementioned source, particularly the use of bold type for the Mixtec forms, the addition of a colon “:”, semi-colon “;”, and comma “,” delimiters between form and sense, different senses, and separate glosses in a single sense all rendered the contents much more amenable to the application of GROBID, as the contents of entries are much more clearly demarcated. Issues specific to GROBID will be discussed in more depth in the following section.

⁷Note in Figure 2, the Spanish gloss of the entry **yodza cay noondi** contains content that was not in the original shown in Figure 1; this was apparently taken from elsewhere in the in the original dictionary as *yodza cay noondi* has been given as the gloss for multiple Spanish terms. One of the major achievements of Jansen and Perez Jiménez (2009) was the consolidation of this information.

3. GROBID-Dictionaries

3.1 System overview

GROBID-Dictionaries (Khemakhem et al., 2017; Khemakhem et al., 2018a) is a machine learning infrastructure for parsing and structuring digital dictionaries based on CRF models (Lavergne et al., 2010). The infrastructure has been tested with digitized and born digital dictionaries in several languages, and is still under development. In the following section, we present the part of the tool’s up-to-date architecture reflecting the logical (lexicographic) structure of the present dictionary.

3.2 Cascading CRF models for lexical information parsing

The lexical information extraction in GROBID-Dictionaries relies on a cascade parsing of the structures in an input dictionary. At each parsing level a CRF model, being trained on samples of the target dictionary, has the goal of predicting a set of labels representing TEI structures. In Figure 3 we present the architecture of different models and labels recognized by the system in the case of the present dictionary.

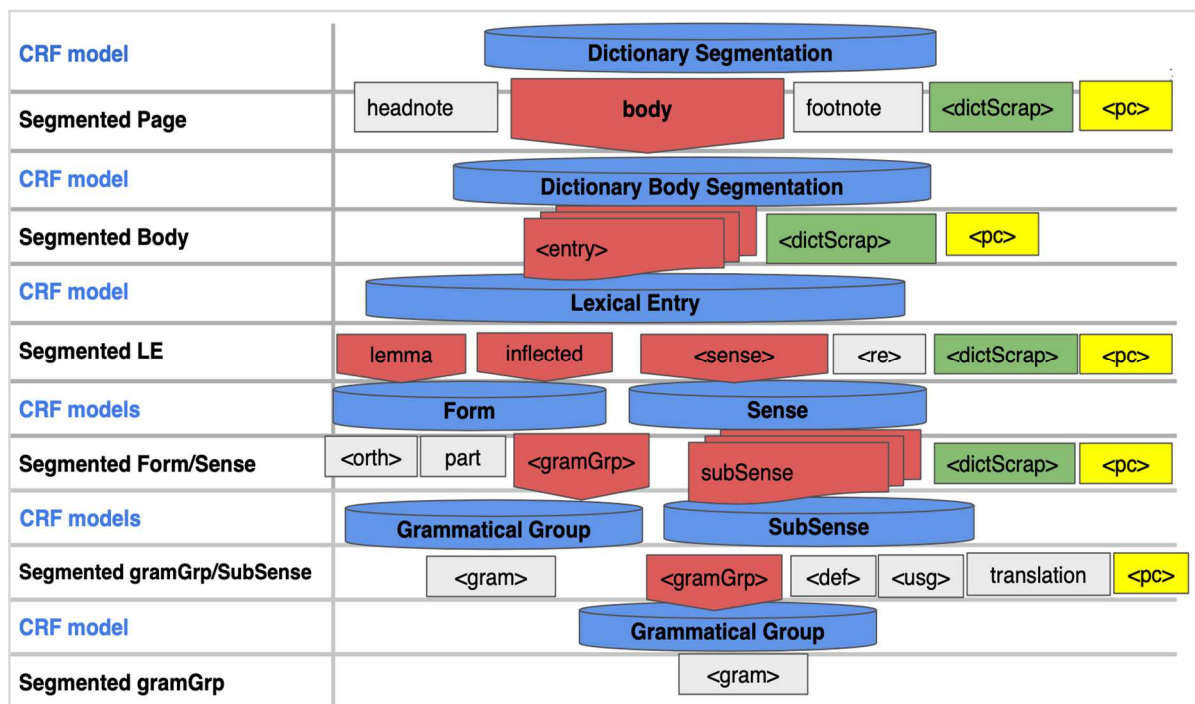


Figure 3: Parts of GROBID-Dictionaries’ architecture activated for parsing the Mixtec dictionary

As a reminder, a text cluster recognized by a CRF model could be either directly wrapped into a valid TEI structure – represented in Figure 3 with angle brackets – or into a pivot XML element – represented in Figure 3 without brackets. Pivot elements are implemented when a TEI construct is typed, such as the `<form>` element that could be typed with either “lemma” or “inflected”, or as for definitions which are serialized in TEI using `<def>` construct. We have used pivot elements just for the training stage, which are then rendered in the final output as a valid TEI construct. `<pc>` is present at almost all segmentation levels, as marking up such information is useful for the machine learning model to learn field limits in a continuous sequence of tokens. A simple find/replace post-processing can remove such valid TEI tags if needed.

Compared to what has been already achieved in Khemakhem et al. (2018a), several improvements have been carried out to cover more lexical features encountered both in this dictionary as well as in other samples of similar lexical description depth:

1. Forms in lexical entries are differentiated into lemma and inflected.
2. The form model parses morphological and grammatical information of different forms, replacing the old model which was designed to extract the main information related to the lemma .
3. After being extracted and segmented into sub-senses, if semantic nesting needs to be reflected then senses can be parsed by a SubSense model to recognize definitions, usage, grammatical information and translation equivalents.

3.3 Experiment

In training the different models required in the architecture we have encountered several challenges related to the logical and physical (typographic) structure of the dictionary. We detail in this section the major obstacles, the implemented solutions, the impact on the annotation process and the results of the experiment.

3.4 Automatic parsing: features and challenges

The logical structure of the dictionary has been affected by the fact that dictionary had been re-compiled from an earlier version, which, as mentioned above, greatly improved the quality and organization of the resource in many ways. While the dictionary looks fairly simple in structure, due to a mixture of issues related to the original vocabulary collection in combination with some conventions in the updated formatting which are not clearly specified by Jansen and Jiménez Perez (2009), it contains some complexities which pose some serious obstacles to parsing, and most of these features are described below.

3.4.1 Forms and related entries

While thanks to the revisions by Jansen and Jiménez Perez (2009) the form section is nicely delimited from the sense by the use of the bold type, there are nonetheless quite a few different features present in that section with unique conventions for demarcation.

The most common supplemental feature in the forms is the inclusion of an inflected form (which is only a single part of the verb phrase), which is delimited by the combination of a comma, followed by the grammatical feature in italic type.

yosico ini tnahandi, futuro cuico: aficionados estar dos

Figure 4: Entry with inflected (future) form *cuico*

There are several different conventions used in related entries and variant forms, none of which have enough instances sufficient for automatic recognition and structuring:

huau ndaha / saha: artejo

Figure 5: Entry for “knuckle” specific to “hand” *ndaha* or “foot”

In the entry *huau ndaha/ saha* the form *ndaha* is “hand” while *saha* is “foot”, thus the content is a related entry and is only part of the full form of the second lexical entry (as the lexical items for knuckles in Classical Mixtec are equivalent to “hand knuckles” and “foot knuckles”). In the entry below, it appears that there are alternate terms which translate into Spanish as *en buen tiempo* (“in good time”) and these alternative phrases are separated by the first comma with the grammatical feature (either verbal tense or mood) preceding the inflected form.

quevui iñe huaha, quevui iñe huii, futuro cuiñe: en buen tiempo

Figure 6: Entry with variant term for part of phrase *huaha ~ huii*.

In only a few instances, where the entry itself is an inflected form of another entry this information is stated in square brackets within the form (bold) portion of the entry. However, this content is mixed between the feature (below *imperativo*), Spanish translation, and then the Mixtec form (below *yosa cahindi*). These instances are actually duplicates of existing entries.

qua cahi [imperativo de yosa cahindi]: ir por algo generalmente

Figure 7: Entry whose form is an inflected form of a separate (related) entry.

3.4.2 Sense information

In entries with multiple gloss-like definitions but which are to be considered a single sense, commas separate the contents⁸:

quevui yahui: feria, mercado

Figure 8: Entry with two glosses of a single sense.

In some cases, the entry is divided into multiple senses (which themselves have one or more gloss), these separate senses are delimited by semicolons.

ñuhu nisitu: cavada tierra; labrada tierra

Figure 9: Entry with two distinct senses.

There are cases of exceptions to these, for instance, while a comma usually delimits different glosses, in a few examples one is used in a normal grammatical way, delimiting clauses. In the following example, the definition is *fofa cosa* “soft thing” and the content after the comma states “such as dirt”.

ñuhu tisaha: fofa cosa, como tierra

Figure 10: Entry showing a comma delimiting separate glosses of same sense.

3.4.3 Usage, etymology grammatical information

In many entries there is supplemental information about the sense given by the original author which generally specifies some aspect of the usage. This is represented in the Jansen and Jiménez Perez (2009) version in round brackets.

ama: bien está (otorgando); sí

Figure 11: Example of usage information in sense.

Likewise, there are some entries with supplemental information which is grammatical in nature, and this is also placed in round brackets but is distinguished from the usage information with italics.

amana: ¿cuándo? (*adverbio interrogativo*), ¿en qué tiempo?

Figure 12: Example of grammatical information in sense

⁸ Despite the structuring, there are many cases in which the use of the sense delimiter “;” does not seem to delimit strictly distinct senses.

However, there are certain cases in which there is grammatical information as well as a translation in the round brackets. Though the structure is distinct, in that within the brackets the grammar information is in italics delimited by a colon and the Spanish translation is to the right of the colon, there are not enough instances to train the system to automatically recognize this.

ca nayndo saha qhundo: llevar alguna cosa (*imperativo*: llevarás esto)

Figure 13: Example of grammatical information and translation of inflected form in sense

In some entries Jansen and Jiménez Perez (2009) added notes of where the sense is metaphorical in nature, these also are represented in round brackets within the sense section. The number of these instances is also not sufficient for the system to recognize and structure this content.

ña tuvui nini dzavua yuqua iyondi: vivir pobre (por metáfora); pobre estar

Figure 14: Example of metaphor specified in sense information

3.4.4 Other issues

Hyphenated content which is present in the source due to line breaks and additional varied use of brackets for various lexical content are also present in the data source, and contain too few instances to provide enough training data annotations for ML to create the desired output. Such content (as well as that mentioned above which lacks sufficient quantities for successful training) are structured in the TEI output either manually, semi-manually, or automatically using XSLT, much of which will be described in a later section.

3.5 Annotation

Covering instances of all the aforementioned observations in a few pages is a very hard task for the annotation. And given the multi-stage annotations, where the annotation is focused at each level on marking up all possible variations of specific structures, the number of pages required to be annotated can grow exponentially.

3.6 Page sampling process

As a random sampling was not an option in the case of this dictionary, we tried to cover the variation of logical and physical structures by selecting pages that represent the maximum number of challenges. We selected just a few pages containing related entries as they are sparsely distributed, and we also had to give up the annotation of

some structures, such as “morphological variants”, given their low number and inconsistent typographic representation. More useful information about language, comparison, transcription, etc. can be found in the prose section, which we decided to ignore in the scope of this experiment.

We have selected and annotating 14 pages from different parts of the dictionary: 10 for training and four for evaluation. We detail the annotated instances for each model, except for the first one dealing with the prediction of the main regions of a page, which has less lexical importance with regard to the scope of this work, in Table 1.

<i>Model</i>	Training	Evaluation
Dictionary Body Segmentation	572 <entry>	270 <entry>
Lexical Entry	572 <sense> 572 <lemma> 28 <inflected> 10 <re>	269 <sense> 270 <lemma> 10 <inflected> 4 <re>
Sense	856 <subSense>	302 <subSense>
Form	787 <orth> 31 <part> 31 <gramGrp>	269 <orth> 11 <part> 11 <gramGrp>
SubSense	905 <def> 32 <usg> 7 <gramGrp> 9 <translation>	319 <def> 11 <usg> 8 <gramGrp> 2 <translation>

Table 1: Page Sampling Statistics.

3.7 Results and discussion

The results of the first two models, **Dictionary Segmentation** and **Dictionary Body Segmentation** were almost perfect, with an above 98 F1 score. In the following table we detail the performance of the rest of the models on the field level, in which the evaluation takes into consideration the prediction of all the tokens of field and not only single tokens. We do not show the evaluation of the **Grammatical Group** model as it has only one label.

<i>Model</i>	Label	Precision	Recall	F1
Lexical Entry	<inflected>	90	90	90
	<lemma>	99.26	99.26	99.26
	<pc>	98.94	99.29	99.12
	<sense>	100	100	100
	<re>	0	0	0
Sense	<subSense>	100	100	100
	<pc>	100	100	100
Form	<gramGrp>	100	90.91	95.24
	<orth>	98.18	100	99.08
	<part>	70	63.64	66.67
SubSense	<def>	91.84	95.3	93.54
	<gramGrp>	100	25	40
	<pc>	76.81	88.33	82.17
	<translation>	100	100	100
	<usg>	60	90	72

Table 2: Field Level Evaluation of the Lexical Models.

The evaluation shows the high performance of the models in predicting lexical structures with the exception of related entries, grammatical information, sense usages within sense and orthography of inflected forms.

In the case of related entries, the training and evaluation datasets combined contain just 32 instances representing two logical representations (collocates and non-collocates) and four physical variations (with/without brackets and with/without commas). We consider the quantity of instances used for training is not sufficient for the **Lexical Entry** model to learn the distribution of such a structure.

For usage and grammatical information blocks, both structures are represented within senses as textual sequences wrapped in a round brackets. The only evident physical difference is the italics used to mark the grammatical information. An in-depth

investigation has shown that such a visual variation has not been translated consistently in the layout information associated with each token of the document and extracted by the PDF utilities libraries in GROBID-Dictionaries. Therefore, the **SubSense** model remains unable to differentiate these two physically similar structures. In the case of the `<part>` label, more annotated instances seem to be needed in the training dataset to strengthen the predictions of the **Form** model.

4. TEI structure of output

Because this resource is being converted to TEI in order to be integrated with the TEI-based project on the contemporary Mixtepec-Mixtec variety⁹, the Classical Mixtec dictionary structure is designed to match the former as much as possible. The exception to this is that due to the inexact nature of the Spanish glossing the default element containing the Spanish is the definition element `<def>`, whereas in the Mixtepec-Mixtec TEI dictionary they are represented as pure translations.

4.1 Basic entry structure

```

<entry xml:id="fruit-plantain">
  <form type="lemma">
    <orth xml:lang="mix">nchika</orth>
    <pron xml:lang="mix" notation="ipa">ndʒiká</pron>
  </form>
  ....
  <sense corresp="http://dbpedia.org/resource/Plantain">
    ....
    <cit type="translation">
      <form>
        <orth xml:lang="en">plantain</orth>
      </form>
    </cit>
    <cit type="translation">
      <form>
        <orth xml:lang="es">plátano</orth>
      </form>
    </cit>
  </sense>
</entry>

```

```

<entry>
  <form type="lemma">
    <orth>chita</orth>
  </form>
  <sense>
    <def>plátano</def>
  </sense>
</entry>

```

Figure 15: Left, partial TEI dictionary entry for *nchika* ‘plantain’ in Mixtepec-Mixtec; right, view of (unenhanced) structure of the historically related form in Classical Mixtec.

Note that while the ISO 639-3 language code is applied to the Mixtepec-Mixtec entry and the Spanish 639-2 tag is applied to the translations of the Classical Mixtec entry, there are no ISO or other any other standardized language codes for ‘Classical Mixtec’, nor is there any documented modern Mixtec variety attributed to Teposcolula.

⁹ For an in depth detailing of the structure and content in the Mixtepec-Mixtec TEI dictionary, see Bowers and Romary (2018a)

4.2 Inflected forms

The entries with inflected forms are shown below in the TEI output. Note that since these are mostly multi-word expressions verb phrases.

yosico ini tnahandi, futuro cuico: aficionados estar dos

Figure 16: Entry with inflected form

```
<form type="lemma">
  <orth>yosico ini tnahandi</orth>
</form>
<pc>,</pc>
<form type="inflected">
  <gramGrp>
    <gram>futuro</gram>
  </gramGrp>
  <orth extent="part">cuico</orth>
</form>
```

Figure 17: TEI encoding of entry with inflected form

4.3 Senses

Entries with multiple senses and multiple glosses/definitions were generally handled well by GROBID, examples of the output of each case (unenhanced) are shown below with the source from the PDF entry above.

ñuhu nisitu: cavada tierra; labrada tierra

```
<entry>
  <form type="lemma">
    <orth>ñuhu nisitu</orth>
  </form>
  <pc>:</pc>

  <sense>
    <def>cavada tierra</def>
  </sense>
  <sense>
    <def>labrada tierra</def>
  </sense>
</entry>
```

ñuhu tisaha: fofa cosa, como tierra

```
<entry>
  <form type="lemma">
    <orth>ñuhu tisaha</orth>
  </form>
  <pc>:</pc>

  <sense>
    <def>fofa cosa</def>
    <def>como tierra</def>
  </sense>
</entry>
```

Figure 18: TEI encoding of entries with multiple senses and multiple definitions.

5. Post-editing: Modifications and enhancements

In terms of the source to target structure, the GROBID process was able to create a conversion of the PDF form of the resource into TEI which represented the majority of the features present in the dictionary. However, in the case of several features, further manual and semi-manual encoding enhancements were necessary in order to create a more dynamic and refined structure. These modifications were necessary due to either: a lack of sufficient tokens required for the machine learning process or to make it more compatible with the Mixtepec-Mixtec TEI corpus. These changes are described in this section.

Other key enhancements made to the output include the following:

- Spanish ISO 639-2 language tag added to all <def> elements
- Unique id's (@xml:id) are added to each <entry> and <re> which are based on the Spanish value (with underscores and token numbers added as needed)
- English translations are being added according to certain categories (at least for those which are sufficiently clear, as not all items can easily be translated)
- Domain tag (<usg type="domain">) is added in certain entries (some of the vocabulary in the source which are initially given <usg type="hint"> can be changed to domain)
- Records of normalizations and assumed phonetic equivalencies made by Jansen and Jiménez Perez (2009) are manually added in the header.

Below we discuss the formatting of content which was not sufficiently structured by GROBID and/or which needed additional structuring to bring it in line with best practice in TEI. In the examples we show both the output of GROBID and the revised TEI structure.

5.1 Related entries

In most cases related entries were correctly identified as such by GROBID, however because there are a number of different types of related entries most of which lack sufficient instances to train the system automatically recognize and encode then in detail, these items are manually refined in TEI.

tay huasi cana / cay idzi yuhu: mozo que comien-za a barbar

Figure 19: Entry with related entry in source

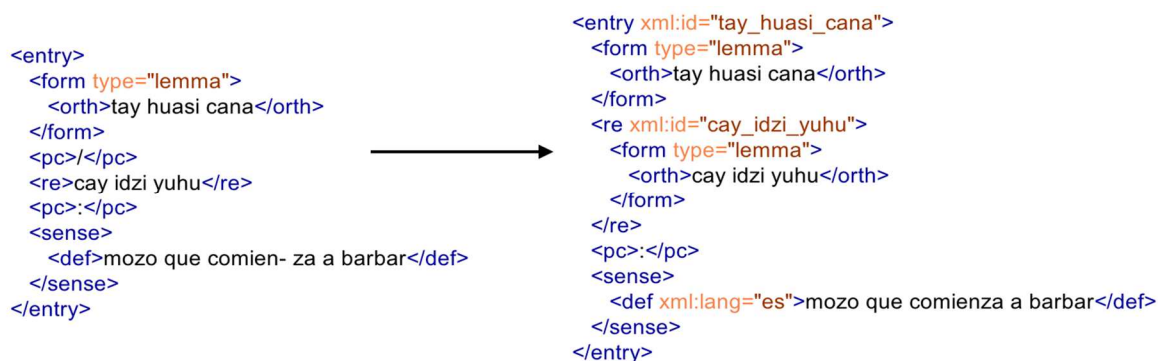


Figure 20: GROBID output (left) with revised TEI structure of form with related entry (right)

5.2 Collocate phrases in the form

In a small number of cases, there is collocate information included in the form. In TEI this is encoded using the <colloc> element.

caa ndodzo ninondi (nuu sito): echado estar (en la cama)

Figure 21: Collocate of headword in source

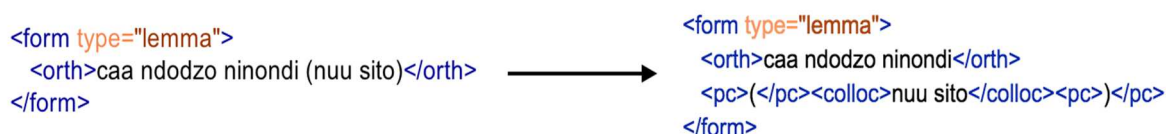


Figure 22: GROBID output (left) with revised TEI structure of form with collocate (right)

5.3 Modern Spanish Translations

There were a number of modernized Spanish translations added by Jansen and Jiménez Perez (2009) which were placed in square brackets.

da queyeni: aprisa; incontinenti [luego]; y luego; luego a la hora; temprano

Figure 23: Entry with modernized Spanish translation in source

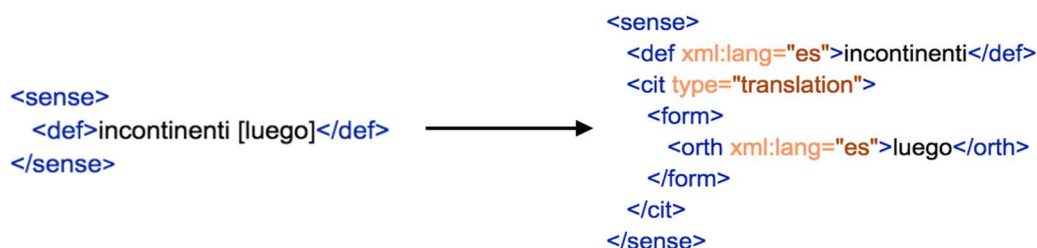


Figure 24: GROBID output (left) with revised TEI structure of modernized Spanish translation (right)

5.4 Inflected forms

In certain cases, even though the source did not have a given feature explicitly labelled, what they did include could be used to infer this, and then key features added in order to enhance the content and bring it in line with general lexicographic practice. One area where this was possible is where there were inflected forms.

yosico ini tnahandi, futuro cuico: aficionados estar dos

Figure 25: Entry with inflected form in source

In these entries with the feature *futuro*, there are two inferable features: first, that the entry is a phrase, (which mostly do not seem to have had simple lexicalized items in Mixtec), second, that given that the feature *futuro* is a feature of tense, that the part of speech verb can be inferred. The example below shows how these features are represented in the revised TEI structure by adding the @type="phrase" to <entry>, adding <pos>verb</pos> to the entry level, and by changing the generic <gram> to <tns> in the inflected future form¹⁰. This enhancement process is done using XSLT.

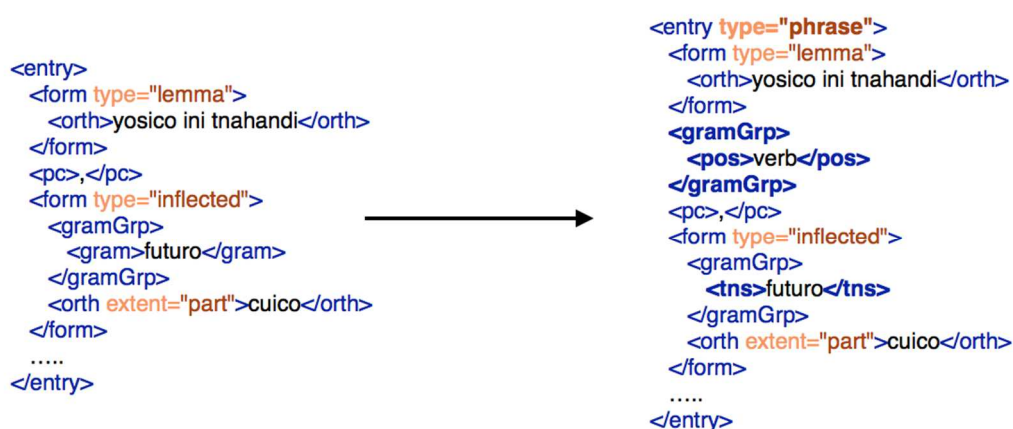


Figure 26: GROBID output (left) with revised TEI structure of phrasal entry with inflected form (right)

¹⁰ The reason that we did not train GROBID to automatically annotate the feature *futuro* as <tns> is that there are some instances of imperative forms listed in the same way. Thus, each of these features is further treated using XSLT specifically targeted, with the imperative forms being output in TEI as: <gram type="mood">.

5.5 Addition of original prologue

As there were different PDF versions created of this resource, some of them included content from the original that were not included in the others. Notably, in the version of Jansen and Perez Jiménez (2009) the original prologue content published in the original was not included; this content was thus added manually and is easily represented in the TEI output.

<p>YO FRAY Gabriel de Sancto Ioseph, Prior Prouincial desta Prouincia de Sanctiago de la Nueva España Ordinis Predicatorum. Auiendo visto el examen, y aprouacion del Vocabulario Misteco, hecho por los Padres de aquella nacion, aquiennes por mi fue cometido . Y siendo vtil y prouechoso como consta por la dicha aprouacion, por la presente doy licencia al Padre Fray Francisco de Aluarado, Vicario de Tamaçulapa : para que pueda imprimir el dicho Vocabulario: con las censuras y notas de los dichos examinadores, juntamente con el Arte que de la dicha lengua Misteca compuso el Padre Fray Antonio de los Reyes , Vicario de Tepusculula . En fee de lo qual di las presentes letras firmadas de mi nombre, y selladas con el sello menor de mi officio.</p> <p style="text-align: right;">Fray Gabrl el de S. Ioseph Prouincial.</p>	<pre><div> <ab xml:lang="es">YO <persName>FRAY Gabriel de Sancto Ioseph</persName>, Prior Prouincial desta Prouincia de Sanctiago de la Nueva España Ordinis Predicatorum. Auiendo visto el examen, y aprouacion del Vocabulario <lang>Misteco</lang>, hecho por los Padres de aquella nacion, aquiennes por mi fue cometido. Y siendo vtil y prouechoso como consta por la dicha aprouacion, por la presente doy licencia al <persName>Padre Fray Francisco de Aluarado</persName>, Vicario de <placeName>Tamaçulapa</placeName> : para que pueda imprimir el dicho Vocabulario: con las censuras y notas de los dichos examinadores, juntamente con el Arte que de la dicha lengua <lang>Misteca</lang> compuso el Padre <persName>Fray Antonio de los Reyes</persName>, Vicario de <placeName>Tepusculula</placeName> . En fee de lo qual di las presentes letras firmadas de mi nombre, y selladas con el sello menor de mi officio.</ab> <signed> <persName>Fray Gabriel de S. Ioseph Prouincial.</persName> </signed> </div></pre>
--	--

Figure 27: Left a PDF version of part of the original prologue and right its TEI encoding.

5.6 Etymology

In the Jansen and Perez Jiménez (2009) source there are roughly 70 instances which are labelled as being metaphorical in nature. These are labelled as follows:

yosa ndehe ichi: fenecer, acabar el que muere (por metáfora)

Figure 28: Entry for metaphorical term ‘yosa ndehe ichi’ as formatted in the source PDF.

Due to a lack of sufficient quantity for training, these items had to be manually identified and annotated as follows:

```
<entry xml:id="yosa_ndehe_ichi">
  <form type="lemma">
    <orth>yosa ndehe ichi</orth>
  </form>
  <sense>
    <def xml:lang="es">fenecer</def>
    <def xml:lang="es">acabar el que muere</def>
  </sense>
  <etym type="metaphor">
    <seg type="desc">por metáfora</seg>
    <cit type="etymon">
      <form>
        <orth>ichi</orth>
      </form>
      <def xml:lang="es">camino</def>
    </cit>
  </etym>
</entry>
```

Figure 29: TEI (partially enhanced) encoded entry for metaphorical term ‘yosa ndehe ichi’

While at present we do not have enough of the Classical Mixtec language to provide full analyses of the majority of the instances of metaphor, this information is nonetheless encoded in the TEI structure as per the recommendations of Bowers and Romary (2016), and Bowers et al. (2018). A partial structured analysis is provided for the phrase “yosa ndehe ichi”, of which only the portion *ichi* ‘path’ is discernible and which is represented as an etymon within the <etym> block in TEI. At a later stage, researchers who are more familiar with the language can enhance this content as needed.

6. Later steps

A logical and needed future aim would be to create a searchable TEI version of the grammar of the language published in 1593 *Arte en Lengua Mixteca* by Fray Antonio de los Reyes. Given that this resource is a grammar and not a dictionary-like text, this would not be a job for GROBID but another, general OCR tool. The text in the PDF available is of low quality, and it is likely that significant manual work would be necessary to carry out this task.

Furthermore, according to Mesolore, much of the Alvarado Classical Mixtec vocabulary was based on entries in the ‘Molina Vocabulario’ Castilian-Nahuatl dictionary (1571), a Castilian-Zapotec dictionary compiled by Juan de Cordova in the Valley of Oaxaca (1578), and Antoni de Nebrija’s Castilian-Latin Dictionarium (1553). Thus, many of these resources have common content and it would be a natural and beneficial next step to create TEI versions of these to expand all of the benefits described in this work with regard to the current Classical Mixtec vocabulary to these other indigenous languages.

7. Conclusion

This project has shown that GROBID can handle the vast majority of the work needed to create a highly structured TEI dictionary from PDF resources. However due to certain issues pertaining to the source document used, its structure and the sample size of certain structures, significant further manual and semi-manual work is required in creating a maximally representative version of the content. Given the richness of the resource, in order to effectively achieve these enhancements it is essential that they are carried out by humans who understand certain details that are only accessible through detailed study.

Converting this resource into TEI brings the data into a highly structured extensible machine-readable format which can be systematically searched, extracted and exported into other data formats using simple XQuery and/or XSLT.

In creating this iteration of the historical resource, we have continued the work of previous scholars (specifically Jansen and Perez Jiménez, 2009) who worked to make this resource available to researchers and Mixtec communities. As this work was carried

out in order to integrate the important resource into an ongoing linguistic and lexicographic project dealing with the Mixtepec-Mixtec variety, we hope to demonstrate how the Alvarado resource can be used as both an etymological and comparative cross-reference between different varieties of Mixtec as well as how TEI is a highly beneficial data format.

8. References

- Alvarado, F. de. (1593). *Vocabulario en Lengua Mixteca. Hecho por los Padres de la Orden de Predicadores*. En México: Con Licencia, en casa de Pedro Balli.
- Bowers, J. & Romary, L. (2016). Deep Encoding of Etymological Information in TEI. *Journal of the Text Encoding Initiative*, (Issue 10). <https://doi.org/10.4000/jtei.1643>
- Bowers, J. & Romary, L. (2017). Language Documentation and Standards in Digital Humanities: TEI and the Documentation of Mixtepec-Mixtec. In A. Kawase (ed.) *Proceedings of the 7th Conference of Japanese Association for Digital Humanities*. Kyoto, Japan: Doshisha University, pp. 21–23.
- Bowers, J. & Romary, L. (2018a). Bridging the gaps between digital humanities, lexicography and linguistics: a TEI dictionary for the documentation of Mixtepec-Mixtec. *Dictionaries: Journal of the Dictionary Society of North America*, 39(2).
- Bowers, J. & Romary, L. (2018b). Encoding Mixtepec-Mixtec Etymology in TEI. *Presented at the TEI Conference and Members Meeting*, Tokyo, Japan.
- Jansen, M. E. R. G. N., & Pérez Jiménez, G. A. (2009). *Voces del Dzaha Dzavui (mixteco clásico). Análisis y Conversión del Vocabulario de fray Francisco de Alvarado (1593). Colegio Superior Para La Educación Integral Intercultural de Oaxaca*.
- Khemakhem, M., Foppiano, L. & Romary, L. (2017). Automatic extraction of TEI structures in digitized lexical resources using conditional random fields. In I. Kosem et al. (eds.) *Proceedings of eLex 2017 conference, Netherlands, Leiden*. Brno: Lexical Computing Ltd., pp. 598–613.
- Khemakhem, M., Herold, A. & Romary, L. (2018a). Enhancing Usability for Automatically Structuring Digitised Dictionaries. *GLOBALEX Workshop at LREC 2018*. Presented at Miyazaki, Japan. Retrieved from <hal-01708137v2>
- Khemakhem, M., Romary, L., Gabay, S., Bohbot, H., Frontini, F. & Luxardo, G. (2018b). Automatically Encoding Encyclopedic-like Resources in TEI. In *The annual TEI Conference and Members Meeting*.
- Khemakhem, M., Romary, L., Gabay, S., Bohbot, H., Frontini, F. & Luxardo, G. (2018c). Automatically Encoding Encyclopedic-like Resources in TEI. In *The annual TEI Conference and Members Meeting*.
- Lavergne, T., Cappé, O. & Yvon, F. (2010). Practical very large scale CRFs. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 504–513.
- Maxwell, M. & Bills, A. (2017). Endangered data for endangered languages: Digitizing

- print dictionaries. *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pp. 85–91.
- Mesolore. Accessed at: <http://www.mesolore.org/tutorials/learn/9/Introduction-to-the-Alvarado-Vocabulario> (24 April 2019)
- Proyecto de indicadores sociolingüísticos de las lenguas indígenas nacionales*. (2015). Accessed at: http://site.inali.gob.mx/Micrositios/estadistica_basica/estadisticas2015/estadisticas2015.html (5 July 2017)
- Ranaivo-Malançon, B., Sae, S., Othman, R. M. & Busu, J. F. W. (2017). Transforming Semi-Structured Indigenous Dictionary into Machine-Readable Dictionary. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, 9(3–11), pp. 7–11.
- Reyes, A. de los. (1593). *ARTE EN LENGVA MIXTECA COMPUESTA*. Accessed at: https://books.google.at/books?hl=en&lr=&id=Vbh9oGk-YKwC&oi=fnd&pg=PA4&dq=%22Arte+en+Lengua+Mixteca%22+Antonio+de+los+Reyes&ots=yPfYu574TP&sig=S3ro_9u4ZAp1-7wGnPVoA4WPS4U (15 April 2019)

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



TEI Lex-0 In Action: Improving the Encoding of the Dictionary of the *Academia das Ciências de Lisboa*

Ana Salgado¹, Rute Costa¹, Toma Tasovac², Alberto Simões³

¹ NOVA CLUNL, Universidade NOVA de Lisboa

² Belgrade Center for Digital Humanities, Serbia

³ 2Ai – Instituto Politécnico do Cávado e do Ave / Algoritmi, Universidade do Minho

E-mail: anasalgado@campus.fcsh.unl.pt; rute.costa@fcsh.unl.pt; ttasovac@humanistika.org;
asimoes@ipca.pt

Abstract

This paper describes some experiments made while encoding the first complete dictionary of the *Academia das Ciências de Lisboa* (DACL) in the context of TEI Lex-0, a community-based interchange format for lexical data aimed at facilitating the interoperability and reusability of lexical resources. Even though the original encoding of the DACL was based on TEI, we decided to switch to TEI Lex-0 because it allowed us to streamline our encoding. Our experiments show that even though TEI Lex-0 is stricter than TEI itself (allowing fewer elements and imposing certain constraints that are not present in plain TEI), it is fully capable of representing the complexities of the entry structure of the DACL. In the paper, we discuss the TEI Lex-0 encoding of the DACL, as well as the conversion methodology and the tools used for the automatic conversion from the original encoding. We are currently focusing on the macrostructural level, more precisely on the types of lexical units and on the written and spoken forms of the lemma, providing a set of modelling principles and representation forms of every type of entry in the DACL. This paper is part of ongoing work and a contribution to the efforts of the DARIAH-ERIC Lexical Resources working group.

Keywords: dictionary encoding; lexicography; TEI; XML; TEI Lex-0

1. Introduction

The digital revolution has transformed the way we conceptualize, plan and implement lexicographic projects. While print dictionaries are slowly going out of fashion, retro-digitized and born-digital dictionaries are increasingly taking advantage of the available technologies. At the same time, however, many dictionaries continue to be designed and implemented following the typographical and editorial conventions of the print medium (Tasovac, 2010: 1). According to Trap-Jensen (2018: 34), “it is necessary that lexicographers shift their focus away from the concrete end product and towards a lexical database”.

The task of updating the first complete dictionary – from A to Z – of the *Academia das Ciências de Lisboa* (DACL), published in 2001, provides the basis for this work. Its great historical value for European lexicographical heritage and the institution’s willingness to update the content of the dictionary dictated the need to convert the print edition into digital format, with the ultimate goal of making this lexical resource available on the web and as a mobile app.

This dictionary – available in print and as a PDF document – was converted into XML using the P5 schema of the Text Encoding Initiative (TEI) (Simões et al., 2016). This process – as described in detail in Section 2 – was conducted with a formal format in mind, and therefore the group focused on the conceptual structure of the dictionary and not on its visual aspect. Nevertheless, the TEI format, although very complete and accompanied by comprehensive documentation, presented some challenges when encoding the DACL. Parts of the original structure diverged from the TEI proposed structure, which led to some adaptations of the official schema. This problem, coupled with the fact that TEI allows multiple solutions for encoding the same type of information, made us look into TEI Lex-0¹ (Romary & Tasovac, 2018), a streamlined version of the TEI standard for dictionaries. In Section 3, we discuss and compare these two standards.

Before we could work on the conversion between these two formats, we had to analyse the TEI Lex-0 schema and create maps from the original structure used in the DACL. Section 4 refers to this analysis, a contribution to the work developed by the DARIAH-ERIC Lexical Resources working group². Section 5 discusses the technological approach used to experiment with the conversion of the DACL into TEI Lex-0, trying to accommodate every change that the TEI Lex-0 working group has published. As the standard has yet to be concluded, our technological architecture is prepared to aim at a moving target, adapting the encoding as the standard evolves. Finally, Section 6 draws some conclusions from the work that has been carried out so far.

2. Dictionary of the *Academia das Ciências de Lisboa*

As previously mentioned, the first complete edition of the DACL was only published in 2001 in a two-volume paper version (the first volume from A to F, and the second from G to Z). At the time, the computational side of the project included a Microsoft Access database and a reporting tool that could generate a Word file for the dictionary, which was then manually edited and formatted before printing. Eighteen years later, the only surviving digital data source of the published dictionary remains the final PDF file. For the Portuguese Academy to move forward and produce a new edition of the dictionary³ using digital tools and structured data, the PDF file had to be reverse engineered in order to convert PDF strings and their typographic features into a

¹ <https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html>

² <https://www.dariah.eu/activities/working-groups/lexical-resources/>

³ The original digital version of the DACL is not publicly available, but the first author of this paper is the coordinator of the new digital edition. The Natural Language Processing group of the Computer Science Department of the University of Minho has been developing the technological support of the new digital edition of DACL, counting on the participation of Alberto Simões from IPCA (Instituto Politécnico do Cávado e do Ave), responsible for the technological support, José João Almeida, and the consultancy of Álvaro Iriarte Sanromán, both from University of Minho. The participation of NOVA CLUNL (Linguistic Research Center of NEW University of Lisbon) is related to its transition into the TEI LEX-0 format.

conceptual structure. A mapping from different font typefaces and font sizes was made to specific structures (e.g., phonetic transcriptions or synonyms). Because the same font typeface and font size were used for different types of information, a heuristic procedure had to be employed, taking into account string content and string order, to infer their semantics. The TEI schema was used as the target format, since it is a well-known and documented format. Nevertheless, as already stated, some specific constructions of the standard had to be changed in order to enable the encoding of some of the dictionary entries. This process was iterative and interactive, with human interaction to fix minor issues on some entries where the default behaviour was not able to correctly determine the entry structure.

To allow the quick edition of the database, the TEI dictionary was split into thousands of small XML documents (one per dictionary entry) that were imported into a native XML database (eXist-DB). Using the eXist-DB ecosystem based on XQuery, LeXmart⁴, a tool framework for lexicographic work, was developed to allow the edition, deletion and creation of new dictionary entries, as well as to validate their structure and overall dictionary coherence (Simões et al., 2016).

3. TEI Guidelines for Dictionary Encoding

The use of open formats based on standards is a crucial aspect of digital humanities initiatives. TEI is a *de facto* standard for the digital encoding of all types of written texts, ranging from novels and poetry to mathematical formulae or music notation⁵. It also defines how specific humanities resources, such as speech, morphological annotated monolingual and parallel corpora, dictionaries and other structures should be encoded.

All TEI documents must include a metadata section, named TEI header, and share a set of common annotation features, defined in the standard as the core module (Chapter 3)⁶. This set includes structural elements, such as paragraphs, lists or bibliographic references.

For dictionaries, Chapter 9⁷ of the TEI Guidelines starts by defining the structure of the dictionary as a book – front matter, body or back matter. It also describes three main elements to encode dictionary entries: `entry`, `entryFree` and `superEntry`. While the document describes precisely when each should be used (`entry` forces a structure; `entryFree` provides a flat representation and allows unstructured entries that should be avoided but may be necessary for some dictionaries; and `superEntry` as a mechanism that can group other entries, such as homonyms), this freedom makes

⁴ <http://www.lexmart.eu/>

⁵ See, e.g., Music Encoding Initiative: <https://music-encoding.org/>

⁶ Elements Available in All TEI Documents: <https://tei-c.org/Vault/P5/1.3.0/doc/tei-p5-doc/es/html/CO.html>

⁷ Dictionaries: <https://www.tei-c.org/release/doc/tei-p5-doc/en/html/DI.html>

it difficult for different authors to keep their dictionaries coherent in terms of structure. To these three tags we can add the `re` element, which allows the encoding of related entries (Bański et al., 2017: 485) and `hom` (homograph) which can be used for encoding homographs.

Flexibility is both a virtue and a shortcoming of TEI. For instance, to create cross-references, the preferred way is to use the `xr` tag. But it is also possible to create links using `anchor/ptr` or `link`. In order to reduce this freedom and define a specific format for dictionaries forcing dictionary encoders to follow the same structural rules, the lexicographic and dictionary-encoding communities are currently discussing a new format to encode dictionaries – TEI Lex-0⁸ – a fully TEI-compliant but streamlined format for facilitating interoperability.

This new format does not intend to replace the Dictionaries Chapter in the TEI Guidelines. Instead, it is framed as a target format that can help uniformize the existing heterogeneously encoded lexical resources and is currently being tested by numerous dictionaries⁹. Given the fact that it is still a work in progress, it can be changed in order to accommodate relevant dictionary structures.

a⁵ [v]. *prep.* (Do lat. *ad* 'para' ou *ab* 'de'). **A** Valores semânticos: **I.** Na expressão de valores locativos, indica: **1.** Direcção para um lugar (real ou virtual). *O navio rumou a oriente. Levar a uma situação embaraçosa. Foi a casa dos sogros. Eu apenas fui a Paris; o meu irmão é que foi para Paris.* Obs. Quando introduz um complemento do verbo *ir* ou do nome *ida*, indica que a permanência no lugar de destino é breve; inversamente o uso da preposição *para* indica permanência prolongada. **2.** Termo de um movimento. *Chegou a casa.* **3.** Afastamento. \approx **DE.** *Esquivar-se a trabalhos.* **4.** Distância medida em unidades de espaço ou tempo. *Há uma estação de comboio a quinhentos metros daqui. A minha casa fica a cinco minutos do mercado.* **5.** Localização, situação precisa ou aproximada. *Ela mora num palacete a São Bento. Pôs as cadeiras a todo o comprimento da sala.* **6.** Adjunção. *Amarrou o cão a um poste. A uma asneira seguiu-se outra.* **II.** Na expressão de valores temporais, indica: **1.** Tempo em que uma coisa acontece (pontual ou habitualmente); concomitância. *Tenho aulas a meio da tarde.* **2.** Distância. *O jogo está a dez minutos do intervalo. A cinco horas do desembarque.* **3.** Progressão para um tempo (em correlação com a *prep. de*). *De mês a mês. De cinco dias a esta parte. A exposição estará aberta ao público de Junho a Setembro.* **4.** Intervalo regular ou duração periódica. *Ele trabalhava a tempo inteiro. Há muitos contratos a prazo.* **III.** Na expressão de outros valores, indica: **1.** Causa. \approx **POR.** *Fez isso a solicitação dos parentes.* **2.** Instrumento, meio e modo. *Pintura a óleo. Navegava a todo o vapor. O móvel apresentava entalhaduras a canivete. Há quem aguente muito tempo a pão e água.* **3.** Finalidade. \approx **PARA.** *O patrão deu-lhe vinho a beber. Pôs*

Figure 1: *a* preposition (DACL).

⁸ To secure interoperability, the Working Group “Retro-digitised Dictionaries”, lead by Toma Tasovac and Vera Hildenbrandt, as part of the COST Action European Network of e-Lexicography (ENeL) started the establishment of TEI Lex-0. Then, TEI Lex-0 was taken up by the DARIAH Working Group “Lexical Resources” which is co-chaired by Laurent Romary and Toma Tasovac. Currently, the work on TEI Lex-0 is conducted by the DARIAH WG “Lexical Resources” and the H2020-funded European Lexicographic Infrastructure (ELEXIS).

⁹ TEI is the basis for a large number of current lexicographic projects, such as Nénufar, ARTFL, or VICAV.

Although we followed TEI in the DACL encoding, we could not find solutions in the Guidelines that covered all the microstructural elements of the dictionary (e.g., the entry *a*, preposition, contains different types and levels of information – grammatical, semantic, pragmatic) – which made us adapt the standard features.

Considering the example referred to above, as can be seen in Figure 1, the sections that begin with an “A” or “I.” are not actually ‘definitions’ of the headword. The information “Valores semânticos” [*Semantic values*] or “Na expressão de valores locativos [...]” [*When expressing locative values [...]*] indicates the properties of the preposition. In the original encoding we used the `def` element to encode the description and created a grouping mechanism (named `group`) that can be used recursively to create as many levels as needed.

4. TEI Lex-0 encoding of the DACL

In order to have an interoperable lexical database and aiming at dictionary content reusability, we intend to convert the DACL into TEI Lex-0 encoding, especially if it allows us to encode the complete extension of the dictionary structure without any kind of adaptation. Therefore, we present some experiments on the encoding of specific parts of the dictionary entries.

It is important to stress that the TEI Lex-0 working group is aiming at a standard that is able to encode a dictionary taking into account its structure and semantic meaning for each specific part of the entry, and not how it looks visually. While the authors agree that there may be cases where the latter approach is useful (namely for the digital preservation of ancient documents), the development of new lexical resources should take into account their own structure. This is crucial if the goal of the lexical resource is not only to be used by humans but also by Natural Language Processing algorithms.

For our experiment, we started by identifying every element in the dictionary. A typical entry includes the following elements: headword, pronunciation, usually followed by some linguistic information (e.g., part-of-speech), the different meanings, usage information, synonyms, antonyms, collocations, etymology, and notes. Examples of usage, cross-references, etc., may also be present.

In a TEI-style encoding, each of these or even other elements of an entry must be distinguished as clearly as possible.

4.1 Macrostructural level: different types of lexical items

In order to be able to define a valid approach to annotate all the entries of the dictionary, we performed an analysis of the different types of lexical units that can be headwords, so that a sample entry for each type was chosen and encoded, enabling us to understand the versatility of the standard. Thus, we first worked on the macrostructural level of the dictionary.

At an initial stage, we listed all the entries of the DACL and identified all the types of lexical units that are summarized in Figure 2: monolexical unit, polylexical unit, affix and abbreviation.

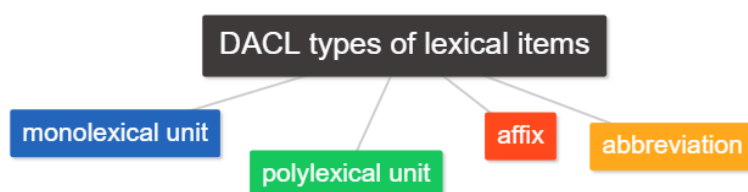


Figure 2: Formal representation of lexical entries (DACL).

In the following chapter, we will illustrate each type of lexical item found in DACL.

4.2 TEI encoding of different types of lexical items

In TEI encoding, the outermost structural level of an entry is marked with the `entry` element that begins with information about the form of the headword – `form` element – i.e., information on the written and spoken forms of one headword related to the description of its spelling and phonetics. The different types of entries are currently being marked with the attribute `type` into the `entry` element. As of this writing, there is no complete agreement within the TEI Lex-0 community on where to encode this information. Currently, as shown below, we are still adding this property to the whole entry. Nevertheless, as this is grammatical information, it should probably be encoded together with the morphologic information.

To illustrate the application of TEI Lex-0, we present the original encoding of the lemma and the conversion to TEI Lex-0 of some entries of the DACL for each of the entry types illustrated in Figure 2.

4.2.1 Monolexical units

Monolexical units can be divided into two types: lexical units, such as nouns, adjectives, verbs and grammatical units, such as conjunctions, determiners, prepositions, and pronouns.

palácio [pelásju]. *s. m.* (Do lat. *palatium*). **1.** Edifício sumptuoso, de grandes dimensões, geralmente construído num espaço urbano e destinado a residência da família real, de personalidades nobilitadas, de dignidades eclesiásticas ou altas individualidades. + *ducal, episcopal, presidencial, real*. **2.** Edifício sumptuoso, de dimensões significativas, onde se encontram sediados determinados organismos públicos. **palácio da justiça**, edifício, em cada localidade, onde funcionam os serviços judiciais e se realizam os julgamentos. *Um advogado seu amigo trabalha no palácio da justiça*. **3.** Casa solarenga, ampla, sumptuosa que lembra um palácio. **olhar para alguma coisa como boi para palácio**. **1.** Não perceber nada de alguma coisa. **2.** Não ligar importância; não dar valor, apreço a. Dim. palacete.

Original encoding

```
<entry id="palácio">
  <form>
    <orth>palácio</orth>
    <pron>pel'asju</pron>
  </form>
  <gramGrp>s. m.</gramGrp>
  <!--etc. -->
</entry>
```

Conversion to TEI Lex-0

```
<entry type="monolexicalUnit" xml:lang="pt"
xml:id="palácio"> <form type="lemma">
  <orth>palácio</orth>
  <pron>pel'asju</pron>
</form>
<gramGrp>
  <gram type="pos" norm="NOUN">s.</gram>
  <gram type="gen">m.</gram>
</gramGrp>
<!--etc. -->
</entry>
```

Example 1: DACL monolexical unit – original encoding and conversion from TEI to TEI Lex-0.

As can be seen in Example 1, in TEI Lex-0, `entry` is used to encode the basic element of the dictionary microstructure and requires the attributes `xml:id` and `xml:lang` in compliance with ISO Standard 16642 for terminological data.

Note that TEI Lex-0 schema only allows `entry` to be used to typeset entries – the `entryFree`, `superEntry` and `re` elements of the TEI Guidelines are not allowed. As for the DACL itself, only `entry` and `re` were being used, and therefore little adaptation was needed at this point.

Lexicographical articles always start with a lemma (headword), which is a non-inflected unit considered as the canonical form. The lemma is encoded using the `form` element with the attribute `type` and value “lemma”. The `orth` element (orthographic form) gives the orthographic form of the headword.

Sometimes the lemma is a borrowed word. In TEI encoding, a unit borrowed from a foreign language is identified within the TEI element `etym`, where etymologic information is encoded, and labelled with the attribute `type` and the value “borrowing” (Bowers & Romary, 2017), as exemplified in Example 2.

workshop [wórkʃɔp]. *s. m.* (Ingl.). Reunião destinada à discussão ou realização de trabalho prático sobre um assunto específico, em que é feita uma aprendizagem através da troca de conhecimentos e experiências. «*Durante o 'workshop' sobre a articulação dos hospitais com os tribunais, foi visível a desconfiança de algumas pessoas*» (DN, 21.2.1992). Pl. workshops.

Original encoding

```
<entry id="workshop">
  <form>
    <orth>workshop</orth>
    <pron>w'orkʃɔp</pron>
  </form>
  <gramGrp>s. m.</gramGrp>
  <etym>Ing.</etym>
<!--etc. -->
</entry>
```

Conversion to TEI Lex-0

```
<entry type="monolexicalUnit" xml:lang="en"
xml:id="workshop">
  <form type="lemma">
    <orth>workshop</orth>
    <pron>w'orkʃɔp</pron>
  </form>
  <gramGrp>
    <gram type="pos" norm="NOUN">s.</gram>
    <gram type="gen">m.</gram>
  </gramGrp>
  <etym type="borrowing"><lang>Ing.</lang></etym>
<!--etc. -->
</entry>
```

Example 2: DACL borrowed word – original encoding and conversion from TEI to TEI Lex-0.

The lexical units formed from other units or bases – derivative lexical units (e.g. *infeliz* [unhappy]; *ensonado* [sleepy]) – are also classified as monolexical units, as shown in Example 3.

ensonado, a [ẽsunádu, -v]. *adj.* (De *en-* + *sono* + suf. *-ado*). Que tem ou está com sono. ≈ SONOLENTO. «*Sertório assoma à porta do quarto: vem, ensonado, a esfregar os olhos.*» (D. MOURÃO-FERREIRA, *Gaivotas em Terra*, p. 139).

Original encoding

```
<entry id="ensonado">
  <form>
    <orth fem="a">ensonado</orth>
    <pron>ẽsun'adu, -e</pron>
  </form>
  <gramGrp>adj.</gramGrp>
<!--etc. -->
</entry>
```

Conversion to TEI Lex-0

```
<entry type="monolexicalUnit" xml:lang="pt"
xml:id="ensonado">
  <form type="lemma">
    <orth>ensonado</orth>
  </form>
  <form type="inflected">
    <orth>ensonado</orth>
    <pron>ẽsun'adu</pron>
  <gramGrp>
    <gram type="gen">m.</gram>
  </gramGrp>
</form>
  <form type="inflected">
    <orth>ensonada</orth>
    <pron>ẽsun'ade</pron>
  <gramGrp>
    <gram type="gen">f.</gram>
  </gramGrp>
</form>
  <gramGrp>
    <gram type="pos" norm="ADJ">adj.</gram>
  </gramGrp>
<!--etc. -->
</entry>
```

Example 3: DACL monolexical lexical units – original encoding and conversion from TEI to TEI Lex-0.

This last example also shows that, when a specific inflected form is featured in the entry, it should be clearly defined as an independent form, and have enough information about the inflected type (in this case, that the item is a feminine form).

For the grammatical information, the TEI Lex-0 standard suggests the use of the `gramGrp` tag. This element can be used in two different places: as a sibling of the `form` element, when the annotation is referring to all the forms present in the entry, or as a child of the `form` element, when the information is specific for that form.

As XML is verbose enough, for DACL annotations will appear mostly following the `form` element, and when used inside it, it will describe only the properties that differ for that form. This way, in the example above, we do not repeat the information about the part-of-speech.

4.2.2 Polylexical units

Polylexical units are present in almost every dictionary. Under this classification, we have included compounds and all kinds of lexical combinations, such as collocations or phrasemes. By compounds we mean every lexical unit formed by two or more elements with autonomy within the language that together form a new lexical unit with a new meaning. By definition, in a general-language dictionary we can only find compounds and more rarely fixed combinations in an entry.

The encoding of compounds can be seen in Example 4:

decreto-lei [dɨkɾetulɛj]. *s. m. Dir.* Acto normativo proveniente do Governo da República. *Atualmente, os decretos-leis são publicados na primeira série-A do Diário da República.* Pl. decretos-leis.

Original encoding

```
<entry id="decreto-lei">
  <form>

    <orth>decreto-lei</orth>

    <pron>dɨkɾetul'ej</pron>
  </form>
  <gramGrp>s. m.</gramGrp>
  <!--etc. -->
</entry>
```

Conversion to TEI Lex-0

```
<entry type="polylexicalUnit" xml:lang="pt"
xml:id="decreto-lei">
  <form type="lemma">
    <orth>decreto-lei</orth>
    <pron>dɨkɾetul'ej</pron>
  </form>
  <gramGrp>
    <gram type="pos" norm="NOUN">s.</gram>
    <gram type="gen">m.</gram>
  </gramGrp>
  <!--etc. -->
</entry>
```

Example 4: DACL polylexical unit – original encoding and conversion from TEI to TEI Lex-0.

In the DACL, Latin phrases, i.e., fixed combinations, appear as headwords too (see Example 5):

fiat lux *loc. lat.* Exprime o desejo de que se torne clara alguma coisa importante.

Original encoding

```
<entry id="fiat lux">
  <form>
    <orth>fiat lux</orth>
  </form>
  <gramGrp>loc. lat.</gramGrp>
```

Conversion to TEI Lex-0

```
<entry type="polylexicalUnit" xml:lang="pt"
xml:id="fiat_lux">
  <form type="lemma">
    <orth>fiat lux</orth>
  </form>
```

```

<!--etc. -->
</entry>
<gramGrp>
  <gram type="pos">loc.</gram>
</gramGrp>
<etym type="borrowing"><lang>lat.</lang></etym>
<!--etc. -->
</entry>

```

Example 5: DACL polylexical unit – original encoding and conversion from TEI to TEI Lex-0.

In this example, even if “locução latina” [latin phrase] is not a part-of-speech, for now we decided to keep it encoded that way. While we are trying to use the Universal Dependencies Part-of-Speech Tagset¹⁰, we needed to add our own tags for specific cases due to the lack of accurate tags for our purpose.

4.2.3 Affixes

In certain dictionaries, such as the DACL, affixes appear as headwords, as shown in Example 6¹¹. The DACL uses bracketed hyphens as visual clues of the position the given affix takes in relation to the lexical unit it is attached to: the headword *(-)carpo(-)* indicates that *carpo* can be used as both a suffix and a prefix. Bracketed hyphens play the role of labels signalling the morphological property of the affix, but are not part of the affix itself. We therefore encode the affix itself as `<orth>carpo</orth>`, while using the element `<lbl>` to reflect the positional labels used in the dictionary.

(-)carpo(-) *elem. de form.* (Do gr. καρπός 'fruto'). Expri-me a noção de *fruto*. *Mesocarpo*, *carpologia*, *pericarpo*.

Original encoding

```

<entry id="carpo">
  <form>
    <orth>(-)carpo(-)</orth>
  </form>
  <gramGrp>elem. de form.</gramGrp>
<!--etc. -->
</entry>

```

Conversion to TEI Lex-0

```

<entry type="affix" xml:lang="pt" xml:id="carpo">
  <form type="lemma">
    <lbl>(-)</lbl><orth>carpo</orth><lbl>(-)</lbl>
  </form>
  <gramGrp>
    <gram type="pos">elem. de form.</gram>
  </gramGrp>
<!--etc. -->
</entry>

```

Example 6: DACL affix headword – original encoding and conversion from TEI to TEI Lex-0.

¹⁰ When labelling entries with part-of-speech appropriate linguistic terminology is crucial, mainly when we are talking about interoperability between lexical resources. This information must be one of the values from the Universal Dependencies Part-of-Speech Tagset: @norm attribute. See <https://universaldependencies.org/u/pos/>.

¹¹ Even if “*elemento de formação*” [affix] is not a part-of-speech, for now we decided to keep it encoded that way.

4.2.4 Abbreviations

Concerning abbreviations, the DACL registers different types of these: abbreviation (*Cf.*), alphabetism (*AAC*), acronym (*VIP*), symbol (*Ag*), contractions (*do* [of the]) and clipped forms (*metro* [metropolitan]).

Ag *símb.* (De *a<r>g<entum>* 'prata'). *Quím.* Símb. da *prata*.

Original encoding

```
<entry id="Ag">
  <form>
    <orth>Ag</orth>
  </form>
  <gramGrp>símb.</gramGrp>
<!--etc. -->
</entry>
```

Conversion to TEI Lex-0

```
<entry type="abbreviation" xml:lang="pt"
xml:id="Ag">
  <form type="lemma">
    <orth>Ag</orth>
  </form>
  <gramGrp>
    <gram type="pos">símb.</gram>
  </gramGrp>
<!--etc. -->
</entry>
```

Example 7: DACL abbreviation – original encoding and conversion from TEI to TEI Lex-0.

VIP [víp]. *s. m. e. f.* Sigla de *Very Important Person* (Pessoa Muito Importante).

Original encoding

```
<entry id="VIP">
  <form>
    <orth>VIP</orth>
    <pron>víp</pron>
  </form>
  <gramGrp>s. m. e. f.</gramGrp>
<!--etc. -->
</entry>
```

Conversion to TEI Lex-0

```
<entry type="abbreviation" xml:lang="pt"
xml:id="VIP">
  <form type="lemma">
    <orth>VIP</orth>
    <pron>víp</pron>
  </form>
  <gramGrp>
    <gram type="pos" norm="NOUN">s.</gram>
    <gram type="gen">m.</gram>
    <lbl>e</lbl>
    <gram type="gen">f.</gram>
  </gramGrp>
<!--etc. -->
</entry>
```

Example 8: DACL abbreviation – original encoding and conversion from TEI to TEI Lex-0.

In these examples, we would like to call attention to the usage of the `pos` element to annotate this type of abbreviation. Again, these are not proper part-of-speech attributes and might change in the future.

Finally, clipped forms are usually treated as nouns, as shown in Example 8:

metro² [métru]. *s. m.* (Red. de *metropolitano*). **1.** Sistema de transporte urbano efectuado por comboios de tracção eléctrica, em linhas parcial ou totalmente subterrâneas. \approx METROPOLITANO. *Encontraram-se na estação de metro. O metro está em greve. boca* de metro.* **2.** Comboio que assegura esse sistema de transporte. *Apanhar, perder o +.*

Original encoding

```
<entry id="metro:2">
  <form>
    <orth>metro:2</orth>
    <pron>m'etru</pron>
  </form>
  <gramGrp>s. m.</gramGrp>
<!--etc. -->
</entry>
```

Conversion to TEI Lex-0

```
<entry type="abbreviation" xml:lang="pt"
xml:id="metro_2" n="2">
  <form type="lemma">
    <orth>metro</orth>
    <pron>m'etru</pron>
  </form>
  <gramGrp>
    <gram type="pos" norm="NOUN">s.</gram>
    <gram type="gen">m.</gram>
  </gramGrp>
<!--etc. -->
</entry>
```

Example 9: DACL abbreviation – original encoding and conversion from TEI to TEI Lex-0.

This example shows yet another detail regarding visual information. There is more than one entry for the lexical unit *metro*. Therefore, as usual, the dictionary includes a superscript number, near the headword, to differentiate each entry. To encode this information we do it two ways: first, the entry identifier has the entry number following the headword, separated by a underscore. As this information is also important to the reader, it is encoded as the attribute `n` (number) in the entry element.

From these examples it is clear that TEI Lex-0 is going in a good direction, making the encoding more verbose but more structural, allowing machines to process this information better.

From these examples it is clear that TEI Lex-0 is going in a good direction, making the encoding more verbose but more structural, allowing machines to process this information better.

5. Automatic conversion of the original TEI schema to TEI Lex-0

Given that we are not dealing with a standard but with the process of creating it, the schema is not fixed. Therefore, our present goal is not to have the dictionary in TEI Lex-0 only, but to keep the original version in our own interpretation of TEI and have another version that can be used for tests and to promote the discussion with the TEI Lex-0 community.

Also, as our entries are stored independently in the XML database, our goal is not to produce a complete XML document for the dictionary, but a set of small XML files per dictionary entry. Therefore, details about the TEI header are deliberately being ignored at this stage, and thus we are not using the complete schema but only the entry portion, considering the entry tag as the document root element. In the future, the header can be stored in an independent record in the database, and a simple tool can be used to construct a TEI/TEI Lex-0 file with the complete dictionary, validating the complete schema.

The conversion between structured formats is not difficult as long as the information is somehow annotated in the source document. This is the case for most of the encoding changes needed in the dictionary, with a few exceptions.

If we were only dealing with structural changes, an interesting approach would be to use the eXtensible Stylesheet Language Transformations (XSLT) language. This would allow the transformation to run on top of eXist-DB, and could even be performed on demand for any desired entry. Nevertheless, to allow us more control when dealing with partially structured content, our approach was to use a generic high-level programming language (Perl).

In order to allow progressive validation, we chose to edit our schema in order to accommodate TEI Lex-0 recommendations, one at a time. For each of these changes, a new part of the script was added to perform the desired changes.

Two main changes needed human intervention: grammatical groups and etymology:

- While TEI allows the grammatical information (under the `gramGrp` element) to be unstructured (i.e., only the visual information, such as “n. m.” for masculine noun), TEI Lex-0 enforces the tagging of the part-of-speech information using specific tags. In order to guarantee the accuracy of this conversion, a list of the complete possibilities for the content of that tag was computed, and the desired annotation was manually added with part-of-speech. Taking the opportunity, we also normalized situations where the entry lexical unit had more than one grammatical analysis — e.g. *vegano* [vegan] whose morphological information is “adj., n. m.” [adjective and masculine noun].



Figure 3: *vegano* [vegan] (DACL new edition).

In these cases, the `gramGrp` element stores a list of possible gram entries, one for each analysis. This mapping was defined manually as a table, and the conversion script simply replaced the existing information with the new one.

- The other tricky conversion is the entry's etymology. It is challenging mainly because, when the PDF document was converted to TEI, not every detail of the etymology was properly annotated. While no information was lost, some portions were stored simply as plain content (text) without proper XML annotation. Unlike grammatical information, the creation of a list of all the possibilities is unthinkable, as the amount of entries that completely share their etymological information is close to zero. Thus, the process for etymology conversion had to be based on an approach that is similar to the one executed during the PDF to TEI conversion: a definition of a set of regular expressions to detect clear portions of the etymology (that do not include any ambiguity), which are annotated first, as anchors. Then, new rules and heuristics are applied using these anchors to detect other bits of information. This process is currently being done and it is expected that 95 % of the entries can be completely automated. The remaining ones might need direct manual intervention. This is a work in progress, and, just like most of the TEI Lex-0 encoding, further discussion on how to encode most of the information properly is still needed.

6. Conclusions and future work

In this paper, we focused on encoding information of different types of lexical units, providing examples, and thus contributing to a more consistent encoding of lexicographic data, constraining the variety of possibilities offered by the TEI Guidelines.

The results obtained are useful for the discussion and definition of the TEI Lex-0 standard. The definition of a standard is very important, as it allows resources or tools to be used interchangeably, but it is also a complex task, as the resulting standard should be able to encode different types of dictionaries, and not just for different languages, but with different purposes as well.

7. Acknowledgements

Research financed by Portuguese National Funding through the FCT – Fundação para a Ciência e Tecnologia as part of the project Centro de Linguística da Universidade NOVA de Lisboa – UID/LIN/03213/2019, and by the European Union’s Horizon 2020 research and innovation programme under grant agreement No 731015 (ELEXIS).

8. References

- Bański, P., Bowers, J. & Erjavec, T. (2017). TEI-Lex0 guidelines for the encoding of dictionary information on written and spoken forms. In I. Kosem et al. (eds.) *Electronic Lexicography in the 21st Century: Proceedings of eLex 2017 Conference*, pp. 485–94.
- Bowers, J. & Romary, L. (2017). Deep encoding of etymological information. In *TEI. Journal of the Text Encoding Initiative*, TEI Consortium, 2017, [⟨https://jtei.revues.org/1643⟩](https://jtei.revues.org/1643). [⟨10.4000/jtei.1643⟩](https://doi.org/10.4000/jtei.1643). [⟨hal-01296498v2⟩](https://hal.archives-ouvertes.fr/hal-01296498v2).
- Gouws, R. H. (2018). Internet lexicography in the 21st century. In *Wortschatz: Theorie, Empirie, Dokumentation*, pp. 215–236.
- ISO 16642:2017 *Computer applications in terminology – Terminological markup framework*
- Khemakhem, M., Foppiano, L. & Romary, L. (2017). Automatic Extraction of TEI Structures in Digitized Lexical Resources using Conditional Random Fields. In *Proceedings of eLex 2017*, Leiden, Netherlands, September.
- Romary, L. & Tasovac, T. (2018). TEI Lex-0: A Target Format for TEI-Encoded Dictionaries and Lexical Resources. In *Proceedings of the 8th Conference of Japanese Association for Digital Humanities*, pp. 274–275. Available at: https://tei2018.dhii.asia/AbstractsBook_TEI_0907.pdf.
- Simões, A., Almeida, J. J. & Salgado, A. (2016). Building a Dictionary using XML Technology. In *5th Symposium on Languages, Applications and Technologies (SLATE'16)*, vol. 51 of Open Access Series in Informatics (OASICS). Germany: Dagstuhl. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, pp. 14:1–14:8.
- Tasovac, T. (2010). Reimagining the Dictionary, or Why Lexicography Needs Digital Humanities. In *Digital Humanities 2010*, pp. 254–256.
- Trap-Jensen, L. (2018). Lexicography between NLP and Linguistics: Aspects of Theory and Practice. In *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*. Ljubljana: Ljubljana University Press, Faculty of Arts, pp. 25–37.

Websites:

TEI Consortium, eds. TEI P5: Guidelines for *Electronic Text Encoding and Interchange*. [Version 3.5.0]. [Last updated on 29th January 2019, revision 3c0c64ec4]. TEI Consortium. <http://www.tei-c.org/Guidelines/P5/> ([13.07.2019]).

DARIAH WG: Lexical Resources and the H2020-funded European Lexicographic Infrastructure (ELEXIS), <https://github.com/DARIAH-ERIC/lexicalresources/tree/master/Schemas/TEILex0>.

Dictionaries:

DACL: *Dicionário da Língua Portuguesa Contemporânea* (2001). João Malaca Casteleiro (coord.), 2 vols. Lisboa: Academia das Ciências de Lisboa & Editorial Verbo. New digital edition under revision.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



Aggregating Dictionaries into the Language Portal

Sõnaveeb: Issues With and Without Solutions

Kristina Koppel, Arvi Tavast, Margit Langemets,

Jelena Kallas

Institute of the Estonian Language, Estonia

E-mail: kristina.koppel@eki.ee, arvi@tavast.ee, margit.langemets@eki.ee, jelena.kallas@eki.ee

Abstract

In this paper we present Sõnaveeb, a new type of language portal of the Institute of the Estonian Language containing data from a growing number of dictionaries and termbases. Sõnaveeb currently displays a total of 200,000 Estonian headwords, obtained from many databases, with many new types of lexicographic information: collocations, etymology, multi-word expressions, etc.

The paper reports on problems encountered so far: the consistency of information and avoiding duplicates when unifying the dictionaries, turning dictionary-specific information into customizations of the central service, deciding on deliberate ambiguities, parsing data fields containing more than one data element, including textual condensation, moving from annotating form (e.g. italics) to annotating content (e.g. a citation), moving from (near) duplicates to sensible information fragments, deciding between an app and a responsive web page, and possible legal problems regarding the authorship of the new central resource, as it may become difficult to show who authored which part of the published resource.

The development of Sõnaveeb continues in the direction of both the tighter aggregation of existing datasets and the addition of new data from other dictionaries and termbases, as well as compiling new data in the new DWS Ekilex.

Keywords: lexicographic database; data aggregation; unified dictionary; Dictionary Writing System; user needs; Estonian

1. Introduction

Sõnaveeb¹ is the new language portal of the Institute of the Estonian Language containing the linguistic information from a growing number of dictionaries and databases. Sõnaveeb was released in February 2019 and presented with the publishing of two new dictionaries, The Dictionary of Estonian 2019 (DicEst) and Estonian Collocations Dictionary 2019 (ECD). In addition, The Basic Estonian Dictionary 2019 (BED) (1st ed. 2014), intended for beginner and advanced language learners, can be used here, as well as two bilingual dictionaries, the Estonian-Russian Orthographic Dictionary for Students 2019 (1st ed. 2011) and the Estonian-Russian Dictionary 2019 (1st ed. 1997–2009), updated with 10,000 new headwords. Special morphological

¹ <https://sonaveeb.ee/> (20 May 2019). Sõnaveeb can be translated into English as Wordweb. It is important to emphasize that it is the language portal, not an ontology.

datasets serve to present morphophonological data for Estonian. The portal contains about 200,000 words and phrases in Estonian and about 70,000 words and phrases in Russian.

The information displayed in Sõnaveeb comes from Ekilex² (Tavast et al., 2018), a Dictionary Writing System maintained and developed by the Institute in collaboration with the software company TripleDev. As of May 2019, Ekilex contains over 50 lexical datasets: general as well as specialized dictionaries. Databases are constantly updated and edited, including changes that are made upon receiving feedback from users. Created data is stored in Ekilex's PostgreSQL database. Ekilex is hosted in the Estonian Scientific Computing Infrastructure (ETAIS) cloud. Archive copies of data are also stored in the Center of Estonian Language Resources' repository Entu³. The metadata on created resources is available in the META-SHARE⁴ repository. Upon creating a metadata entry in META-SHARE, a DOI is assigned to each resource.

A new version of the portal is created and archived once a year. Each version is marked by the year and has the date of its creation, e.g. Sõnaveeb 2019 (14.02.2019).

In the next sections we discuss the list of issues, whether they are already solved, in the process of being solved, or lack a known solution. Undoubtedly, there will be more exciting challenges in the near future as we continue to import new data. Several issues are very much in line with the objectives and outcomes of the Horizon 2020 project ELEXIS (European Lexicographic Infrastructure)⁵ developing strategies for extracting, structuring and linking of lexicographic resources.

2. Internet skills and organizing the presentation of data

The Sõnaveeb user interface has two different modes of information display for different types of users: advanced and simple. Robert Lew (2013) has stated that web users tend to resort to very simple strategies for internet-based information retrieval, and that users' general tendency is to gravitate towards natural-language queries. The bad news is that "end-users tend not to change the default settings of an information retrieval system" (Markey, 2007: 1077, cited by Lew, 2013). Online dictionaries should somehow cope with unsophisticated strategies of general web use. We agree with Lew (2013: 29):

This is a conclusion that many lexicographers find hard to accept, and an argument can be made that a minority of expert users (such as language professionals) are worth catering for as well. Ideally, an online dictionary interface

² <https://ekilex.eki.ee/> (20 May 2019).

³ <https://entu.keeleressursid.ee/> (20 May 2019).

⁴ <http://www.meta-share.org/> (20 May 2019).

⁵ <https://elex.is/objectives/> (20 May 2019).

will combine simplicity (for those who cannot be bothered) with sophistication (for those who can). A reasonable way to achieve this is to offer a simple default interface with an optional advanced alternative.

In Sõnaveeb, we try to combine simplicity with academic sophistication and trustworthiness. As the system has mostly been developed in cooperation with lexicographers, not laymen, we tend to prefer lexicographers' cultivated taste. However, we have conducted some user interviews on particular topics, e.g. synonyms, parts of speech and web sentences, and we are willing to use this information to present our data in a better, i.e. more flexible way.

2.1 Advanced mode vs. simple mode

Modes are used to filter data. The user can currently choose between two modes of information display: advanced or simple. The advanced mode is intended primarily for native speakers. It displays all the information on a word that comes from different sources. The advanced mode is a sophisticated view that might require more options for further filtering. At present we are working on the inclusion of prescriptive data (from the prescriptive Dictionary of Standard Estonian (ÕS 2018), in order to present both descriptive and prescriptive data. This is a challenge, as there have been quite a number of data conflicts from the user's perspective in parallel separate online dictionaries (the descriptive DicEst vs. prescriptive ÕS).

The simple mode is intended primarily for learners at the A2–B1 proficiency levels. It shows 5,000 basic Estonian words (headword list of the Basic Estonian Dictionary (BED); see Kallas et al., 2014) and information is presented in a simpler way: the definitions are shorter, knowledge is organized using controlled vocabulary, there is explicit information about the most frequent morphological forms, etc.

2.2 Choosing languages

As of May 2019, lexical data is available for two languages: Estonian and Russian, each as both source and target language. The list of languages is planned to be increased as there are more bilingual databases available at our Institute.

2.3 Mobile app or responsive web page?

Sõnaveeb.ee is a responsive web page with the same information content for both mobile and desktop resolutions. Around 73% of traffic is desktop, while 25% is mobile and 2% is tablet usage. There are around 22,000 monthly and 2,000 daily active users. 56% are new and 44% returning visitors (Google Analytics, 30 May 2019).

The most frequent question since opening the Sõnaveeb website in February 2019 has been: Will there also be an app? No, for the following reasons:

- The web is better for reaching a wider audience, especially if dictionary use is as sporadic as shown by the high ratio of new visitors. Users cannot be expected to install an app that they will only use once.
- As apps are platform-specific, their development and maintenance are currently beyond our financial means.
- Dictionary content is visually simple enough to be presented using web technologies.
- Lexical resources in the form of a website are more easily indexable by search engines. Although we haven't achieved it yet, it is possible to show up in search results for individual words.

3. Aggregation issues

3.1 The Ekilex data model and the unification of dictionaries

The data model of Ekilex has been described in Tavast et al. (2018). For the purposes of the current paper, it is sufficient to note that we have a many-to-many (i.e. n:m) relations between words and meanings. The link table between these two entities is called a lexeme and is defined as “this word in this meaning, as described in this dataset”. Words and meanings are dataset-agnostic, allowing a gradual transition from the initial condition of several independent datasets to the end goal of a single Ekilex resource containing all lexical information known about the Estonian language.

The initial import of the separate datasets resulted in massive duplication of both words and meanings (see Figure 1). Each word had at least as many homonyms in the Ekilex resource as there were imported datasets.

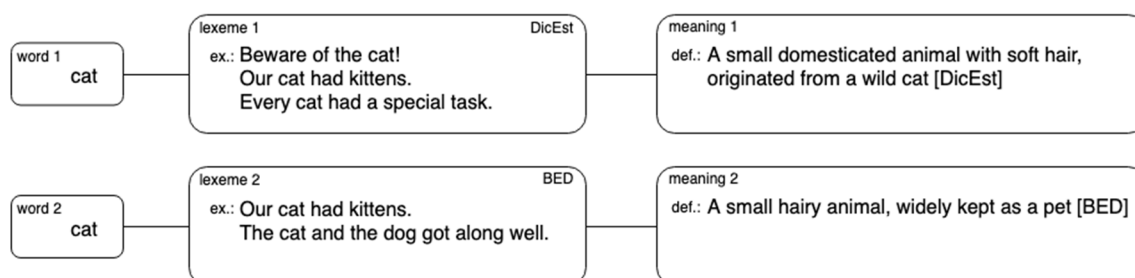


Figure 1. Initial condition: separate dictionaries with duplicate words and meanings

The first step in the transition was the unification of homonyms. Lexicographers manually decided which homonyms were legitimate, and the rest were unified automatically. The result was that there were no longer too many homonyms, but now each word had at least as many senses as there were imported datasets (see Figure 2). The manual effort of unifying the words was relatively small, as there are only about 1,500 legitimate homonyms in Estonian.

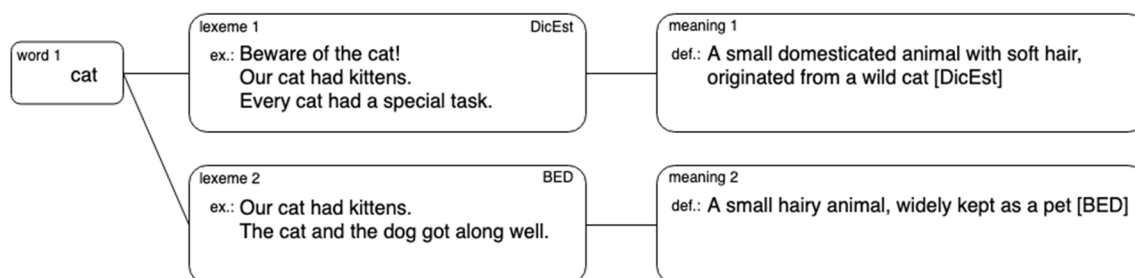


Figure 2. First step in unification: words are unified, but lexemes and meanings are still duplicated

The next step was the manual unification of meanings. The difficulty here is that datasets differ in their sense divisions, often deliberately, depending on the target audience and purpose of the dictionary, so there are no direct correspondences between meanings across datasets. As of May 2019, this work is still ongoing, even for clear cases, and there is no known solution for the unclear ones, unless the solution is to alter the sense divisions of the original datasets. The result for the successfully unified meanings is that there are two lexemes between a word and a meaning, or two statements about the same word-meaning correspondence, see Figure 3.

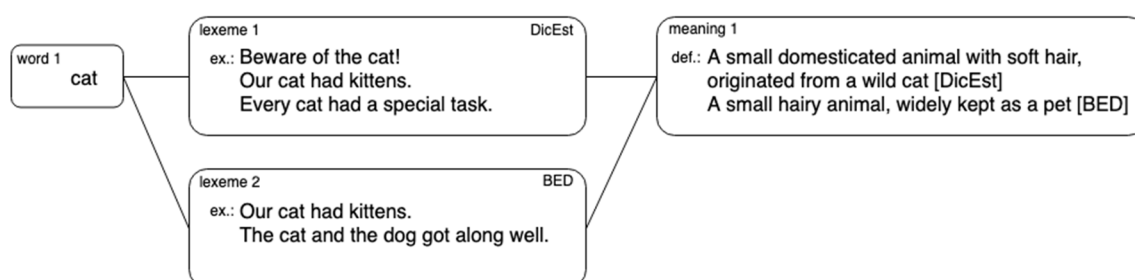


Figure 3. Second step in unification: Words and meanings are unified, each dataset still has its own lexeme between them

Since the Ekilex data model is flattened for display in Sõnaveeb by aggregating lexemes and meanings (this aggregation corresponds to the traditional understanding of word sense), this stage of unification resulted in a very unclear display of information in Sõnaveeb. There were still as many “senses” as imported datasets, but meanings (mainly represented by definitions) were first added together and then repeated under every sense. This was so counter-intuitive for readers that we temporarily disabled version updates of Sõnaveeb, displaying the previous stage instead.

The final stage of unification is still in development. To dispose of the duplicate lexemes they will be added up, with the sum lexeme containing a union of all data elements in lexemes between the same word and the same meaning (see Figure 4).

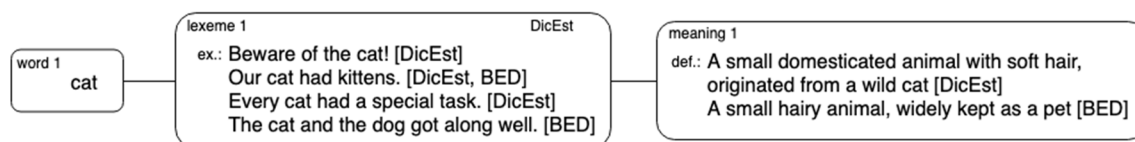


Figure 4. Unified datasets: there are no duplicates in any of the three major entities of the data model

Clear cases of duplication among the first imported datasets will be solved at this step. Work with less clear cases, and especially with more specialized datasets, will continue for a long time. It is also not known yet to what extent such unification is even possible.

3.2 From duplicates to sensible information fragments

Separate datasets have brought into the Ekilex resource several duplicates (or near duplicates) that require special attention. We have to decide whether it is useful or possible to adapt them into information fragments for reuse in other contexts, or if they should just be avoided.

Weitzman (2014) stressed that content management systems must support sensible fragments of information that can be presented in different contexts, e.g. in Ekilex we have “duplicate” information from different datasets for the same meaning (definitions and domain indicators). The task is to be able to describe these different user situations, in which each has its own requirements on the information. These fragments cannot be automatically derived; instead they have to be carefully designed. As the separation of content and presentation has been implemented in Ekilex, we try to reuse the information in the most sensible way, e.g. information from BED has been presented in the simple mode intended primarily for learners at the A2–B1 proficiency levels (see Chapter 2.1).

Sense division in the source datasets (DicEst, ECD and BED) has been manually disambiguated using a specially developed tool. After unifying the senses, we get both long definitions (from DicEst) and short definitions (from BED) that might be presented in different modes: long definitions in the advanced mode, and short/simple definitions in the simple mode. The ongoing migration of senses from separate datasets to a single resource creates several questions about how to merge the pieces of information (e.g. definitions) and to what extent is data provenance important for the users. Answers might depend on different perspectives: lexicographers are protective of their wordings, while user preferences are yet to be seen.

Collocations were found in the same three sources (DicEst, ECD and BED). In Ekilex we faced the problem of overlapping: some multi-word units (MWU) from DicEst were collocations in ECD, e.g. *punane vein* ‘red wine’ and *kollane kaart* ‘yellow card’, and some collocations in ECD were usage examples in DicEst, e.g. *kodune aadress* ‘home address’, *isiklik elu* ‘personal life’ and *ebaväärikas käitumine* ‘undignified behaviour’. To avoid duplicates, the authors of the ECD deleted the collocations that were MWUs in DicEst prior to import into Ekilex to avoid duplicates in both Ekilex and Sõnaveeb. This might have been solved differently by building a connection from the collocation to be presented both ways: as collocation and an MWU.

Concerning usage examples, the authors of DicEst (Langemets et al., 2018) stated that they have added all kinds of usage examples: full sentences, collocations and phrases. However, the research conducted by Kristina Koppel (2019, forthcoming) showed that neither language learners nor lexicographers themselves considered collocational phrases (e.g. *kangesti palav ilm* ‘very hot weather’, *väljapaistva arhitektuuriga ehitis* ‘a building with extraordinary architecture’) to be suitable examples. This is an issue to be solved in the future: it might be reasonable to move towards presenting phrases as MWEs or collocations, rather than usage examples.

Senses and collocations have occasionally been presented also in other datasets, e.g. in the prescriptive dictionary ÕS. One of the lessons learned so far is: do not import the dictionary as a whole. Extract valuable pieces of information instead. In this case, there is no need to analyse the dictionary database once again to fully understand for what purpose any fragment of (duplicate) information has been included in the dictionary. It is sufficient to import the pieces of information that undoubtedly add value.

3.3 From dictionaries to information layers

Some of the source datasets are focused on specialized information, such as morphology, word formation, collocations, etymology, language planning or language proficiency levels. They have been authored as separate dictionaries with varying degrees of autonomy from each other. In moving towards a single database, these datasets are turned into information layers and applied to the central “backbone” of headwords already present in the database, removing the need to specify variations of the same information again in separate dictionaries.

Morphology is a case in point. Declination patterns of Estonian words are well established and rarely debated among lexicographers, and morphological information has been centralized into The Estonian Morphological Database of the Institute of the Estonian Language 2019. This database is considered as a central service for all datasets.

Figure 5 shows aggregated information in Sõnaveeb for *diskussioon* ‘discussion’ from

different datasets: definition (from DicEst), collocations (from ECD), inflected forms (from the morphological database), etymology (from DicEst) and web sentences (external data from etSkELL via the Corpus Query System KORP⁶ API).

Esti Keele Instituut

diskussioon

EESTI KEEL → EESTI KEEL

LETA TÄNA

LIIHTRE

diskussioon

nimisõna

1 arvamuste vahetamine (nt koosolekul, trüki sõnas), arutlus või vaidlus

mille üle

Uus seaduseelnõu tekitas elava diskussiooni

Esti keele sõnaraamat 2013

1

Naabersõnad

KASUTUSNAITED

OMADUSSÕNAGA

avallik diskussioon

elav

poliitiline

sisuline

õhiskondlik

tõsine

plikk

äge

huvitav

tuline

suur

põhjalik

teaduslik

terav

akadeemiline

lalaphjaline

aktiivne

konstruktiivne

avatud diskussioon

argumenteeritud

TEGUSÕNAGA

diskussioon toimub

tekib

kaib

jätkeb

algab

puhkeb

kestab

vallandub

vaibub

diskussiooni tekitama

algatama

alustama

pidama

jatkama

arendama

argitama

juhtima

modereerima

avama

käivitama

vallandama

edendama

valitma

diskussiooni kaasama

astuma

sekkuma

laskuma

diskussioonis osalema

diskussioonist osa võtma

NIMISÕNAGA

diskussiooni teema

objekt

küsimus

algatamine

tekitamine

arendamine

diskussiooni keskmes

diskussiooni tulemusena

Sõnavormid

ainsus

diskussioon

diskussiooni

diskussiooni

diskussiooni

diskussioonis

diskussioonis

diskussioonist

diskussioonile

diskussiooni

diskussioonilt

diskussioonis

diskussioonis

diskussioonis

diskussioonita

diskussioonga

mitmus

diskussioonid

diskussioone

diskussioone – diskussioonisid

diskussioonesse

diskussioonesse

diskussioones

diskussioonidest – diskussioone

diskussioonidele – diskussioonile

diskussioonidelt – diskussioonelt

diskussioonideks – diskussiooneks

diskussiooniden

diskussioonidena

diskussioonideta

diskussioonidega

Sõna seosed

(seda kirjeldust ei ole)

Päritolu

LAENSÕNA

< saksa *Diskussion*

< ladina *discussio* 'arutamine, porutamine', hillisladinas ka 'uurimine, labivaatus' (sonast *discutere* 'purustama; laiali ajama; korvaldama, nurja ajama', hillisladinas 'valja uurima, arutama')

Sama sõna e-keelenõus

Veebileaseid

Kirjanike poolt alatatud diskussioon muutus öldrahvalikuks.

Figure 5. Aggregated information for diskussioon ‘discussion’ in Sõnaveeb (advanced mode)

Since all lexicographers trust the morphological database, it was agreed that morphology would only come from there, and any morphological information manually added to other dictionaries would be ignored during import. However, not all differences between dictionaries were inconsistencies. Rather than all possible forms from the database, we have chosen to present a subset: only most frequent forms in the simple mode, only approved forms for prescriptive language advice, only corpus-attested forms in advanced mode, and only forms that distinguish homonyms in most other dictionaries.

It would be ideal if inflected forms were labelled accordingly in the morphological database. The problem is that they are not. All target groups see either the full theoretically possible paradigms or trivially filtered subsets (e.g. learners only see the first of alternative forms). For lexicographers, this is a step in the wrong direction. They feel they already had the correct manually selected forms in their dictionary.

⁶ <https://korp.keeleressursid.ee/> (20 May 2019).

which are now gone. Tagging is planned and can be partially automated based on these same datasets: if a form is listed in a learner's dictionary, it can be labelled as suitable for learners, in addition to attaching corpus frequencies to forms.

The situation is similar with collocations. BED and ECD were compiled as separate dictionaries, and BED was the first dictionary where collocations were presented explicitly. The manually selected learner-level collocations from BED were not imported to Ekilex. Instead, all collocations were imported from ECD and then filtered. The simple mode in Sõnaveeb only shows collocations consisting entirely of words included as headwords in BED. As a result, there are many more collocations for a headword in the simple mode than there were originally in BED, including collocations where the collocate as a word is included in BED but the sense is not, for example there is a collocation *liblika nukk* ‘butterfly pupa’ under headword *liblikas* ‘butterfly’, although *nukk* ‘doll’ is only defined as a toy in BED. Again, the solution would be semi-automatic labelling of collocations for the language level, which is planned but has not been started.

Concerning prescriptive data, the preparatory phase of the new normative dictionary (ÕS 2025) started in 2019. It has already been agreed that prescriptive statements will be a layer on top of the otherwise descriptive backbone, rather than a separate dictionary. This will constitute a major change for the prescriptive ÕS, and issues may arise.

3.4 Linking and reuse of data

Ekilex treats all word-like entities as words, including ones that were unstructured character strings in previous systems. The objective is to improve data quality by replacing character strings with entity references. A practical problem is that this inevitably requires manual disambiguation, the additional workload of which comes as an unpleasant surprise to the lexicographer. More importantly, such linking exposes inconsistencies. Some of these may be deliberate, and in any case the lexicographer is understandably not happy about this. Notable examples of this type of issue are synonyms, equivalents, collocations, usage examples and definitions.

The representation of synonyms and equivalents was mixed in the earlier systems that Ekilex imported data from. They were word entities in termbases, but character strings in general lexical datasets. Of the latter, DicEst authors had manually ensured that synonyms were all valid, symmetrical ($A=B$ and $B=A$) and unambiguous (the homonym number and sense number of the target word were also given), and other datasets contained few synonyms, so these were easy to import.

Russian equivalents, on the other hand, were completely ambiguous character strings. If the same string was given as an equivalent more than once, we had no way of knowing if these were the same meaning, a polysemous word or separate homonym.

The current solution has been to import them all as one polysemous word waiting to be manually disambiguated, resulting in the most frequently used Russian words having over 20 meanings. This result can be seen when searching in the Russian-Estonian direction, and was so unexpected for both users and lexicographers that we had to display a special warning about searching in that language direction.

The same problem was in the collocations dictionary database, where the headword, its collocates and possible context were added as character strings. In preparation for importing into Ekilex, the lexicographers semi-manually disambiguated the collocates so that they were easy to interpret as references to word entities. The contexts remained ambiguous and we applied automatic disambiguation where possible.

The Ekilex data model, and also for end users in Sõnaveeb, represents collocations so that one is always a relation between two or more lexeme entities. It is not necessary to specify one of them as the headword or otherwise superior component. The import did give asymmetrical information about the components, because the collocation's relation with the headword, unlike other components, also contained information about which part of speech group and grammatical relation group that collocation belongs to from the point of view of the headword. The following combinations were present in the dictionary, with the following issues:

- The collocation was listed under only one component. Due to the symmetry of the Ekilex model, it also appeared when viewed from the opposite direction, which was unexpected for the lexicographers, who had deliberately only included it in one direction.
- The collocation was listed under the headword, as well as under other collocates. Symmetry was expected here, but another issue emerged. As the collocation was edited separately in each direction, possibly by different lexicographers, it was possible that the information given was different, for example the same collocation could be in plural under one collocate and in singular under the other. This problem was also evident in example sentences. If the importer found identical examples, it imported them only once. Problematic were the cases when one of the lexicographers had edited the sentence for clarity, so the examples were no longer identical, resulting in the collocation having two very similar examples in Ekilex.

The authors of dictionaries currently imported into Ekilex do not have a common understanding of what a usage example is, as mentioned in Chapter 3.2. The shortest examples are word-like entities, making them candidates for being treated as word entities instead of usage examples. We adopted the practical heuristic that we imported an example as a word entity if it was either one word, or was included in the DicEst as a MWU. This is in addition to the issue of the same phrase being described as a MWU/example/collocation across the imported dictionaries (see Chapter 3.2. on duplicates).

Likewise, definitions in the imported dictionaries were sometimes word-like, or consisted of a comma-separated list of word-like strings. The lexicographers agreed that these were more like synonyms or synonym lists than definitions, but we decided not to attempt parsing them during import. If lexicographers consider it necessary, they can manually change those definitions in Ekilex.

While most commas between word-like strings were indeed separators, there were exceptions, e.g. *tee ruttu, muidu jääd hiljaks* ('hurry up, otherwise you'll be late') where the comma was part of the expression. Especially among Russian equivalents and usage examples, the strings often further contained textual condensations that were too underspecified to expand automatically.

1. Examples resulting in two expansions:

ET olgu peale(gi) = olgu peale / olgu pealegi 'well and good'

RU женатый [мужчина] = женатый / женатый мужчина 'married man'

обыденная ~ разговорная речь = обыденная речь / разговорная речь 'colloquial speech'

2. Examples resulting in more than two expansions:

RU смесь ~ раствор соединяет ~ связывает строительные камни = смесь соединяет строительные камни / смесь связывает строительные камни / раствор соединяет строительные камни / раствор связывает строительные камни 'the mixture connects building stones'

RU подорожник снижает ~ понижает опухлость ~ отёчность = подорожник снижает опухлость / подорожник снижает отёчность / подорожник понижает опухлость / подорожник понижает отёчность 'plaintain reduces puffiness'

3. Examples where the expansion requires linguistic knowledge:

ET ta on töö peale ~ tööle laisk = ta on töö peale laisk / ta on tööle laisk 'he/she is too lazy to work'

RU в дальнейшем ~ впредь будь осторожнее = в дальнейшем будь осторожнее / впредь будь осторожнее 'be more careful in the future'

Due to the third group, we decided not to attempt automatic expansion, but to leave the corrections to be done manually in Ekilex.

The condensations have been used for conserving space in print dictionaries. In electronic form, space limitations are replaced by the need to search for items. It would of course be possible to create an index that would refer all full forms to the condensed form, but indexing the third group would require exactly the same linguistic

knowledge that expanding them would. We have yet to reach a decision on what to do with such condensations.

Source datasets contained annotations of form (bold, italic, subscript and superscript) using several different markup notations. The use of italic was especially ambiguous. Two frequent meanings of italic script were citations and metalanguage (the “or” between alternatives, for example). We set out to enforce marking up of content, not form, so that the italic would be replaced with a citation or metalanguage as necessary. This was straightforward, thanks to the limited nomenclature of italicized metalanguage items.

Where we ran into a wall, however, was with subscript and superscript. The orthodox way would have been to distinguish between their meanings in mathematics, chemistry, legislation, etc., mark each up with its correct meaning, and then display all of those meanings as subscript or superscript as before. While that would have been the correct way to do it, we decided to take the easier route and leave them marked up as subscript and superscript. After all, it is highly unlikely that mathematics or chemistry would change their notation so that we would have to replace the superscript with some other formatting. So we decided to tolerate an inconsistency in Ekilex that is theoretically messy, but very convenient in practice.

3.5 Authorship of separate dictionaries

Firstly, as mentioned in Chapter 3.1, in the Ekilex data model the words (i.e. headwords) and meanings (i.e. definitions and domain indicators) are dataset-agnostic. Secondly, after having processed, systematized, unified, supplemented, edited, etc. the information across datasets, the Ekilex resource receives the status of a single database containing all lexical information known about the Estonian language, protected by the Copyright Act.

We will make it possible to “(re-)derive” separate datasets from the Ekilex resource if there is a demand for them, e.g. from the owner of the economic rights (the government or a company), or from the authors of previous datasets or government regulations (e.g. from 2006 in Estonia, the literary norm is supposed to be based on the most recent printed (!) prescriptive dictionary ÕS issued by the Institute of the Estonian Language)⁷.

Since starting working in Ekilex, the work on separate dictionaries will develop into the work on specific information layers. Again, several questions might arise, for instance the following. Should we show explicitly the origin/authorship of every piece of information after unification of the datasets? Who is the author of a “(re-)derived” dictionary if we use unified information fragments available in Ekilex for free but

⁷ <https://www.riigiteataja.ee/akt/114062011003> (20 May 2019).

compiled by several other lexicographers? Will the authors develop into content renters rather than owners (Bego, 2018)? These are issues to be solved.

4. External data in Sõnaveeb

4.1 Audio pronunciation, speech synthesis and speech recognition services

In Sõnaveeb, users can listen to the pronunciation of about 5,000 of the most frequent headwords, as well as their most important inflected forms, and of about 7,000 unadapted loan words. The information on pronunciation has been aggregated from different datasets: from BED (headwords and inflected forms) and the dictionary of Foreign Words (VL, unadapted loan words). In the case of unadapted loan words, we used Estonians who speak foreign languages (Italian and Spanish) at high proficiency levels. For the pronunciation of the most frequent words and their inflected forms, we used professional actresses.

Text-to-Speech synthesis⁸, developed by the Institute of the Estonian Language, is used for reading out the example sentences chosen by lexicographers. The same application is quite widely used by Estonian newspaper publishers: users can listen to all articles on the internet, as well as on Estonian Public Broadcasting for reading out subtitles⁹.

Speech recognition¹⁰, developed by the Department of Cybernetics of the Tallinn Technological University, is used when dictating words. Speech recognition operates in real time. For optimum quality, users have to pronounce the search word clearly and steadily.

4.2 Web sentences

In Sõnaveeb, authentic example sentences from the corpus are displayed. They have been automatically selected and they have not been edited.

The example sentences are queried from the Estonian Corpus for Learners 2018 (etSkELL)¹¹ (250 million words) via the Corpus Query System KORP API. etSkELL corpus was compiled using the GDEX tool (Kilgariff et al., 2008; Kosem et al., 2019) in Sketch Engine, and consists of sentences from various media texts, fiction, scientific texts, Estonian Wikipedia and Estonian textbooks. The example sentences for Russian

⁸ <http://www.eki.ee/heli/> (20 May 2019).

⁹ <https://heliraamat.eki.ee/> (20 May 2019).

¹⁰ <http://bark.phon.ioc.ee/webtrans/> (20 May 2019).

¹¹ DOI: 10.15155/3-00-0000-0000-0000-07335L

are queried from the ruSkELL 1.6 corpus via Sketch Engine JSON API. In Sõnaveeb, up to 26 web sentences per lemma are shown. In many cases, especially for low-frequency words, these are the only usage examples for a headword (Koppel, 2019, forthcoming).

Although all sentences in the corpus meet the criteria of good dictionary examples (Koppel, 2017), some of them are still incorrect. In many cases, this is due to errors in corpus annotation (lemmatization and part of speech tagging); polysemous words and homonymy also cause problems. (Koppel et al., 2019, forthcoming) Users assume that all information included in Sõnaveeb is compiled or edited by lexicographers, and hence is error-free. Web sentences, on the other hand, are authentic and unedited. After receiving user feedback that some users find some of the web sentences inappropriate, the editors of Sõnaveeb decided to use the same strategy as in Merriam-Webster's¹² and Collins's¹³ dictionary portals and added an explicit note saying that the sentences were chosen automatically, they are unedited and they might contain errors. An evaluation of the Estonian GDEX configuration was carried out in 2019. The results show that according to lexicographers and Estonian language learners at the B2-C1 proficiency levels, 85% of the GDEX-selected examples were actually rated as suitable dictionary examples (Koppel, 2019, forthcoming).

5. Issues for the future

The future challenges involve compiling new data in the Ekilex, as well as the addition of new data from other dictionaries and termbases to be presented in Sõnaveeb.

- 1) **Prescriptive and descriptive data.** Concerning prescriptive data, the preparatory phase of the new normative dictionary (ÕS 2025) started in 2019. It has already been agreed that prescriptive statements will be a layer on top of the otherwise descriptive backbone, rather than a separate dictionary. This will constitute a major change for the present prescriptive dictionary (ÕS 2018), and issues may arise. Langemets et al. (2020, forthcoming) mention upcoming controversial cases where data from a descriptive dictionary (e.g. DicEst 2019) is opposed to data from a prescriptive dictionary (e.g. ÕS 2018).
- 2) **Synonyms.** At the moment only synonyms from DicEst are displayed in Sõnaveeb. We initiated the project for a synonyms database in 2019. Synonym candidates will be automatically extracted from different resources for importing into Ekilex, using word embeddings and semantic mirroring methods.
- 3) **Etymological data.** Dealing with etymology is an especially complicated area

¹² <https://www.merriam-webster.com/> (20 May 2019).

¹³ <https://www.collinsdictionary.com/> (3 June 2019).

in the data model. Etymological data is an information layer for all dictionaries, currently only consisting of the etymological information contained in DicEst. For importing, etymologies were structured by creating and linking word entities for all the source languages: automatically where possible, but leaving several types of corrections to be done manually. We also plan to import the academic Estonian Etymological Dictionary (ETY), which will add more complexity.

- 4) **Information on different language levels according to language proficiency.** About 13,000 headwords will have indications of language proficiency level (A1-C1). The data on proficiency levels comes from etLex¹⁴: a database of vocabulary of different proficiency levels compiled in the Institute.
- 5) **Frequency information.** We plan to visualize frequency information in Sõnaveeb. The information comes from the Estonian National Corpus (crawled every two years since 2017). Periodic renewals of the corpus will also make it possible to present language change information.
- 6) **Terminological data.** Ekilex contains and supports both semasiological and onomasiological termbases. Only general dictionaries have been published so far in Sõnaveeb, however. Publishing termbases is planned for 2019 and involves the decision of whether to display their information onomasiologically, as is traditional for such termbases as IATE¹⁵, or semasiologically, to be consistent with the current Sõnaveeb. Terminologists are convinced it should be onomasiological, but evidence suggests that users don't really understand the difference, and proper user research is needed.
- 7) **Bilingual data.** We plan to continue providing Russian equivalents to Estonian headwords (approx. 10,000 per year). We plan to increase the list of languages as there are more bilingual databases available at our Institute, e.g. Estonian-Latvian/Latvian-Estonian, Estonian-Finnish/Finnish-Estonian, Estonian-Chinese.

6. Conclusions

In this paper, we have described principles of aggregating and presenting of information in Sõnaveeb: a new language portal of the Institute of the Estonian Language, released in February 2019. The user can choose between two modes of information display: advanced or simple. The advanced mode is intended primarily for native speakers. The simple mode is intended primarily for learners of Estonian L2 at the A2–B1 proficiency levels. There are (so far) two language options in Sõnaveeb: it is

¹⁴ <http://www.eki.ee/keeletase> (20 May 2019).

¹⁵ <https://iate.europa.eu> (20 May 2019).

possible to choose between Estonian (monolingual) and Russian (bilingual). Users are provided with both the desktop and the responsive mobile design.

The project started in 2017 and so far the main focus has been on the development of a unified data model and on the import of different lexicographic and terminological databases from the earlier used DWSs.¹⁶ The final goal is to develop a single source of lexicographic and terminological data in order to avoid duplication of data, to improve accessibility and to force the reuse of data.

This paper reported on problems encountered so far while aggregating the data into the single source, together with the solutions we have elaborated. When unifying the dictionaries, we have paid special attention to (near) duplicates, considering their possible usefulness for different user perspectives or an empty duplication to be avoided. We have parsed and are still parsing data fields containing more than one data element.

In centralizing data from separate dictionaries and databases, we consider different information layers as specific central services. These are multimedia files (audio services and pictures), morphology, etymology, collocations, synonyms, etc. We also provide access to different kinds of external sources: corpora sentences (through Corpus Query System's API), speech synthesis and speech recognition.

We have started user research on specific information layers to get a better understanding of users' wishes and needs. We are aware that, while developing the user interface to be more and more intuitive, internet skills still need to be improved.

We will make it possible to “(re-)derive” separate datasets from the Ekilex resource if there is a demand for them. We are trying to be very careful about the authorship of different pieces of information after unification of the datasets.

The development of Sõnaveeb continues both towards tighter aggregation of existing datasets and the addition of new data from other dictionaries and terminological databases, as well as compiling new data in Ekilex. In the near future, we foresee the compilation of prescriptive data, synonyms, Estonian L2 data, neologisms, other bilingual data, terminological data, etc.

7. Acknowledgements

The creation and development of the portal was funded by the Digital Focus programme of the Ministry of Education and Research (2018–2021) and by the EKI-ASTRA programme (2016–2022). The creation of the dictionary and terminology

¹⁶ EELex <http://eelex.eki.ee/>, Termeki <https://term.eki.ee/> and Multiterm <https://www.sdl.com/software-and-services/translation-software/terminology-management/sdl-multiterm/> (20 May 2019).

database Ekilex was funded by the EKI-ASTRA programme (2016–2022). Software development has been provided by OÜ TripleDev.

The research received funding from the European Union's Horizon 2020 research and innovation programme, under grant agreement No 731015.

8. References

- BED: *Eesti keele põhisõnavara sõnastik 2019. [The Basic Estonian Dictionary 2019]* Eesti Keele Instituut. Sõnaveeb 2019. Available at: <https://sonaveeb.ee> (14.2.2019).
- Bego, K. (2018). *Ten challenges for the Internet*. Available at: <https://www.ngi.eu/news/2018/10/22/ten-challenges-for-the-internet/> (30 May 2019).
- Collins Dictionary*. Accessed at: <https://www.collinsdictionary.com/> (20 May 2019)
- DicEst: *Eesti keele sõnaraamat 2019. [The Dictionary of Estonian 2019.]* Eesti Keele Instituut. Sõnaveeb 2019. Available at: <https://sonaveeb.ee> (14.2.2019).
- ECD: *Eesti keele naabersõnad 2019. [The Estonian Collocations Dictionary 2019]* Eesti Keele Instituut. Sõnaveeb 2019. Available at: <https://sonaveeb.ee> (14.2.2019).
- Eesti Keele Instituudi eesti keele morfoloogiline andmebaas 2019. [Morphological database of Estonian 2019]* Eesti Keele Instituut. Sõnaveeb 2019. Available at: <https://sonaveeb.ee> (14.2.2019).
- Eesti Keele Instituudi vene keele morfoloogiline andmebaas 2019. [Morphological database of Russian 2019]* Eesti Keele Instituut. Sõnaveeb 2019. Available at: <https://sonaveeb.ee> (14.2.2019).
- EELex: Langemets, M., Loopmann, A. & Viks, Ü. (2006). The IEL dictionary management system of Estonian. In G.-M. de Schryver (ed.). *Proceedings of the Fourth International Workshop on Dictionary Writing Systems: Pre-EURALEX workshop: Fourth International Workshop on Dictionary Writing System*. Turin: University of Turin, 2006, pp. 11–16.
- Eesti-vene sõnaraamat 2019. [Estonian-Russian dictionary 2019]* Eesti Keele Instituut. Sõnaveeb 2019. Available at: <https://sonaveeb.ee> (14.2.2019).
- Eesti-vene õpilase ÕS 2019. [The Standard Estonian Dictionary for Russian School Students 2019]* Eesti Keele Instituut. Sõnaveeb 2019. Available at: <https://sonaveeb.ee> (14.2.2019).
- Ekilex*. Accessed at: <https://ekilex.eki.ee/> (20 May 2019)
- etSkELL 2018: *Sketch Engine for Estonian Language Learning 2018*. Accessed at: <https://etskell.sketchengine.co.uk/> (15 May 2019)
- ETY: *Eesti etümoloogiasõnaraamat (2012) [Estonian etymological dictionary (2012)]*. Tallinn: Eesti Keele Sihtasutus.
- Kallas, J., Tuulik, M. & Langemets, M. (2014). The Basic Estonian Dictionary: the first Monolingual L2 learner's Dictionary of Estonian. In A. Abel, C. Vettori & Ralli, N. (eds.) *Proceedings of the XVI EURALEX International Congress: The*

- User in Focus*, Bolzano/Bozen, Italy, pp. 1109–1119.
- Koppel, K. (2017). Heade näitelausete automaattuvastamine eesti keele õppesõnastike jaoks [Automatic detection of good dictionary examples in Estonian learner's dictionaries]. *Eesti Rakenduslingvistika Ühingu aastaraamat*, 13, pp. 53–71. DOI:10.5128/ERYa13.04.
- Koppel, K. (2019) (forthcoming). Leksikograafide ja keeleõppijate hinnangud automaatselt tuvastatud korpuslausete sobivusele õppesõnastiku näitelauseks [Suitability of automatically selected example sentences for learners' dictionaries as tested on lexicographers and language learners]. *Lähivõrdlusi. Lähivertailuja*, 29.
- Koppel, K., Kallas, J., Khokhlova, M., Suchomel, V., Baisa, V. & Michelfeit, J. (2019) (forthcoming). SkELL corpora as a part of the language portal Sõnaveeb: problems and perspectives. *Proceedings of eLex 2019*.
- KORP: Accessed at: <https://korp.keeleressursid.ee/> (15 May 2019)
- Kosem, I., Koppel, K., Kuhn, T. Z., Michelfeit, J. & Tiberius, C. (2018). Identification and automatic extraction of good dictionary examples: the case(s) of GDEX. *International Journal of Lexicography*, 32(2), pp. 119-137. <https://doi.org/10.1093/ijl/ecy014>.
- Langemets, M., Tiits, M., Udo, U., Valdre, T. & Voll, P. (2018). Eesti keel uues kuues: Eesti keele sõnaraamat 2018. *Keel ja Kirjandus*, 12, pp. 942–958.
- Langemets, M., Kallas, J., Norak, K. & Hein, I. (2020) (forthcoming). New Estonian Words and Senses: Detection and Description. *Dictionaries*.
- Lew, R. (2013). Online dictionary skills. In I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets & M. Tuulik (eds.) *Proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia*. Available at: http://eki.ee/elex2013/proceedings/eLex2013_02_Lew.pdf.
- Merriam-Webster Dictionary*. Accessed at: <https://www.merriam-webster.com/> (20 May 2019)
- ruSkELL1.6: *Sketch Engine for Language Learning (SkELL) for learners of Russian*. Accessed at: <https://www.sketchengine.eu/ruskell-examples-and-collocations-for-learners-of-russian/> (15.5.2019).
- Sketch Engine*. Accessed at: <https://www.sketchengine.eu/documentation/api-documentation/> (15.2.2019)
- Sõnaveeb: *Sõnaveeb 2019 [Wordweb 2019]*. Accessed at: <https://sonaveeb.ee> (20.5.2019)
- Tavast, A., Langemets, M., Kallas, J. & Koppel, K. (2018). Unified Data Modelling for Presenting Lexical Data: The Case of EKILEX. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (eds.) *Proceedings of the XVIII EURALEX International Congress: EURALEX: Lexicography in Global Contexts*. Ljubljana, Slovenia, pp. 749–761.
- VL: *Võõrsõnade leksikon. [The Dictionary of Foreign Words] 8., põhjalikult ümber töötatud trükk*. Eesti Keele Instituut, Kirjastus Valgus, 2012. Available at: <http://www.eki.ee/dict/vsl/> (30.5.2019).
- Weitzman, L. (2014 [2004]). Meta-design for “sensible” information. *Interactions*, Vol.

11, Issue 2, March, April, pp. 71–73. Updated by author in 2014. DOI: 10.1145/971258.971284.

ÕS 2018: *Eesti õigekeelsussõnaraamat ÕS 2018*. [*The Dictionary of Standard Estonian ÕS 2018*]. Eesti Keele Instituut. Tallinn: Emakeele Sihtasutus, 2018). Available at: <http://www.eki.ee/dict/qs2018/> (30.5.2019).

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



LeXmart: A Smart Tool for Lexicographers

Alberto Simões^{1,2}, Ana Salgado³, Rute Costa³,

José João Almeida²

¹ 2Ai – Instituto Politécnico do Cávado e do Ave

² Algoritmi, Universidade do Minho

³ NOVA CLUNL, Universidade NOVA de Lisboa

E-mails: asimoes@ipca.pt; anasalgado@campus.fcsh.unl.pt; rute.costa@fcsh.unl.pt;
jj@di.uminho.pt

Abstract

The digital era has brought some challenges to lexicographers, but it has also brought new opportunities as part of the rise of information technology and, more recently, the emergence of digital humanities. This paper provides a description of LeXmart, the framework that supports the digital development of the Portuguese Academy of Sciences Dictionary. LeXmart is a smart tool framework to support lexicographers' work that offers different types of tools, ranging from a structural editor to a set of validation tools.

Given that the dictionary is stored in eXist-DB, LeXmart is developed on top of its ecosystem, using W3C standard languages, and offering default functionalities offered by eXist-DB, namely a RESTful API.

Keywords: e-lexicography; dictionary; lexical databases; lexicographic framework; XML

1. Introduction

The digital era has brought both challenges and new opportunities to lexicographers on the back of Information Technology and the recently developed Digital Humanities. Most e-dictionaries are now embedded into websites, mobile applications, and digital products, besides being also offered as services. Lexicographers have been using a number of computational tools, e.g., word processors, spreadsheets, and in a few cases, databases for their work. Large publishing houses have developed their own in-house systems, but few have made their applications freely or even commercially available. In these new settings, the lexicographic work had to change its course so as to prepare resources and create formats to achieve the main goals of this era: sharing and reusing dynamic data enabling interoperability by using standards and compatible formats.

This paper provides a description of the LeXmart¹ framework to support lexicographers' work, which underlies the digital development of the Portuguese Academy of Sciences Dictionary (DACL), and focuses particularly on its implementation, database support, structural editor, and reporting tools, which have

¹ <http://www.lexmart.eu/>

proven to be useful for lexicographers to edit the entries and run control checks on them.

As mentioned above, the concept of a dictionary and its production process has undergone major changes on the back of new technologies. Although we can say this also holds true for Portugal, the fact is that these digital resources continue to be designed and implemented according to the same typographic and editorial conventions of the former print editions, “We still consult dictionaries by going to a particular web site. Dictionaries do not come to us” (Tasovac, 2010: 1), without exploring the possibilities of the digital context (Tarp, 2009; Trap-Jensen, 2018).

This paper is structured as follows: Section 2 presents a small introduction to the DACL and its background, and summarizes the process of its conversion from PDF to the structured format of the Text Encoding Initiative (TEI) Dictionaries Chapter. Section 3 presents LeXmart in detail – this section focuses on three main aspects of the framework: tools for lexicographic work, tools supporting website development and information availability, and a brief discussion of the current RESTful API. Finally, in Section 4, we draw some conclusions about the functionalities of the tool, and conclude with further research avenues, both for the specific case of the DACL and of LeXmart.

2. The Portuguese Academy Dictionary

In Portugal, in spite of the successive attempts of the Academy of Sciences (ACL), only in the 21st century (more precisely in 2001) did the ACL publish a complete dictionary (from A to Z), *Dicionário da Língua Portuguesa Contemporânea*, in a two-volume paper version (the first volume from A to F and the second from G to Z). At that time, the authors decided, for a computational approach, to develop a database using Microsoft Access, and a reporting tool to generate a Word file for the dictionary, which was subject to some minor changes both in content and format before printing. Although the database, or even the work file, would be the best source for future developments, the only media that survived these 18 years was the PDF file that originated those same printed versions. In 2015, some preparatory work for an online Portuguese Academy Dictionary was performed through the *Instituto de Lexicologia e Lexicografia da Língua Portuguesa* (ILLP) and a database was developed by a team working in Natural Language Processing at the University of Minho², which now draws on the participation of IPCA³ and NOVA CLUNL⁴.

The DACL is a general language contemporary dictionary with a descriptive nature and a normative concern. It had a synchronous printed edition and it is addressed to a

² The team works with Alberto Simões and José João Almeida (Natural Language Processing of the Computer Science Department), and the consultancy of Álvaro Iriarte Sanromán.

³ Alberto Simões from IPCA is responsible for the technological support of the new digital ACL dictionary.

⁴ The participation of NOVA CLUNL is related to the DACL’s transition into the TEI Lex-0 format.

vast audience whose mother tongue is Portuguese. A typical entry includes the following elements: headword, pronunciation, followed typically by some linguistic information (e.g., part of speech), the different meanings, usage labelling, synonyms, antonyms, collocations, etymology, and notes. Examples of usage labelling, cross-references, etc., may also be present. In order to guarantee the interoperability and reusability of dictionary content, during the DACL encoding process, the authors have been participating in the TEI Lex-0 discussion⁵, a streamlined version of the TEI Guidelines, simplified and enhanced for regular use.

2.1 Reverse engineering: from a PDF to a structured TEI document

The project started with the automatic conversion of a PDF file into a text format, where each string was annotated with its position on the page and the font face and font size used in the original document. A list of pairs containing font faces and sizes was computed and analysed manually. For example, small caps were used to indicate synonyms and antonyms; very large fonts corresponded to the opening letter of each section of the dictionary; a specific font list was used for phonetic transcription. Unfortunately, most of the document uses the same font face and font size, making it impossible to detect automatically what their role in the entry is. Using this information, a superficial and very rough annotation was performed on the PDF transcription.

The next step resulted in the detailed annotation using a set of rewriting rules. These rules, instead of being applied to font information, were applied to the annotated parts of the document and their content. As a case in point, to detect synonyms and antonyms, rewriting rules searched for the asymptotically equal (\simeq) or the not asymptotically equal ($\not\simeq$) signs. For other finite lists (e.g., grammatical information), a list of the allowed values was prepared manually. For other annotations, positional information (relating to the other already annotated portions of the document) was used.

In order to make this process easier, and as the headwords of the dictionary entries were easy to detect (with a few exceptions that were fixed manually), the full dictionary was divided into thousands of small documents, one for each dictionary entry. This was useful to ensure that the rewriting process was not applied to entries that had already been validated by the TEI schema.

2.2 XML Database

Different approaches were analysed in order to allow lexicographers to edit each dictionary entry cooperatively. The first option considered was the storage of each XML file in a version control system, such as Subversion or GIT. Lexicographers would use

⁵ A contribution to the work developed by the DARIAH-ERIC Lexical Resources group: <https://www.dariah.eu/activities/working-groups/lexical-resources/>.

an IDE (Integrated Development Environment), such as oXygen's XML Editor⁶ or Altova XML Spy⁷, in order to create, edit, delete and validate entries. Two main issues were behind the decision not to follow this direction: lexicographers had to use the version control system directly (although it would not have been difficult to teach them how to use it, as there are very intuitive clients for these systems, such as GitKraken⁸ or Atlassian SourceTree⁹, since there was no regular staff, but rather a dynamic team of volunteers, training sessions would have been very hard to schedule); and the difficulty of making the IDE work in a transparent way, without the need for deep XML knowledge. Although not the main issue, the need to index and search the XML files also made us look for other ways of managing XML files.

The second option was to store the documents in a database. For that, and after searching for some options, the eXist-DB¹⁰ database was chosen. Although there are other interesting databases, eXist-DB developers work closely with oXygen XML Editor developers, which makes it easy to connect and use oXygen to edit files stored in eXist-DB. While we do not intend to have all the lexicographers using oXygen, the fact that both the developers and the project coordinators can use it is a valuable asset.

The choice of using eXist-DB paid off, as it is not just an XML aware database, but a feature rich platform to develop XML based applications, allowing the development of websites entirely with W3C standard XML technologies, e.g., XPath, XQuery and XForms. This was the beginning of LeXmart, as small tools started to be developed on top of eXist-DB and, from tool to tool, an interesting and useful framework was developed.

3. LeXmart: a smart framework to support the lexicographers' work

LeXmart is an open-source web platform created to allow lexicographers to easily edit and publish lexical resources. As noted at the end of the previous section, LeXmart started as a set of small independent tools developed on top of eXist-DB. These tools were later compiled in a common interface, resulting in the framework we are presenting here.

This section starts by discussing other tools available to lexicographers to develop their work; it follows with the description of the tools developed on top of the eXist-DB platform, starting with the end-user features (searching), lexicographic support tools

⁶ <https://www.oxygenxml.com/>

⁷ <https://www.altova.com/>

⁸ <https://www.gitkraken.com/>

⁹ <https://www.sourcetreeapp.com/>

¹⁰ <http://exist-db.org/>

(creating, deleting and editing entries, validating entries, detecting inconsistencies in the whole dictionary), and content management tools; it then provides a brief description of the available API offered by eXist-DB and what will be made public very soon.

3.1 Dictionary editing tools

With the advent of personal computers, publishers started using software applications to help their work on preparing the material for printed dictionaries. While in some situations authors simply used a standard tool (such as a database management system) to help store the information about each dictionary entry, some large companies developed their own dictionary management tools. There is little information regarding these, as such tools were developed in-house to support the publisher's editorial work, and not as commercial tools.

Using the Internet as the backend for a dictionary management system is not new. The DEB (Horák & Rambousek, 2007) was one of the first examples. At that time, Web 2.0 was already a reality, but the DEB was still developed as a typical CGI (Common Gateway Interface) application. Its entries were stored in a Berkeley DB XML database that although XML-aware lacked most of the new XML database functionalities. The interface was also complex and not easy to use. This project evolved (Rambousek & Horák, 2015), implementing SOAP Web Services to interact between a server (DEB) and a set of clients. The server is responsible for the management of the data, using W3C standards, and specifically its dissemination as linked data. DEBWrite is one of the clients, and acts as a front-end application for lexicographers. In order to offer higher customization on the structure of the dictionary entries, DEBWrite provides an online editor for the dictionary micro-structure that parameterizes the dictionary editor. The resulting editor for the dictionary entries is now more versatile, but the interface stills lacks some usability.

LeXmart has been developed since 2016 (Simões et al., 2016a). More recently, Lexonomy (Měchura, 2017) is a good example of what modern dictionary editing software can look like. Lexonomy, offered both as a service and as a software package, also uses Xonomy as the XML editing software, while SQLite is used as the data backend.

3.2 End-user tools

The DACL is not yet publicly available for end-users. Nevertheless, searching the dictionary is crucial for end-users and lexicographers alike. Therefore, two different approaches were implemented to perform searches: one to search by headword and thus quickly find a definition; and another search by entry content (any part of the entry) enabling a broader search (named *reverse* search), and allowing the user to use the DACL almost as an onomasiological dictionary (Simões et al., 2016b).

The implementation of such queries is quite simple in XQuery, as it allows the search for XML elements containing specific words. Therefore, in the first search type, the query is performed looking up the content of `orth` elements, while in the second search type, the query is performed for the textual content of the whole entry.

The only relevant detail is that eXist-DB uses Lucene as its document database, and therefore the convenient definition of search indexes can make queries much faster.

Presenting the search results is even simpler. With the advent of HTML5, all modern browsers support HTML documents with XML fragments inside (or with HTML with custom tags, if you prefer). Thus, the XQuery script just outputs the entry's XML directly to the browser, which renders it with a custom-defined Cascading Style Sheet (CSS) file. If a user searches, for example, the word *golfinho*, they may obtain all the results where the word *golfinho* occurs, not only in the lemma, but in any section of the lexicographic articles (see Figure 1).

The screenshot shows the search results for the term 'golfinho' in the Dicionário da Academia das Ciências de Lisboa. The interface includes a header with the logo and name of the dictionary, a search bar with 'palavra' and a magnifying glass icon, and a 'Páginas' dropdown. The results are displayed in three separate boxes, each with a title, a revision status, and a list of entries. The first box is for 'Delfim:4', the second for 'Golfinho:2', and the third for 'beluca , beluga'. Each entry includes its part of speech, domain, and a detailed description. The database URI and revision status are also shown at the bottom of each result box.

Entry	Part of Speech	Domain	Description	Database URI	Revision Status
Delfim:4	n. m.	Astron.	Constelação boreal, que ocupa uma superfície de 189 graus quadrados, situada nas proximidades do Equador celeste, entre a constelação de Águia e a do Cavalo Menor. (Do lat. <i>delphin</i> , - <i>inis</i> < gr. δελφίς 'golfinho', relativo ao cetáceo que levou Anfitriote a Neptuno para ser desposada)	/db/academia/Delfin_4.xml	Importado
Golfinho:2	n. m.	Astron.	O m. que Delfim:4 (Do gr. δελφίς, -ίως, pelo lat. <i>delphin</i> , - <i>inis</i> , alterado por influência de <i>golfo</i>)	/db/academia/Golfinho_2.xml	Importado
beluca , beluga	n. f.	Zool.	1. Esturção branco (<i>Acipenser huso</i> ,) que atinge mais de cinco metros de comprimento e de cujas ovas se faz caviar. 2. Golfinhos (<i>Delphinapterus leucas</i> ,) de pele branca, das regiões árticas. (Do rus. <i>belukha</i>)	/db/academia/beluca.xml	Importado

Figure 1: Result of the reverse search for *golfinho* [dolphin] – first three hits.

As can be seen, the entries are shown sorted (proper names are shown first – *Delfim:4*, *Golfinho:2*, and then common names – *beluca*). Meta-information about the entry is also shown (the database URI, e.g. `/db/academia/Delfin_4.xml`, for the entry document and its revision status, in this case “Importado” [imported]).

3.3 Lexicographic support tools

3.3.1 Entry editor

While using a dedicated XML editor such as oXygen can boost productivity as it contains quite interesting features, it is not user-friendly, and its usage can be rather complex in some situations. In order to allow faster editing, an online editor was developed on top of eXist-DB, based on the Xonomy¹¹ JavaScript editor. This editor can be accessed by all authenticated users after a headword search. Figure 2 shows the interface presenting the entry for *arrulho* (cooing). Note that there are two buttons, one for editing the entry, and another one for deleting it.

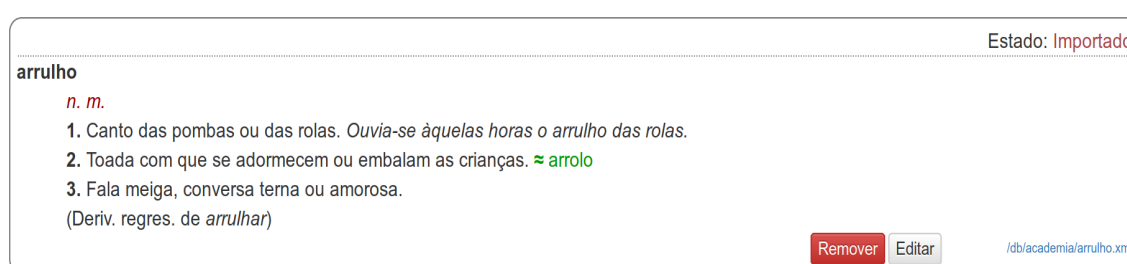


Figure 2: Entry for *arrulho* with authenticated interface for entry editing.

Xonomy is configured by a JavaScript data structure, annotated with some JavaScript functions, that specifies the allowed XML structure, and enables the configuration of drop-down menus to insert, remove, or adapt parts of the entry. The documents edited by Xonomy are fetched and stored using AJAX (Asynchronous JavaScript and XML) calls to the eXist-DB RESTful API.

While Xonomy has its own limitations to support some validation aspects, the XML is internally rewritten to a non-standard XML format, which Xonomy is able to understand and manage correctly. When the lexicographer saves the entry, this non-standard XML format is again converted into valid TEI.

Figure 3 shows Xonomy working. While its appearance is quite similar to an XML document, it is presented without the visual noise of the opening/closing tags. The elements can also be configured with actions. That same figure shows the menu that pops up when the user clicks on a sense tag. This menu allows adding some metadata to the entry (revised or as a new meaning), adding a new sense after the current one, removing completely the selected sense, or marking it as digital only. This flexibility of Xonomy that can thus define different actions directly on tags allows the lexicographer to work without the need to know the TEI structure, or the need to directly write XML elements.

¹¹ Available at <https://github.com/michmech/xonomy/>.

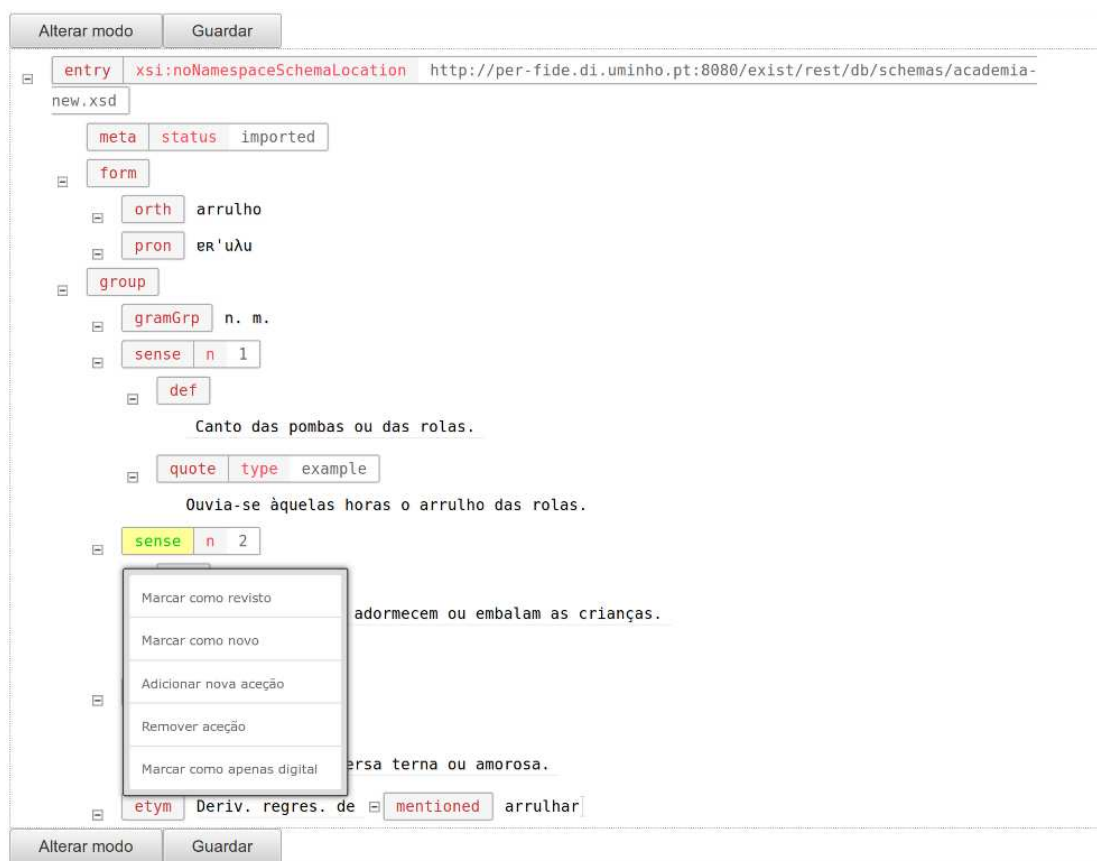


Figure 3: Xonomy XML editor on top of eXist-DB.

There is also an option to create a new entry in the dictionary. It validates a word that has not been included yet (and if it has, it requires the lexicographer to force its creation with a different entry number), creating the basic XML structure.

3.3.2 Entry creation and deletion

As shown in the previous section, when searching an entry the result list shows a button to remove that entry. When this button is used, a pop-up asks the user to confirm the deletion and removes the entry if requested. Given that the eXist-DB data is being exported to the filesystem as a collection of XML documents that are being stored in a GIT repository (once a day), there is a comprehensive backup of all the changes and deletions, allowing us to recover any mistakenly deleted entry.

Regarding the creation of new entries, there is a small form asking only for the headword. The system automatically searches to see if the word already exists in the dictionary and, if it does, the user is requested to rethink the entry creation, or to explicitly indicate the entry number. If the word is not included in the dictionary, then a new file is created with a boilerplate XML document, with the headword already filled in, and enough structure for the lexicographer to start writing the definition right away.

3.3.3 Meta annotation

Although not directly a developed tool, the annotation of entries or parts of entries with metadata is extremely relevant in order to allow the lexicographers to organize their work.

For example, marking the editing status of an entry is extremely important. For this reason, the possibility of adding this kind of annotation was created. Initially, an entry has the status “imported” (from the original PDF). New entries are created with the “new” tag. Then “revised” is used when the entry has been revised (it is a completed entry) and, finally, “edited”, when only a sense or part of the entry has been edited. These statuses can be inserted at the level of the entry or at the level of the different elements that compose the microstructure (usually, senses).

Another important notation is the “digital only” tag, which only appears at the level of the entry or sense, and signals the senses or entries that will only appear in the digital version of the dictionary (and will be excluded from any paper versions).

3.3.4 Filters and statistics

Dictionaries contain information from different sources: different countries or regions of a country, different domains of knowledge, different register types (colloquial, formal, etc.). All this information needs to be codified in the dictionary, and needs to be coherent across the dictionary.

It is easy to find examples of hand-made dictionaries where different abbreviations are used for the same word, different words are used to catalogue different senses in the same domain, and these are only a couple of very simple examples. Using computer tools to assist on the development of a dictionary means these tools should enable some form of consistency check. In part, consistency can be easily guaranteed by using pick-up lists in the editor, but when the work stems from an existing dictionary, other tools need to be developed to find already existing inconsistencies.

In order to allow the lexicographer to control precisely this kind of information, LeXmart has tools to create lists of entries for each type of annotation, and to view graphically the distribution of that information about use.

To provide an example of how these tools are used, consider the work on a specific domain of knowledge, such as biochemistry. While lexicographers are able to construct the entries, and check their structure and completeness, they might not be apt to evaluate the quality of the definitions, or even to write them in the first place. The possibility of filtering the dictionary by a specific area of knowledge allows the lexicographer to export all the entries from that area into a PDF file and send it to an expert in that area. This same type of approach can be used for geographic variants. It is not likely that a Portuguese lexicographer is completely sure about information regarding words imported from Brazil, Angola or Mozambique.

As for statistics (see Figure 4), LeXmart allows the lexicographer to look at the list of possible values for a specific type of markup and understand if there are duplicates (with different forms) or look at a graph and realize whether a specific area of knowledge has insufficient entries to be considered as independent (for example, the printed DACL dictionary has a single entry in the cutlery domain).

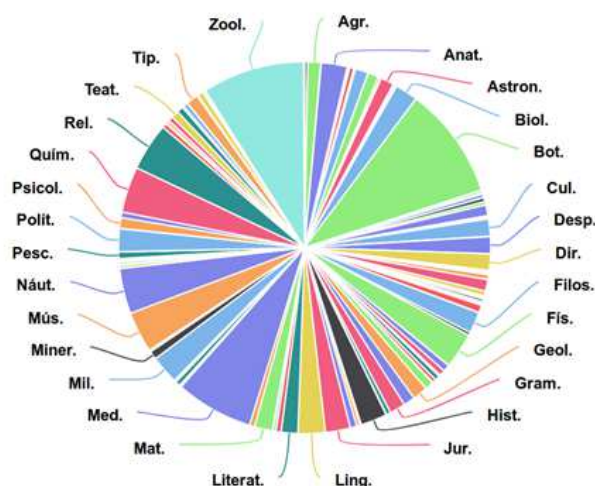


Figure 4: Distribution of areas of knowledge in the DACL.

A final filtering functionality is also available: exporting the entries from a specific text file. Basically, it is possible to upload a text file, where each line includes a headword, and the system will output the entries for those terms.

3.3.5 Reports

In order to understand the evolution process of the lexicographic work, LeXmart allows users to export reports. Currently implemented reports include listing all the new entries that were not present in the printed version of the DACL, the entries from the DACL that were already edited inside LeXmart, and the list of the entries that are marked as finished.

3.3.6 Validation

The eXist-DB database validates the *well-formedness* of the XML syntax, and only allows the storage of valid files. Although it is also possible to configure the database to validate the XML according to a specific schema, that was not the choice as it would limit the storage of files that are being modified, or it would break the full database whenever a minor change was made to the schema. Nevertheless, it is important to know which entries need to be edited and corrected to comply with the defined schema. For that, we created an XQuery validation script. As simple as this script may seem, it took some time to understand the different approaches available for eXist-DB to validate schemas. It takes about 3 minutes to validate the 69K entries outputting an XML document with a report for each failing file. To make reports easier to read, the

XQuery script was tuned to output only the invalid entries. Without it, a full report for all the files would be created.

3.4 Beyond the lexicographic work: content management system

Not directly related to the lexicographic work, a minimal content management system was created in order to allow the creation of ad-hoc pages with relevant information about the dictionary. This system is based on an independent collection where an XML page is created for each page to be published. The pages are edited using TinyMCE¹² 14, a well-known WYSIWYG editor based on JavaScript.

3.5 Portuguese Academy Dictionary RESTful API

Given that the dictionary is stored in eXist-DB, it comes by default with a RESTful API. While the API is currently private, we are working on making the DACL freely available on the web and as soon as that work is finished the API will also be made available. The existence of this API makes bulk editing possible.

In some situations, bulk editing was needed: either some error from the conversion process was detected, or the schema changed to accommodate some new data, or even some changes needed to be made to the entire dictionary. This is still true at the moment as the DACL is progressively being converted from the TEI standard to TEI Lex-0.

For those situations, a practical way to edit each and every document in the database or edit every document that matches a specific pattern is highly relevant. Although the edition can be done entirely in XQuery, having access to a rich language with powerful regular expressions was crucial. With that in mind, a new Perl module was developed (XML::eXistDB::REST) that allows the query of the dictionary, retrieval of documents, and updating their content. This module is under work, but a beta version is already available at the Comprehensive Perl Archive Network (CPAN).

This type of approach has the major disadvantage of not being completely integrated with LeXmart. Nevertheless, its importance makes it worth mentioning.

4. Conclusions and future work

The challenge of converting a paper dictionary into an electronic dictionary is not a new one. This has been done by different teams, and we did it for the *Dicionário Aberto* (Simões et al., 2016b) and for the *Dicionário de Sinónimos do Galego* (Gómez Clement et al., 2016). Although we have presented the process of reverse engineering the PDF file and converting it into an electronic dictionary, that is not our main goal. We intend

¹² <https://www.tiny.cloud/>

to use that dictionary to bootstrap it to live electronically and allow snapshots to paper whenever necessary.

The process of creating a dictionary from scratch or using a previous version as a base can lead to similar problems: how to allow concurrent editing, how to force coherence, how to guarantee regular backups, and other issues. Therefore, we have discussed our approaches to these problems, and how our system was prepared to help lexicographers in their tasks.

Although an interesting set of tools has already been developed, some other requirements made by the lexicographers need to be addressed in the near future:

- Instead of creating HTML reports of each week's work, we intend to create daily and weekly reports of editions, generated as XML documents, imported into another collection. This is a very interesting resource to have, in order to monitor the activity in the dictionary, and to have a log of every change performed.
- Currently, our web application is restricted to authenticated users. In the future, an open interface needs to be available to end-users. Although the simple mechanisms to search for entries are already developed (although restricted), we think there are a couple of other interesting approaches. For example, synonym and antonym annotation can be used to present the dictionary as a graph/WordNet-like structure.
- Formats – either as eBooks or a print version. For that, we expect to create a set of exporting tools, both to ePub format and to PDF. For the latter, we expect to use LaTeX¹³ or XSL-FO¹⁴, as these tools enable the automation of the exporting process. This could even allow the dictionary to be exported as different volumes by knowledge area.
- Regarding the framework, LeXmart needs some polishing and should be translated into English. We intend to have the current version available in a GIT repository very soon.

5. Acknowledgements

Research financed by Portuguese National Funding through the FCT – Fundação para a Ciência e Tecnologia as part of the project Centro de Linguística da Universidade NOVA de Lisboa – UID/LIN/03213/2019, and by the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015 (ELEXIS).

¹³ <https://www.latex-project.org/>

¹⁴ Extensible Stylesheet Language Formatting Objects - <https://www.w3.org/TR/xslfo20/>

6. References

- Gómez Clemente, X. M., Guinovart, M. G. & Simões, A. (2016). *Dicionário de Sinónimos do Galego*. Xerais.
- Horák, A. & Rambousek, A. (2007). DEB Platform Deployment – Current Applications. In RASLAN 2007: Recent Advances in Slavonic Natural Language Processing, Brno, Czech Republic. Masaryk University, pp. 3–11.
- Měchura, M. B. (2017). Introducing Lexonomy: an open-source dictionary writing and publishing system. In I. Kosem et al. (eds.) *Electronic Lexicography in the 21st Century: Lexicography from Scratch. Proceedings of the eLex 2017 conference, 19-21 September 2017, Leiden, The Netherlands*. Brno: Lexical Computing Ltd., pp. 662-679.
- Rambousek, A. & Horák, A. (2015). DEBWrite: Free Customizable Web-based Dictionary Writing System. In I. Kosem, M. Jakubiček, J. Kallas & S. Krek (eds.) *Electronic lexicography in the 21st century: linking lexical data in the digital age*. Ljubljana/Brighton: Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd., 2015, pp. 443–451. ISBN 978-961-93594-3-3.
- Simões, A., Almeida, J. J. & Salgado, A. (2016a). Building a Dictionary using XML Technology. In *5th Symposium on Languages, Applications and Technologies (SLATE'16), vol. 51 of Open Access Series in Informatics (OASICS)*. Germany: Dagstuhl. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, pp. 14:1–14:8.. DOI: <http://dx.doi.org/10.4230/OASICS.SLATE.2016.14>.
- Simões, A., Iriarte, A. & Almeida, J. J. (2016b). Dicionário-Aberto: construção semiautomática de uma funcionalidade codificadora. In A. Lemaréchal, P. Koch & P. Swiggers (eds.) *Actes du XXVIIe Congrès international de linguistique et de philologie romanes (2013)*, pp. 201–300, Nancy, july. ALTIF. Section 16: Projets en cours; ressources et outils nouveaux.
- Tarp, S. (2009). Beyond Lexicography: New Visions and Challenges in the Information Age. In H. Bergenholtz, S. Nielsen & S. Tarp (eds.) *Lexicography at a crossroads. Dictionaries and Encyclopedias Today, Lexicographical Tools Tomorrow*. Bern: Peter Lang AG, International Academic Publishers, pp. 17–32.
- Tasovac, T. (2010). Reimagining the Dictionary, or Why Lexicography Needs Digital Humanities. In *Digital Humanities 2010*, pp. 254–256.
- Trap-Jensen, L. (2018). Lexicography between NLP and Linguistics: Aspects of Theory and Practice. In *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*. Ljubljana: Ljubljana University Press, Faculty of Arts, pp. 25–37.

Websites:

- TEI Consortium, eds. TEI P5: Guidelines for Electronic Text Encoding and Interchange. [Version 3.5.0]. [Last updated on 29th January 2019, revision 3c0c64ec4]. TEI Consortium. <http://www.tei-c.org/Guidelines/P5/> ([13.07.2019]).

Dictionaries:

DACL: *Dicionário da Língua Portuguesa Contemporânea*. (2001). João Malaca Casteleiro (coord.), 2 vols. Lisboa: Academia das Ciências de Lisboa & Editorial Verbo. New digital edition under revision.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



Make My (Czechoslovak Word of the) Day

Michal Škrabal¹, Vladimír Benko²

¹ Charles University, Institute of the Czech National Corpus,
Panská 7, 110 00 Praha 1, Czech Republic

² Slovak Academy of Sciences, E. Štúr Institute of Linguistics,
Panská 26, 811 01 Bratislava, Slovakia

E-mail: michal.skrabal@ff.cuni.cz, vladimir.benko@juls.savba.sk

Abstract

Our paper introduces an experiment aimed at creating a database to be used as the source for a *Word of the Day* (*WotD*) application. Using a database of translation equivalents derived from a Czech-Slovak parallel corpus as a point of departure, semi-automated procedures are described that would preprocess the raw data so that the size of the lexicon to be processed manually is minimized. A by-product of this experiment is a list containing Czech to Slovak translation equivalents of differing levels of similarity, which could be an interesting source of information for Czech and Slovak contrastive studies.

In the last chapter the lexicographical application of acquired data is described. The criteria for selecting individual headwords remain an open question at the moment. Personally, we lean towards a combination of different aspects so that the final selection is as diverse and user-attractive as possible. The intended microstructure of the *WotD* dictionary entry is also presented. Its first peculiarity is the dual metalanguage, making it two explanatory dictionaries in one rather than a translation dictionary. Secondly, the content of the entries is closely related to the digital-born and corpus-based nature of the dictionary. Thus, some elements presented in traditional explanatory dictionaries are reduced or completely omitted in our microstructure – while others are highlighted.

Keywords: Word of the Day; translation equivalent; Czech; Slovak; *Treq* database

1. Introduction

Many online dictionaries and other lexicographic/didactic resources have their *Word of the Day* (*WotD*), a feature that on a daily basis focuses on a chosen lexeme, giving users a wide range of varied information about it. For example, Merriam-Webster's *WotD*¹ presents the profile of the selected word every day. It makes reference to the pronunciation of the expression, its definition, a brief commentary on the origin of the word and connection with other, related words and, eventually, two or three examples of its usage, most often from current media, sometimes also from older literary works. To make *WotD* even more interactive and entertaining, it also contains numerous links to additional materials concentrated on the Merriam-Webster web portal (such as *Test Your Vocabulary*, *Word Games*, *Trending Now*, *Words at Play*, etc.).

¹ <https://www.merriam-webster.com/word-of-the-day>.

Another example, *One Hungarian Word a Day* (OHWaD)², aims at a different target group: being written in English, it is primarily intended for L2 students of Hungarian. At the beginning they are asked to guess the meaning of the selected word from three possibilities, whereupon the correct English equivalent is revealed. Subsequently two or three example sentences are given as well as a short glossary of semantically close words and phrases with their English counterparts. This way students learn six new words from Monday to Saturday, whereas Sunday is dedicated to revision in the form of a quiz: students are supposed to choose the correct equivalent for six newly learned words and to use each word in the made-up Hungarian sentence.

Another concept hidden under a similar name can be found in, for example, the Polish project *Słowa dnia*³. These “words of the day” are based on the relative frequency of words in daily newspapers that is clearly higher than their frequency in the comparative period of the previous year (cf. also Meriam-Webster’s *Word of the Year*⁴ based on the frequency with which each word has been searched for in the dictionary in the past year). Of course, frequency may be one of the criteria for selecting such “prominent” words (see also chapter 4 below), nonetheless, our project is closer to the first two projects mentioned above.

Since, at least to our knowledge, there is no such project for either Czech or Slovak, we thus propose a simple database to help generate individual parts of such a series for either of these languages. It would be a rudimentary automated system open to extra modules that could facilitate lexicographers’ work and utilize the corpus data (that are available for both languages in abundant volume) as much as possible. Besides, it combines a modern, quantitative approach with traditional lexicographical practice (definitions taken from older printed dictionaries, etymological information, etc.) and incorporates the long-standing and very popular tradition of so-called “linguistic columns” (called *jazykové koutky* in Czech / *jazykové kútiky* in Slovak) into a lexicographical project.

2. The data

Though a bilingual Czech to Slovak dictionary (Horák et al., 1979; cf. also Gašparíková & Kamiš, 1967; Nečas & Kopecký, 1964) was available in machine-readable form, we decided not to use it for this project, mainly for two reasons. Firstly, its paper version was published four decades ago and therefore does not reflect recent developments in either the Czech or Slovak lexis, especially after the political changes in our societies since 1989. Secondly, as it had been compiled in the pre-corpus era, many translation equivalents are not sufficiently attested, or are even simply wrong (cf. Ripka &

² <https://www.catchbudapest.com/one-hungarian-word-day>.

³ <http://slowadnia.clarin-pl.eu>.

⁴ <https://www.merriam-webster.com/words-at-play/word-of-the-year-2018-justice>.

Skladaná, 1980). Moreover, we *could* use a resource that is much more up-to-date, with translation equivalents attested in a parallel corpus and supplemented with frequency data.

2.1 The *Treq* database

The *Treq*⁵ application serves for querying the Czech to foreign language(s) dictionaries that have been automatically created based on data derived from the *InterCorp* parallel corpus (Čermák & Rosen, 2012). This parallel corpus also includes a Czech-Slovak component (Nábělková & Vavřín, 2018) that currently (in version 11) comprises the following text types:

- fiction (the so-called *Core* [of the corpus])⁶ – 10.5 million tokens;
- legal texts of the European Union from the *Acquis Communautaire* corpus – 23.3 million tokens;
- proceedings of the European Parliament dated 2007-2011 from the *Europarl* corpus – 14.8 million tokens;
- movie subtitles from the *Open Subtitles* database – 7 million tokens.

The overall size of the whole *InterCorp* v11 corpus is more than 1.7 billion running words / 2.14 billion tokens⁷, of which more than 45.4 million running words / 56.2 million tokens accounts for a Czech-Slovak component (i.e. less than 3%). Nevertheless, this amount of data is sufficient for our purposes.

Access to the extracted data⁸ is mediated by the *Treq* online search interface (<http://treq.korpus.cz/>). The application provides a list of all translation candidates of a given word (or even multi-word expression) found in *InterCorp* that are, by default, sorted by decreasing frequency. The more often the equivalent of the search term occurred compared to other equivalents, the higher the probability that it is plausible.

2.2 The *Treq* dump format

Besides the online access, the *Treq* database was available for use in the framework of our experiment in a simple three-column text format, containing the *frequency*, *Czech*

⁵ The acronym *Treq* stands for *Translation Equivalents*.

⁶ Only fiction texts have been manually corrected in terms of OCR and sentence alignment. All other texts were processed automatically only. For the list of tools used, see <http://wiki.korpus.cz/doku.php/en:cnk:intercorp:verze9#acknowledgements>.

⁷ For information about the exact composition of the corpus and the size of its components, see <http://wiki.korpus.cz/doku.php/en:cnk:intercorp>. For general information about the InterCorp project, see Čermák & Rosen (2012) or Rosen (2016).

⁸ For detailed information about the automatic processing of data, see Škrabal & Vavřín (2016).

word, and *Slovak word*, respectively. Lists for both lemmatized and raw word forms derived from the four basic *InterCorp* components were provided in eight separate files, with total word counts as shown in Table 1.⁹

Treq component	Word forms	Lemmas
Core	433,962	198,346
Acquis	808,812	438,023
Europarl	716,703	348,963
Subtitles	489,324	292,231

Table 1: *Treq* source data.

All files were sorted by descending frequency. The first 20 lines of the two *Acquis* files are shown in Table 2.

Lemmas			Word forms		
Freq	cs	sk	Freq	cs	sk
807,038	.	.	806,355	.	.
764,487	,	,	752,456	,	,
473,280	a	a	478,142	A	a
345,762	v	v	343,364))
343,458))	254,050	V	v
249,537	((249,578	((
215,721	na	na	207,633	na	na
190,750	být	byť	141,178	se	sa
190,027	článek	článok	124,192	O	o
144,811	se	sa	112,538	1	1
140,941	s	s	112,514	nebo	alebo
132,084	nařízení	nariadenie	92,848	;	;
130,075	který	ktorý	90,316	Článek	Článok
125,418	o	o	88,897	"	"
116,925	z	z	87,046	2	2
113,391	nebo	alebo	85,123	:	:
113,238	komise	komisia	84,227	S	s
112,617	1	1	83,258	Č	č
112,416	stát	štát	76,867	pro	pre
109,491	společenství	spoločenstvo	70,417	-	-

Table 2: *Treq* source data (*Acquis*).

⁹ We are grateful to Martin Vavřín for kindly providing us with this data.

The source of the data is easily recognizable by the nouns present in the list: *článek/článok* ‘article’, *nařízení/nariadenie* ‘regulation’ or *komise/komisia* ‘commission’, clearly indicating the EU legal discourse.

We decided to use the lemma files for further processing only. The data in the following text are based on the *Acquis* file.

3. Preprocessing

Czech and Slovak are languages belonging to the West Slavic group that are very close and to a large extent mutually intelligible. There exist, nonetheless, some differences at the phonetic, orthographic and lexical levels¹⁰ that are targeted by our *WotD* project.

It is obvious that the list of candidate entries should not contain only identical or “similar” lexical items. They should predominantly consist of equivalents that are “sufficiently different”. As the resulting list will have to be eventually validated by a linguist, the preprocessing should aim to eliminate as many “similar” words as possible, so that the list to be processed manually is not too long. The frequency information is naturally another indication to take into account.

3.1 The pipeline

The preprocessing was performed by means of simple Linux tools: *egrep* utility for regex-based filtrations, *sed* batch editor for character substitutions, and *cut* and *paste* utilities for column manipulations.

The processing pipeline consisted of the following steps:

- adding rank numbers to lemmas;
- removing items containing non-alphabetical characters (66,568 lines removed);
- removing items containing uppercase letters (mostly proper names and abbreviations; 35,736 lines removed);
- removing single-letter items;
- removing items with identical source and translation (42,660 lines removed) – here is the respective regex trick:

```
egrep -v "[[:space:]]([[:alpha:]]+)[[:space:]]\1$" input >output
```

- deleting diacritics that denote the lengths of vowels (*á* > *a*, *é* > *e*, etc.), as well

¹⁰ See e.g. Sokolová, Musilová & Slančová (2005: 5), who refer to F. Uher’s and M. Sokolová’s older research from the 1980’s. According to them, there is a formal and semantic agreement between the Czech and Slovak texts in 38% of lexemes, a partial agreement even in 46%, while 16% are problematic in terms of communication. Out of the 500 most frequent lexemes in Czech and Slovak, 230 (46%) were completely identical, 154 (31%) were partially identical and 116 (23%) were completely different.

as the palatalization of consonants (e.g., $d' > d$, $l' > l$, etc.); removing identical items after this filtration using the same regex trick (4,899 lines removed);

- deleting all vowels; removing identical items after this filtration (19,397 lines removed).

At this point, we still had 95,572 candidate translations that could finally be reduced by applying a frequency threshold. After some experimentation, we decided to set it to 100. The sizes of all four resulting lists are shown in Table 3.

Treq component	Lemmas (original list)	Lemmas (filtered list)	%
Core	433,962	1,867	0.43
Acquis	808,812	5,007	0.62
Europarl	716,703	3,517	0.49
Subtitles	489,324	910	0.19

Table 3: Preprocessed data

The next table shows the first 20 (out of more than five thousand) *WotD* candidates filtered from the *Acquis* list.

Rank	Freq	cs	sk	Rank	Freq	cs	sk
12	132084	nařízení	nariadenie	45	44348	moci	môť
16	113391	nebo	alebo	49	37488	vzhledem	keďže
19	112416	stát	štát	51	34376	ohled	zreteľ
27	84698	evropský	európsky	59	31823	smlouva	zmluva
29	75711	tento	toto	62	29039	země	krajina
30	72069	český	č	65	27721	jenž	ktorý
32	62315	tento	táto	66	27326	odstavec	odsek
38	53562	pro	na	70	25467	všechn	všetok
39	52703	být	sa	71	24945	zejména	najmä
43	44726	být	by	73	23678	jiný	iný

Table 4: *WotD* candidates based on the *Acquis* list.

The Rank column contains rank values from the original list, which makes it apparent how many words have been deleted during the step-by-step filtration (i.e., lemmas with rank 1-11, 13-15, 17-18, 20-26, etc., were deleted). The resulting list still contains a certain amount of noise (e.g., the item ranked 30 is most likely a result of different lemmatization policies for abbreviations being used by the various taggers), yet even among the first 20 items, we can find very good *WotD* candidates. In general, lists preprocessed in the described way not only can save a lot of time for linguists, but can

put the whole enterprise into the “doable” category.

3.2 The data filtered out

As an interesting by-product of the above procedure, we also got three lists of translation equivalents that are equal or “reasonably similar”. These data can be of some interest not only to linguists in the areas of contrastive studies, language typology, phonology, etc., but also to translators – it is a known fact that translation between close languages is straightforward only in a deceptive sense. Some examples are given in Tables 5 and 6; however, this is beyond the purview of our current paper.

Rank	Freq	cs	sk	Rank	Freq	cs	sk
8	190750	být	byť	151	12697	činnost	činnosť
42	46147	příloha	príloha	166	11496	předpis	predpis
69	25842	případ	prípád	180	10390	změna	zmena
75	23441	příslušný	príslušný	189	10040	před	pred
82	22021	hospodářský	hospodársky	196	9767	část	časť
89	19945	den	deň	200	9308	agentura	agentúra
100	18508	oblast	oblasť	207	9042	veřejný	verejný
102	18290	třetí	tretí	220	8607	stanovit	stanoviť
115	16531	při	pri	247	7757	další	ďalší
147	13068	měnit	meniť	260	7411	přístup	prístup

Table 5: Most frequent translation equivalents differing in quantity of vowels and soft consonants only

Rank	Freq	cs	sk	Rank	Freq	cs	sk
9	190027	článek	článok	55	33263	soulad	súlád
10	144811	se	sa	68	25978	podle	podľa
13	130075	který	ktorý	72	23803	informace	informácia
17	113238	komise	komisia	76	23010	podmínka	podmienka
20	109491	společenství	spoločenstvo	80	22439	muset	musieť
28	79802	pro	pre	86	20407	výrobek	výrobok
33	59929	směrnice	smernica	97	18814	svůj	svoj
40	52366	opatření	opatrenie	99	18720	žádost	žiadost
41	49881	rozhodnutí	rozhodnutie	103	18080	společnost	spoločnosť
50	37213	mít	mať	104	18001	společný	spoločný

Table 6: Most frequent translation equivalents differing in combination of vowels and soft consonants only

4. Lexicographic application

The *WotD* application is meant to be the first step in a broader *WotD* project, ideally one involving both Czech and Slovak lexicographers – as both Czechs and Slovaks form the target group of users. Confronting the dual view of the same topic would certainly be beneficial to both nations, which once lived together within one country. The Czech and Slovak languages would once again stand side by side, as they did before. While they are mutually intelligible to the older generation who remembers the Czechoslovak federation, for the youngest generation this is far from being the case – quite often using English as a mediating language.

4.1 List of headwords

The question of choice of words for the *WotD* project is crucial and deserves an elaborated conception. Nonetheless, whatever criteria are chosen, the point is that preselection of the candidates is taken care of by a computer, and a lexicographer only revises automatically generated drafts of entries. Our application generates a further editable draft version of the given entry, relying primarily on corpus data (frequency, most common collocations, exemplification using the *GDEX* tool (Kilgarriff et al. 2008), etc.), complemented by a lexicographic description taken from existing dictionaries and by other features. Such a draft would be subsequently edited by a lexicographer who would also write a brief commentary – a usage note or even an essay (the Czech lexicographer would comment on a Czech word whereas the Slovak lexicographer would comment on a Slovak word – or, occasionally, even vice versa). As these feuillets on the various linguistic subjects are rather popular in both countries, we believe a broad audience would become interested in the project. After all, the public can be actively involved in it – e.g. by commenting on individual words on the project website, by voting for the most popular word(s) or for words to be processed in the future, or in other ways.

The criteria for selecting individual headwords remain an open question at the moment. The pipeline described above eliminated formally similar words from the candidate list. However, even these may appear in the final inventory – although being words common to both languages, they are still potentially different in their use (including cases of false friends), frequency, etc. However, the largest group of words will naturally be those specific for one of the languages – with the most common equivalent(s) in the second language, including pairs that are the source of the linguistic humour¹¹. Personally, we lean towards a combination of different aspects so that the final selection

¹¹ In the Czech environment it has long been believed that Czech *veverka* ‘squirrel’ is called *drevokocúr*, literally ‘tree-tomcat’, in Slovak. Such a word, however, does not exist in Slovak at all, as a formally similar word *veverička* is used. See Nábělková (2008: 219-232) for the description of this inter-language myth in detail.

is as diverse (both semantically and grammatically) and user-attractive as possible. The aim is to educate the audience in an engaging form: we want readers to realise on the one hand the interconnection of these two languages (lexicon inherited from a common Slavic basis, mutual reciprocal loanwords, commonly used internationalisms), on the other hand their diversity, deepening after the break-up of Czechoslovakia in 1993.

4.2 The microstructure of the WotD entry

With regard to the microstructure of individual *WotD* entries, the whole project has at least two specifics. The first one is the dual metalanguage – Czech and Slovak, making it, de facto, two explanatory dictionaries in one rather than a translation dictionary. Mutual equivalents here serve only as a secondary means to emphasise a contrastive nature. A top-down layout with a vertical partition seems to be ideal: the left half will be reserved for the Czech part of the entry, the right half for the Slovak part, while the individual elements of the microstructure will be horizontally aligned side by side.

In addition, it is a born-digital project that would result in an electronic dictionary that can be augmented in the event of public interest by any number of items. We take processing a set of 365 dictionary entries as a suitable beginning, provided that a new entry is published daily for the time span of one year. The inventory would then gradually expand and in the final stage it would cover, albeit in an unbalanced way, the whole alphabet. In fact, there would be two lists of entries – a Czech one and a Slovak one, both easily searchable. A close connection between the dictionary and corpora in the form of numerous links is commonplace.

The content of the entries will also be closely related to the born-digital and corpus-based nature of the dictionary. Some elements presented in traditional explanatory dictionaries would thus be reduced or completely omitted in our *WotD* microstructure – while others would be highlighted. For example, in traditional dictionaries the lemma is most often followed by morphological/grammatical information. In *WotD*, the emphasis would be laid on frequency data and usage specifics (typical genre/text-type, communication situation, etc.). This should be demonstrated by some suitable examples which, in the spirit of the famous Firthian dictum “You shall know the word by the company it keeps” (Firth, 1957: 11), would illustrate the meaning(s) of the word, but also its creative alterations in specific texts (e.g. fiction) or in spoken language. The difference between the spelling and the pronunciation of Czech/Slovak words is not as large as in English, therefore the sound recording of the word could move from the heading of the entry to the exemplification part – and indicate, among other things, different semantics and usage within spoken and written language (which is, at least in Czech, close to diglossia – Bermel, 2014). In addition, the exemplification section should include a direct link to the corpora concerned, providing additional examples to a potentially interested person.

The example part would be followed and supplemented by the lexicographer's commentary, the imaginary central part of the whole entry. It should be written in a popular, entertaining style and should aptly reflect the place of the given word in the lexical system of language, along with the differences from the second language. These may appear on the diachronic level, as a variance in the development of semantics and/or the use of the same word in both languages. Therefore, basic etymological information should be provided as well.

Only at the end of the entry can explanatory definitions from existing Czech and Slovak dictionaries be cited. Although this is the central part of the entry in traditional dictionaries, we perceive them instead as an interesting appendix providing a contrast to the modern, corpus-based approach to lexicography.

The microstructure of the dictionary entry is far from definitive; on the contrary, it is a mere suggestion that should underline the specificity of our project and which will need to be properly tested by compiling several sample entries.

5. Acknowledgements

This work has been, in part, funded by the Slovak VEGA Grant Agency, Project No. 2/0017/17. It was also supported by the European Regional Development Fund-Project "Creativity and Adaptability as Conditions of the Success of Europe in an Interrelated World" (No. CZ.02.1.01/0.0/0.0/16_019/0000734). During its creation we used the tools developed within the Czech National Corpus project (LM2015044) funded by the Ministry of Education, Youth and Sports of the Czech Republic within the framework of Large Research, Development and Innovation Infrastructures.

6. References

- Bermel N. (2014). Czech Diglossia: Dismantling or Dissolution?. In J. Árokay, J. Gvozdanović & D. Miyajima (eds.) *Divided Languages? Diglossia, Translation and the Rise of Modernity in Japan, China, and the Slavic World*. Cham: Springer.
- Čermák, F. & Rosen, A. (2012). The case of InterCorp, a multilingual parallel corpus, *International Journal of Corpus Linguistics* 13 (3), pp. 411-427.
- Firth, J. R. (1957). *A synopsis of linguistic theory 1930-1955*. In J. R. Firth (ed.) *Studies in Linguistic Analysis*, Special volume, Philological Society. Oxford: Blackwell, pp. 1-32.
- Gašparíková, Ž. & Kamiš, A. (1967). *Slovensko-český slovník*. Praha: Státní pedagogické nakladatelství.
- Horák, G. et al. (1979). *Česko-slovenský slovník*. Bratislava.
- Kilgariff, A.; Husák, M.; McAdam, K.; Rundell, M. & Rychlý, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus. In E. Bernal & J. DeCesaris (eds.) *Proceedings of the 13th EURALEX International Congress*. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu

- Fabra, pp. 425-432.
- Nábělková, M. (2008). *Slovenčina a čeština v kontakte – Pokračovanie príbehu*. Bratislava: Veda.
- Nábělková, M. & Vavřín, M. (2018). *Korpus InterCorp – slovenština, verze 11 z 19. 10. 2018*. Ústav Českého národního korpusu FF UK, Praha 2018. Accessed at: <http://www.korpus.cz> [17 July 2019].
- Nečas, J. & Kopecký, M. (1964). *Slovensko-český, česko-slovenský slovník rozdílných výrazů*. Praha: Státní pedagogické nakladatelství.
- Ripka, I. & Skladaná, J. (1980). Česko-slovenský slovník. *Slovenská reč* 45(6), pp. 364-372.
- Rosen, A. (2016). InterCorp – a look behind the façade of a parallel corpus. In E. Gruszczyńska & A. Leńko-Szymańska (eds.) *Polskojęzyczne korpusy równoległe. Polish-language Parallel Corpora*. Warszawa: Instytut Lingwistyki Stosowanej, pp. 21-40.
- Sokolová, M., Musilová, K. & Slančová, D. (2005). *Slovenčina a čeština (Synchrónne porovnanie s cvičeniami)*. Bratislava: Filozofická fakulta Univerzity Komenského v Bratislave.
- Škrabal, M. & Vavřín, M. (2017). The Translation Equivalents Database (Treq) as a Lexicographer's Aid. In I. Kosem et al. (eds.) *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference*. Leiden: Lexical Computing.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



Collecting Collocations for the Albanian Language

Besim Kabashi

Corpus und Computational Linguistics,
Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
Albanology, Ludwig-Maximilians-Universität München, Germany
E-mail: [besim.kabashi@{fau,lmu}.de](mailto:besim.kabashi@fau.lmu.de)

Abstract

The presented paper describes the collecting of data from different sources to build a collocation data set with the aim of compiling the first contemporary collocation dictionary for the Albanian language. The work is based (1) on the analysis of empirical data, i. e. linguistic corpora, using the computational methods and tools, as well as (2) on traditional dictionaries. As empirical data we use the AICo (Albanian Text Corpus), the AICoPress 2017-2019, N-Grams extracted from both, methods like Log-likelihood and Dice coefficient using the IMS Open Corpus Workbench (CWB) and the Corpus Query Processor, Web version (CQPweb). Despite the enormous support, an unsupervised automated compilation of a collocation dictionary of high quality, like those created by lexicographers, seems to be impossible without intervention. In order to complete the collection of the data we additionally use lexical information extracted from traditional dictionaries. The primary goal is to create a language resource that can be used among others also for Natural Language Processing purposes. The presented work is still in progress and, of course, will change until its final version.

Keywords: Albanian, collocations; NLP lexicography; corpus linguistics; language resources

1. Collocations as lexical data

Linguistic data are important or even necessary for numerous applications to make the communication easier, or at least as support data for building further linguistic datasets as resources for natural language processing.

Collocations may serve one of two purposes in dictionaries. On the one hand, they are used as standalone data in collocation dictionaries. On the other hand, they serve as “additional” information in other types of dictionaries, e.g. definition dictionaries. They are not only important for non-native language learners, but also for native speakers, who sometimes need to find established ways of combining lexical items. Typical examples of collocations that can cause problems for foreign language learners are combinations such as *strong tea* (e.g. instead of *powerful tea*) in English. Many researchers distinguish between various related concepts of collocation, sometimes labelled “significance-oriented” and “statistically-oriented”, e.g. Herbst (1996). The former are often semantically restricted and are thus particularly difficult to learn. But even items that just frequently co-occur without being conventionalized may be relevant for dictionary users, since such combinations are often differentiated in usage style or are only common in specific domains. Those data can be used not only while translating from one language into another, but also for writing or speaking in a specific field, working on a desktop computer, or simply searching on a smartphone for a specific word usage.

2. Collecting collocations for a dictionary of Albanian

Currently no collocation dictionary exists for Albanian. The aim of this project is to fill this gap in Albanian lexicography by collecting collocation data for such a dictionary. For the speakers of a language, a collocation dictionary, e.g. Benson et al. (2010), Quasthoff (2010), or Häcki Buhofer et al. (2014), offers the possibility to select fine-grained collocations, to express oneself idiomatically in a conversation or text.

As this work is very data intensive, and a basic data set, e.g. for extending a given resource using NLP tools, is not available, there is a need of elementary work. For this reason we decided to take an approach consisting of three steps.

In order to collect the lexical data, we use a lexicon for NLP, presented in Kabashi (2019), and an automatic morphology for the Albanian language, presented in Kabashi (2015), to lemmatize the word forms, because Albanian is an inflected language and has a rich morphology. We also use tagged texts and the *AlCo* presented in Kabashi (2017), using the tagset presented in Kabashi & Proisl (2018). Since there is no syntactic parser available for Albanian, the extraction process is based on surface-oriented methods, e.g. n-grams and distance-based cooccurrences, see for example Evert (2013) and Proisl (2019).

3. Selecting the data sources

For an empirical data driven approach, selecting data sources also determines the quality of the data. One of the data sources is the *AlCo* (An Albanian Text Corpus), cf. Kabashi (2017). The corpus contains 100 million words and covers different domains of language and contains different text types. Additionally, another recently compiled corpus of press texts serves as a further data source – it is a reference corpus named *AlCoPress* (2017–2019) that contains approximately 32 million text words, taken from seven newspapers and a news agency. Around 70 million words are currently raw data. All in all, the data sources are around 200 million words. The amount of data, from an empirical point of view, compared to similar corpora of other languages, is still too small, but it allows extracting valuable information in most search cases and also profiling the knowledge derived from the data.

4. Methods and tools for exploring the data sources

To explore the linguistic data we use n-grams (2- to 10-), IMS Open Corpus Workbench (CWB)¹, and the Corpus Query Processor, Web version (CQPweb). This information is then complemented by the traditional selection of lexical entries from different dictionaries and lexicons, e.g. Kostallari et al. (1980), Samara (1998), Thomai et al. (2004), Dhrimo et al. (2007), and Thomai et al. (2006).

¹ Cf. <http://cwb.sourceforge.net/>.

4.1. N-Grams

With the n-grams technique it is possible to extract the data from raw text without² any preparation, e.g. tagging or formatting. Frequency lists of n-grams allow the researcher to find words that often occur together by aggregating common combinations, cf. the 4-grams listed below. Example *10044 për herë të parë*, eng. *for the first time*, shows this effect – the accumulation of frequent word sequences. This very simple method is very useful, but a lot of cases remain, i.e. entries with low frequency, which may still have valuable collocation information, and are not listed on the top of the frequency list, but towards the end. See for example the second part of the list, after the frequency 62, where the frequencies of the word *çaj*, eng. *tea*, are listed. In this case the word-forms (of *çaj*) are not lemmatized, so the word-forms *çaj*, *çaji*, *çajin*, *çajit*, ... (with the properties case, number, gender, and definiteness for nouns), are listed separately as they originally occur in texts.

10044	për herë të parë	7	një <u>çaj bimor</u>
6599	. Nga ana tjetër	7	. Çaji i gjetheve
2999	do të thotë që	6	rastet <u>çaji zihet</u> derisa
2659	një kohë të gjatë	6	ta përzieni me çajin
2598	. Për këtë arsye	6	përbërjen e çajrave .
2304	<i>pjesën më të madhe</i>	6	me <u>çajin nga kamomili</u>
2241	gjithnjë e më shumë	6	lugë <u>çaji të kanellës</u>
2083	<i>pjesa më e madhe</i>	6	lugë <u>çaji me piper</u>
437	një kohë të shkurtër	6	Ky çaj përdoret për
433	të nivelit të lartë	5	shumë <u>çaj të ftohtë</u>
<hr/>		5	i <u>pemës së çajit</u>
		5	e <u>çajit të koprës</u>
		5	1 lugë <u>çaji pluhur</u>
62	Si përgatitet çaji:	4	vinte <u>çaji i darkës</u>
23	të <u>çajit të gjelbër</u>	4	<u>tufë çaji</u> në tregun
22	e <u>çajit të gjelbër</u>	4	shumë <u>çaje qetësuese</u> ,
15	i <u>çajit të gjelbër</u>	4	të prodhimit të çajit
16	një lugë çaji me	4	rigoni e çaji i
11	Nga <u>çaji i përgatitur</u>	4	<u>qese të çajit të</u>
10	që nuk <u>pinë çaj</u>	4	që <u>çaji i kajsisë</u>
10	për të <u>bërë çajra</u>	4	përgatisni <u>çajrat me fruta</u>
10	një <u>filxhan me çaj</u>	4	përdorni <u>çaj të tharë</u>
9	<u>filxhan çaj jeshil</u> .	4	ose <u>çaj me sheqer</u>
9	<u>bërë çajra kundër sëmundjeve</u>	4	monopolin e çajit kinez
9	<u>bërë çajra kundër sëmundjeve</u>	4	pini <u>çaj pa sheqer</u>
8	se <u>çaji i zi</u>	[...]	
8	një <u>çaj të ngrohtë</u>	1	përgatisni një <u>çaj frutash</u>
8	nga një <u>gotë çaji</u>	1	një <u>çaj para gjumit</u>
8	<u>çaji i malit dhe</u>	1	Një çaj pa avull
7	të <u>çajit të zi</u>	1	me çaj para buke

List 1: The list of some of the most frequent 4-grams and the 4-grams of çaj.

Not all occurrences of the words are collocations of the word *çaj*, e.g. *një lugë çaji me*, eng. *teaspoon*, where the word *tea* is a collocation of the word *spoon*. At the same time,

² In this case, driven by a script running on a Linux operating system.

not all collocations of the word *čaj* can be found in the n-gram lists. In this case, the types and/or the amount of text do not cover all collocations of the word. Additional texts from certain domains and increasing the overall amount of text would increase the probability of covering them.

4.2. CWB & CQPweb

The IMS Open Corpus Workbench (CWB) is “a collection of open-source tools for managing and querying large text corpora [...] with linguistic annotations”.³ CQPweb (Corpus Query Processor) is a software package, a web-based corpus analysis system, to explore corpus data, cf. Hardie (2012).

In contrast to the n-gram method, CWB and CQPweb, and particularly CQPweb, offer a lot of functions for calculating collocations. As the tools support the processing of linguistic data, also based on linguistic annotations, the data can be explored on more dimensions, e.g. by searching based on POS-tags or within certain domains.

To find the collaboration candidates, CQPweb can use the *Conservative LR*, *Dice coefficient*, *Log-likelihood*, *Log Ratio* (filtered), *MI2*, *Mutual Information*, *T-score*, *Z-score*, and as well as the simple *rank by frequency*. For each lexical entry the different measurement results help to find words which can be added to the respective lexical entry.

In the example above (cf. Figure 1) a list of collocations of the word *puně*, eng. *work*, is shown, calculated based on Log-likelihood by CQPweb. Below in the section *Example Entries* we present a detailed entry (as a working version) for this word. Each method, depending on different criteria, can offer different collocation candidates, e.g. using the *Log-likelihood* results is different than the *Dice coefficient* results. The statistics offered by CQPweb e.g. observed collocate frequency, the number of texts, and Log-likelihood (in figure 1) make selecting collocate candidates easier. Through using all of them, it is possible to gather more collocation candidates. Some of those candidates cannot be used, e.g. because they are function words in a sentence and not, for example, prepositions that are associated with the collocation candidate. An example is the word *ně*, eng. *in*, listed in Figure 1, which depending on the concrete context can be a collocation, or not. In positions 7 and 8 (in Figure 1), the words *njě*, eng. *one*, and *kjo*, eng. *this*, are listed very high, but in the sense of collocation they are both irrelevant. The list in this abbreviated version does not show the “typical” collocations, as may be expected from a native speaker. The fact that the data sources are only written texts might explain this.

Due to those problems we found it necessary to complement the results from CQPweb with the information contained in traditional dictionaries.

³ Cf. <http://cwb.sourceforge.net/> .

4.3. Dictionaries

Traditional dictionaries like definition dictionaries, i.e. Kostallari (1980) and its newer editions, collect information based on long-term observation of language. They do not offer intentionally typical examples of collocations within their usage examples, but as the goal is different to the collocation dictionaries, they are not listed separately either. As a result, only a small number of collocations can be found within the entries, mostly within the examples. The number of collocations in those cases is higher if the lexical entry has more lexical meanings.

Collocation controls						
Collocation based on:	Word form	Statistic:	Log-likelihood			
Collocation window from:	3 to the Left	Collocation window to:	3 to the Right			
Freq(node, collocate) at least:	5	Freq(collocate) at least:	5			
Filter results by:	specific collocate:	and/or tag:	(none)	Submit changed parameters	Go!	
Extra information: Log-likelihood (LL) scores collocations by significance: the higher the score, the more evidence you have that the association is not due to chance. More frequent words tend to get higher log-likelihood scores, because there is more evidence for such words.						
There are 6,728 different words in the collocation database for this query (Query "(?longest) punë" returned 8,304 matches in 5,062 different texts)						
[0.384 seconds - retrieved from cache]						
No.	Word	Total no. in whole corpus	Expected collocate frequency	Observed collocate frequency	In no. of texts	Log-likelihood
1	bërë	38,675	60.775	479	434	1149.436
2	në	1,026,294	1612.754	3,094	2002	1117.321
3	shumë	105,378	165.595	626	543	750.415
4	këtë	100,779	158.368	580	513	667.906
5	vite	12,753	20.041	216	109	638.976
6	për	531,048	834.507	1,627	1233	601.649
7	një	450,538	707.991	1,397	1161	531.709
8	Kjo	60,674	95.345	323	300	334.808
9	pa	39,544	62.141	250	189	321.917
10	siguruar	2,631	4.134	74	24	289.181
11	aftësisë	254	0.399	38	20	276.965
12	vështirë	6,929	10.889	102	88	275.551
13	gjejnë	1,510	2.373	57	55	255.208
14	bëjnë	11,549	18.149	121	108	254.553
15	mbarë	2,310	3.63	64	57	248.243
16	ende	12,115	19.038	119	107	237.286
17	madhe	18,353	28.841	143	136	230.56
18	hyn	972	1.527	45	33	219.549
19	mirë	35,207	55.326	196	180	215.447
20	ngarkuara	139	0.218	27	19	212.12
21	prish	383	0.602	34	28	210.546
22	kolosale	121	0.19	26	26	210.1
23	fillova	279	0.438	30	29	197.698
24	vjetërsisë	48	0.075	20	8	193.139

Figure 1: The collocation results for the word *punë*, eng. *work*, calculated based on Log-likelihood, by the CQPweb.

Another dictionary type that offers collocation candidates is a dictionary of synonyms, e.g. Thomai et al. (2004) for Albanian. Below – as an example, in Figure 3, marked with the sign ♦ and the name of the dictionary – are listed collocations which are taken

from the mentioned dictionary. If the data are available in electronic form, then a fast search and extraction of any lexical information is possible, otherwise the only remaining option is to gather it manually.

Combining all three methods, i.e. n-grams, CQPweb and extracting information from traditional dictionaries, makes it possible to collect a lot of lexical information for certain lexical entries which can be used for compiling a lexicon of collocations

5. Selecting the lexicon entries and their types

The collected lexical information needs to be organized in lexical entries. The selection of lexical entries is in some cases very difficult, especially the decision of whether to select an entry at all. Furthermore, the selection of collocation candidates (for the explication part of the entry) is not easy and depends on many criteria. We start with the entries that are semantically related and continue with those that are very frequent. But not every frequent cooccurrence is chosen for a lexical entry, as explained in the above case of *një* and *kjo*, in the section CWB & CQPweb.

Based on other collocation dictionaries, the lexical entries are differentiated according to their part-of-speech. Also, the grammatical relations between collocated words are important to organize the mezzo- and micro-structure of an entry, e.g. Noun–Adjective or Verb–Noun. This is described in the followed sections based on examples.

6. Example Entries

6.1. Nouns

Noun entries are organized as in Figure 2. The entry begins with its head, followed by its part-of-speech. The collocations, without an example of where they occur in texts, are listed as a sequence, separated by commas. This sequence can be separated by a pipe (|) and *<jo, pa, i, ... [i.e. negations]>* for words of opposite meaning and bullets (• i.e. very strong, · i.e. less strong), if the collocations can be grouped/assorted/-separated in sense of meaning. The letter *i*, e.g. in *i shkurtër* (eng. *short*) is the article (determiner) of the masculine gender of the adjective *shkurtër*. The feminine gender is *e*, which is not written. The explication part contains also the collocations with its prepositions, which are listed after the mark \blacksquare , e.g. *para ~i* (i.e. *para afati*), eng. *before the deadline*. The next sign ♦♦ marks the word compounds, e.g. *afatgjatë* (from *afat* + *~gjatë*), eng. *long term*.

Some dictionaries, e.g. Häcki Buhofer et al (2014), also list examples of authentic use for each word. For the work presented here, currently no examples are taken into account, i.e. no such examples are contained in the draft version of the dictionary. Example sentences may still be included in a future on-line version of the dictionary.

afat E (eng. *deadline*, Noun)

+MBE (eng. *Adjective*)

~ *i shkurtër* | ~ *i gjatë* · ~ *mesatar* · ~ *i kaluar* · ~ *i tejkaluar*, ~ *i skaduar* · ~ *i <pa>mbaruar* · ~ *i shtyrë* · ~ *i vazhduar* · ~ *i zgjatur* | ~ *i shkurtuar* · ~ *i <pa>përshtatshëm* · ~ *i <pa>caktuar* · ~ *i <pa>detyrueshëm* · ~ *i shlyer* [...]

+F (eng. *Verb*)

shtyej ~ · *respektoj* ~ · (*tej*)*kaloj* ~ · *mbaron* ~ · *vjen* ~ · *afrohet* ~ · *caktoj* ~ · *vazhdoj* ~ [...]

▣ (i. e. used with prepositions, e.g. *me*, *pa*, ...)

me ~, *pa* ~, *para* ~*it*/~*ës*, *pas* ~*it*/~*ës*, *përtej* ~*it*/~*ës* [...]

◆ (i. e. word-formation)

~*caktim*, ~*vënie* · ~*shtyrje* · ~*kalim* [...]

>MBE

~*shkurtër*, ~*mesëm*, ~*gjatë*, ~*caktuar*, ~*shtyrë* [...]

Figure 2: The working version of the lexical entry for the collocation *afat*, eng. *deadline*, as a result of evaluating the n-grams, using the CQPweb, calculated based on Log-likelihood, and informed by the traditional dictionaries.

Polysemic collocations, those with different senses, are listed separately, as shown in the following entries:

bar, ~i 1 E (eng. *grass*)

+MBE

i njomë | *i tharë* · *i thatë* · *i rritur* · *i gjelbërt* · *i mbirë* · *i mbjellur* · *i prerë*, *i kositur* · *i mbledhur* · *i rrëzuar* · *i mirë* · *i keq* [...]

bar, ~i 2 E (eng. *medicament*)

+MBE

shërues · *qetësues* · *i mire* [...]

▣

kundër dhimbjes · *kundër kollës* · *kundër ftohjes* [...]

bar, ~i 3 E (eng. *bar*)

+MBE

i <pa>njohur · *i frekuentuar* (*shumë* | *pak*) [...]

▣

~*i më i afërt* · *i të rinjve* [...]

Figure 3: The working version of the lexical entry for the collocation *bar*, eng. *grass*, *medicament*, *bar*, as a result of evaluating the n-grams, using the CQPweb statistics, and informed by the traditional dictionaries.

Putting together all lexical data gathered from n-grams, cf. figure 1, through CQPweb and extracted from traditional dictionaries, the following entry can be created:

çaj, ~i **E** (eng. *tea*)

+MBE
i nxehtë | i ftohtë • i ëmbël | i hidhur • i fortë, i rëndë | i lehtë, i lig • i zi • i gjelbër • mjekësor [...]

+ E_{ABL}
mali • bjeshke • frutash • trëndafili • kaçeje • bliri • dafine • murrizi • rozmarine • borzilloku • eukalpti • kanelle • alku [...]

■
me sheqer • me mjaltë • me limon [...]

Figure 4: The working version of the lexical entry for the collocation *çaj*, eng. *tea*, as a result of evaluating the n-grams, using the CQPweb statistics, and informed by the traditional dictionaries.

6.2. Verbs

Verb entries have the same structure as the noun entries. The head of the entry has – like most Albanian dictionaries – the grammatical information on aorist and participle, in addition to the part-of-speech information. The following example shows an entry of a verb.

kry|ej ~eva, ~yer **F** (eng. *end, complete, finish*)

+E
një punë • një punim • një detyrë • një detyrim • një porosi • një vepër • një veprim • një aksion • një shkollë • një studim • një udhëtim • pushimin_{DET} [...]

+NDF
mirë | keq • shpejt | ngadalë [...]

Figure 5: The working version of the lexical entry for the collocation *kryej*, eng. *complete successfully*, as a result of evaluating the n-grams, using the CQPweb statistics, and informed by the traditional dictionaries.

6.3. Adjectives

The number of these entries is smaller than the numbers for nouns and verbs. Adjectives are more often listed as collocates of the nouns. Lexical entries of adjectives can be created from a reverse index of them. A lexical entry of an adjective looks as follows:

privat ~e MbE (eng. *private*)

+E

punë • *çështje* • *interes* • *lidhje* • *jetë* • *shtëpi* • *banesë* • *makinë* • *pajisje* • *udhëtim* [...]

Figure 6: The working version of the lexical entry for the collocation *privat*, eng. *private*, as a result of evaluating the n-grams, using the CQPweb statistics, and informed by the traditional dictionaries.

6.4. Adverbs

The number of adverb entries is currently very small. Similar to the adjective entries, the adverb entries can be gathered from the verb entries, in addition to the mentioned methods used for the extraction of information to create the noun and verb entries. Most adverbs, such as *good*, can occur with a large number of verbs. A lexical entry for an adverb looks as follows:

mirë NdF (eng. *good*)

+ F

jam • *jetoj* • *kaloj* • *ndihem* • *bëj* • *di* • *kuptoj* • *kujtoj* • *shikoj* • *rri* • *pushoj* • *ha* • *flas* • *vishem* • *dukem* • *njoh* • *shkoj* | *vij* • *mendoj* • *luaj* • *mësoj* • *veproj* • *informoj* • *përshtat* • *zgjohem* • *gdhihem* • *dalloj* • *mbroj* • *ruaj* • *kujdesem* • *pagua* • *pres* • *shfrytëzoj* • *funksionon* • *arsyetoj* • *përmbledh* • *laj* • *pastroj* • *pjek* • *gatuaj* • *siguroj* • *këqyr* • *mbikëqyr* • *sillem* • *eci* • *udhëzoj* • *shkruej* • *them* • *filloj*, *nis* | *mbaroj*, *përfundoj* • *punoj* • *hap* | *mbyll* • *shpjegoj* • *këshilloj* • *ndaj* • *bashkoj* | *largoj* • *jap* • *përgatit* • *drejtoj* • *realizoj* [...]

Figure 7: The working version of the lexical entry for the collocation *mirë*, eng. *good*, as a result of evaluating the n-grams, using the CQPweb statistics, and informed by the traditional dictionaries.

6.5. The detailed form of an entry

The following entry shows in detailed form the entry of the noun *punë*, eng. *work* and the verb *punoj*, eng. *to work*. The possible combinations are listed: N-V, e.g. *nis punën*, eng. *to begin with the work* and N-ADJ/ADV, e.g. *punë e mirë*, eng. *good work*; One important extension would be to add the prepositions and the case information as given with *punë* PREP+DAT *sipas ligjit*, or *filloj/nis* ACC (=punën) | *një* NOM (=punë).

pun|ë ~a ~ë ~ët E

+ F

(e) *nis* ~ <+ACC>, *filloj* • *bëj* • *ndërpres* • *vazhdoj*, *rifilloj* • *kryej*, *mbaroj* • *harroj* • *kërkoj* • *siguroj* • *pëlqej*, (ia) *pëlqej* ~ <+OBJ+DAT> • *dua* | *urrej* • *pengoj* | *nxit* • *lavdëroj* | *përqesh* • *kujtoj* • *ngadalësoj*, *zhagit*, *prolongoj* • *pezulloj* • *udhëheq*, *drejtoj* • (ia) *mbështet* ~ <e dikujt> • (ia) *këshilloj* ~ <dikujt> • (ia) *mohoj* ~ <dikujt> • (ia) *ndaloj* ~ <dikujt> [...] *shtyj* ~ *përpara*, • (i) *fle* ~ <dikujt> [...]

+ MbE

e mirë | *e keqe* • *e vështirë* • *e rëndë* | *e lehtë* • *e shpejtë* | *e ngadaltë* • *e shumë* | *e paktë* • *e ndryshme* | *e njëjtë* • *e gjatë* • *kujdestare* • *kujdestarie* • *nate* | *dite* • *mbrëmjeje* | *mëngjesi* • *legale* | *ilegale* • <jo> *serioze* • *tinëzare* • *publike* | *private* • *e madhe* | *e vogël* • *e lirë* | *e shtrenjtë* • *e* <pa> *ndershme* • *e* <pa> *ngutshme*, *e*

<pa>nxitueshme, e <pa>nxituar • vullnetare • e paparë | e mirënjohur • <jo>profesionale • amatore • kuptimplote • e <pa>këndshme • e pistë • e ndyrë | e pastër • e <pa>ditur • e <pa>njohur • e <pa>vlefshme • e mirëfilltë, e kënaqshme • e <pa>vëmendshme • e <pa>kuptueshme • <jo>detyruese • e <pa>përshtashme • <e parë, dytë, e tretë, ...> • e qetë • e <pa>ndërprerë • e <pa>kryer • e <pa>rregullt • e ligë • e <pa>shëndetshme • e mundimshme • e <pa>rrezikshme • e frikshme • e <pa>parëndësishme • e <pa>drejtë • e gatshme • e dështuar • e <pa>zakonshme | e jashtëzakonshme • <jo>precize • <jo>sistimore • e kotë • <in>formale • <jo>normale • e <pa>përfunduar, e posapërfunduar • e përkryer • <jo>humane • inxhinierike • edukuese • sezonale • e nisur, e posanisur • e lënë përgjysmë • marramendëse • e marrëzishme • e lodhshme, lodhëse • bujqësore • blegtorale • ndihmëse • plotësuese, mbështetëse • kryesore, kyçe • dytësore • banale • fisnike • tregtie • zeytare • aktive | pasive • madhështore, e mrekullueshme • poshtëruese, e poshtër, • patriotike, atdhetare, atdhedashëse, frikësuese • rraskapitëse • eksploatuese • e zezë • e mbarë • e prapë • diletante • minimale | maksimale • e dënueshme • përfituese • përgatitore • përmbyllëse • intensive, e sistemuar • artistike • sociale • mendore • motivuese [...]

+ PREP

+NOM

<me | pa> ligj • <me | pa> normë • <me | pa> rregull • <me | pa> vullnet • <me | pa> hamendje • <me | pa> dyshim • <me | pa> marrëveshje • me orë të shumta • pa u lodhur • <me | pa> përtësë • <me | pa> shije • <me | pa> dinjitet • <me | pa> kuptim • në të zezë • <me | pa> letra • <me | pa> dokumente • <me | pa> detyrim • <me | pa> leverdi • <me | pa> plan • pa fund • <me | pa> fat • <me | pa> rëndësi • <me | pa> pagesë • <me | pa> para • <me | pa> kujdes, (GgK. pa lidhje) • pa ide • <me | pa> nxitim • <me | pa> ngut • <me | pa> nder [...]

+DAT

sipas ligjit • sipas rregullave • sipas normës • sipas planit [...]

+E

fillestari/eje • amatori/eje • profesionisti/eje • diletanti/eje • dreqi • gomari • fëmijësh • të rinjsh • djemsh | vajzash • burrash | grash • dimri • pranvere • vere • vjeshte • fshati • ndërtimtarie • hajduti, hajni • rrugaçi • dembeli, përtaci • pasioni • qejfi • trimash, trimërie [...]

+ F

filloj, nis +ACC+DET, një +NOM+INDET • mbaroj +ACC+DET • humb ACC+DET • ndërpres ACC+DET • kujdesem për NOM+INDET • kërkoj (një) NOM+INDET • dua (një) NOM+INDET, nuk dua NOM+INDET • mendoj për (një) NOM+INDET [...]

-dhënës/-je • -marrës/-je • -kërkues/-im • -prishës/-je • -ndreqës/-je • -gjetës/-je • -kryes/-erje [...]

pun|oj ~ova ~uar F

+ NdF

mirë | keq • shpejt | ngadalë • shumë | pak • ndryshe • kështu • ashtu • lirë | shtrenjtë • gjatë • kot • si i çmendur • vetëm • <i>legalisht • <jo>seriozisht • tinëzisht • privatisht | publikisht • falas • natën | ditën • mëngjesve | mbrëmjeve • të dielave • së mbari, së prapthi <jo>sistematikisht • intensivisht • rëndë • pastër • qetësisht • i <pa>stresuar • i <sh>qetësuar • i <pa>pakoncentruar • i vetmuar • fizikisht • vullnetarisht [...]

+ PREP

+NOM

deri vonë • gjatë festave • <me | pa> kujdes, (GgK. pa lidhje) • pa ide • <me | pa> nxitim • <me | pa> ngut • <me | pa> ligj • <me | pa> normë • me plan • <me | pa> rregull • <me | pa> para, <me | pa> pagesë • <me | pa> vullnet • <me | pa> hamendje • <me | pa> dyshim • me orë • me ditë • <me | pa>

marrëveshje • *me orë të shumta* • *<me | pa> përtesë* • *<me | pa> vëmendje* •
nën tarifë • *me qetësi, në qetësi* • *<me | pa> kokë/krye* • *<me | pa> mend* •
<nën | pa> presion • *tërë ditën | tërë natën* • *<me | pa> dëshirë* • *në <ndërtimtari*
...> • *për <dikë+ACC>* • *si <inxhinier, mjek ...>* • *si i pavarur* • *si udhëheqës* •
si ndihmës [...]
+DAT
sipas ligjit • *sipas rregullave* • *sipas normës* • *sipas planit* [...]
+ FInf
<pa / duke> u ngutur • *<pa / duke> u nxituar* • *pa u lodhur* [...]
IDM:
si gomar • *si kalë* • *sa (për) <dy, ...> vetë* [...]
FRZ:
ia _{DAT+ACC} *punoj (keq/\$mirë) <dikujt* _{DAT} *>* [...]

Figure 8: The working version of the lexical entry for the collocation *punë*, eng. *work*, in the detailed, more extensive version, as a result of evaluating the n-grams, using the CQPweb statistics, and informed by the traditional dictionaries.

The example above contains all data collected for the entries and the current state of the work on lexical entries. In general, the aim is to keep the entries shorter, but they can “grow” to be very detailed.

7. Conclusions

Currently, the number of lexical entries is around 2000, with 40 to 110 entries for each letter of the Albanian alphabet. A number of entries are not yet complete with all their possible information, i.e. the work on these entries is not finished yet. A few of them will be deleted, while some new entries will presumably be added in the ongoing process of reviewing the lexical entries during the work with data. The first results have been encouraging.

8. References

- Benson, M., Benson, E. & Ilson, R. F. (2010). *The BBI Combinatory Dictionary of English. Your guide to collocations and grammar*. Third edition revised by Robert Ilson. Amsterdam, Philadelphia, John Benjamins.
- Evert, S. (2013). Tools for the acquisition of lexical combinatorics. In R. H. Gouws, U. Heid, W. Schweickard & H. E. Wiegand (eds.) *Dictionaries. An International Encyclopedia of Lexicography. Supplementary volume: Recent Developments with Focus on Electronic and Computational Lexicography (HSK 5.4)*, 104. Berlin & New York: Mouton de Gruyter, pp. 1415–1432.
- Dhrimo, A., Tupja, E. & Ymeri, E. (2007). *Fjalor sinonimik i gjuhës shqipe (= Dictionary of Synonyms of the Albanian Language)*. Tiranë: Toena.
- Hardie, A. (2012). CQPweb – combining power, flexibility and usability in a corpus analysis tool. In: *International Journal of Corpus Linguistics* 17 (3), pp. 380–409.
- Häcki Buhofer, A., Dräger, M., Meier, S. & Roth, T. (2014): *Feste Wortverbindungen des Deutschen. Kollokationenwörterbuch für den Alltag*. Tübingen: Francke. Online: <https://kollokationenwoerterbuch.ch/>.

- Herbst, T. (1996). What are collocations: sandy beaches or false teeth. *English Studies*, 1996, pp. 379–393.
- Kabashi, B. (2015). *Automatische Verarbeitung der Morphologie des Albanischen*. Erlangen: FAU University Press.
- Kabashi, B. (2017). “AlCo – një korpus tekstesh i gjuhës shqipe me njëqind milionë fjalë” (= AlCo – a hundred million word corpus of the Albanian language). In: *Seminari XXXVI Ndërkombëtar për Gjuhën, Letërsinë dhe Kulturën Shqiptare* (= *The XXIV International Seminar for Albanian Language, Literature and Culture*). Universiteti i Prishtinës, Kosovo, Universiteti i Tiranës, Albania. Nr. 36/2017, pp. 123–132.
- Kabashi, B. & Proisl, T. (2018). “Albanian Part-of-Speech Tagging: Gold Standard and Evaluation”. In N. Calzolari et al. (eds.) *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan. European Language Resources Association (ELRA) Paris, pp. 2593–2599.
- Kabashi, B. (2019). A lexicon of Albanian for Natural Language Processing. In R. H. Gouws, U. Heid, T. Herbst, S. Schierholz & W. Schweickard (eds.) *Lexicographica, International Annual for Lexicography, Vol. 34*, pp. 233–242.
- Kostallari, A. (kryeredaktor), Thomaj, J., Lloshi, X., & Samara, M. (1980). *Fjalor i gjuhës së sotme shqipe* (= *Dictionary of Contemporary Albanian Language*). Tiranë: Akademia e Shkencave e RPS të Shqipërisë.
- Proisl, T. (2019). *The Cooccurrence of Linguistic Structures*. Erlangen: FAU University Press.
- Quasthoff, U. (2010). *Wörterbuch der Kollokationen im Deutschen*. Berlin, etc.: de Gruyter.
- Samara, M. (1998): *Fjalor i antonimeve në gjuhën shqipe* (= *Dictionary of Antonyms in the Albanian Language*). Shkup: Shkupi.
- Sinclair, J. M. (1991) *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
- Thomai, J., Samara, M., Shehu, H. & Feka, T. (2004). *Fjalori sinonimik i gjuhës shqipe* (= *The Dictionary of Synonyms of the Albanian Language*). Tiranë: Akademia e Shkencave e Republikës së Shqipërisë.
- Thomai, J., Samara, M., Haxhillazi, P., Shehu, H., Feka, T., Memisha, V. & Goga, A. (2006). *Fjalor i gjuhës shqipe* (= *Dictionary of the Albanian Language*). Tiranë: Akademia e Shkencave e Republikës së Shqipërisë.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



Investigating Semi-Automatic Procedures in Pattern-Based Lexicography

Laura Giacomini^{1,2}, Paolo DiMuccio-Failla

¹ Institute for Translation and Interpreting (IÜD), University of Heidelberg,
Plöck 57a, D-69117 Heidelberg

² Institute for Information Science and Natural Language Processing (IwiSt), University of
Hildesheim, Universitätsplatz 1, D-31141 Hildesheim

E-mail: laura.giacomini@iued.uni-heidelberg.de, paolodimuccio@gmail.com

Abstract

In this contribution we present existing pattern description models with different degrees of computerization, discuss their potential from the perspective of the creation of an e-lexicographic resource for language learners, introduce the parameters of pattern accuracy and ontology reliability for a qualitative evaluation of the results, and make some proposals for a future quantitative evaluations. The models discussed are a) Hanks's CPA and the Pattern Dictionary of English Verbs (PDEV), b) methods employed by Tecling (Technologies for Linguistic Analysis, Pontifical Catholic University of Valparaiso, Chile) and Verbario, a pattern database of Spanish verbs, and c) an ongoing lexicographic project for the compilation of a learner's dictionary of Italian linked to a conceptual ontology. These approaches are founded in the tradition of theories focussing on the connection between lexis and grammar, especially in John Sinclair's view of *normal patterns of usage* as the true bearers of meaning of a language.

Keywords: pattern-based lexicography; semi-automatic procedures; ontology; pattern of usage; learner's dictionary

1. Introduction

Linguistic approaches covering, to different degrees, the interplay between lexical patterns and grammatical frameworks, or, in John Sinclair's words, "the meeting of lexis and grammar" (Sinclair, 1991: 81), have a quite long tradition ranging from lexicogrammar theories (cf. Halliday, 1992), to Gross's *classes d'objets* (1994) and Herbst's notion of *Konstruktikon* (2016). This tradition is largely intertwined with corpus-based and corpus-driven methods. In the context of pattern-based lexicography, especially in the sense of Sinclair (1991) and Hanks's Theory of Norms and Exploitations (Hanks, 2013), much research has been done to integrate notions of lexical semantics into the study of (phraseological) word combinations, giving birth to pioneering dictionaries such as the Collins COBUILD English Language Dictionary (COBUILD 1987) and the New Oxford Dictionary of English (Hanks & Pearsall, 1998).

However, methods for the computerization of the lexicographic process have been only recently taken into consideration as an essential part of pattern-centred dictionary research. In this contribution, we would like to compare existing semi-automatic pattern description models, discuss their potential from the perspective of the creation of an e-lexicographic resource for language learners, and make some proposals for

improving work in the future.

This study belongs to the initial phase of our lexicographic project for the compilation of a learner’s dictionary of Italian, for which the description of syntactic and semantic patterns of language has been chosen as the core microstructural criterion (DiMuccio-Failla & Giacomini, 2017a, 2017b). In the following, we will refer to the project as the IFL (Italian as a foreign language) project.

In the next section we first introduce the three models we are comparing in our study (Section 2). We then move to the relevant steps in the lexicographic process and corresponding solutions offered by the three models (Section 3). Finally, we discuss the impact of semi-automatic procedures on the lexicographic workflow and propose parameters for qualitative evaluation (Section 4).

2. Models

In this study we take into consideration three models for pattern-based lexicographic description. All these approaches originate from Sinclair’s notion of *normal patterns of usage* as the true lexical units of a language: according to Sinclair, in general each major normal sense of a word can be associated with a distinctive pattern of usage determined by collocation, colligation, semantic preference and semantic prosody (e.g. Sinclair, 1996, 2004).

- Hanks’ CPA and the Pattern Dictionary of English Verbs (PDEV) as its lexicographic result,
- Methods employed by Tecling (Technologies for Linguistic Analysis, a group of research in computational linguistics and NLP affiliated to the Pontifical Catholic University of Valparaiso, Chile) to automatically induce a taxonomy of nouns and generate patterns from corpora, and
- Experiments carried out within our lexicographic project for learners of Italian.

The Pattern Dictionary of English Verbs (PDEV) is the practical result of the application of Hanks’ Theory of Norms and Exploitations (Hanks, 2013) and the technique of Corpus Pattern Analysis (CPA, Hanks, 2004b). The Pattern Dictionary of English Verbs is primarily intended as a resource for use in computational linguistics, due to its pattern formalization, but also in language teaching and cognitive science. It presently includes 1,423 complete verbs out of a total of 5,392. For each verb, a set of patterns is provided in which semantic types or semantic roles are indicated for each argument. Arguments in a pattern are linked to nodes in the CPA Ontology, a shallow semantic ontology which contains 253 semantic types.

Researchers at Tecling have taken Hanks’ theory and the CPA’s approach as a starting point to develop methods to automatically induce taxonomies of nouns and patterns of verbs from corpora. The language of application is Spanish. In the framework of the

Verbario project (2014-2017, www.verbario.com), a database of Spanish verbs was semi-automatically created. Verbario currently features two versions, one with manually created patterns, another with automatically generated patterns.

The IFL project is presently carried out by a group of researchers at Heidelberg University (Germany), Hildesheim University (Germany) and the University of Modena and Reggio-Emilia (Italy). We aim, on the one hand, at describing patterns of verbs and other word classes in a dictionary for learners of Italian and, on the other hand, at developing an ontology-like conceptual network in which semantic fillers (semantic types and roles) are collected, and on which lexicographic pattern description can be based. Our model has a clear cognitive orientation, in that it attempts to define word meanings by first identifying prototypical concepts and then finding and logically arranging related concepts. In the current, initial stage of the project, we are mainly concerned with studying patterns of different word classes, especially working on semantically homogeneous verbs. We also make some experiments in other languages (English, German, French), to test the validity of our method (cf. Orlandi, Giacomini & DiMuccio-Failla, 2019) and refine the results obtained for Italian.

3. Pattern-based lexicographic process and semi-automatic procedures

The models we intend to compare share, on the one hand, a common theoretical background, which has found application in different languages. On the other hand, they develop different strategies for the implementation of the core steps within the pattern-based lexicographic process: (a) detecting patterns in corpora, (b) selecting semantic types, (c) formally or informally expressing patterns, and (d) building taxonomies/ ontologies for semantic types. Moreover, they organize these procedural steps in different ways and choose different principles for sorting the meanings of polysemous words.


In this section, we describe and compare all the different strategies, especially from the perspective of computerization. For the purpose of this contribution, we will concentrate on verbal patterns only, since this is the main focus of the three models.

3.1 Identification of patterns and semantic fillers

PDEV, Verbario and the IFL dictionary all record data from corpora. For the PDEV, the British National Corpus has been used as the main reference corpus. Different to the PDEV, web corpora (esTenTen and itTenTen) have been used in Verbario and the IFL project for Spanish and Italian. Web corpora have the advantage of being large and heterogeneous enough to offer a broad spectrum of contexts, covering many different text genres and text types. On the negative side, at least for what concerns itTenTen in the focal project, the relatively great amount of noise and the imbalance in the distribution of text sources posed some problems. For these reasons, the IFL

project also integrates in the lexicographic process a comparison of corpus data with existing general language and collocation dictionary data.

Table 1 summarizes the steps that enable the assignment of concordance lines to patterns:



CPA	Tecling	IFL project
Concordance sampling	Concordance sampling	Collocation extraction and concordance sampling
	Syntactic structures extraction	
	Semantic information extraction	
Sample analysis and pattern identification	Sample analysis and pattern identification	Sample analysis and pattern identification

Table 1: Process of pattern identification in the three models (blue field: manual step, grey field: semi-automatic step, white field: automatic step)

In the three models, concordance analysis delivers syntactic and semantic information about a verb in its contexts. However, in the IFL project collocation analysis is the starting point of investigation on which the analysis of concordances is based.

From the beginnings of the Sinclairian tradition in pattern-based lexicography, concordances have played a crucial role. An important issue concerns the appropriate number of concordance lines to be taken into consideration. Sinclair makes the case for *small samples*, a “screenful” of around 25 lines, which should be enough to get a first overview of the patterning of a node (2003: xiii-xiv). Analysis then continues in two possible directions: the main patterns can be confirmed by subsequently adding new small samples from the same dataset until no new information is obtained, or data can be refined if the initial search results are not satisfactory.

Hanks suggests that detailed analysis requires the selection of a random sample of up to 1,000 concordance lines, usually starting with a small sample of 200-250 lines (2004a: 255, PDEV). Tecling uses subsequent samples of around 100 corpus lines each, and note that a maximum of three samples is usually sufficient to identify all major patterns of a verb. Concordances are automatically generated, whereas the association of concordances with patterns is a manual step, usually carried out in an iterative way (see Figure 1 for the usual procedure). Hanks points out that “the identification of a syntagmatic pattern is not an automatic procedure: it calls for a great deal of lexicographic art” (from the PDEV website).

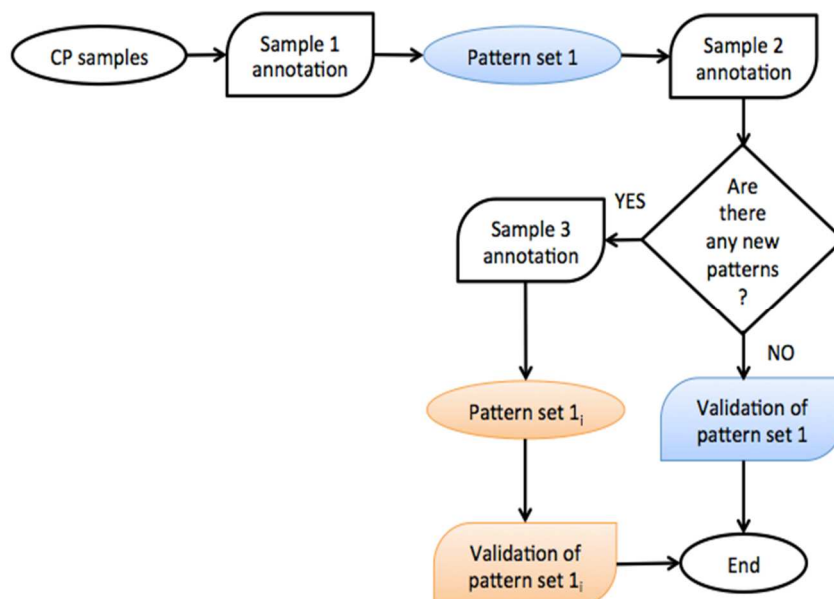


Figure 1: Iterative analysis of corpus concordances for pattern identification

CPA:

In CPA, one starts with concordance lines extracted by the SketchEngine concordancer (Kilgarrieff et al., 2004), and groups them into semantic homogeneous sets, whereas “associating a ‘meaning’ with each pattern is a secondary step, carried out in close coordination with the assignment of concordance lines to patterns” (Hanks, 2004b: 88).

Each concordance line is manually annotated with a pattern number, exploitations and non-relevant data (e.g. errors or quotations) (Figure 2).

letters or telephone calls, or who were	adopted	1	or unable to give information about bowel
little girls being sold as prostitutes. `She	adopts	1	them. Everybody is going to love this film
delight as a good-hearted youngster who	adopts	1	vagrant Hume Cronyn in Christmas On Division
businesses and societies are encouraged to `	adopt	1.a	' one or more panes of glass. One pane will
the bell and say that you would like to	adopt	1.a	an old person. Do not only go and see them
naturally, and we `naturalize' those whom we	adopt	1.a	fully into our own community -- those who
outside of Ireland may be automatically `	adopted	1.f	' here with the problem and its solution
local radio. Still uncertain of what tone to	adopt	2	, the campaigners brought six people dressed
Buddhism in her middle years, and more recently	adopted	2	a congenially fellow-travelling stance

Figure 2: Annotation of concordance lines according to CPA

Once patterns have been identified, semantic values (types and roles) are manually attributed to the arguments of the input word in each pattern by referring to the CPA Ontology (cf. Section 3.3). One of the main issues of this step concerns the choice of the appropriate semantic values: “among the most difficult of all lexicographic decisions is the selection of an appropriate level of generalization on the basis of which senses are to be distinguished” (from the PDEV website). As already mentioned in Section 2,

each entry in the PDEV consists of a formalized pattern with its semantic fillers, an *implicature* expressing the pattern meaning in natural language (Hanks, 2004b: 88), a usage example and frequency indication (cf. Figure 3 for the core microstructural items).



Figure 3: Entry example in PDEV

Tecling:

Concordances of a verb are extracted from the esTenTen corpus by means of Jaguar, a tool for corpus exploitation (<http://www.tecling.com/jaguar>). Concordance analysis is complemented with dependency analysis carried out by using Syntaxnet, Google's open-source parser. Semantic analysis also plays a role at this stage: named entities are classified through POL, a NER-tool for detecting and classifying names of geographical places, persons and organizations (<http://www.tecling.com/pol>), while common names are classified through a previously generated taxonomy (cf. Section 3.3). Patterns are identified on the basis of syntactic functions and semantic types. Experiments have been carried out to compare manual and automated pattern identification with the aim of improving automation in order to support lexicographers' work (Renau, et al., 2019; Renau & Nazar, 2016). Manual analysis of a set of verbs has been used as a gold standard to test the results of automatic analysis, in which semantic fillers are obtained from the available taxonomy. The main problem with the automatic output overspecification of semantic values for the arguments is that the implemented algorithm selects the first available semantic type by proceeding bottom-up in the taxonomy, frequently producing too specific and too many patterns (ibid.: 895-897).

The CPA orientation of this work is reflected by the entry structure in Verbario, for instance for the Spanish verb *aburrir* (to bore) (Figure 4):

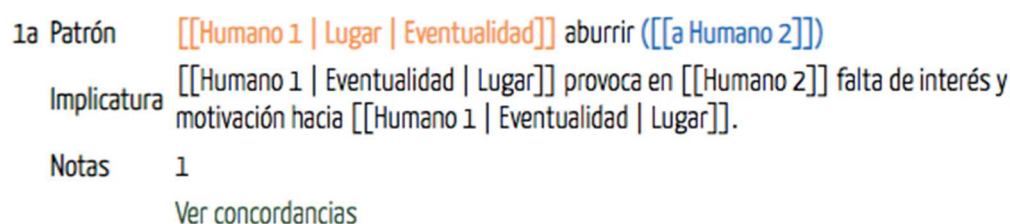


Figure 4: Entry example in Verbario

IFL project:

In the IFL project, we collect collocations of a node verb from the itTenTen corpus through the Sketch Engine word sketch tool, and then extract concordance lines referring to these collocations in order to validate them. Random corpus samples

filtered by using the GDEX function (Kilgariff et al., 2008) are analysed according to the meaning of the node verb and patterns are thus gradually identified.

Lexicographic work in the IFL project has a clear phraseological orientation: not only do we firmly believe that patterns are phraseological in nature (Sinclair, 1991, 1996), but we also explore collocations to identify, validate and refine our patterns and, in general, we use phraseological disambiguators to cluster patterns with close meanings (DiMuccio-Failla & Giacomini, 2017b). Collocation analysis sheds light on the syntactic, semantic, and phraseological features of verbs at the same time. Figure 5 shows the informal pattern description in a possible data presentation mode of the planned IFL dictionary.

! seguire¹ [attività] qn./qs. v. trans. <seguo, ho seguito>
 \se'gui:re\
[SPAZIALMENTE]
! Seguire qn. (che sta andando da qualche parte) : Andare dietro a qn., mantenersi dietro qn.; spostarsi in modo da rimanere in presenza di qn. ES.: Abbiamo seguito la guida turistica fino al castello. COLLOC.: s. qn. da vicino, s. qn. a distanza, s. un corteo. Segua quella macchina! **S. qn. in qc. luogo [est.] : Andare in qc. luogo dietro o subito dopo qn. (di solito per raggiungerlo lì).** ES.: I tifosi seguirono la squadra dentro gli spogliatoi. **S. qn. a ruota [sport.] : Seguire qn. con la propria ruota anteriore vicina alla sua ruota posteriore. [...]** ➔ **SINONIMI:** inseguire; accompagnare; pedinare. **CONTRARI:** guidare.

Figure 5: Entry example (*seguire, to follow*) in the planned IFL dictionary

Challenges typically encountered at this stage are:

- In the case of the IFL project, first grouping collocations into semantically homogeneous sets, each identifying a pattern.
- Distinguishing primary patterns from secondary patterns.
- Assigning semantic values to argument slots: selection restrictions and determining the appropriate degree of generalization.

3.2 Pattern sorting

Interestingly, the sorting of patterns in the final application of the three models (PDEV, Verbario, IFL project) complies with different principles. In the PDEV, the patterns of a verb are sorted according to their cognitive salience. Also in the IFL project, senses identified by patterns are sorted according to their cognitive relevance: we start from the idea of a conceptual network in which the related senses of a word are organized in a radial set around one or possibly more prototypical concepts. This assumption verifies the cognitivist account of polysemy proposed by Brugmann and Lakoff (cf., among others, Brugmann & Lakoff, 1988). The fundamental meaning of a verb is thus followed by other senses linked by metonymy, abstraction, and metaphor relations (cf. examples in DiMuccio-Failla & Giacomini, 2017b).

In the Tecling project, patterns are sorted by decreasing order of frequency (Renau & Nazar, 2016: 827). This sense enumeration approach has been criticized in the past, not least in the context of the COBUILD Dictionary, because it often leads to unnatural results (cf. Lew, 2013; DiMuccio-Failla & Giacomini, 2017b). We think that the cognitive criterion of sense disambiguation is the most suitable way of presenting meanings of polysemous words to language learners, since it logically guides the dictionary users from a prototypical meaning towards all related senses (e.g. figurative senses).

3.3 Ontology building

The role of an ontology of semantic types and semantic roles in pattern-based lexicography is of crucial importance: the systematic conceptual classification of these items guarantees consistency in their use throughout the dictionary and potentially simplifies pattern formulation. In this contribution, for reasons of simplicity, we will use the term *ontology* to refer to a typically hierarchical structure of entities or concepts, irrespective of its complexity and degree of expressiveness, therefore also including taxonomies. The three discussed models show clear differences with regard to

- the method for ontology building and
- the way in which the ontology interfaces with pattern identification.

Table 2 shows the role of the ontology within the process of pattern identification in the three models:

CPA	Tecling	IFL project
CPA Ontology	Taxonomy	
Concordance sampling	Concordance sampling	Collocation extraction and concordance sampling
	Syntactic structures extraction	
	Semantic information extraction	
Sample analysis and pattern identification	Sample analysis and pattern identification	Sample analysis and pattern identification
		Conceptual ontology

Table 2: Ontology and pattern identification in the three models (blue field: manual step, grey field: semi-automatic step, white field: automatic step)

CPA:

The CPA Ontology is based on work done by Pustejovsky et al. (2004). It is a shallow semantic ontology created by progressively compiling and organizing a list of semantic types (El Maarouf, 2013). As previously indicated, in the PDEV each argument of each

pattern is linked to a node in the CPA Ontology, which can be accessed via the dictionary website. Here is a brief example of a hierarchy:

Anything > Entity > Physical Object > Inanimate > Artefact > Building

Final nodes of the ontology may be very specific, but sibling concepts are still missing. For instance, the only two available subcategories of Building are Cinema and Theatre, whereas for Food only the subcategory Meat is given.

Tecling:

Tecling uses a statistically-based taxonomy induction algorithm to generate a taxonomy of Spanish nouns from a corpus. Different quantitative approaches are simultaneously applied, among which the computation of similarity coefficients to identify sibling words and of asymmetric co-occurrence to find parent-child nodes (Nazar & Renau, 2016). As pointed out in Renau and Nazar (2016), this procedure relies on an existing taxonomy structure. In fact, semantic types contained in the CPA Ontology provide the conceptual architecture into which around 35,000 Spanish nouns are automatically inserted. The results are compared with the Spanish WordNet 1.6 (Atserias et al., 2004), which serves as a gold standard (for a brief discussion on the use of wordnets as sources for semantic types, see further down in this section). Insights into the automatically induced taxonomy are provided by the ontology webpage (www.tecling.com/kind). For instance, if we search for the category Comida (Food), we get a full list of four hypernyms and 157 hyponyms. The taxonomy induction algorithm employed by Tecling can detect both symmetric and asymmetric relations, and achieves an estimated average of 77.86% precision and 33.72% recall on the total results (Nazar & Renau, 2016).

IFL project:

In our project, a conceptual ontology is developed alongside the process of pattern identification. It is important to note that we presently work on patterns without employing an external ontology. Instead, we build a new conceptual network according to a bottom-up procedure, in which semantic types (and lexicalized semantic roles) selected for patterns are progressively fed into the ontology.

For instance, one of the patterns of the verb *seguire* (*to follow*) is

seguire il racconto, la spiegazione o l'argomentazione di qn.

We insert the semantic types *Racconto* (Narration), *Spiegazione* (Explanation) and *Argomentazione* (Argumentation) into our ontology and link them to other concepts, e.g. synonyms such as *Narrazione* (Narration) and hypernyms, in this case via a polyhierarchical structure, in the sense that the three semantic types have two different hypernyms, *Evento comunicativo* (Communicative event) and *Rappresentazione formale di un evento comunicativo* (Formal representation of a communicative event)

(Figure 6):



Figure 6: Excerpt from the ontology, with semantic types derived from pattern formulation

In order to systematically detect relevant types, we analyse clusters of semantically close verbs (e.g. synonyms, converses, or troponyms), which display some meaning overlap and are likely to share a number of semantic fillers. As we are still at an initial stage of the project, we only have a very small number of items in our ontology, corresponding to a small number of words in the dictionary. As the ontological structure is being configured, its items are used in a top-down procedure to fill argument slots of new verbs. Being dependent on their usage as semantic fillers in argument slots, the hierarchy of types only has to be as systematic and coherent as normal language usage.

The ontology is not only a repository of semantic types, it also provides a clear overview of the lexical domains we intend to cover and facilitates consistent dictionary definitions. The upper part of the ontology draws on the EuroWordNet model (Vossen et al., 1998), which, in turn, is based on Lyon’s (1977) tripartite entity categorization. In the lower part of the ontology entities are further classified into types (cf. DiMuccio-Failla & Giacomini, 2017a). Tests performed on ItalWordNet and experiments carried out on English, German and French using the Princeton WordNet, GermaNet and WoNeF, reveal that wordnets have a limited reliability with regard to semantic types: they pose major problems for meaning disambiguation (for instance, synsets are not always clearly distinct from each other). Moreover, they often introduce scientifically motivated subcategorizations “that are not in ordinary usage” (Jezek & Hanks, 2010) and therefore not useful for lexicographic purposes¹ (wordnets’ drawbacks in this sense have also been described by Hanks and Pustejovsky (2005) and Renau et al. (2019)). As pointed out by Polguere, the Princeton WordNet’s ontological structure “is not as cognitively relevant as it was expected to be by its designers [...], [since] the focus of the project shifted at an early stage from psycholinguistics to computer applications”

¹ Jezek & Hanks (2010) also see a problem in the attempt to force all items of a language into a taxonomic hierarchy.

(2014: 397). This aspect, which appears to be common to all wordnets, is crucial to our approach to lexicography, which, instead, has a strong cognitive orientation.

The challenge typically encountered at this stage is:

- No conceptual ontology is already available from which semantic types can be reliably obtained.

4. Semi-automatic procedures and lexicographic workflow: qualitative analysis

We will now concentrate on the impact of automatic procedures on time efficiency and the quality of the lexicographic results, for instance on the accuracy of patterns and reliability of the underlying ontology. In the previous sections we introduced the set of automatic steps used either in all three models or only in some of them:

- taxonomy induction (Tecling)
- concordance sampling (CPA, Tecling, IFL project)
- syntactic structures extraction (Tecling)
- semantic information extraction (Tecling)
- collocation extraction (Tecling, IFL project)
- pattern extraction (IFL project)

We attempt to assess the potential of these methods specifically for the production of a learner's dictionary, which is the main goal of the planned IFL project but not of the two other models. The results of CPA and Tecling research, namely PDEV and Verbario, are in fact rather to be understood as databases in which formal data representation can serve as a possible source for a learner's lexicography.

Due to the differences between the described models (e.g. language, degree of computerization, intended goal), at the moment we cannot rely on any metrics for a quantitative evaluation of the results. Even within the same model, a quantitative evaluation is a difficult goal to achieve. As pointed out by Renau et al. (2019: 897) in the case of automatic pattern generation, for example, it is even impossible to establish a baseline, since we are not dealing with a classification system in which a certain chance of success with a random selection or a trivial method is given.

We therefore provide a primarily qualitative analysis based on the examination of the achieved results (pattern accuracy, also in comparison to monolingual dictionaries, and ontology reliability), and observations made by the involved researchers about their own work. The parameters we chose for assessing the quality of the final results are pattern accuracy and ontology reliability. Details regarding final results according to these parameters will now be presented, followed by remarks on time efficiency and source data.

4.1 Pattern accuracy and ontology reliability

	PDEV	Verbario	IFL project
Sample	<i>follow, need, choose, adopt, eat</i>	<i>abrir, aburrir, acentuar, activar, cortar</i>	<i>seguire, inseguire, accompagnare, pedinare, incalzare</i>
Pattern uniqueness	Patterns are distinct from each other	Patterns are not always distinct from each other	Patterns are distinct from each other
Pattern expressiveness	Heterogeneous degree of expressiveness: several semantic fillers appear to be too generic	Generally limited degree of expressiveness	High degree of expressiveness: semantic fillers are as specific as possible
Semantic coverage	Large semantic coverage, almost all dictionary senses corresponding to normal usage match a pattern (Dictionaries: COBUILD, ODE)	Large semantic coverage, almost all dictionary senses corresponding to normal usage match a pattern (Dictionaries: DAELE, SALAMANCA)	Each dictionary sense corresponding to normal usage matches a pattern (Dictionaries: TRECCANI, DE MAURO)
Ontology depth	Shallow ontology with limited inheritance levels	This kind of taxonomy appears to have a greater depth than the CPA Ontology	The depth of the ontology depends on normal language usage (bottom-up approach)
Relation patterns-ontology	Top-down approach: coherent usage of semantic types in patterns according to the depth of the ontology	Top-down approach: coherent usage of semantic types in patterns according to the depth of the ontology	Bottom-up approach: the ontology is systematically filled with semantic types selected during pattern identification

Table 3: Pattern accuracy and ontology reliability in the three models

Pattern accuracy is tested by selecting a small verb sample from each dataset and considering, for each verb, the uniqueness of patterns (each pattern identifies one

distinct meaning), the expressiveness of semantic types used as argument slot fillers (degree of generalization), and semantic coverage in comparison to meaning presentation in existing monolingual learner's dictionaries². Ontology reliability is tested by the conceptual depth of the ontology and the way in which the ontology interfaces with the building of dictionary patterns. Table 3 shows the results of our analysis. Verbario has been considered in its manual version, since the automatically generated verb entries cannot be presently accessed online.

The clearest difference between the lexicographic results obtained by the three models concerns the expressiveness of patterns, and the depth of the ontology (for the important factors here see the examples mentioned in Section 3.3). Pattern expressiveness seems to be more dependent on the chosen approach rather than on process automation, which explains the similarity between data in the PDEV and Verbario as opposed to the IFL project data. These observations hint at the fact that the correlation between computerization, on the one hand, and pattern accuracy and ontology reliability on the other, should not be overrated in any direction.

4.2 Time efficiency and initial data

Some remarks need now to be made on time efficiency: generally speaking, time efficiency is enhanced by the application of any automated procedures. However, the balance between the amount of time saved thanks to automatic data extraction and the amount of time spent to correct and prepare data for presentation in a dictionary should also be taken into account (cf. also Renau et al. (2019) on the comparison between manual extraction and automatic extraction of patterns).

In our experience, manual work for pattern identification and ontology building requires a considerable amount of time, but this process can be significantly accelerated as soon as targeted initial data are available, for instance complex collocations, or semantic and pragmatic information found in discourse (e.g. stage-dependent conditions for a verb's meanings, cf. Kratzer (1995)). We are presently investigating methods for creating automatic procedures that are able to provide this kind of raw data. In addition to this, cross-language experiments with English, German and French help us refine both ontological and lexicographic data (a multilingual approach has been partly adopted in the context of CPA as well, cf. Baisa et al. (2016)). The manually compiled conceptual ontology in the IFL project may serve in the future as a gold standard for a quantitative evaluation of automatically obtained results, not only for Italian but also for other languages.

² There is no learner's dictionary for Italian yet, thus we had to use general monolingual dictionaries.

5. Conclusions

The idea of a direct comparison with similar methods originated in issues encountered during our empirical work on patterns. These issues can be summarized as follows:

- Detecting patterns in corpora and validating them against the content of existing dictionaries requires a considerable amount of time, especially for extracting relevant data such as syntactic structures and collocations.
- Formulating patterns (either informally or formally) is a conceptually complex activity; especially the choice of adequate semantic types and roles for argument slots would be easier if a corresponding ontology was already available.
- Building such an ontology is closely related to the building of patterns. Due to the impossibility of using existing ontologies or wordnets as a source for cognitive conceptual information (cf. Section 3.3), this task is also particularly challenging, especially for what concerns the selection of the appropriate generalization level.

All these issues greatly affect the lexicographic workflow in the IFL project, and will presumably remain at the heart of the discussion also in the future. We assume that automation can only improve the workflow in a satisfactory way as long as it does not require much manual effort to correct data at a later stage. As shown in the previous sections, much depends on the theoretical approach to pattern description. For the moment, the degree of pattern expressiveness and cognitive consistency aimed at in the IFL project can only be achieved by native speaker introspection (on the importance of introspection and intuition, cf. Sinclair (1991, 2004)). Introspection, however, benefits from the availability of suitable, automatically extracted initial data and will be further enhanced as soon as the linked conceptual ontology reaches a sufficient level of completeness to be used to automatically detect patterns of similar verbs in corpora.

6. References

- Atserias, J., Villarejo, L. & Rigau, G. (2004). Spanish WordNet 1.6: Porting the Spanish WordNet across Princeton versions. In N. Calzolari, K. Choukri & T. Lino et al. (eds.) *Proceedings of the Fourth International Conference on Language and Resources Evaluation (LREC)*, pp. 161-164.
- Baisa, V., Može, S., & Renau, I. (2016). Multilingual CPA: Linking Verb Patterns across Languages. In T. Margalitazde & G. Meladze (eds.) *Proceedings of the XVII EURALEX International congress. Lexicography and linguistic diversity*, pp. 410-417.
- Brugman, C. & Lakoff, G. (1988). Cognitive topology and lexical networks. In S. Small, G. Cottrell & M. Tanenhaus (eds.) *Lexical Ambiguity Resolution*. San Mateo, California: Morgan Kaufmann, pp. 477–507.
- COBUILD (1987): *Collins COBUILD English Language Dictionary*. Collins.
- COBUILD: *Collins COBUILD Advanced Learner's Dictionary (2014)*. Harper Collins.

- DAELE. *Diccionario de aprendizaje del Español como Lengua Extranjera*. Accessed at: <http://www.iula.upf.edu/rec/daele> (10 June 2019)
- DE MAURO (2019). *Il Nuovo De Mauro*. Accessed at: <https://dizionario.internazionale.it> (10 June 2019)
- DiMuccio-Failla, P. V. & Giacomini, L. (2017a). Designing an Italian learner's dictionary based on Sinclair's lexical units and Hanks's corpus pattern analysis. In I. Kosem et al. (eds.) *Proceedings of the Fifth eLex Conference Electronic Lexicography in the 21st Century*. Leiden, Netherlands.
- DiMuccio-Failla, P. V. & Giacomini, L. (2017b). In M. Mitkov (ed.) *Computational and Corpus-Based Phraseology. Second International Conference, Europhras 2017*, LNAI 10596. Springer, pp. 290-305.
- El Maarouf, I. (2013). Methodological Aspects in Corpus Pattern Analysis. *ICAME Journal*, 37, pp. 119-148.
- Gross, G. (1994). Classes d'objets et description des verbes. *Langages*, pp. 15-30.
- Halliday, M. A. K. (1992). Some lexicogrammatical features of the zero population growth text. *Discourse Description: Diverse Linguistic Analyses of a Fund-Raising Text*. Amsterdam: Benjamins, pp. 327-358.
- Hanks, P. (2004a). The syntagmatics of metaphor and idiom. *International Journal of Lexicography*, 17(3), pp. 245-274.
- Hanks, P. (2004b). Corpus pattern analysis. In *Proceedings of the XI EURALEX International Congress*, Vol. 1, pp. 87-98.
- Hanks, P. (2013). *Lexical analysis: Norms and exploitations*. MIT Press.
- Hanks, P. & Pearsall, J. (eds.) (1998). *New Oxford Dictionary of English*, 1st ed. Oxford University Press, Oxford
- Hanks, P., & Pustejovsky, J. (2005). A pattern dictionary for natural language processing. *Revue Française de linguistique appliquée*, 10(2), pp. 63-82.
- Herbst, T. (2016). Wörterbuch war gestern. Programm für ein unifiziertes Konstruktikon. In S. J. Schierholz, R. H. Gouws, Z. Hollós & W. Wolski (eds.) *Wörterbuchforschung und Lexikographie*. Berlin/Boston: de Gruyter.
- Jezek, E., & Hanks, P. (2010). What lexical sets tell us about conceptual categories. *Lexis*, 4(7).
- Kilgariff, A., Rychly, P., Smrz, P., & Tugwell, D. (2004). Itri-04-08 The sketch engine. *Information Technology*, pp. 105-116.
- Kilgariff, A., Husák, M., McAdam, K., Rundell, M. & Rychlý, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus. In *Proceedings of the XIII EURALEX International Congress*, pp. 425-431.
- Kratzer, A. (1995). Stage-level and individual-level predicates. In G. N. Carlson & F. J. Pelletier (eds.) *The generic book*, Chicago. University of Chicago Press, pp. 125-175.
- Lew, R. (2013). Identifying, ordering and defining senses. In H. Jackson (ed.) *The Bloomsbury Companion to Lexicography*. Bloomsbury, pp. 284-302.
- Lyons, J. (1977). *Semantics*. Cambridge University Press.
- Nazar, R. & Renau, I. (2016). A taxonomy of Spanish nouns, a statistical algorithm to

- generate it and its implementation in open source code. In N. Calzolari et al. (eds.) *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16)*. European Language Resources Association (ELRA).
- ODE 2019: *Oxford Dictionary of English*. Accessed at: <https://en.oxforddictionaries.com>. (05 June 2019)
- Orlandi, A., Giacomini, L. & DiMuccio-Failla, P. V. (2019). I disambiguatori fraseologici nella lessicografia di apprendimento: una proposta per l'italiano e il francese. *Repères-Dorif*, 18.
- PDEV. *Pattern Dictionary of English Verbs*. Accessed at: <http://pdev.org.uk> (10 June 2019)
- Polguère, A. (2014). From writing dictionaries to weaving lexical networks. *International Journal of Lexicography*, 27(4), pp. 396-418.
- Pustejovsky, J., Hanks, P., & Rumshisky, A. (2004). Automated Induction of Sense in Context. *COLING 2004*. Geneva, Switzerland.
- Renau, I. & Nazar, R. (2016). Automatic Extraction of Lexical Patterns from Corpora. In T. Margalitazde & G. Meladze (eds.) *Proceedings of the XVII EURALEX International congress. Lexicography and linguistic diversity*, pp. 823-830.
- Renau, I., Nazar, R., Castro, A., López, B., & Obreque, J. (2019). Verbo y contexto de uso: un análisis basado en corpus con métodos cualitativos y cuantitativos. *Revista Signos. Estudios de Lingüística*, 52(101).
- SALAMANCA (2006) Cuadrado, J. G. *Diccionario Salamanca de la lengua española:[español para extranjeros]*. Santillana.
- Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.
- Sinclair, J. (1996). The search for units of meaning. *Textus: English Studies in Italy* 9(1), pp. 75-106.
- Sinclair, J. (2003). *Reading concordances: An introduction*. Pearson Longman.
- Sinclair, J. (2004). *Trust the text: Language, corpus and discourse*. Routledge.
- TRECCANI (2019). *Vocabolario Treccani*. Accessed at: <http://www.treccani.it/vocabolario> (05 June 2019)
- Verbario. *Sémantica de los verbos en contexto*. Accessed at: <http://www.verbario.com> (05 June 2019)
- Vossen, P. et al. (1998). The EuroWordNet Base Concepts and Top Ontology. Document LE2-4003, D017, D034, D036, WP5. Accessed at: <http://dare.ubvu.vu.nl/bitstream/handle/1871/11130/D017.pdf> (20 March 2019)

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



ELEXIFINDER:

A Tool for Searching Lexicographic Scientific Output

Iztok Kosem, Simon Krek

Jožef Stefan Institute, Ljubljana, Slovenia
E-mail: iztok.kosem@ijs.si, simon.krek@ijs.si

Abstract

Access to lexicographic research is highly important for lexicographers when conceptualizing and compiling dictionaries, and preparing their publications for presentation to the lexicographic community. There have been several attempts to offer a systematic record of lexicographic scientific output, and advanced search of it, but most of them are no longer updated, focus only on bibliographic data, and do not include works from other fields related to lexicography. The tool called Elexifinder has been developed within the European Infrastructure for Lexicography (ELEXIS) project in order to facilitate knowledge exchange in the lexicographic community and promote open access culture in lexicographic research. In this paper, we present the first version of the tool that contains 1,755 publications and 78 videos in 11 different languages, and offers various search options to users. We describe the Elexifinder architecture, the process of including content, and present the interface's features. The paper concludes with the presentation of future plans, including the various publications that will be included in the next version of the tool.

Keywords: Elexifinder; lexicographic research; ELEXIS; lexicography; online tool

1. Introduction

In state-of-the-art lexicography, it is paramount that lexicographers have access to resources such as corpora and other dictionaries, and tools such as dictionary-writing systems and corpus query systems. Yet, it is equally important that lexicographers have constant access to scientific output in lexicography and disciplines related to lexicography, so they can follow the projects and research of their colleagues around the world, develop new ideas, conceptualize dictionaries, understand and address linguistic problems, and position their own work in the lexicographic community.

One of the issues faced by lexicographers is that lexicographically-relevant scientific output is very scattered. Journals focused on lexicography (the *International Journal of Lexicography*, *Dictionaries*, *Lexicographica*, *Lexikos* etc.) are published by different publishers. Then, each lexicographic association has its own proceedings, and moreover, their availability varies – some associations have all their proceedings freely available on their website (e.g. EURALEX), while others only the more recent ones (e.g. ASIALEX). Accessibility is especially an issue with older literature and books, as in most cases these resources are available only in print (although many of these publications have been digitized and can be searched in Google Books).

Further difficulty in finding lexicographically-relevant scientific output lies in the fact that many fields are of relevance to lexicography, for example lexical semantics, pragmatics, corpus linguistics, and more recently natural language processing. This means that lexicographers need to constantly follow journals, proceedings and other resources covering those fields for any relevant papers.

An additional obstacle to this form of knowledge exchange is language. Namely, lexicographers are usually very familiar with lexicographic research in their native language, and possibly in other languages they are fluent in. They can also fairly easily find lexicographic research in English, not only because there is an abundance of literature available but also because it is much better covered by search engines. However, to identify the relevant literature or authors in other languages is much more difficult. This can lead to isolation of researchers or communities, especially the ones that do not (also) publish in English. Thus, their work, as relevant and innovative as it may be, stays unnoticed in other communities.

In this paper, we present the Elexifinder tool which addresses these issues and has been developed within the European Infrastructure for Lexicography (ELEXIS) ¹, a H2020 project funded by the European Commission. First we conduct an overview of some existing efforts in collecting and recording lexicographic research, and their relevance for the development of the tool. This is followed by the description of the tool, the technology behind it and the procedure for including research papers. Next, we present parts of the tool interface and current contents. We conclude by discussing a few potential use cases and presenting plans for the future, both in terms of content and features.

2. Past and current efforts

There have been several known attempts to make an inventory of lexicographic literature, which have been focussed on collecting research publications, bibliographic information on publications and/or dictionaries, or both. We make an overview of them in this section.

Internationale Bibliographie zur germanistischen Lexikographie und Wörterbuchforschung by Wiegand (2012) is a five-volume bibliography that covers the largest number (33,339) of lexicographic works of all resources listed here. The shortcomings of this resource are that it is available in print only and that it is mainly focused on the field of German studies. Similarly limited in scope is Ahumada's (2016) collection of 6,560 items mainly from Hispanic (meta)lexicography.

¹ <https://elex.is/>

The second largest bibliography of lexicography is the one by Córdoba Rodríguez², which contains 10,192 items published between 1940 and 2003, including relevant newspaper articles. The search can be conducted by thematic blocks or by authors (alphabetically). One shortcoming of this bibliography is that it is no longer updated.

Also in need of an update is the International Bibliography of Lexicography, initiated by the European Association for Lexicography (EURALEX).³ It contains approx. 2,000 entries⁴ that can be viewed thematically or alphabetically. It also includes links to lists of reference portals and lexicographic bibliography collected by R.R.K. Hartmann (2007). As it is stated on the resource website, the bibliography has not been updated since 2012. In addition, the website offers rather limited search options.

The Online Bibliography of Electronic Lexicography (OBELEX) is an ongoing project at the Institute for the German Language (Möhrs & Töpel, 2011), which has started in 2008 and consists of two databases. The first database, called OBELEXdict includes over 17,000 online dictionaries (Möhrs, 2016), but as the authors point out on the resource website,⁵ “the term ‘dictionary’ [...] has a broad interpretation, i.e. all word-related reference works were included, without the quality of the content having been checked”. The users can search the database by type of dictionary, title, language (family), limit the search to dictionaries with audio or video files, illustrations, etc. The second database, called OBELEXmeta,⁶ contains bibliographic information on around 2,000 entries, which cover articles, monographs, anthologies and reviews. Most of the works in the database have been published from 2000 onwards, but some older relevant works are also included. Advanced search options, such as searching by title, author, year of publication, language and keywords, are provided.

A more recent large-scale bibliographic project proposal is LexBib by Lindemann et al. (2018) that aims to create

"a domain-specific online bibliography of lexicography and dictionary research (i.e. metalexicography) which offers hand-validated publication metadata as they are needed for citations, and which in addition is complemented with the output of an NLP toolchain." (ibid: 699)

In addition to ensuring a comprehensive coverage of lexicographic literature,⁷ the importance of this proposal lies in enabling easy citation extraction and the introduction of automatic keyword indexation (and evaluation). In the first, testing

² Accessible at <http://www.udc.es/grupos/lexicografia/bibliografia/index.html>.

³ Accessible at <http://euralex.pbworks.com/w/page/7230036/FrontPage>.

⁴ The exact number is not provided on the resource website.

⁵ <https://www.owid.de/obelex/dict/en?info>

⁶ <https://www.owid.de/obelex/meta/en>

⁷ <https://www.zotero.org/groups/1892855/lexbib/items>

phase, the authors propose including only items in English, published between 2000 and 2017.

One of the advantages of LexBib is that the focus is not only on recording bibliographical data but also on indexing full-text publications. And when talking about large collections of full-text lexicographic publication, we must definitely mention an impressive lexicographic corpus of over 5,000 lexicographic articles and books (29.2 million tokens), compiled by Gilles-Maurice de Schryver and used in studies such as Lew and de Schryver (2014) and, part of it, de Schryver (2009, 2012). The corpus is not publicly available but all efforts should be made to make further use of all the manual labour that has been put into preparing the texts of this corpus.

We can conclude that existing resources on lexicographic research are mainly focused on bibliographical aspects, especially the information about titles, authors, and keywords. Furthermore, some of the resources are no longer updated or have limited coverage. An additional problem is accessibility, as certain resources are available only in print or are private collections. One thing that none of the existing or planned resources address is the fact that scientific output is no longer limited to articles and books. It has become multimodal; there are now many video presentations on important lexicographic topics available.

3. ELEXIFINDER

The European Infrastructure for Lexicography (ELEXIS) is a project running from 2018-2022, with the aim to build a sustainable infrastructure for lexicography (cf. Krek et al., 2018). The objectives emphasized in ELEXIS are the following: the infrastructure will (1) foster cooperation and knowledge exchange between different research communities in lexicography in order to bridge the gap between lesser-resourced languages and those with advanced e-lexicographic experience; (2) establish common standards and solutions for the development of lexicographic resources; (3) develop strategies, tools and standards for extracting, structuring and linking of lexicographic resources; (4) enable access to standards, methods, lexicographic data and tools for scientific communities, industries and other stakeholders; (5) and promote an open access culture in lexicography, in line with the European Commission recommendation on access to and preservation of scientific information.

Fostering knowledge exchange in lexicography is thus one of the main objectives of ELEXIS, and improving access to lexicographic scientific output falls very much under this description. This solution will be provided in the form of a tool called Elexifinder⁸ that aims to become some sort of lexicographic Google and will not only help lexicographers in finding the relevant literature, but also allow contributions of papers or suggestions for further inclusion in the tool. The tool also addresses another objective

⁸ <http://er.elex.is>

of the project, namely promoting open access culture, as open access publications are given priority in the inclusion process.

3.1 Elexifinder architecture

Elexifinder has been built using some of the elements of the Event Registry system architecture (Leban et al., 2014; Leban et al., 2016a). Event Registry⁹ is a system used for identifying world events from news articles. Articles in different languages are collected as soon as they are detected by the Newsfeed service, then semantically enriched, and clustered to detect events (i.e. articles covering the same event). Semantic enrichment includes the identification and disambiguation of so-called concepts, which include named entities (people, places, locations) as well as non-entities or topics. Concepts are identified by wikification, “a process of entity linking that uses Wikipedia as the knowledge base” (Leban et al., 2016b).

Elexifinder uses only a portion of this system, namely the semantic enrichment (to enable various search options) and the interface. For the first version, the decision was made not to make significant modifications to the interface, as we wanted to have some content to be able to properly evaluate the usefulness of its functionalities.

3.2 Data collection and preparation

Preparing a publication for insertion into Elexifinder consists of two steps: the preparation of metatextual information and the preparation of publication content. The following metatextual information is recorded:

- Publication title
- Publication authors¹⁰
- Publication keywords (if available)
- Publication source. This can be the name of a conference or a journal, usually the year and/or the number of the issue is also included, e.g. EURALEX 2016 or Lexikos 2013-13.

⁹ <http://eventregistry.org/>

¹⁰ One of the issues that has been identified only after the launch of the first version of Elexifinder was multiple variants of authors’ names, such as John Sinclair and John McHardy Sinclair, or Danie Prinsloo, Danie J. Prinsloo and D.J. Prinsloo. This will be corrected for the second version of the tool by establishing the links between these variants and choosing one of them as the canonical form for Elexifinder. Slightly problematic will be authors that have changed surnames, e.g. Annette Klosa and Annette Kückelhaus, as both forms can actually be considered canonical.

- Publication language. ISO3 language codes are used.
- Publication URL. If the publication is not available online (e.g. in case of a book), the link points to the publisher website where the book is presented.
- Publication date. The format used is YYYY-MM-DDThh:mm:ss. For conference proceedings, the first day of the conference is used if no other date is provided. For journal issues, the last day of the issue scope is used, for example if the journal has four issues per year, the date used for the first one is at the end of the first quarter (i.e. the end of March).
- Location of the source. Recorded as a URI (Uniform Resource Identifier) of the city of the publication publisher or conference, which can be found using the Autosuggest location service by the Event Registry (<http://eventregistry.org/documentation?tab=suggLocations>).
- Location of the first author. Recorded as a URI of the city (of the affiliation) of the first author.

At the moment, the metatextual information is recorded in an Excel spreadsheet. This has the advantage of easy copying of repeating information such as publication name or location URI. It is planned to later offer an online form for individual contributions where the metadata entry would be even easier.

The second part is content preparation. Publications are usually obtained in PDF format, although if DOC(X) format is available it is preferred. The first step consists of converting the files into the TXT format, followed by checking the files and correcting any conversion errors. At this point, a copy of TXT versions – which at this point still reflect the PDF originals – is archived. This is to ensure that they can be used for any (corpus) analyses that require entire texts. The next step is the removal of content not needed for semantic enrichment: header and footer information (often repeated on every page), page numbers, publication title, author information, abstract, keywords, and references. In addition, figures, tables (and titles), footnotes and appendices are often removed, although in the case of tables that often depends on their content. For example, tables containing (only) statistics are removed but tables containing textual information are not. Similarly, footnotes containing only URLs are removed but footnotes containing remarks are not. In general, the focus is on maintaining the content that can be most informative about the topic(s) of the publication.

We have also decided to include videos of presentations with lexicographically relevant content. The same metatextual information is recorded, and the content used for semantic enrichment is an abstract or accompanying text (e.g. an abstract of the presentation at the conference).

Both metatextual information and content are then transferred into a JSON file which is needed for Elexifinder preparation.

3.3 Current contents

At the time of writing, Elexifinder included 1,755 publications and 78 videos in 11 different languages. The contents were:

- EURALEX conference proceedings from 1983 to 2016 (1,552 papers in total).
- eLex conference proceedings from 2009 to 2017 (203 papers in total)
- 21 video presentations from the eLex 2011 conference
- 33 video presentation from EURALEX 2018 conference
- 18 video presentations from various symposia in Slovenia (seven in English, 11 in Slovene)
- six video presentations from the WNLEX Workshop 2018 in Ljubljana

Importantly, all the content in Elexifinder at the moment is open access. Open access publication will continue to be prioritized, and when we start including books and monographs later we will make an appeal to publishers to publish the PDF versions of the publications, especially older ones, somewhere on their website.

3.4 Elexifinder interface

In this section we present the Elexifinder interface, including various search options available to the users. Elexifinder consists of a search window, filter line and result window. The search window offers users the option to search by keywords, either in publication title or body, or by concepts (named entities or topics). The auto-suggest functionality facilitates searching (see Figure 1). In addition, advanced search commands are supported, for example by using a – sign before a keyword one can limit the search results not to include a certain keyword (e.g. dictionary –thesaurus returns publications with a keyword “dictionary” but not including “thesaurus”).

The line with filters enables filtering by Locations (of the first author), Sources (journal or conference, and/or a specific author), Category, Time of interest (from/to specific date or period), Language, and data type (text or video). These filters can be used independently or in combination with the keywords in the search window. For example, by leaving the search window empty and using the Locations filter, one can search for all the publications coming from authors from a certain country.

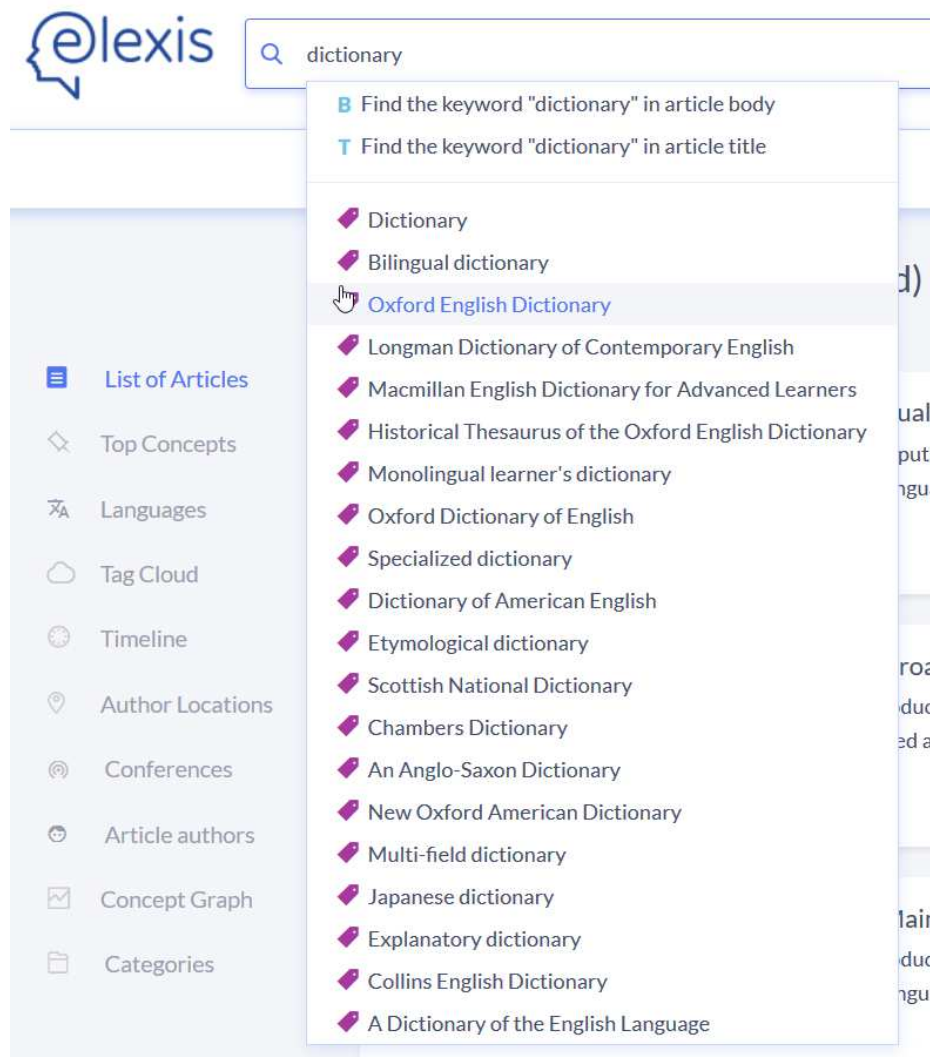


Figure 1: Auto-suggest functionality in search

The results part of the Elexifinder interface is dedicated to showing the search results. On the homepage, before any search is conducted, a map with the locations of all the authors of all the publications in Elexifinder is shown by default (Figure 2). Once any search is conducted, the results window offers a list of publications found (right-hand panel) and a left-hand menu with different visualization options. For each result in the list, the information on title, author(s), source and date of publication is provided, and in the default List view, first few lines of the text are also shown. The other types of view are Grid (each type of information is clearly named), Compact (List view but without the first few lines of text), and Details (List view + a list of most relevant semantic concepts).

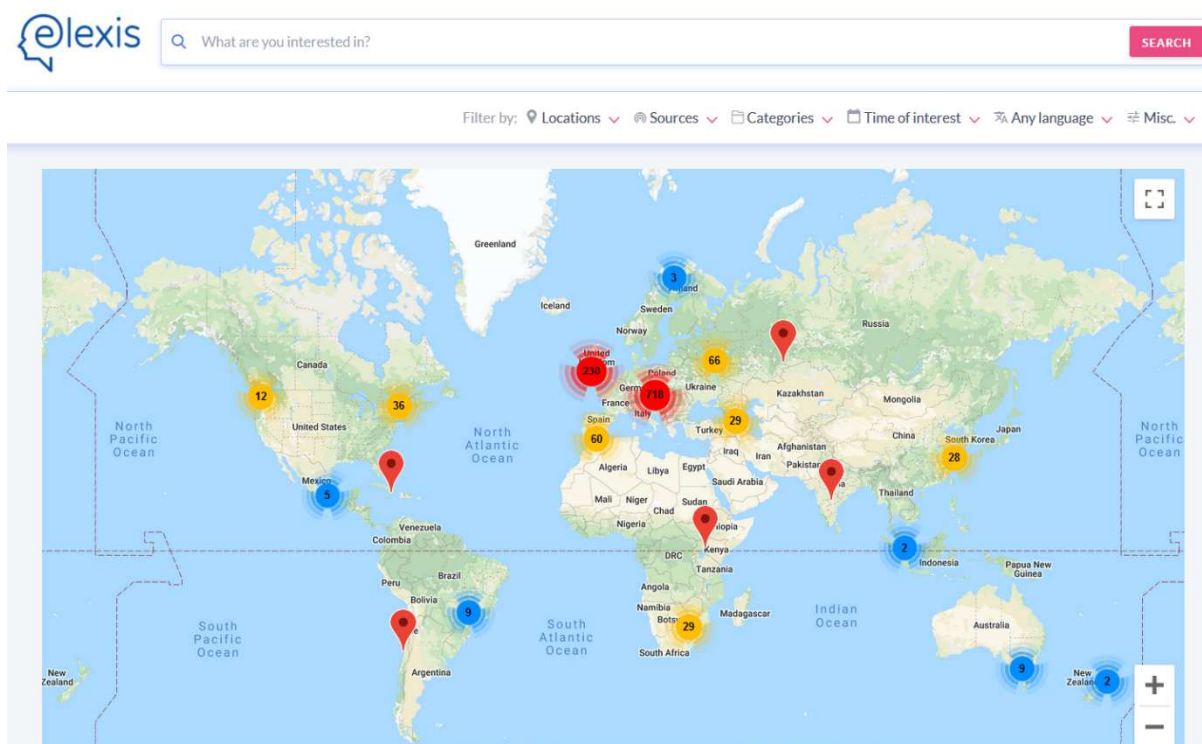


Figure 2: Homepage of Elexifinder

In the left-hand menu, the users can obtain more information on the results, and filter them (by clicking directly on maps or diagrams). Available options:

- **Top Concepts** provides a list of most relevant concepts found in the results. Concepts can be listed by relevance (default setting), frequency or uniqueness. By clicking on any concept, it is possible to limit the search further (in addition to the search condition).
- **Languages** displays a list of languages in which the publications in the results are written.
- **Tag Cloud** of top keywords in the results. Any keyword is clickable to further limit the search results.
- **Timeline** shows a distribution of results on a timeline. Daily, weekly or monthly view is available.
- **Author Locations** offers a map of locations of the first authors, and a diagram showing the number of publications per country (based on first author information).

- **Conferences**¹¹ includes three features: a diagram of all the conferences, with the number of publications; a map with locations of sources and number of publications, and a diagram showing numbers of publications per country.
- **Article Authors** offers a list of authors (not only first authors) that have authored the most publications in the results.
- **Concept Graph** shows a graph of most frequent concepts, with links between them if they exist.
- **Categories** is a visualization of automatically assigned categories and sub-categories to the results. At this moment, these are still general categories (taken from DMOZ¹²) rather than categories adapted to lexicography.

A useful feature is the option to download an image of every diagram, map or cloud shown in Elexifinder. Moreover, certain features such as Top Concepts also offer the option to download the data displayed in the diagram in the TSV format.

4. Future plans

Elexifinder was launched at the beginning of 2019, and an extensive list of publications and video recordings for further inclusion has already been prepared. This includes journals such as *Dictionaries*, *IJL*, *Lexikos*, *Lexicon*, *Lexicographica*, *Nordiske Studier i Leksikografi*, *Slovenščina 2.0* and others, and proceedings of Asialex, LexicoNordica, GLOBALEX workshops, etc. Also on the list are collective volumes, monographs and similar works. As far as videos are concerned, we aim to include video presentations from all the relevant conferences (e.g. eLex, EURALEX, Asialex) and other specific international or national events. Moreover, videos of interviews with (famous) lexicographers (e.g. the FutureLearns interview with Michael Rundell)¹³ will be included.

It is important to note that many items mentioned above will likely not be collected anew, but will be obtained from Gilles-Maurice de Schryver and the LexBib team. This will prevent the duplication of effort and enable the focus of further work on missing content, i.e. content currently not covered by any of the existing bibliographic or textual resources. This is particularly the case with publications that are not in English.

Also, special attention needs to be paid to research works in non-lexicographically dominated publications. As already mentioned in the beginning, there are many fields that produce research relevant for lexicographers, and tracking down such papers can

¹¹ This feature will be renamed after journal papers and other publications are added.

¹² <https://en.wikipedia.org/wiki/DMOZ>

¹³ <https://www.youtube.com/watch?v=5NO2YfJIXOA>

be challenging. The first obvious step would be to track down lexicographically-related special issues, but much more work is needed to identify individual papers. To improve the coverage of Elexifinder and to ensure quick updating of its database, it is envisaged that members of the lexicographic community will be able to directly contribute to the resource, either by suggesting relevant publications or videos for inclusion, or by providing the content and its metatextual information directly. Besides the obvious benefits of recently published works being immediately available to the community, there are benefits for editors, reviewers and other people involved in publication preparation as they will be able to search for any related publications of the same author(s) with the same or similar content.

In addition to enhancing the Elexifinder database with new content, improvements of the frontend are planned. The first part of the improvements is connected to searching. This includes cross-lingual searching, which would enable automatic translation of search terms into all other languages of publications found in Elexifinder. Such a feature is already part of the Event Registry system and its identification of events, so the aim is to adapt it to the needs of Elexifinder.

Partly linked to cross-lingual searching is the introduction of a new and more lexicographically-oriented categorization of publications, or freshly devised “ontology for lexicography”, which would replace the existing DMOZ-based categorization. For this, we will work together with the LexBib team on keyword indexation (and evaluation) in order to devise a common solution. Also, existing ontologies such as the META-SHARE ontology (McCrae et al., 2015) will be used as a starting point.

Other improvements will be done on the homepage, where we plan to include additional content that would help promote lexicography and attract users to the website more regularly. Such content includes the list of most searched/clicked publications or videos, alerting users to the most recent inclusions, presenting a list of publications (in different languages) on a certain topic of interest, listing conference and journal calls, etc.

In sum, Elexifinder will become an integral part of ELEXIS infrastructure that will be complementary to other resources and tools developed within ELEXIS, and continuously improved long-term with the help of the lexicographic community.

5. Acknowledgements

The research received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 731015.

6. References

- Ahumada, I. (2016). Metalexicografía del español: clasificación orgánica y tipología de los diccionarios en el Diccionario Bibliográfico de la Metalexicografía del Español (DBME). In *Anuario de estudios filológicos*, (39), pp. 5–24.
- De Schryver, G.-M. (2009). Bibliometrics in Lexicography, *International Journal of Lexicography*, (22,4), pp. 423–465.
- De Schryver, G.-M. (2012). Trends in Twenty-Five Years of Academic Lexicography. *International Journal of Lexicography*, (25,4), pp. 464–506.
- Euralex: International Bibliography of Lexicography. Accessed at: <http://euralex.pbworks.com>. Date of access: 15th May 2019.
- Hartmann, R.R.K. (2007). *Bibliography of Lexicography*. Accessed at: <http://euralex.pbworks.com>. Date of access: 15th May 2019.
- Krek, S., Kosem, I., McCrae, J. P., Navigli, R., Pedersen, B. S., Tiberius, C. & Wissik, T. (2018). European Lexicographic Infrastructure (ELEXIS). In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (eds.) *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*. Ljubljana: Ljubljana University Press, Faculty of Arts, pp. 881–891. Available at: <http://euralex.org/wp-content/themes/euralex/proceedings/Euralex%202018/118-4-2986-1-10-20180820.pdf>.
- Leban, G., Fortuna, B. & Grobelnik, M. (2016a). Event Extraction from Media Texts. In C. Sammut & G. I. Webb (eds.) *Encyclopedia of Machine Learning and Data Mining*. Boston: Springer, pp. 1–7. Available at: https://link.springer.com/content/pdf/10.1007%2F978-1-4899-7502-7_901-1.pdf.
- Leban, G., Fortuna, B. & Grobelnik, M. (2016b). Using news articles for real-time cross-lingual event detection and filtering. *First International Workshop on Recent Trends in News Information Retrieval (NewsIR'16), Padova, Italy*. Available at: <https://pdfs.semanticscholar.org/f917/c0cff24fed1af45f94c53b74ca0229874966.pdf>.
- Leban, G., Fortuna, B., Brank, J. & Grobelnik, M. (2014). Event Registry: Learning About World Events from News. *Proceedings of the 23rd International Conference on World Wide Web*, pp. 107–110.
- Lew, R. & de Schryver, G.-M. (2014). Dictionary Users in the Digital Revolution, *International Journal of Lexicography*, 27(4), pp. 341–359. Available at: <https://doi.org/10.1093/ijl/ecu011>.
- Lindemann, D., Kliche, F. & Heid, U. (2018). LexBib: A Corpus and Bibliography of Metalexicographical Publications. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (eds.) *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*. Ljubljana: Ljubljana University Press, Faculty of Arts, pp. 699–711. Available at: <http://euralex.org/wp-content/themes/euralex/proceedings/Euralex%202018/118-4-2929-1-10-20180820.pdf>.

- McCrae, J., Labropoulou, P., Gracia, J., Villegas, M., Rodríguez-Doncel, V., Cimiano, P. (2015). One ontology to bind them all: The META-SHARE OWL ontology for the interoperability of linguistic datasets on the Web. In: *Proceedings of 12th Extended Semantic Web Conference (ESWC 2015)*. Portorož, Slovenia. DOI: 10.13140/RG.2.1.3233.6244
- Möhrs, C. & Töpel, A. (2011). The "Online Bibliography of Electronic Lexicography" (OBELEX). In: I. Kosem & K. Kosem (eds.) *Electronic lexicography in the 21st century: New applications for new users. Proceedings of eLex 2011, Bled, 10 - 12 November 2011*. Ljubljana: Trojina, Institute for Applied Slovene Studies, pp. 199-202.
Available at: <http://www.trojina.si/elex2011/Vsebine/proceedings/eLex2011-26.pdf>.
- Möhrs, C. (2016). Online Bibliography of Electronic Lexicography. The Project OBELEXmeta. In T. Margalitadze & G. Meladze (eds.) *Proceedings of the 17th EURALEX International Congress: Lexicography and Linguistic Diversity. Tbilisi, Georgia 6-10 September 2016*, pp. 906-909. Available at: http://euralex.org/wp-content/themes/euralex/proceedings/Euralex2016/euralex_2016_100_p906.pdf.
- Wiegand, H.E. (2012). *Internationale Bibliographie zur germanistischen Lexikographie und Wörterbuchforschung. Mit Berücksichtigung anglistischer, nordistischer, romanistischer, slavistischer und weiterer metalexikographischer Forschungen*. Berlin, Boston: De Gruyter Mouton.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



Lexicographic Practices in Europe: Results of the ELEXIS Survey on User Needs

Jelena Kallas¹, Svetla Koeva²,

Margit Langemets¹, Carole Tiberius³, Iztok Kosem⁴

¹ Institute of the Estonian Language, jelena.kallas@eki.ee, margit.langemets@eki.ee

² Institute for Bulgarian Language, Bulgarian Academy of Sciences, svetla@dcl.bas.bg

³ The Dutch Language Institute, carole.tiberius@ivdnt.org

⁴ Jožef Stefan Institute, iztok.kosem@ijs.si

Abstract

The paper presents the results of a survey on lexicographic practices and lexicographers' needs across Europe (and beyond) both for born-digital and retrodigitized resources. The survey was conducted during the period from 11 July to 1 October 2018 in the context of the Horizon 2020 project ELEXIS (European Lexicographic Infrastructure). The survey was completed by 159 respondents from a total of 45 countries, comprising 36 European countries and nine countries outside Europe.

Looking in detail at the results of the survey, the paper focusses on determining what constitutes a job description of a modern lexicographer, including the training needed. One of more notable findings is that lexicographic training is still in most cases provided by the employer rather than obtained through formal education programmes. Furthermore, a list of various dictionary-writing systems and corpus-query systems is provided, including their features currently most often used by lexicographers. Accompanying this is information about the features lexicographer want or need in their tools. Also, the paper offers insights into current trends in lexicography and what lexicographers see as the most important emerging trends that will affect lexicography in the future. Overall, these results provide a detailed insight into what is needed in terms of tools and training and thus feed back into the ELEXIS project and will help to fine-tune resources within ELEXIS.

Keywords: e-lexicography; lexicographers' needs; survey; lexicographic practices

1. Introduction

In lexicography, there is a lot of research available on methods of dictionary compilation, dictionaries, and dictionary users and their needs. On the other hand, until recently at least, there has been little literature on lexicographers, their practices and needs. This has become even more important in the age of significant changes in lexicography, brought about by technological progress and the move from the print medium to the digital one. The need to bring lexicographers from different countries together in order to tackle the challenges of modern-day dictionary making has thus become even greater.

The first steps towards addressing this issue were made in the European Network of e-

Lexicography (ENeL)¹, a COST Action funded by the European Union that brought together nearly 300 lexicographers from 30 different countries. Other than enabling the exchange of knowledge and expertise, ENeL produced highly valuable results such as various surveys among lexicographers and their institutions and a European survey on dictionary use (Kosem et al., 2019), the largest dictionary user survey to date.

One of the most important outcomes of ENeL has been ELEXIS (European Lexicographic Infrastructure)², a Horizon 2020 infrastructure project dedicated to lexicography. This new infrastructure aims to enable efficient access to high quality lexicographic data, and to bridge the gap between more advanced and less-resourced scholarly communities working on lexicographic resources. ELEXIS activities have used the results of ENeL, however further research was needed to obtain a detailed insight into current lexicographic practices and the needs of lexicographers. Consequently, two surveys have been carried out within ELEXIS focussing on various aspects of the lexicographic workflow such as software and tools, publication, retrodigitization, metadata and data formats. The first survey was targeted specifically at individual lexicographers. The second survey focussed on institutions and targeted senior lexicographers and IT specialists from eleven ELEXIS lexicographic partner institutions. The survey for institutions is not part of this paper (the results are presented in detail in ELEXIS Deliverable 1.1, see Kallas et al., 2019), however we include relevant findings from the survey when appropriate.

In this paper we initially provide the background of the survey. Then, in Section 3, we introduce the general principles, aims, structure and the implementation of the survey, followed by the presentation of the results in Section 4. The last two sections are dedicated to the discussion on the implications of the survey findings for lexicographic practices and ELEXIS efforts, and conclusions about the overall value of the survey and future plans in the ELEXIS project.

2. Background

The information on lexicographic practices has often been generalized based on selected project(s) or researchers' experience (cf. Hartmann, 2003; Atkins & Rundell, 2008; Klosa 2013). While such works are very important for lexicography and manage to show the state-of-the-art of the discipline, they do not point out the differences and similarities between lexicographic practices in different countries or even at different institutions.

One of the main projects/initiatives that helped gather a great deal of information on lexicographic practices across Europe, and thus fill some of these information gaps, was

¹ <http://www.elexicography.eu/> (1 July 2019)

² <https://elex.is> (1 July 2019)

the COST action European Network of e-Lexicography (ENeL). Within the Action, a number of surveys have been carried out, and we summarise the most relevant here.

The first survey, conducted in 2014, focused on the workflow of corpus-based lexicography. Six general monolingual dictionaries, one bilingual dictionary, and seven specialized dictionaries and databases were covered. All 14 resources were published online, one also in print (at the time). The main findings stated in the report (Tiberius & Krek, 2014) were that the role of computers in lexicography is continuously increasing, but the compilation of dictionaries is still a highly labour-intensive task. Most projects followed to a certain extent the phases of the lexicographical process proposed by Klosa (2013), with the analysis phase taking by far the most time. Lack of IT support was one of the problems mentioned by the majority of projects. Some attention was also paid to user involvement, with the main finding being that users need to be involved in the later stages of a lexicographical project (afterlife, etc.); crowdsourcing was mentioned as one option of earlier involvement, but it was concluded that more research (and a separate survey) was needed on this subject.

The second survey, conducted in 2014/2015, focused on Dictionary Writing Systems (DWS) and Corpus Query Systems (CQS) (Krek et al., 2014). It consisted of 94 questions and was completed by 69 lexicographers and computational experts (computational linguists, software developers, etc.) from 35 different institutions in 25 different countries. The part of the report dealing with DWSs showed that 10 institutions used off-the-shelf products, 12 institutions developed their own software, whereas 16 institutions used customized software (XML editors, databases, wikis, etc.). In terms of functionality, most DWSs supported validation and consistency checking, and offered the use of templates for common dictionary structures. On the other hand, many DWSs did not include a spellchecker or integration with a CQS. As far as CQSS were concerned, 65% of the respondents reported that their institutions used them – by far the most widely used was (no)Sketch Engine³ (11 institutions), followed by IMS Open Corpus WorkBench⁴ (4). The evident trend was that open-source and commercial CQS at the time met the needs of lexicographic projects, while this was not the case for DWS considering the share of institutions that developed or had been developing in-house solutions.

The third survey, also conducted in 2014/2015, dealt with automatic knowledge acquisition for lexicography (Tiberius et al., 2015). It consisted of 134 questions and was completed by 51 respondents (lexicographers, software developers, computational linguists, etc.) from 20 different countries. Thirteen different types of lexicographic data were proposed on the list of data types that could be automatically acquired from a CQS. The results revealed that more commonly extracted types of lexicographic data

³ <https://www.sketchengine.eu/nosketch-engine/> (1 June 2019)

⁴ <http://cwb.sourceforge.net/> (1 June 2019)

were lemma lists, frequency information, example sentences, grammatical patterns, and multiword expressions (ranging from collocations to idioms), while other types such as form variants, neologisms, translation equivalents, lexical semantic relations, word senses, linguistic labels, definitions and Knowledge-Rich Contexts⁵ were automatically extracted by only a few institutions. Given the state-of-the-art of lexicography at the time, it is slightly surprising that, for example, translation equivalents, lexical semantic relations (e.g. synonyms) and linguistic labels were automatically extracted only by a few institutions. This finding is particularly interesting in the case of translation equivalents and lexical semantic relations, as they were reported, after lemma lists, frequency information and example sentences, to be among the types of data that were integrated in the published dictionaries without intervention.

The aforementioned three surveys (Tiberius & Krek, 2014; Krek et al., 2014; Tiberius et al., 2015) provided a great deal of insight into lexicographic practices around Europe. Still, in some cases requesting more elaborate answers from the respondents should perhaps have given better results. Moreover, it would be better if all the surveys were conflated into one so that a more general picture per institution or respondent could be obtained, and that the questions could be connected. Finally, the number of institutions, and to a lesser extent countries, could be greater.

The survey conducted in the ELEXIS project and presented in this paper aimed to address some of these shortcomings. Also, due to rapid changes in lexicography and related disciplines an update to this overview of lexicographic practices was very much needed. One aspect that we wanted to add to this overview was the education and training of lexicographers, and their needs related to this.

3. Survey of Lexicographers' Needs

3.1 General principles and aims

The main aim of the survey was to get a good overview of lexicographic practices across Europe both for born-digital and retrodigitized resources, different tools and methods used by lexicographers around Europe, as well as the needs that they have now or anticipate to have in the short-term and long-term future. However, the survey was also disseminated outside Europe, as we were also interested in lexicographic practices around the world. In order to get as many responses as possible, we limited the length of the survey.

Many different channels were used for disseminating the survey, e.g. international and national mailing lists, social networks (e.g. ELEXIS Facebook and Twitter profiles),

⁵ In terminography, a sort of hybrid of a good example and a definition, illustrating the meaning characteristics of a term, but not being a formal definition.

group or individual emails, a booth at the EURALEX conference, etc. It was important to get a good coverage of countries to enable comparisons, and more importantly, to help us in preparing more targeted activities with the ELEXIS project, such as training workshops and materials, and help to fine-tune resources developed within the project.

Equally important was the attempt to get several respondents from the same country, in terms of institution, age, role in the team, dictionary project, etc. to ensure that the data would be representative of a country and not of a single institution, generation, project and so forth. As a result we managed to obtain answers from a rather heterogeneous group of respondents in terms of their experience, employment status, projects they are involved in (types of dictionaries, language etc.), and the country in which they are based (see Section 4.1). This to some extent ensures that the results can be generalized to the lexicographic community as a whole.

3.2 Structure and implementation

The method chosen for the survey was an online questionnaire. Several survey tools were considered, and in the end Google Forms⁶ was chosen as it is simple to use and manage, and it covered the majority of our needs. The survey was publicly announced on various mailing lists on 13 July 2018 and was closed on 1 October 2018.

The survey⁷ contained 44 questions divided into six sections, i.e. (1) General information; (2) Ongoing work; (3) Software and tools; (4) Publication; (5) Retrodigitization; (6) Past and future. There were three different types of questions used in the survey: (1) "yes/no" questions, (2) multiple choice questions, and (3) open-ended questions. Not all questions were obligatory.

4. Results

4.1 Respondents' background and projects

The survey was completed by 159⁸ respondents from a total of 45 countries, comprising of 36 European countries (140 respondents) and nine countries outside Europe (19 respondents). We decided to categorise under European countries those nations with close cultural ties to Europe (and inclusive status in EU-funded initiatives such as

⁶ <https://www.google.com/forms/about/> (1 July 2019)

⁷ The survey for institutions was more detailed, containing 86 questions but divided into the same six sections. Both questionnaires can be found in the appendix of ELEXIS Deliverable 1.1, see Kallas et al., 2019). Because of data privacy issues the raw data cannot be shared.

⁸ As some questions were optional, not all questions were answered by each respondent. For this reason, we provide the number of responses for each question (i.e. N = number_of_responses) in the results.

COST Actions) and with active partners in the ELEXIS consortium.

Figure 1 illustrates that the majority of the respondents work as full-time or part-time in-house employees and less than one quarter as freelancers (mainly the respondents from Northern Europe and the USA).

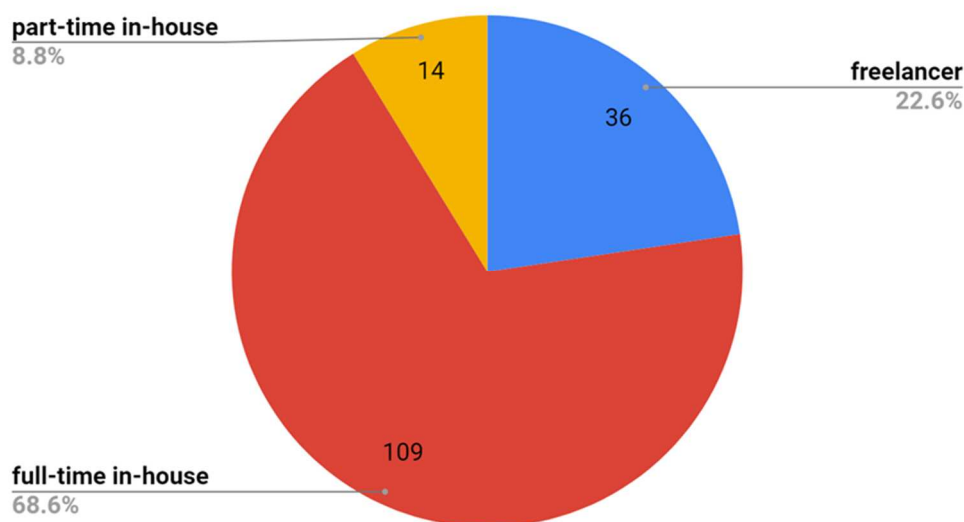


Figure 1: Employment (N=159)

A total of 77.9% of in-house lexicographers (123 respondents) work at public institutions or non-governmental organizations, while 17.2% at a university. There were only a small number of respondents (4.9%) working for private/commercial companies in Europe.

The respondents were thus quite representative of European (monolingual) dictionary-making community, considering that in the European survey on dictionary use and culture (Kosem et al., 2019: 96) it was reported that in the majority of the countries participating in the survey monolingual dictionaries are published solely or mainly by public institutions funded by the government.

A total of 58.5% of the respondents were involved in compilation of monolingual dictionaries or databases, either general, specific or dictionaries for learners. Much fewer respondents were involved in compiling bilingual (15.1%), multilingual (13.2%) and dialectal (8.8%) dictionaries or databases.

Sixty-one percent of the respondents have a PhD and the majority have an MA or BA degree in language/linguistics (81.1%). More than one third of respondents (35.8%) have more than 20 years of work experience in the field of lexicography, 24.5% have 10-20 years of work experience and roughly every fifth respondent (20.1%) has 5-10 years of work experience. These responses may be an indication that people who start

working as lexicographers stay in the field for a long time.

More than one third of the respondents (34.6%) have been trained within their own institute, usually by a tutor or senior lexicographer. Roughly every fourth respondent (25.8%) has attended special courses or several courses since starting working in lexicography. Other forms of training attended by the respondents were workshops or summer schools. Only a small number (11.3%) of the respondents reported studying lexicography at university, either as part of an MA course on lexicography or as a special course.

The respondents reported working in teams of different sizes, with relatively similar shares being reported across all team sizes. Overall, the majority of our respondents work in teams consisting of under 10 members, and the predominant team size was 3-6 people (27.4%). More than half of the respondents (56.6%) reported working in a team that consisted only of people from their own institution, and 43.4% reported working together with people outside their institution.

A total of 122 different projects were mentioned by the 158 respondents. Fifty-three of these are permanent projects; these are mainly voluminous monolingual contemporary dictionaries, Wiktionaries, etymological and dialectal dictionaries, as well as a few bilingual dictionaries. Another 18 projects have a duration of 15-20 years; these are also mainly voluminous monolingual contemporary dictionaries, etymological and dialectal dictionaries, as well as bilingual dictionaries.

150 respondents answered the question on dictionary publication format. Out of the 122 reported dictionary projects, 100 (82%) would be published online – 55 of them online only, 45 also in print. For four projects, the respondents reported that dictionaries would also be available as apps. Only 24 projects out of the 122 mentioned in the survey would appear in print only. A reason for publishing in print is tradition; the dictionary is part of a larger project and previous volumes have appeared in print. These results are also in line with what was reported by Kosem et al. (2019) on the status of lexicography (types of dictionaries being compiled and their format) in the 26 countries involved in their study.

The majority of the project databases on which the respondents are working are organized from word to meaning (word-based databases, 87.3%). Databases organized from meaning to word (concept-based, 8.9%) are used mainly in terminological projects. There is also a small number of projects (3.2%) that combine both word-based and concept-based organization of the database.

4.2 Software and tools

Eighty-nine out of 159 respondents answered the question about software and tools. More than half of them reported that they use both a DWS and a CQS in their work.

Altogether 15 DWSs and 22 CQSs were mentioned by the respondents. The tools can be divided into three main categories: commercial, open-source and in-house. General purpose editors, dictionary publishing platforms and App Builders were considered as a separate category.

There are mainly three types of DWS+CWS combinations used by the lexicographers:

- 1) in-house DWS and commercial CQS (e.g. Ekilex⁹ and Sketch Engine¹⁰)
- 2) commercial DWS and commercial CQS (e.g. IDM¹¹ and Sketch Engine)
- 3) in-house DWS and in-house CQS (e.g. LexDF¹² and IMS Open Corpus Workbench).

The first combination listed is also the most common model. Altogether 54.8% of the respondents reported using Sketch Engine as CQS, other CQSs¹³ used were, for example, IMS Open Corpus Workbench, CoRes¹⁴, Korp¹⁵, NoSketchEngine, AntConc¹⁶, and COSMAS II¹⁷. Generally, the lexicographers in our survey reported using one CQS and one DWS, but some respondents use several DWSs (e.g. iLex¹⁸, Lexonomy¹⁹ and TLex²⁰) and several CQSs (mostly Sketch Engine in combination with other CQSs such as KonText²¹, Lexpan²² or Korp) at the same time. The following reasons were given for using more than one system: (a) moving from commercial or in-house to open-source; (b) different project needs or needs of lexicographers, e.g. one system is more suitable for retrodigitized dictionaries, another one for born-digital dictionaries; one for word-based, another for concept-based lexicography, etc.

⁹ <https://ekilex.eki.ee> (1 July 2019)

¹⁰ <https://www.sketchengine.eu/> (1 July 2019)

¹¹ <http://dps.cw.idm.fr/index.html> (1 July 2019)

¹² The product is not publicized, but registered with Inven2, The UiO patent and IPR organization, since 2014.

¹³ For the full list see Kallas et al. 2019.

¹⁴ <https://korpus.dsl.dk/corest/index.htm> (1 July 2019)

¹⁵ <https://spraakbanken.gu.se/eng/korp> (1 July 2019)

¹⁶ <http://www.laurenceanthony.net/software/antconc/> (1 July 2019)

¹⁷ <https://www.ids-mannheim.de/cosmas2/> (1 July 2019)

¹⁸ https://issuu.com/jens.erlandsen/docs/ilex_brochure_120dpi (1 July 2019)

¹⁹ <https://lexonomy.eu/> (1 July 2019)

²⁰ <https://tshwanedje.com/tshwanelex/> (1 July 2019)

²¹ https://kontext.korpus.cz/first_form?corpname=syn2015 (1 July 2019)

²² <http://www1.ids-mannheim.de/lexik/uwv/lexpan.html> (1 July 2019)

Relevant for these results are also the findings of the survey for institutions. All but one institution participating in this survey use one or more DWSs and it is still quite common²³ for the institutions to develop an in-house system (five institutions indicated that they use an in-house DWS). It is also not uncommon for the institutions to use more than one DWS because of different project needs. About half of the partner institutions indicated that they did make some adaptations/customizations to an off-the-shelf DWS to make it more suitable for their project(s). The following customizations were mentioned: customization of schemas, DTDs and menus; customization of view options (e.g. for getting an overview of the entry); customization of search and extraction options. All but two institutions use one or more CQSs, often combining a commercial system with an in-house or open-source system.

4.3 Compiling methods and automatic knowledge extraction

All respondents answered the question about compilation methods. As shown in Figure 2, the majority of the respondents reported compiling their dictionaries manually (57.9%).

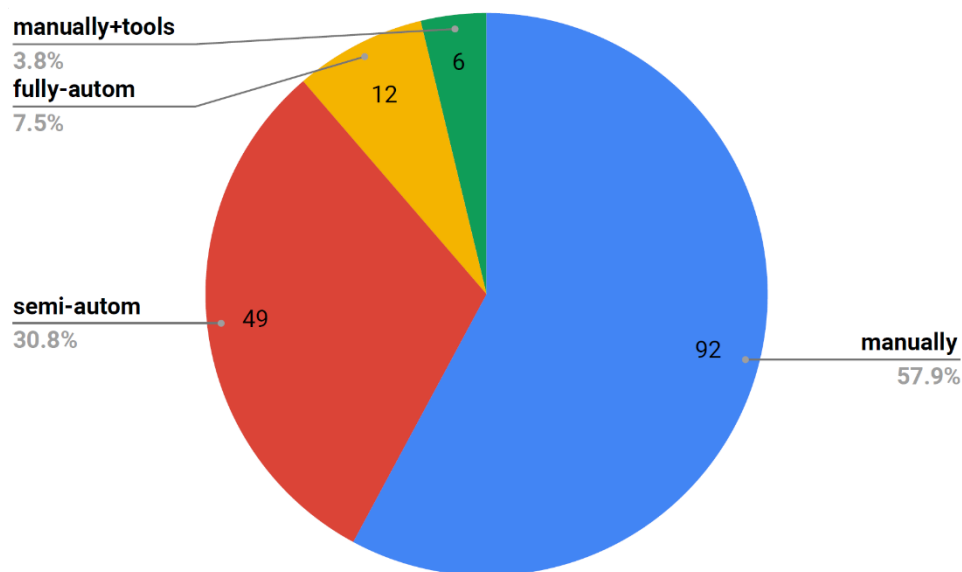


Figure 2: Compiling methods for all projects (N=159)

Based on other answers provided, the respondents perceive the manual method as analysing the data (often by using CQS) and then inserting the information into their DWS or some other tool manually. Nearly one third of the respondents (30.8%) work with semi-automatically collected data, and only 7.5% (12 respondents) are using fully-automatically collected data. It is interesting that the respondents who marked their

²³ As was the case in the COST ENeL survey (Tiberius & Krek, 2014).

project to be born-digital²⁴ (N=65) mentioned using different compiling methods: mainly semi-automatic (43.1%) and manual (!) (33.8%). The most common types of data for which the respondents reported using automatic extraction methods include headword list (20.8%), collocations (12.7%) and frequency information (11.3%). Automatic extraction of multi-word expressions (8%), dictionary examples (7.5%) and form variants (6.1%) is fairly common, too. Less than 5% of the respondents, respectively, reported using automatic extraction for patterns (4.7%), neologisms (3.8%), lexical-semantic relations (3.8 %), domain information (4.4%), multilingual data from parallel/comparable corpora (3.8%), definitions (3.3%) and audio data from speech corpora (2.4%).

4.4 Various aspects of the lexicographer's job

In this part we have chosen four different aspects of the lexicographer's job that feed back into the ELEXIS projects in terms of training, IT support, user involvement and tools for retrodigitization. Namely, all these aspects are important in state-of-the-art lexicography as the job of a lexicographer has changed – most lexicographers do not just edit dictionary entries anymore, but are also involved in various other aspects of dictionary-making.

4.4.1 Involvement of lexicographers in the online publication process and user

research

Lexicographers were asked to specify what kind of work they are doing when they are involved in online publication or user research. It was an open-ended question, but three options were proposed: 1. evaluating the user interface and providing new ideas; 2. creating add-on materials (e.g. blogs, slideshows, videos, quizzes, word games); 3. communicating with IT persons / user experience designer (UX) / interface designer (IX).

Just over a quarter (27%) of 63 respondents answered that they are not involved in online publication, while the 33.9% who were involved in online publication dealt with user interface evaluation, and communication with IT specialists, including user experience designers and interface designers. In addition to user interface evaluation

²⁴ Furthermore, terms such as “born-digital” and “IT support” seem to have been interpreted in different ways by different respondents, even although a definition of “born-digital” was provided. For example, the share of respondents who answered the question whether they work on born-digital dictionaries affirmatively was unusually high, especially considering the information they provided for related questions about the types of projects, compilation methods and the format of publication, which suggest a different interpretation of the term “born-digital”. This experience shows not only that all terms should be defined in future surveys, but also that there is a need for a discussion of the term in the lexicographic community, something that the ELEXIS project should also pay attention to.

and communication with IT specialists, 16.9% of the respondents were involved in the production of add-on materials. Just 11.9% are involved only in user interface evaluation and 8.5% only in IT communication. Other tasks mentioned include project management, updating user guides, organizing and testing new editions (or updates of existing editions), working on promotional activities (e.g. media interviews, presentations, Word of the Day), analysis of user feedback, answering user questions, etc.

The respondents were also asked if they are involved in user research for their projects, and if so what kind of user research they do. The options proposed were, for example, analysing user logs or interviewing end users. Just under two thirds (62.5%, or 55 out of 120 respondents) revealed that they are not involved in user research. A total of 59% of those lexicographers who do user research conduct analyses of user logs, 33.2% also conduct interviews with end users (mostly before and during the conceptual phase of the dictionary). Other tasks mentioned include the analysis of data from language-related advisory services and Google Analytics, the analysis of user feedback, mostly proposals and corrections (the feedback is gathered through mail or online feedback forms), conceiving and supervising user studies carried out by others, and informal consultation.

4.4.2 IT support

As expected, IT support is an important part of lexicographer's job. Over 80% of the respondents answered this question and reported to have either basic (43.9%) or good (37.8%) IT support. We did not look into the dynamics between lexicographers and IT staff in more detail in this survey, but it definitely deserves more attention, particularly the way in which IT staff are perceived by lexicographers, and whether there are differences in the way the lexicographers perceive IT staff or computational linguists or NLP experts. IT tasks are also the only tasks that seem to be outsourced in dictionary projects, ranging from designing the online interface of the dictionary to developing and/or offering support in the use of DWS or CQS. Trustworthy experts, efficiency and another view of the data and content (which might help to identify some lexicographic problems) were mentioned as positive experiences. The cost (too expensive, lack of (regular) funding), more additional work (to teach and explain lexicographic details), delays and communication problems were mentioned as negative experiences when outsourcing.

4.4.3 Crowdsourcing, gamification and data enrichment

The results of the survey show that crowdsourcing and gamification are not yet common practices in the lexicographic projects that the respondents are involved in. Nonetheless, the wish for tools for crowdsourcing was put down by several respondents in the survey. These results are not that surprising, as crowdsourcing has become a hot topic in

lexicography only in the last five years, so it is understandable that many projects (and lexicographers) are still cautious about using the wisdom of the crowd.

Of particular interest are the results of the question related to data enrichment (i.e. adding additional linguistic and non-linguistic information to the data, such as normalizing values, geo-locating, expanding content, etc) which not only concerns retrodigitized dictionaries, but also born-digital dictionaries which can be enriched with various types of information. Different forms of data enrichment were mentioned in the context of retrodigitization by the respondents, e.g. text normalization, expanding abbreviations, adding grammatical information as well as adding internal and external links. Relatedly, the survey for institutions showed that data enrichment is not yet very common in current lexicographic projects within the ELEXIS consortium. Only two institutions indicated that they include images and/or videos in their dictionaries.

4.4.4 Retrodigitization

As retrodigitization of older/printed dictionaries (i.e. the process of converting a dictionary published in paper into a digital, computer-readable format, which involves not only scanning and OCRing but also data encoding and enrichment) is an emerging trend in modern e-lexicography, we asked the respondents about their involvement in different phases of the retrodigitization process. The aim was to get an overview of the software used in this process and to collect lexicographers' opinions on which dictionaries should be retrodigitized. The number of the respondents, 16, that answered these questions was rather low. This may be due to the fact that some parts of the retrodigitizing activities (image and text capturing) are not directly related to the lexicographic work. If we look at the individual phases of retrodigitization, we see that the 16 respondents reported to be mainly involved in the activities which require lexicographic competence, such as data encoding (15 responses) and data enrichment (13 responses).

5. Discussion

5.1 Lexicographer's training and job description

The information on the experience and training of the respondents of the survey points to another potential issue in lexicography. Although the respondents reported having quite a lot of experience in lexicography, they all had to be trained by their employer; very few of them actually had a formal education in lexicography. Consequently, dictionary-makers have to be prepared for extra costs related to training of their staff, and need to plan projects accordingly.

This situation makes degree programmes such as EMLex (European Master in

Lexicography)²⁵ very important for the training of young generations of lexicographers, and the development of the field in general. At the same time, it is essential that lexicographers are provided with different types of quality training materials, something that ELEXIS is also dedicated to provide as part of Work Package 5.

Education and training of lexicographers will need to become more and more interdisciplinary, as the findings of the survey indicate that a lexicographer's job is far from being monotonous. Modern lexicographers need to possess much more than just linguistic skills; other skills in their repertoire need to include project management, communication with computational staff, promotional activities, responding to user questions and feedback, etc.

It is noteworthy that most lexicographers are not involved in the final dictionary publication (of an online dictionary) or user research. The former finding is to be expected, as normally this job is left to web/interface designers; however, one cannot help wonder whether dictionaries are really better for it. On the other hand, the lack of at least some involvement into user research is worrying, especially considering the current lack of user research in most European countries (and around the world) (Kosem et al., 2019: 96). Knowing the users is important; as Atkins and Rundell (2008: 5) rightly point out, “the content and design of every aspect of a dictionary must, centrally, take account of who the users will be and what they will use the dictionary for”. Part of the solution might be in conducting regular European- or world-wide surveys, such as Kosem et al. (2019) and Müller-Spitzer (2014), as this brings lexicographers together and also promotes the discipline (and dictionaries) among the general public.

5.2 Existing tools and lexicographers' wishes for the future

As shown in Sections 4.2 and 4.3, the lexicographers use a wide variety of DWSs and CQSs, in different combinations. Moreover, they often use more than one DWS or CQS, mostly because of the needs of specific dictionary projects. The finding that an in-house solution is the predominant form of DWS used is in line with the findings of the ENeL survey (Krek et al. 2014). It thus seems that existing off-the-shelf DWS often still do not meet (all) the needs of lexicographic projects.

As the development of new open-source tools is an important part of the ELEXIS project, it was also important to learn about the respondents' wishes regarding DWS and CQS, in other words what would be their ideal tool. The majority of the respondents mentioned that their ideal DWS should be free, online, open-source, browser independent, fast, intuitive, and easy to maintain. This supports the view of ELEXIS and reaffirms our aims to develop online open-source tools such as Lexonomy.

²⁵ <https://www.emlex.phil.fau.eu/> (1 July 2019)

Other features, such as supporting real-time collaborative input, real-time saving, localization, customizability both in terms of functionalities and interface, online publishing of the results, and proper documentation (i.e. it should not be a black-box system) were also listed. While many existing DWSs already have most of these features listed by the respondents, it seems that all of them are mandatory as far as lexicographers are concerned.

The respondents also believe it is important that their DWS is interoperable with other resources, operating systems and tools. Thus, API and script support is expected. This was mentioned both in connection with the possibility of automatic pre-compilation of entries and the possibility to integrate lexicographic information automatically from CQS into DWS. Similar findings were observed in the survey for institutions, where most partner institutions felt that the integration of DWS and CQS would be beneficial, especially for the linking, selection and retrieval of examples, collocations, etc. Again, this is something that the ELEXIS project is working on addressing as, at the time of writing this paper, the beta version of the Sketch Engine pull feature in Lexonomy was already available. The feature enables quick search and import of examples, collocations, synonyms, and even definition candidates (for some corpora) from Sketch Engine into Lexonomy.

In terms of CQS, the answers from the institutional survey are also important to mention here, as the respondents listed some features that they missed in existing CQSs, such as sense clustering (clustering concordances against senses)²⁶, implementation of syntactic and semantic annotation, detection of neologisms, automatic acquisition of translation equivalents, diachronic analysis, etc. The topicality of these features is also evidenced by the fact that the ELEXIS project contains various activities focussed on these.

Relatedly, several respondents also pointed out the need for better tools for retrodigitization. Such tools include automatic processes where the quality of output highly influences the amount of manual labour needed to prepare the digital version of the dictionary, for whatever purpose it is then used.

5.3 Current trends and looking ahead

It can be said that automatic knowledge extraction in lexicography is definitely on the increase, and the findings of this survey are very much similar to the findings of the ENeL survey in 2014-2015 (Krek et al., 2014). Also, headword lists, frequency information and multiword expressions (collocations in particular) are still the most commonly extracted types of information. Less common automatic extraction of

²⁶ The respondents might have been influenced by the formulation of the question, as this was one of the suggestions listed to help them understand the question.

information that is more semantically-based, such as senses, definitions, lexical relations, etc., can be attributed to the fact that lexicographers do not seem to think that existing tools already perform these tasks satisfactorily enough; this is evidenced in the respondents' answers to the question on their needs in the next 10-15 years, where the most mentioned topic was the need for better tools for extraction and automatic processing of data from corpora.

Moreover, lexicographers seem to be well aware of the potential of the Semantic Web, Linked Open Data, and Artificial Intelligence for lexicographic purposes.

One of the things that the respondents reported had improved was the interaction between the users and the dictionary, since users can now directly contact lexicographers online about words they are looking for, technical issues, etc. At the same time, the respondents called for more and better tools to analyse user behaviour. Considering the poor status of user research in many countries (Kosem et al., 2019) and lack of lexicographer involvement in user research (reported in the survey presented in this paper), such tools and probably training to help facilitate research into dictionary use should definitely be provided.

Two of the emerging trends in lexicography are crowdsourcing and gamification; however, at the moment their use is largely limited to user feedback (e.g. mistakes in entries or suggestions for new words). The use of crowdsourcing during dictionary compilation is used by only a few lexicographic institutions and projects, for example the Thesaurus of Modern Slovene (Arhar et al., 2018), the Collocations Dictionary of Modern Slovene (Kosem et al., 2018), the Estonian project for the dictionary of associations (Vainik, 2018) and the Taalradar project²⁷ at the Dutch Language Institute, but in those cases it has proven to be very effective. Still, progress from the situation reported in the ENeL survey in 2014 (Krek et al., 2014) can definitely be observed. But overall, it seems that lexicographers are still searching for the best ways of including these methodologies in dictionary compilation. Potential issues could be the lack of suitable case studies, and the lack of relevant features in existing DWS or the lack of tools supporting these methods. This need was also reported by several respondents in this survey.

Among other relevant wishes expressed by the respondents that deserve to be mentioned are the need for a common standard for the development of lexicographic resources, the need for a central repository, and the need for tools for harmonization of dictionary formats. The respondents expect a significant change in relation to lexicographic data modelling and publishing policy. The turn towards unified data is expected, with respondents mentioning that publishers will produce a single resource containing all the data that the publisher has about the language, including data traditionally not considered part of a dictionary. Considering that providing solutions

²⁷ <https://taalradar.ivdnt.org/>

to these issues is also part of the ELEXIS agenda, it is good to see that the lexicographers are aware of them.

It is also important to note some of the concerns that were expressed by the respondents. These were often connected to the quality and reliability of lexicographic data in state-of-the-art lexicography, information overload, rapid technology development, and the potentially reduced value of lexicographic skills in digitally oriented projects. Several respondents were concerned about the overestimated value of the presentational component of dictionaries, especially in relation to presentation on smartphones, which may result in neglecting the aspect of the quality and reliability of lexicographic data. Last but not least, a few respondents noted the low status of lexicography in their countries. This echoes events such as the recent discontinuation of important national dictionary projects (e.g. Great Dictionary of Polish; Żmigrodzki, 2018), reports on the absence of teaching dictionary use in schools (Kosem et al., 2019), and acceptance of documents such as the Resolution at EURALEX 2016 Congress, promoting the importance of lexicography.

6. Conclusion

The survey conducted as part of the ELEXIS project has provided useful insights into existing practices and needs of lexicographers around Europe. The survey successfully complements the surveys conducted during the ENeL COST Action, especially in terms of raising awareness of issues such as lexicographer education and training, lexicographers' needs connected with tools, and the latest lexicographic trends. It also points to the importance of regular updating of information about the lexicographic practices, methods, tools and formats used in institutions across Europe and the world.

We intend to collect more data on lexicographic practices in the coming years, e.g. by including the ELEXIS observer institutions and their lexicographers in a follow-up survey. In this way, we intend to devise some form of a lexicographic practice map of Europe so that similarities and differences between practices at different institutions in different countries can be easily analysed. This would facilitate institutional collaboration and the search for common solutions. Finally, all the results have already informed and will continue to inform the preparation of the deliverables of the ELEXIS project, such as tools, resources and training materials that will be produced in the next three years.

7. Acknowledgements

The research received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

8. References

- Arhar Holdt, Š., Čibej, J., Dobrovoljc, K., Gantar, P., Gorjanc, V., Klemenc, B., Kosem, I., Krek, S., Laskowski, C. & Robnik Šikonja, M. (2018). Thesaurus of modern Slovene: by the community for the community. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (eds.) *Proceedings of the 18th EURALEX International Congress, 17-21 July 2018, Ljubljana*. Ljubljana: Ljubljana University Press, Faculty of Arts, pp. 401-410. <http://euralex.org/wp-content/themes/euralex/proceedings/Euralex%202018/118-4-2991-1-10-20180820.pdf>.
- Atkins, S. B. T. & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.
- Hartmann, R.R.K. (ed.). (2003). *Lexicography: Dictionaries, compilers, critics, and users*. London: Routledge.
- Kallas, J., Koeva, S., Kosem, I., Langemets, M. & Tiberius, C. (2019). ELEXIS deliverable 1.1 Lexicographic Practices in Europe: A Survey of User Needs. https://elex.is/wpcontent/uploads/2019/02/ELEXIS_D1_1_Lexicographic_Practices_in_Europe_A_Survey_of_User_Needs.pdf (22 February 2019).
- Klosa, A. (2013). The lexicographical process (with special focus on online dictionaries). In H.R. Gouws, U. Heid, W. Schweickard & H.E. Wiegand (eds.) *Dictionaries. An International Encyclopedia of Lexicography. Supplement Volume: Recent Developments with Focus on Electronic and Computational Lexicography*. Berlin–Boston: de Gruyter, pp. 517–524.
- Kosem, I., Krek, S., Gantar, P., Arhar Holdt, Š., Čibej, J., Laskowski, C. (2018). Collocations Dictionary of Modern Slovene. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (eds.) *Proceedings of the 18th EURALEX International Congress, 17-21 July 2018, Ljubljana*. Ljubljana: Ljubljana University Press, Faculty of Arts, pp. 989-997. <http://euralex.org/wp-content/themes/euralex/proceedings/Euralex%202018/118-4-2939-1-10-20180820.pdf>.
- Kosem, I., Lew, R., Müller-Spitzer, C., Ribeiro Silveira, M. & Wolfer, S. et al. (2019). The image of the monolingual dictionary across Europe. Results of the European survey of dictionary use and culture. *International Journal of Lexicography*, 32(1), pp. 92-114. <https://doi.org/10.1093/ijl/ecy022> (30 May 2019).
- Krek, S., Abel, A. & Tiberius, C. (2014). *Dictionary Writing Systems & Corpus Query Systems. Survey – WG3 ENeL*. http://www.elexicography.eu/wp-content/uploads/2015/04/ENeL_WG3_Vienna_DWS_CQS_final_web.pdf (30 May 2019).
- Müller-Spitzer, C. (ed.) (2014). *Using Online Dictionaries*. Berlin, Boston: De Gruyter.
- Tiberius, C. & Krek, S. (2014). *Workflow of Corpus-Based Lexicography*. Deliverable COST-ENeL-WG3 meeting, July 2014, Bolzano/Bozen. http://www.elexicography.eu/wp-content/uploads/2015/04/LexicographicalWorkflow_DeliverableWG3BolzanoMe

- eting2014.pdf (30 May 2019).
- Tiberius, C., Heylen, K. & Krek, S. (2015). *Automatic Knowledge Acquisition for Lexicography. Survey – WG3 ENeL*. http://www.elexicography.eu/wp-content/uploads/2015/10/ENeL_WG3_Survey-AKA4Lexicography-TiberiusHeylenKrek.pptx (30 May 2019).
- Vainik, E. (2018). Compiling the Dictionary of Word Associations in Estonian: from scratch to the database. *Eesti Rakenduslingvistika Ühingu aastaraamat = Estonian Papers in Applied Linguistics* 14, pp. 229–245. <http://dx.doi.org/10.5128/ERYa14.14> (30 May 2019).
- Žmigrodzki, P. (2018). Methodological issues of the compilation of the Polish Academy of Sciences Great Dictionary of Polish. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (eds.) *Proceedings of the 18th EURALEX International Congress, 17-21 July 2018, Ljubljana. Ljubljana: Ljubljana University Press, Faculty of Arts*, pp. 209-219. <http://euralex.org/wp-content/themes/euralex/proceedings/Euralex%202018/118-4-2973-1-10-20180820.pdf> (30 May 2019).

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



Language Varieties Meet One-Click Dictionary

Egon W. Stemle, Andrea Abel, Verena Lyding

Institute for Applied Linguistics, Eurac Research, Bolzano - Bozen, IT

E-mail: {egon.stemle, andrea.abel, verena.lyding}@eurac.edu

Abstract

The goal of the STyrLogism Project is to semi-automatically extract neologism candidates (new lexemes) for the German standard variety used in South Tyrol, and generally to create the basis for long-term monitoring of its development. We use automatic lexico-semantic analytics for the lexicographic processing, but instead of continuing to develop our independent neologism detection application, we have recently become part of a thriving community of users and developers within the EU infrastructure project ELEXIS, which aims to harmonize efforts that relate to producing and making dictionary resources available, and to develop tools with consistent standards and increased interoperability. Consequently, we moved the development of our neologism application into Lexonomy, one of ELEXIS' promoted open-source projects. In the following, we report on the current state of this ongoing development by describing how we integrate our work with the Sketch Engine and Lexonomy tools, pointing out the challenges involved, and discussing how our work on language varieties can be evaluated.

Keywords: language variety; One-Click Dictionary; web corpus; dictionary of variants; ELEXIS

1. Introduction

The goal of the STyrLogism Project is to semi-automatically extract neologism candidates (new lexemes) for the German standard variety used in South Tyrol (STyrGerman), an autonomous province in Northern Italy where German is an official language. Direct applications for these neologisms are, for example, the consideration for future editions of the *Variantenwörterbuch des Deutschen* (Dictionary of variants of the German language, abbr. VWB) (Ammon et al., 2016) and other dictionaries. In the medium term, the project should additionally serve as an empirical basis for the long-term observation and evaluation of trends in STyrGerman, which also makes it interesting for language policy and language planning measures.

In total, there are up to three official languages (Italian and German - plus Ladin, in the Ladin valleys) and an institutional bi- or trilingualism in the region, which means that the two (or three) languages have the same standing and there is an effective multilingual obligation of the civil servants and the right to address the administration in one of the languages; through personal linguistic socialization and individual biographical constellations and experiences, people in the region usually acquire diverse individual language repertoires, that is (multilingual) dynamic communicative competences. Moreover, from a pluricentric perspective, South Tyrol is a national semi-centre, inhabits a peripheral location in the German-speaking area, and offers an interesting language contact situation, especially with regard to the German and Italian languages (Abel, 2018). All this makes South Tyrol in general and STyrGerman in

particular an interesting object of investigation for linguistic studies.

The completely revised second edition of VWB appeared in 2016, 12 years after the first edition, but STyrGerman could not be analysed to the same extent as the varieties of the full centres (Germany, Austria, Switzerland) and, in addition, methodological decisions led to some developments and phenomena being less represented: Firstly, only corpus data with journalistic prose served as a source for the new edition of the VWB, while for the first edition various text genres had been used, which were not based on digital corpora yet but on text excerpts on paper. However, standard texts from newspaper corpora alone do not unequivocally cover the entire relevant language usage. For example, “*Bar*” has a particular meaning in STyrGerman in the sense that it is used to refer to a place to have coffee, that is, as a synonym for “coffee shop”, whereas in the other German varieties it only has the meaning “night bar”, and it is difficult to extract sentences conveying this STyrGerman meaning of “*Bar*” from newspaper texts. In them, “*Bar*” is often mentioned, for example, together with break-ins, but is hardly described in a way to infer its different usage (e.g. mentioning what people usually do there, drinking coffee, eating a croissant, reading the newspaper). A case in point is the following excerpt from original data: “Zu der Bluttat war es vor dem Eingang der ‘*Bar Pleres*’ in Matsch gekommen” (translation: “The bloody deed took place in front of the entrance of the ‘*Bar Pleres*’ in Matsch”)¹ (Abel, 2018). Furthermore, many relevant linguistic phenomena can be monitored not only with standard text corpora but additionally—and some phenomena even better—with web corpora and corpora of computer-mediated communication, because language changes on social media and the internet can be in public online usage for a while before getting included into mainstream newspapers and other text genres (Androutsopoulos, 2011). However, social media and web corpora were not included in the data for the VWB.

Secondly, in the course of the VWB project, it was not possible (for financial reasons) to check systematically whether new STyrGerman lexemes should be included or obsolete ones should be eliminated. This is a matter of linguistic change that is closely related to the research on neologisms, which in our case also includes variants that are commonly used in STyrGerman but are not yet lexicalised (Abel & Stemle, 2018). We are aware that these are not neologisms in the narrower sense, but we do not need to make this distinction with regard to data processing. The research on neologisms is typically divided into two categories: one category for words used in a new meaning, and another for new lexemes with an unseen graphical representation (Kinne, 1998). In the past, we have concentrated on the detection of neologism candidates of the latter category. As an example, we can mention “*Vollkornpizzetta*” (“very small, round-shaped pizza made of whole-grain”). The particular part of this compound word is “*Pizzetta*” that derives from “*Pizza*” being “-etta”, the diminutive suffix in Italian. But the whole word is not a loan word from Italian; the compound modifier “*Vollkorn*” is the German word for “whole-grain” and not the Italian word “*integrale*”. However,

¹ Dolomitenkorpus, 2001: <http://www.korpus-suedtirol.it>

it would not be the same to talk about a “kleine Vollkornpizza” (“small pizza, minipizza”), because “pizzetta” in Italian refers to a particular type of pizza, usually a very small, round-shaped pizza (with a diameter of around 5 cm), which you offer, for example, at a buffet as finger food.

Lastly, the focus for including variants was on the occurrence of different word forms and on differences in word meanings, but there exist collocations which are not specific for a variety because of their individual words, but because the words are frequently combined and thus represent a collocation. For example, the meaning of “jemanden in die Mobilität entlassen/überstellen” (literal translation “*to release/transfer someone into mobility”; the actual meaning “to let someone go after a company struggled for some time” is a transfer from the Italian “mobilità”) is only specific to STyrGerman (Abel, 2018).

Overall, as reported in earlier work (Abel & Stemle, 2018), the STyrLogism Project changes some of the collection parameters and attempts to remedy some of the aforementioned shortcomings. First and foremost, we use web data as a valuable complement to standard texts (Barton & Lee, 2013), so that we can now observe short-term and fast-moving developments in online media. Overall, we aim to provide semi-automatic support for the detection of new lexemes and lexeme combinations that are more frequent in STyrGerman than in other variants—or even exclusive to STyrGerman—and, finally, we also want to employ methods to detect meaning shift, which previously has been done manually as part of exploratory analyses within the project.

2. Related Work

The approaches for neologism detection can be divided into two groups. One, usually applied to a single set of new data, uses language resources such as word lists or linguistic patterns. The word lists are compiled from existing lexicographic resources such as dictionaries or corpora, combined with filters to eliminate non-words, typographical errors, named entities, and so on, and the linguistic patterns are, for example, markers of lexical novelty like punctuation marks that can signal new words, as shown in O’Donovan and O’Neill (2008) and Paryzek (2008). The other group, usually applied to multiple datasets, uses statistical measures or machine learning to calculate and evaluate the increase in usage or the change in meaning over time or in different registers. Examples can be found in Stenetorp (2010) and Kilgarrieff et al. (2015). Finally, these two approaches can also be combined.

Wortwarte² (Lemnitzer, 2000-2019) is the most relevant previous project in relation to our own, as it is an ongoing project with an online portal that has been regularly collecting and documenting new German words. The system is based on German online-newspaper texts: a web crawler regularly collects data from pre-defined sites, such as

² <http://www.wortwarte.de/>

newspapers and magazines. After the HTML content has been cleaned up, the plain text is used to build a new time slice of a corpus. The selection of neologism candidates is based on short-term evaluations in which the new corpus is compared with the continuously growing German reference corpus (Das Deutsche Referenzkorpus – DeReKo. See Kupietz and Längen (2014) for an overview) with around 42 billion word tokens (status: Q1.2018). In order to avoid “random” errors (e.g. typing errors) and to filter out spelling mistakes, the selection of neologisms is conducted manually after the comparison with DeReKo. The results of these analyses are published online at irregular intervals, but typically about once a week. The results usually include a few words with their exemplary use in a sentence and the reference as to where they came from.

O’Donovan and O’Neill (2008) use a similar idea, but due to the lack of free access to a continuously growing reference corpus for English they use and update their own Chambers Harrap International Corpus (CHIC) web corpus. It consists of more than 500 million words of International English and stands in the tradition of the Bank of English rather than a static, balanced resource like the British National Corpus (BNC). They also use other resources, like lemmatization and morpho-syntactic information, such as a headword list augmented with inflected forms. Kerremans, Stegmayr, and Schmid (2011) also crawl their own reference corpus and additionally use an explicit component to monitor the changed over time for selected terms: they use the commercial search engine Google and regularly crawl the content of search results returned for each “to-be-monitored” neologism.

3. STyrLogism: Evolution

3.1 Initial implementation

The first incarnation of the STyrLogism Project system (Abel & Stemle, 2018) consisted of a list of manually selected URLs from news, magazines and blog websites of South Tyrol, and regular data crawls from the Heritrix³ Internet Archive crawler. The whole content from the crawled web pages was saved in the Web ARChive (WARC) archive format. Then, we used Schäfer and Bildhauer’s (2012) texrex toolkit. This comes already set up to process WARC files, and directly works with the Heritrix output. It removes HTML and scripts, and uses a simplistic heuristic to split paragraphs in the resulting text. So-called boilerplate, that is, navigational elements and menus, date strings, copyright notices, among others, are then identified and quantified as an annotation on a paragraph level. Finally, a two-step duplicate detection is employed: the first removes perfect duplicates, that is, documents that are identical up to the last character; the second step removes near-duplicates. The resulting data was converted into a list of word forms and a corpus for the NoSketchEngine (NoSkE) (Rychlý, 2007). We then made case-insensitive comparisons of the list of word forms with: a) the one from our reference corpora, b) the additional

³ <https://archive.org/projects/>

word lists, which was in practice a simple Named Entity Recognition, and c) with the combination of all formerly crawled data sets. Our reference corpora were DECOW14 (Schäfer & Bildhauer, 2012) with around 60 million word forms, and the South Tyrolean Web Corpus (Schulz et al., 2013) with around 2.4 million word forms; the additional word lists consisted of named entities, terminological terms from the region, and specific terms of the German standard variety used in South Tyrol (altogether around 53,000 word forms). The cleaned data of the latest crawl was then tokenized—but not lemmatized—and converted into a word list. This list of candidate words consisted of those in the latest crawl that appeared less than a predefined number of times in all of the other data. Finally, the candidates were manually checked in a specifically crafted streamlined interface. This interface shows a predefined number of neologism candidates on one page along with the first (and possibly only) results as a KWIC result. The user can then click the candidate to get the whole result page of this candidate’s search query in the NoSkE, where all additional meta information for each search result is available. The user can also click a check-box or enter a comment into a text field (which automatically triggers the check-box) to make a note of this candidate for later curation. Finally, the user can go to the next page, which automatically discards all unmarked candidates from further processing. In a second ‘curation’ step, a user can see all the previously marked candidates with single KWIC results of all occurrences of the candidate in different crawler runs. This stage gives an overview of the currently tracked neologism candidates with quick access to individual occurrences over time.

3.2 Updated Method

Here, we will report on our current work that is conducted as part of our institution’s observer status in the European Lexicographic Infrastructure (ELEXIS) project (Krek et al., 2018). ELEXIS features the One-Click Dictionary toolchain to automatically generate, for example, headword lists, word (and other lexical units) senses, definitions, and corpus-based examples. The toolchain consists of the corpus query system Sketch Engine⁴ (Kilgarriff et al., 2014) and the dictionary writing system Lexonomy⁵ (Měchura, 2017); together they are supposed to support lexicographers along the entire pipeline of producing a dictionary (see Granger & Paquot (2012) for an overview of electronic means in the planning, writing, and dissemination of dictionaries), from corpus to screen, where dictionaries are pre-generated automatically from a corpus (using Sketch Engine) and then post-edited (using Lexonomy).

ELEXIS, among other things, aims to harmonize efforts on a larger European scale that relate to producing and making dictionary resources available, and to develop tools to update existing or new resources with consistent standards and increased interoperability. We hope that through cooperation within ELEXIS more opportunities

⁴ <https://www.sketchengine.eu>

⁵ <https://www.lexonomy.eu>

and desirable developments arise: With access to current methods and tools, and a collective awareness of challenges and information about upcoming solutions for the next generation of online dictionaries, we can integrate our local digital resources into modern workflows and also provide feedback that influences the design of use-cases for tools and workflows.

The One-Click Dictionary is a convenient automation for exchanging lexicographic data between a Sketch Engine corpus and a Lexonomy dictionary, and will eventually cover, for example, the extraction of example sentences, the detection of definitions, descriptions and collocations, and the clustering of word senses. The computations and analyses are carried out by the Sketch Engine, and the results are transmitted to Lexonomy as dictionary entries. The communication is channelled through an Application Programming Interface (API), that is a set of defined functions and procedures that lets computers talk to each other. In Lexonomy, the data can then be edited and eventually published as an online dictionary, ideally under an open-source license, for example, CC0, CC-BY, CC-BY-SA⁶ or ODbL⁷. There will also exist some dedicated features for post-editing an automatically generated dictionary: for example, features for quickly splitting and lumping senses, and for distributing example sentences into senses. Furthermore, Lexonomy as a light-weight, web-based system for writing and publishing dictionaries will also support features like, for example, a mechanism for handling cross-references. In the future, users will be able to include cross-references from one entry to another entry or to a location in another entry (such as a specific sense inside another dictionary entry). Lexonomy will make sure the cross-references are clickable when the entry is formatted for display. Figure 1 shows this relationship on the left: Users interact with the Sketch Engine and Lexonomy web interfaces, and the two processes analyse their respective corpora and dictionaries. The data and functions of the other service are accessed via their API.

It should be noted that Sketch Engine is a subscription-based service, although free access for non-commercial use of Sketch Engine between 2018 and 2022 is funded⁸ by the EU through ELEXIS. Lexonomy, on the other hand, is open-source software, with source code available from a GitHub repository⁹ and licensed under the MIT License, which allows unrestricted re-use even for commercial purposes; so anyone can download and set up a local installation of Lexonomy and customize it to meet specific requirements. In addition, the development of Lexonomy is backed by the sponsorship of Lexical Computing (the company that makes Sketch Engine) and by funding from ELEXIS. This design provides access to the internal data representation of Lexonomy dictionaries and simplifies the task of transferring applications and data to another setup as needed; it also enables on-premise data storage, which retains the ability to

⁶ <https://creativecommons.org/licenses/>

⁷ <https://opendatacommons.org/licenses/odbl/summary/>

⁸ <https://www.sketchengine.eu/elexis/>

⁹ <https://github.com/elexis-eu/lexonomy/>

failover to a different data centre when everything else fails. Additionally, this brings about the possibility of designing one's own applications that rely on Lexonomy without much risk of a possible vendor lock-in. This is illustrated in Figure 1 on the right, where users interact with their own application, which in turn uses the API to access Lexonomy data and functionality while managing its own (private) data.

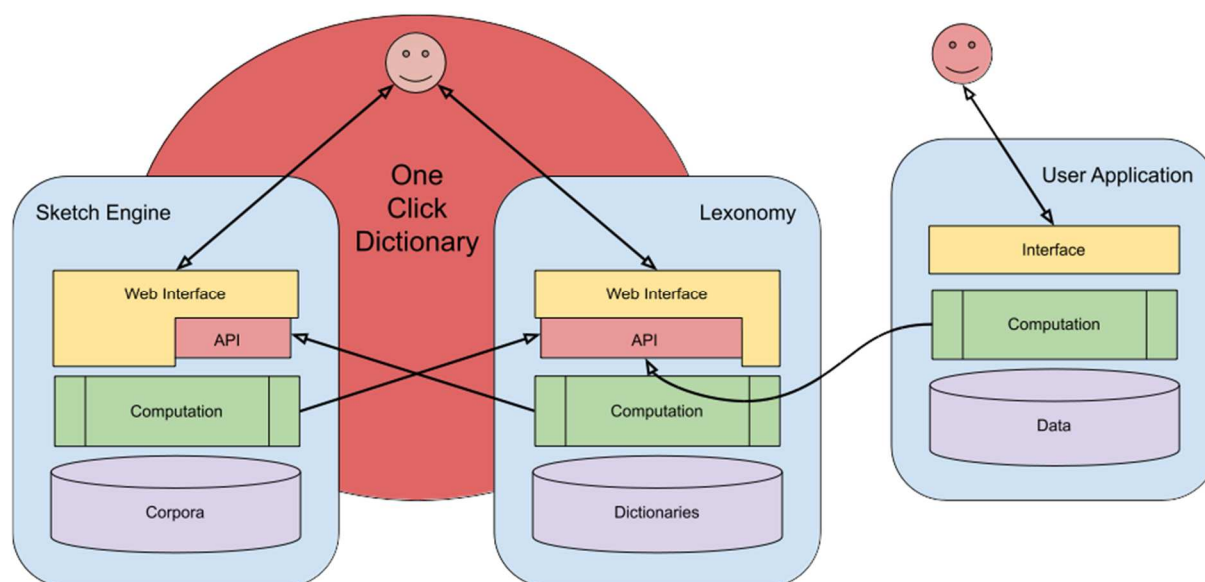


Figure 1: The One-Click Dictionary automatizes the data exchange between Sketch Engine and Lexonomy. The communication is channelled through an API, and users interact with the services via their respective web interfaces. On the other side, users can also design and use their own applications to access data in Lexonomy via an API.

Additionally, there exists another possibility: The development of a user application could also become part of Lexonomy. It is an open-source project with a growing community embedded in an ongoing European Union infrastructure project dedicated to lexicography. The users already include the University of Ljubljana, the Dutch Language Institute (Instituut voor de Nederlandse Taal), and Eurac Research (i.e. the authors of this paper). These users are also active contributors¹⁰ to the GitHub repository, and all have participated in two previous hackathons. Both hackathons lasted approximately 2.5 days, and one was conducted with all participants on-site, the other on scheduled days with scheduled telephone and video conferencing. During these hackathons questions, problems, ideas could be discussed, joint strategies worked out and above all (partially) implemented. The general progress of the development of Lexonomy can be tracked by the contributions in the repository and the activities in the ticketing system but, above all, the development can be influenced by active participation on these channels and the dedicated Google Group¹¹.

¹⁰ <https://github.com/elexis-eu/lexonomy/graphs/contributors>

¹¹ <https://groups.google.com/forum/#!forum/elexis-lexonomy>

For the STyrLogism Project, we have started to use Sketch Engine’s web corpus capabilities, which include on-demand web crawling (also of predefined individual sites), boilerplate removal, deduplication, and tokenization, tagging, lemmatization. The boilerplate removal is applied on crawled texts to remove unwanted portions, namely navigation and menus, advertising, legal text, tabular data and any other types of text unsuitable for linguistic analysis and therefore for inclusion in a corpus. The data then undergoes a deduplication procedure where both perfect duplicates, as well as near duplicates, are removed so that only one instance of each text is preserved, and finally a Natural Language Processing pipeline divides the text into words (tokenization), enriches it with part-of-speech (PoS-tagging) and assigns the base form to each word form (lemmatization). In addition, we have begun to participate in the development of Lexonomy and advance our use-case to adapt Lexonomy as a replacement for our previous interface. We believe that the common ground between the different users will promote rich development and that we will be able to overcome certain difficulties with growing user and development communities.

4. Conclusions and Outlook

For a pending in-depth evaluation, we will use the VWB and an automatically generated “One-Click Dictionary”. This will allow us to check the automatically generated lexicon, but will also allow us to put the VWB to the test with the automatically calculated data. Ideally, by using this approach, we should overcome the previously mentioned shortcomings of the VWB. So far we can at least say that a manual search for meanings of “Bar” on the latest web data—in contrast to the old newspaper data—was successful. That is, we found a use of “Bar” in the sense of “coffee shop”: “In der Bar des Hotels sind auch Tagesgäste gerne willkommen und geniessen köstliche Kuchen und dazu Kaffee” (“Day guests are also welcome at the hotel’s bar to enjoy delicious cakes and coffee”).

Some of the pressing desiderata worth mentioning in conclusion are the availability of appropriate corpora to observe language use (including everyday situations) and detect trends of the local standard variety of STyrGerman, as well as extensive support for automatically extracting relevant data for variety lexicography (e.g. collocations, “new” word forms and meanings).

Cooperation with an international lexicographic infrastructure such as ELEXIS should strengthen the position of local varieties and dialects, provide access to current methods and tools, and also influence their design. In addition, local digital resources will be integrated into modern workflows and jointly tested.

5. References

- Abel, A. (2018). Von Bars, Oberschulen und weißen Stimmzetteln: Zum Wortschatz des Standarddeutschen in Südtirol. In S. Rabanus (ed.) *Deutsch als Minderheitensprache in Italien: Theorie und Empirie kontaktinduzierten*

- Sprachwandels*, pp. 283–323.
- Abel, A. & Stemle, E. W. (2018). On the Detection of Neologism Candidates as Basis for Language Observation and Lexicographic Endeavours: The STyrLogism Project. In J. Čibej, V. Gorjanc, I. Kosem, & S. Krek (eds.) *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*, pp. 535–544.
- Ammon, U., Bickel, H. & Lenz, A. N. (Eds.). (2016). *Variantenwörterbuch des Deutschen: Die Standardsprache in Österreich, der Schweiz, Deutschland, Liechtenstein, Luxemburg, Ostbelgien und Südtirol sowie Rumänien, Namibia und Mennonitensiedlungen*. 2nd, updated and extended edition. Berlin/Boston: De Gruyter Mouton.
- Androutsopoulos, J. (2011). Language change and digital media: A review of conceptions and evidence. In K. Tore & N. Coupland (eds.) *Standard languages and language standards in a changing Europe*, pp. 145–161.
- Barton, D. & Lee, C. (2013). *Language online: Investigating digital texts and practices*. Milton Park, Abingdon, Oxon: Routledge.
- Granger, S. & Paquot, M. (eds.). (2012). *Electronic Lexicography*. Oxford, New York: Oxford University Press.
- Kerremans, D., Stegmayr, S. & Schmid, H.-J. (2011). The NeoCrawler: Identifying and retrieving neologisms from the internet and monitoring ongoing change. In K. Allan & J. A. Robinson (eds.) *Current Methods in Historical Semantics*, pp. 59–96. <https://doi.org/10.1515/9783110252903.59>.
- Kilgarrieff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J. & Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography*, 1(1), 7–36. <https://doi.org/10.1007/s40607-014-0009-9>.
- Kilgarrieff, A., Herman, O., Bušta, J., Kovář, V. & Jakubíček, M. (2015). DIACRAN: a framework for diachronic analysis. In F. Formato & A. Hardie (eds.) *Corpus Linguistics 2015: Abstract Book*. Lancaster, UK: UCREL.
- Kinne, M. (1998). Der lange Weg zum Neologismenwörterbuch. Neologismus und Neologismenlexikographie im Deutschen. Zur Forschungsgeschichte und zur Terminologie, über Vorbilder und Aufgaben. In W. Teubert (ed.) *Neologie und Korpus*, pp. 63–110.
- Krek, S., Kosem, I., McCrae, J. P., Navigli, R., Pedersen, B. S., Tiberius, C. & Wissik, T. (2018). European Lexicographic Infrastructure (ELEXIS). In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (eds.) *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*, pp. 881–891.
- Kupietz, M., & Lungen, H. (2014). Recent Developments in DeReKo. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani & S. Piperidis (eds.) *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*.
- Měchura, M. (2017). Introducing Lexonomy: An open-source dictionary writing and publishing system. In I. Kosem et al. (eds.) *Electronic Lexicography in the 21st Century: Lexicography from Scratch. Proceedings of the eLex 2017 Conference*.

- O'Donovan, R. & O'Neill, M. (2008). A Systematic Approach to the Selection of Neologisms for Inclusion in a Large Monolingual Dictionary. In J. D. Elisenda Bernal (ed.) *Proceedings of the 13th EURALEX International Congress*, pp. 571–579.
- Paryzek, P. (2008). Comparison of selected methods for the retrieval of neologisms. *Investigationes Linguisticae*, 16, 163. <https://doi.org/10.14746/il.2008.16.14>
- Rychlý, P. (2007). Manatee/Bonito – A Modular Corpus Manager. *First Workshop on Recent Advances in Slavonic Natural Language Processing (RASLAN 2007)*, pp. 65–70.
- Schäfer, R. & Bildhauer, F. (2012). Building Large Corpora from the Web Using a New Efficient Tool Chain. In N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani & S. Piperidis (eds.) *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*.
- Schulz, S., Lyding, V. & Nicolas, L. (2013). STirWaC: compiling a diverse corpus based on texts from the web for South Tyrolean German. In S. Evert, E. Stemle & P. Rayson (eds.) *Proceedings of the 8th Web as Corpus Workshop (WAC-8)*, pp. 35–45.
- Stenetorp, P. (2010). Automated Extraction of Swedish Neologisms using a Temporally Annotated Corpus. Master's Thesis. Royal Institute of Technology (KTH), Stockholm, Sweden.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



Identification of Languages in Linked Data: A Diachronic-Diatopic Case Study of French

Sabine Tittel¹, Frances Gillis-Webber²

¹ Heidelberg Academy of Sciences and Humanities, Seminarstraße 3,
D-69117 Heidelberg, Germany

² Department of Computer Science, University of Cape Town, Cape Town, South Africa
E-mail: sabine.tittel@urz.uni-heidelberg.de, fran@fynbosch.com

Abstract

When modelling linguistic resources as Linked Data, the identification of languages using language tags and language codes is a mandatory task. IETF's BCP 47 defines the standard for tags, and ISO 639 provides the codes. However, these codes are insufficient for the identification of diatopic variation within a language and, also, for different historical language stages. This weakness hampers the accurate identification of data, which in turn leads to ambiguity when extending, aggregating and re-using this data—a key notion of Linked Open Data and the Semantic Web. We show the limitations of language identification with a case study of French linguistic data from both a diachronic and a diatopic perspective. Our exemplary data derives from dictionaries of Old French, Middle French, and of Modern French dialects, and from a Modern French linguistic atlas. For each exemplar, we propose a solution using the *privateuse* sub-tag of BCP 47's language tag, staying within the boundaries of existing standards. Using a predefined pattern for the *privateuse* sub-tag, the solutions enable a dialect, a patois, in combination with a time period, to be defined and identified. This can lead to shared agreement of language tags that will increase interoperability within the context of Linked Data.

Keywords: language codes; language tags; language annotation; Linked Open Data; French dialects

1. Introduction

Over the last decade, modelling linguistic data using the Resource Description Framework (RDF), following the Linked Data (LD) paradigm, has become a widespread method for the creation of datasets for a multilingual web of data. It enables machine-readable, cross-resource access to data that are otherwise spread across the web as isolated datasets. However, for the modelling of linguistic resources as LD, the use of language tags is essential: the annotation with language tags whose form adheres to established standards ensures unambiguous language identification of linguistic information, such as lexemes and their graphic and phonetic realizations. Because of the interlinking of lexemes and their different realizations, the LD format can be particularly valuable for linguistic resources that document the diatopic diversity of a given language (i.e., with a spatial reference). Examples are regional dictionaries or linguistic atlases. These resources can be complemented with historical data to

introduce a diachronic perspective to the diatopic variation of the language (i.e., considering evolution through history). This can be, e.g., data from historical dictionaries that indicate regional characteristics. The publication of these resources as LD and the corresponding means of data query can enhance studies that focus on the diatopic richness of modern-day languages and on the evolution of diatopic variation at the same time. The use of language tags is specified by IETF’s BCP 47 (Phillips & Davis, 2009: 1-4) and the required language codes come from ISO 639 (International Organization for Standardization, n.d.). Within our field, however, we observe a lack of language tags and codes hampering the required language annotation. In this paper, we address the issue of language tagging with French linguistic resources combining a diatopic with a diachronic perspective: in a case study, we investigate data of Old-, Middle- and Modern French resources with (regional) dictionary data and linguistic atlas data.

After a short outline of the diachronic-diatopic landscape of French linguistic resources (Section 1.1), we briefly describe RDF, LD (Section 1.2), and the identification of languages (Section 1.3). In the following section, we introduce the use of a pattern for language tags (Section 2). Our case study of French uses exemplary data of historical and modern dictionaries (Section 3) and of a linguistic atlas (Section 4). For each exemplar, we demonstrate a solution for the language tagging, using the pattern described. We evaluate the solutions in Section 5, and in Section 6, we present an interface which can be used to generate (and decode) language tags according to our pattern. We conclude the paper in Section 7.

1.1 Diatopic linguistic resources and a diachronic perspective

The regional varieties, dialects and patois¹ of the French of France are under-represented in linguistic consideration in general and in lexicography in particular (Rézeau, 2001: 7). This is all the more true for the diatopic reflection from a diachronic angle: the historical development of French regionalisms has not been studied in a comprehensive yet detailed way (Gleißgen & Thibaut, 2005: XII). Studies focusing on single topics such as a particular region in a particular time period have been conducted, recently by, e.g., Chauveau (2016), and Rézeau (2016).

There are many resources that can be exploited for diatopic-diachronic studies: for the different language periods of French, dictionaries, corpora, and—for modern French in

¹ We are aware of the discussion of the terms that denote different variations within the diatopic diasystem of French. In this paper, we will use the terms following the French literature, where *régionalité linguistique* (of French) is clearly distinguished from *dialectes*, the first referring to variation within the standard language, the latter to the primary dialects of France that are the successors of the Old French dialects (Gleißgen & Thibaut, 2005: V), and patois typically designating a local variety of a dialect. Note that we use ‘patois’ as a non-pejorative term.

particular—linguistic atlases are available.² Modern resources covering French varieties include dialect or patois dictionaries (e.g., Rézeau, 2001; Varlet, 1896; Vasseur, 1998), linguistic atlases (e.g., Gilliéron & Edmont, 1902–1910; Lanher et al., 1979–1988; Dondaine & Dondaine, 1972–1991), corpora (Thun, 2011)³, and, also, individual studies (e.g., Rézeau, 2007) focusing on regional French, dialects and patois. For the historical language stages however, there are fewer resources with diatopic content. A reason for this is that from ca. 1500 AD—with the constitution of French (evolving from a Parisian scripta⁴ that had occurred around 1250) as a national language (Wolf, 1979: 94f.)—to the beginning of the 19th century, dialects almost exclusively belonged to the oral culture (Berschin et al., 2008: 203–211). Consequently, studies on the subject of regionalisms are scarce for this time period. Earlier however, in medieval times, the primary dialects included in the notion of Old- and Middle French, such as Picard and Anglo-Norman, were used for both oral and written communication. Hence, we look at the transmission of numerous linguistic primary resources (texts in manuscripts, often accessible in scholarly text editions) documenting regional variation during the Middle Ages. For this time period, studies mainly focus on a single primary resource and how to localize its language in a specific region (notably works by J.-P. Chambon, e.g., Chambon, 1997, and G. Roques, cf. the ‘Liste Roques’ in Glessgen & Trotter, 2016: 473–635). There are also many-volumed, comprehensive dictionaries of the historical language stages, in particular the *Dictionnaire étymologique de l’ancien français* (DEAF, Baldinger et al., 1971–) for Old French, the *Dictionnaire du moyen français* (DMF, ATILF – CNRS & Université de Lorraine (2015)) for Middle French, and the *Französisches Etymologisches Wörterbuch* (FEW, von Wartburg, 1922–) for the diachronic description of French until the present day. These dictionaries—although not necessarily conceived as data sources for diatopic linguistics—provide a synopsis of the knowledge of the particular historical language stage. By incorporating the results of historical dialect studies, they thus contribute to our knowledge of regional variation evolving through time.

Digitization of diatopic resources. It is a European consensus that geographic variation of languages needs to be valorized and promoted, particularly online: UNESCO, La Francophonie⁵ and other international organizations emphasize the need for (culturally and) linguistically diverse local content to be published online and for a vitalization of multilingualism on the Web, cf. Vannini & Le Crosnier, 2012: 13–21. A large number of the resources in our focus—word lists, dictionaries, linguistic atlases, texts—are currently only available in print. Only a few are available in digital form, and mostly

² We identified five language periods of French, cf. Gillis-Webber et al. (2019: Section 4 with Fig. 4).

³ Corpus of letters written by prisoners, soldiers, prostitutes, etc., that document the diatopic variation within the French substandard language.

⁴ The written form of a spoken dialect.

⁵ <https://www.unesco.com/>; <https://www.francophonie.org/> [13-02-2019].

as digital images.⁶ Many have yet to be (retro-)digitized. Digitization would allow for “many new approaches to the quantitative comparison of languages, be it for a better understanding of cross-linguistic variation in grammatical structure or for new and improved historical comparative reconstructions” (Bouda & Cysouw, 2012: 15). One such approach is the representation of the resource in RDF, which in turn allows for the extension to LD.

1.2 Enabling resource integration with the Resource Description

Framework and Linked Data

RDF⁷ is a data model that represents knowledge in a graph data structure facilitating data interchange on the (Semantic) Web. It is a fundamental technology of the Semantic Web, in which data is structured and meaning can thus be inferred by machines. RDF expresses data as sets of statements in the form of *subject-predicate-object*-triples. Each *subject* and *object* is a node; the *predicate* (or *property*) forms a relation (edge) pointing from the source node (*subject*) to a target node (*object*). Nodes and edges are identified with URIs (Uniform Resource Identifier, accessible via HTTP), and the object can also be described as a string literal (Cyganiak et al., 2014). LD can be described as a set of recommended practices for publishing RDF as structured data on the Web (Bizer et al., 2009). Applying LD principles (Berners-Lee, 2006) to the modelling of linguistic data comes with significant advantages, such as structural interoperability (cross-resource access by using same format and same query language), conceptual interoperability (through shared vocabularies), accessibility (through standard Web protocols), and resource integration by means of interlinking (Chiarcos et al., 2013). Because of the exploratory nature of LD, URIs identifying, e.g., lexemes, their senses, and their concepts referring to the things denoted, *things* and the usage of their *designations* can be explored in a cultural context without being restricted to the vehicle of a particular language. The integration of resources of different language stages and diatopic variation enables observation through time and space, including, e.g., borrowing and word formation processes, and semantic shift within a large data collection. For Old French, the first steps have been made by modelling exemplary lexicographic data of the DEAF as LD using the OntoLex-Lemon vocabulary⁸, and the modelling of a scholarly text edition of a Middle French medical treatise using RDFa (Tittel & Chiarcos, 2018; Tittel et al., 2018). To the best of our knowledge, there are no other historical linguistic resources of French represented as LD that could be exploited for diachronic-diatopic studies.

⁶ Cf., e.g., the references at https://www.lexilogos.com/lorrain_dictionnaire.htm [10-06-2019].

⁷ RDF 1.1. Primer, 2014, <https://www.w3.org/TR/rdf11-primer/> [10-05-2019].

⁸ <https://www.w3.org/2016/05/ontolex/> [13-05-2019].

1.3 Identification of languages

When modelling linguistic resources in RDF, it is necessary to identify the language of the resource and the information therein (be it a *word*, a *multiword expression*, a *sense*, a *graphical realization*, a *phonetic representation*), and to annotate literals with a language tag. IETF's BCP 47 specifies the Best Current Practice for language tags; the language tag typically begins with a language code and it must conform to established standards (Cyganiak et al., 2014). The language code comes from external resources such as ISO 639, which provides the authoritative list of language codes. Alternatives are catalogues like Glottolog, Ethnologue, and MultiTree.⁹ However, these alternatives do not meet the requirements of BCP 47 for the encoding of languages. They also reveal significant shortcomings concerning registration, hierarchization, diachronic and dialectal criteria, all of which have been discussed in detail in Gillis-Webber and Tittel (2019: 4:6-8) and Gillis-Webber et al. (2019). Lexvo¹⁰ provides dereferenceable URIs only for languages registered by ISO 639 (de Melo, 2015). It is, thus, insufficient for our use.

An exemplary lexical entry in RDF (identified as **E0**), modelled using OntoLex-Lemon and serialized in Turtle¹¹ is:

```

1  @PREFIX :      <http://www.example.com/entry/> .
2  @PREFIX ontolx: <http://www.w3.org/ns/lemon/ontolx#> .
3  @PREFIX lexinfo: <http://www.lexinfo.net/ontology/2.0/lexinfo#> .
4  @PREFIX dct:    <http://purl.org/dc/terms/> .
5  @PREFIX rdfs:   <http://www.w3.org/2001/02/rdf-schema#> .
6  @PREFIX dbpedia: <http://www.dbpedia.org/resource/> .
7
8  :alconorque a ontolx:LexicalEntry , ontolx:Word ;
9      lexinfo:partOfSpeech lexinfo:Noun ;
10     dct:language      <http://lexvo.org/id/iso639-1/pt> ,
11                        <https://iso639-3.sil.org/code/por> ;
12     rdfs:label         "cork oak"@en , "alconorque"@pt ;
13     ontolx:denotes     dbpedia:Quercus_suber .

```

⁹ <https://glottolog.org>, <https://www.ethnologue.com>, <http://multitree.org/> [07-06-2019].

¹⁰ <http://lexvo.org> [07-06-2019].

¹¹ Terse RDF Triple Language, <http://www.w3.org/TR/turtle/> [11-01-2019]. In the following code examples, namespaces are assumed defined the usual way. We include hypothetical URIs using the namespace `<http://www.example.com/entry/>`.

where Lines 10-11 show the applicable language URIs for the lexical entry indicated as ‘Portuguese’, from ISO 639-1 and ISO 639-3 respectively; Line 12 shows the language code ‘English’ (ISO 639-1 ‘en’) for the literal “cork oak”, and the language code ‘Portuguese’ (ISO 639-1 ‘pt’) for the literal “alconorque”.

The ISO 639 standard shows significant shortcomings with respect to regional variation and to historical language stages, as was shown in Gillis-Webber and Tittel (2019: 4:4-5); *cf.* also Figures. 4 and 5. This prevents the unambiguous identification of languages, even more so when modelling multiple ‘snapshots’ of data of the same language through time and space.

2. Pattern for Language Tags

As demonstrated in **E0**, the use of ISO 639 language codes in language tags is straightforward for most modern and well-known languages. However, the problem of missing or inadequate language codes extends to any variety or dialect of a language which is requires representation on the web, and for which an ISO 639 code is simply not available. Language tags, as prescribed by BCP 47, have the syntax:

language-extlang-script-region-variant-extension-privateuse

with each portion, called a sub-tag, separated by a hyphen (Phillips & Davis, 2009: 4). Gillis-Webber & Tittel (2019) propose a pattern for the *privateuse* sub-tag.¹² The pattern for the *privateuse* sub-tag is of the form:

x-language-otherlect-timeperiod-region-uri

where x- is a BCP 47 requirement indicating *privateuse*, and language (a language, dialect, patois or pidgin), otherlect (an ethnolect, sociolect, or idiolect), timeperiod, region, and URI are all parts of the sub-tag, separated by a hyphen (Gillis-Webber & Tittel, 2019: 4:12). Apart from the *privateuse* sub-tag, the sub-tags are specified by BCP 47 as “identified on the basis of its length, position in the tag, and its content”; each sub-tag typically is part of an ISO standard or registry (*ib.*) For the *privateuse* sub-tag, the use of a key (Table 1) is proposed to identify each part, thus allowing for flexibility of content and variable length thereof.

¹² Note that this pattern is not intended to replace any content that would typically be included in other sub-tags. To see the most recent updates to the pattern, please go to: <https://londisizwe.org/language-tags/>.

Part	Key 1	Key 2
language	0	0 = User-defined 1 = Glottocode
otherlect	1	0 = User-defined 1 = Glottocode
timeperiod	2	0 = one year only, BC 1 = one year only, AD 2 = start:BC - end:BC 3 = start:BC - end:AD 4 = start:AD - end:AD
region	3	0 = Geohashed latitude and longitude coordinates – polygon 1 = Geohashed latitude and longitude coordinates – point only 2 = URI to GeoJSON-LD 3 = Code from ISO 3166 4 = Identifier from GeoNames
URI	4	0 = URI shortcode from https://londisizwe.org/language-tags/

Table 1: The key for each part in the *privateuse* tag.

We identified the following set of competency questions (CQs) for the pattern, where [lect] can be replaced by any language, variant, dialect, patois, and scripta.

CQ 1 How to identify a [lect] that has no ISO 639 language code, but whose parent language does?

CQ 2 How to identify a [lect] for which ISO 639 provides a language code that indicates a different time period?

CQ 3 How to identify a [lect] for which ISO 639 provides two language codes?

CQ 4 How to identify a [lect] in space that has neither an ISO 639 code nor a code from an alternative directory?

CQ 5 How to identify a [lect] in time?

CQ 6 How to identify endonyms and exonyms of a [lect]?

When evaluating the pattern, these CQs should be answerable. Using the case study of French, we will revisit the CQs in Section 5 to test the efficacy of the proposed pattern.

3. Modelling of regional variation in dictionary data

For our case study, we will embrace both diachronic and diatopic data of French, with the latter typically mirroring aspects of the former.

3.1 Old French

Old French should be understood as an umbrella term for a number of dialects resulting from the process of settlement and romanization, different substrates, strates, etc. These dialects present distinctive linguistic realities from the beginning of the 12th century, *cf.* Rickard (1974: 54–65; 71–84).

For the Old French period, the contribution of the DEAF to our knowledge of diatopic variation of Old French has been discussed by Möhren (2016) and Tittel (2016). The DEAF allows for the annotation of data with 35 scriptae, including broader categories like ‘Nord-Est’ or ‘Centre’ (*cf.* Figures 4 and 5). For Old French, the ISO 639-3 language code is ‘fro’ («842–ca.1400»), but there are no ISO 639 language codes available for the scriptae except for Anglo-Norman (‘xno’) and Judéo-French (‘zrp’). For the modelling of DEAF data with OntoLex-Lemon, although ‘fro’ has been used as the language tag, this does not allow for the data to be differentiated on scriptae (Tittel & Chiarcos, 2018: 64f.).

An exemplar (**E1**) derived from the DEAF is *jannaie* (designating a terrain covered with gorse), a lexeme marked as Gallo.¹³ It can be modelled as follows:

```
1 :jannaie a ontolex:LexicalEntry , ontolex:Word;
2 ontolex:canonicalForm :jannaie_lemma .
3
4 :jannaie_lemma a ontolex:Form ;
5 ontolex:writtenRep "jannaie"@fro-x-00gallo .
```

In our language tag on Line 5, as an ISO 639 language code does not exist for (Old) Gallo, we have made use of a compiled language tag: *fro* identifies it as from the Old French period, and *00* indicates that it is a user-defined language (i.e., a code from an alternative directory to ISO 639 has not been used).¹⁴

¹³ DEAF J 136,9; <https://deaf-server.adw.uni-heidelberg.de/lemme/jaon#jannaie> [10-05-2019].

¹⁴ For a discussion of further approaches to language tagging Old French dialects, *cf.* Gillis-Webber & Tittel (2019: 4:9-11).

3.2 Middle French

The comprehensive dictionary for the Middle French period is the DMF. With respect to the study of dialectal characteristics of the Middle French lexis, the DMF is a resource of limited value and difficult access (Renders, 2016: 95f.). However, the DMF has the potential for facilitating the study of diatopic variation of late medieval French: the data structure of the DMF entry does not contain a label that specifically tags information as being dialectal (thus, the information cannot easily be accessed in a machine-aided way), but the running (unstructured) text of approx. 1,190 entries (Renders, 2016: 89) includes in effect such information; this can be exploited.

Although the French written standard spread in Middle French time, the dialects still maintained their role in the literature. The DMF defines a list of 29 “*étiquettes régionales*” (Renders, 2016: 86) comparable with the DEAF scriptae list. For Middle French, the ISO 639-3 language code is ‘frm’ («ca. 1400–1600»); this can be utilized to identify the language, but the challenge of codes for its dialects needs to be addressed.

In the following exemplar (**E2**), we model a lexeme that is marked as dialectal: *appreper* v. “s’approcher (d’un lieu)” “Région. (Wallonie)”.¹⁵ The language code from ISO 693-1 for modern Walloon is ‘wa’. But as for the Old French language period, the code should not be used for the Middle French period.

```
1 :appreper  a   ontolex:LexicalEntry , ontolex:Word ;
2 ontolex:canonicalForm :appreper_lemma .
3
4 :appreper_lemma  a   ontolex:Form ;
5 ontolex:writtenRep "appreper"@frm-x-00walloon .
```

In our language tag on Line 5, frm identifies it as from the Middle French period, with 00 indicating that it is a user-defined language (cp. E1).

3.3 Modern French

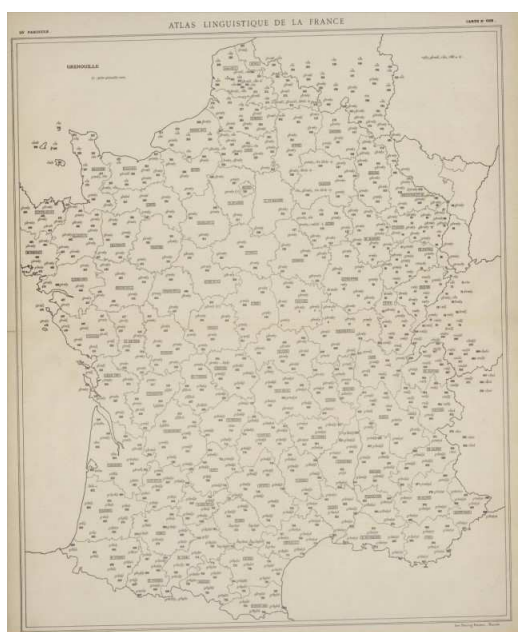
Today, standard French is dominant in all regions of France. Nevertheless, regional variation, dialects and patois characterize its linguistic landscape (Wolf, 1979: 165). This is illustrated, e.g., by the many dictionaries and surveys referenced by Lexilogos for French dialects. Attempts to revive regional varieties gave impetus to the creation of many linguistic atlases of France, beginning as early as 1897-1901 with the *Atlas linguistique de la France* – ALF (Gilliéron & Edmont, 1902–1910, Fig. 1a) and leading

¹⁵ <http://atilf.fr/dmf/definition/appreper> [01-03-2019].

to the many large-sized volumes of the series *Atlas linguistiques de la France par régions* – ALFR (Séguy, 1973: 78).

The language code for Modern French is ISO 639-1 ‘fr’. For the majority of French regional varieties, ISO 639 codes are not available, exceptions being ISO 639-3 ‘nrf’ for the Norman dialect¹⁶, ‘pcd’ for Picard, and ‘wln’ for Walloon.

Given the amount of linguistic resources with diatopic data for modern French, we have selected exemplary data, namely from dictionaries of different patois. We focus on one use



(a)

Reluaint, Reluot, Relure — divers temps du verbe Reluire. — Les étoiles reluaint bein lai neu dan le temps ; c'à signe de plieue beintot. — Ai lai bénédiction, es Vêpe, combien en aivo de cierges ! ci reluot que ci beillo lai beurlue. — Vos regaidez note baissin ? c'à de lai poudre qui ai aichetai que le fait si relure. — Voyez Lu, Luot.

Rémaginai (se) — réfléchir, changer d'idée. — A n'éto pâ diôre du pays qu'à sé rémaginai, et pu al é revenu. — Vos voiras, ailé ; lai neu vos vos rémaginerez ; et ci iré mieux.

Rembrun, Rembrunché ; air sournois, mécontent. — Al é in air to rembrunché àjedeu. — Al é son rembrun ; i ne sai pâ d'ou veint.

Remontrai — à peu près le même sens qu'Recordai
Renoille, Renouille, R'noille — grenouille. — En nô fau ailai pouâchai des renoilles. — Ai lai quoue de l'Etang en y é des renoilles, ci fait pô. — Ce n'a p'encore le Colas Mignotte qu'inventeré lai quoue es renouilles. — En certains cas, comme après un mot finissant par une voyelle, il faut prononcer ainsi : En y aivo ine eurnoille dan lai fontaigne. — Les petiotes eurnouilles faisant in bru... in bru ! en fau entende.

Rentaires — rente ; revenus en nature d'une rente ou maitairie ; ou bien prix du fermage, de l'amodiation. —

(b)

Figure 1: (a) ALF map n° 668 ‘grenouille’. (b) Denizot (1910: 120).

case: the designations for the frog. To model the data simply using ‘fr’ as the language code does not account for the linguistic reality in the regions in our focus: it would render the diatopic variation generic. BCP 47 specifies a region sub-tag that is typically used to indicate (diatopic or diastratic) variation within a country or territory, the standard being a code from ISO 3166. However, ISO 3166 registers administrative (sub-)divisions (in our case, *régions* and *départements* of contemporary France) whose boundaries do not necessarily match the language boundaries.¹⁷ Hence, we make use of the *privateuse* subtag and codes provided by Glottolog, e.g., for Burgundian in **E3** (‘bourg1247’), in line with the pattern in Table 1. However, the patois spoken in Burgundy (and in any other region) differ. It is thus necessary to further distinguish

¹⁶ Falsely described as “Guernésiais, Jèrriais” which excludes the continental area.

¹⁷ <https://tools.ietf.org/html/bcp47#section-2.2.4>;
<https://www.iso.org/obp/ui/#iso:code:3166:FR> [1106-2019].

the language tag on patois. We do this by adding the name of the location where the patois has been recorded. This can be (1) a region or (2) a place name.

To identify a language in a region (1), as a subset of the language denoted by the Glottocode, we use the latitude and longitude coordinates of the location provided by the geographical database GeoNames¹⁸ and we convert the coordinates into a Geohash¹⁹, where Geohash is a system for encoding geographic coordinates as a base32 string, in a syntax acceptable for BCP 47 (Gillis-Webber & Tittel, 2019: 4:10). To identify a place name (2) within the language tag, we refer to its equivalent entry in GeoNames.

3.3.1 Language of Burgundy

E3, from *Dictionnaire de patois de Mancey* (Millot (1905–1922 (edition 1998))):

```

1  @PREFIX pwn:          <http://wordnet-rdf.princeton.edu/id/> .
2
3  :gornaille a    ontolex:LexicalEntry , ontolex:Word ;
4  :rdfs:label "gornaille"@fr-x-01bour1247-342996271 ;
5  ontolex:canonicalForm :gornaille_lemma ;
6  ontolex:sense      :gornaille_sense ;
7  ontolex:evokes     :frog_lexConcept.
8
9  :gornaille_lemma a  ontolex:Form ;
10 ontolex:writtenRep  "gornaille"@fr-x-01bour1247-342996271 .
11
12 :gornaille_sense   a  ontolex:LexicalSense ;
13 ontolex:isLexicalizedSenseOf :frog_lexConcept .
14
15 :frog_lexConcept a  ontolex:LexicalConcept ;
16 ontolex:lexicalizedSense :gornaille_sense ;
17 ontolex:isConceptOf    dbpedia:Frog ;
18 ontolex:definition     "grenouille"@fr ;
19 dct:references pwn:01642406-n .

```

In our language tag on Lines 4 and 10, fr identifies the tag as from the Modern French period, with 01 indicating that the Glottocode for the Burgundy language is used. To

¹⁸ <https://www.geonames.org/> [07-06-2019].

¹⁹ <https://www.movable-type.co.uk/scripts/geohash.html> [07-06-2019].

identify the patois spoken in Mancey, a commune in the Saône-et-Loire *département*, we made use of the equivalent identifier from GeoNames, 2996271, prepending it with 34 as per Table 1.

E4, from the *Vocabulaire patois de Sainte-Sabine et ses environs (Côte-d’Or)* (Denizot (1910), Fig. 1b):

```

1  :renoille a ontolex:LexicalEntry , ontolex:Word ;
2  rdfs:label "renoille"@fr-x-00saintesabine-30u0g6r--
3  u0e36--u07zp--u0sbk--u0t5k--u0u4u ;
4  ontolex:canonicalForm :renoille_lemma ;
5  ontolex:sense :renoille_sense ;
6  ontolex:evokes:frog_lexConcept .
7
8  :gueurnouille_lemma a ontolex:Form ;
9  ontolex:writtenRep "renoille"@fr-x-00saintesabine-30u0g6r--
10 u0e36--u07zp--u0sbk--u0t5k--u0u4u .
11
12 :gueurnouille_sense a ontolex:LexicalSense ;
13 ontolex:isLexicalizedSenseOf :frog_lexConcept .

```

The use of GeoNames to identify the location of Sainte-Sabine, a commune in the Côte-d’Or *département*, would be a wrong approach for this case: the title of the resource clearly indicates that the vocabulary has been recorded in Sainte-Sabine and, also, within its vicinity. Unfortunately, the introduction of the resource gives only a vague description of what it means: “montagnes des environs des Pouilly-en-Auxois et de Blignysur-Ouche”, Denizot (1910: 14). We drew a polygon of the area that is, thus, only an approximation as well (Figure 2a). The geographic coordinates representing the polygon are: (49.62686,4.91473), (48.04287,4.66964), (47.6435,5.59192), (47.88325,6.85844), (48.40865,7.23867), (49.72584,5.81263), (49.62686,4.91473).

The last coordinate is the same as the first, and so we excluded the last one and then converted the latitude and longitude coordinates to a Geohash to a precision of five digits, cf. Gillis-Webber and Tittel (2019: 4:10f.): u0g6r--u0e36--u07zp--u0sbk--u0t5k--u0u4u. Lines 2-3 and 9-10 show the use of these Geohashes, with the pattern 00 defining the language as user-defined and 30 defining a geohashed polygon region.

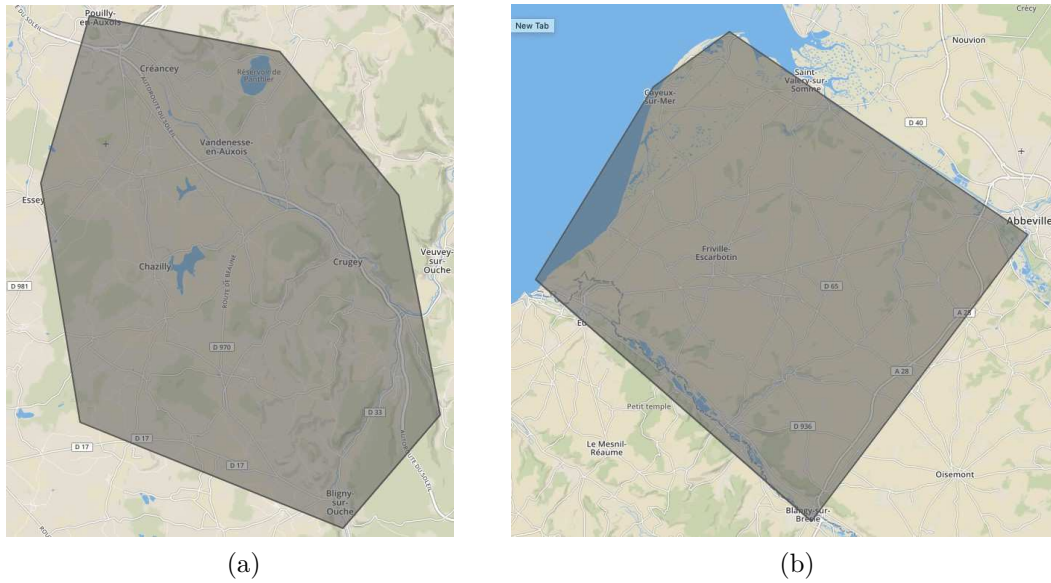


Figure2: (a) Approximate region where the patois of Sainte-Sabine was recorded. (b) Region of Vimeu in Picardy

3.3.2 Language of Picardy

E5, from Dictionnaire des parlers picards du Vimeu (Somme) (Vasseur (1998)):

```

1  :guernouille a    ontolex:LexicalEntry , ontolex:Word ;
2  rdfs:label
3  "guérnouille"@pcd-x-30u0cje--u0cj3--u0buz--u0chj--u0cm1 ;
4  ontolex:canonicalForm :guernouille_lemma ;
5  ontolex:sense      :guernouille_sense ;
6  ontolex:evokes     :frog_lexConcept .
7
8  :guernouille_lemma a    ontolex:Form ;
9  ontolex:writtenRep
10 "guérnouille"@pcd-x-30u0cje--u0cj3--u0buz--u0chj--u0cm1 .
11
12 :guernouille_sense a    ontolex:LexicalSense ;
13 ontolex:isLexicalizedSenseOf :frog_lexConcept .

```

In the language tag on Lines 3 and 10, the language code uses the ISO 639-3 code ‘pcd’ for the modern Picard language. To specify the region of Vimeu in Picardy (Fig. 2b), we have again defined a region, converted into Geohashes.

4. Modelling of regional variation using linguistic atlas data

We modeled a small set of exemplary data from the ALF. It seems clear to us that most of the regional differences manifested in a linguistic atlas concern phonetic variation. However, the regional particularities also concern the lexis, especially in border regions of France. These regions document phenomena of cultural and linguistic contact with other languages, e.g., with German, Franco-Provençal, Occitan, and Breton. These phenomena are of great interest, in particular to researchers in Historical Linguistics and Digital Humanities. With its rich lexical and phonetic data, an atlas could add significant value to the landscape of semantically accessible linguistic data sets.

For the transformation of linguistic atlas data into LD, the information on a map needs to be turned into points. This leads to two issues: dealing with (a) the geographic data acquisition points (which, in the context of ALF, is place names) and (b) the phonetic transcription indicated for each point.

For (a), Gally et al. (2013: 188f.) describe that they semi-automatically provided each of the 992 data acquisition points of the digitized ALF with geographic coordinates. For (b), typically, the data sources for the linguistic atlases are surveys where interviewees pronounced words and phrases and interviewers transcribed the phonetic realizations using a phonetic alphabet. For the ALF, Abbé Rousselot and Jules Gilliéron established a phonetic alphabet in 1891 which then was also used by the makers of the atlases of the series ALFR. The transcriptions were written onto the maps by hand. To ensure the structural interoperability of atlas data within the Semantic Web, the transcriptions need to be re-encoded using the standard *International phonetic alphabet* (IPA, International Phonetic Association, 2005), cp. Moran (2012) who uses IPA as an interlingual pivot for different transcription systems.

4.1 Exemplary data for Lorraine

We have used data from the ALF map n°668 (Fig. 1a). In **E6**, for the lexeme *grenouille* “frog”, we model the phonetic realizations of three acquisition points taken from the Meurthe-et-Moselle *département* in Lorraine (Table 2) using the `phoneticRep` property of the OntoLex-Lemon vocabulary.

- n° 162 (Sexey-les-Bois)
- n° 170 (Moncel-sur-Seille)
- n° 171 (Mailly-sur-Seille)

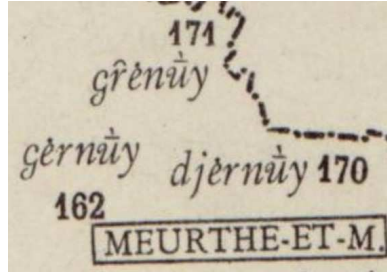


Table 2: Extract from ALF map n° 668.

E6, from *Atlas linguistique de la France* (Gilliéron & Edmont, 1902–1910):

```

1  :grenouille  a   ontolex:LexicalEntry , ontolex:Word ;
2  rdfs:label    "grenouille"@fr ;
3  ontolex:canonicalForm :grenouille_lemma ,
4  ontolex:sense    :grenouille_sense ;
5  ontolex:evokes   :frog_lexConcept .
6
7  :grenouille_lemma  a   ontolex:Form ;
8  ontolex:writtenRep "grenouille"@fr ;
9  ontolex:phoneticRep "gK@nu-:j"@fr-fonipa-x-01lorr1242-342996683 ,
10                             "g@rnu-:j"@fr-fonipa-x-01lorr1242-342974669 ,
11                             "dZ@rnu-:j"@fr-fonipa-x-01lorr1242-342993415 .
12
13 :grenouille_sense  a   ontolex:LexicalSense ;
14 ontolex:isLexicalizedSenseOf :frog_lexConcept .

```

In Lines 9-11, we have re-encoded the phonetic transcription (*cf.* Table 2) using IPA characters. To identify the phonetic characters of the string literals, we include the subtag *fonipa*, which is compliant with BCP 47 (Phillips & Davis, 2009: 43). In the *privateuse* portion, 01 indicates a code from Glottolog has been used. As with **E3**, the place name for each geographic acquisition point has been represented by its equivalent GeoNames identifier, prepended with 34. E.g., the phonetic representation of the lexeme recorded in Sexey-les-Bois (n° 162, Line 10) is identified as 2974669.²⁰

²⁰ <http://www.geonames.org/2974669/sexey-les-bois.html> [06-06-2019].

5. Discussion

Revisiting the CQs, all questions, with the exception of **CQ6**, are answerable with the available data from our case study.

CQ1 is answered by **E1–E4** and **E6**. For **E1** and **E2**, codes exist in alternative directories, but they do not reflect the correct time periods. Hence, we opted to identify the language using a user-defined code, indicated by 00 from Table 1. **CQ2** is, thus, also answered by these two exemplars. For **E3**, **E4** and **E6**, a Glottocode is available, indicated by 01 from Table 1.

CQ3 is answered by our Modern French exemplars. Although different language codes are available for Modern French in each ISO 639 part, we make use of ‘fr’ from ISO 639-1; as per the BCP 47 specification, the shortest language code available has to be used.

CQ4 is answered by **E3–E6** showing two solutions: (1) **E3** and **E6** make use of an identifier from GeoNames, indicated by 34 from Table 1, (2) **E4** and **E5** both make use of a user-defined language (defined with pattern 00) and of Geohashes that represent the geographic coordinates for a polygon shaped region (defined with pattern 30 and with -- serving as an internal delimiter between each Geohash). A detailed description of associating a geographic area with a language is discussed in Gillis-Webber and Tittel (2019), which also addresses **CQ5**.

Although the pattern allows for a more precise definition of the language in question, for **E4** and **E5** the language tags intuitively feel too long: the Geohash, while useful, is opaque, and may require further annotation in order to be human-readable. While the proposed pattern serves as an interim solution for language-tagging lesser-known or less-discussed languages, the problem still remains that the dependency of a language tag on an ISO standard or registry is a flaw of language tags and the RDF specification. As an alternative to a language tag, we should be able to encode a URI in the vein of "jannaie"@deaf:fro/gallo, where deaf: is the namespace.

Gillis-Webber and Tittel (2019) suggest exploring the creation of a sub-datatype for `rdf:langString`, which would thus allow for the datatype URI to be encoded, as an alternative to the language tag. However, doing this presents challenges. A literal consists of two elements: a lexical form and a datatype URI (Cyganiak et al., 2014). If the datatype URI is `http://www.w3.org/1999/02/22rdf-syntax-ns#langString`, then a third element is introduced to the literal: namely “a non-empty language tag as defined by BCP 47”, *ib.* All other datatype URIs are mapped to RDF-compatible XSD types, none of which would allow the introduction of a custom URI in the place of a language tag, *ib.* To allow for an alternative datatype URI, the RDF specification would have to be amended. However, as a sub-datatype of `rdf:langString`, the constraints of BCP 47 would still apply. It thus seems easier to propose a change to BCP 47: namely to

allow, for the *privateuse* sub-tag only, the following characters: [-:/a-zA-Z0-9]. This would then render a language tag of the form "jannaie"@x-deaf:fro/gallo. To be RDF-compatible, the namespace for x-deaf: would have to be defined in the same RDF document in which the language tag is used.

We considered creating a user-defined simple XML Schema datatype, as a restriction on an existing datatype (Carroll & Pan, 2006). Although it would not render a language tagged string literal, it would render a string literal with an encoded URI: "jannaie"^^<http://example.org/simpleTypes#froGallo>. However, the URI, although it clearly identifies the language, would not be dereferenceable which is in opposition to one of the principles of LD. Furthermore, it would not be appropriate for use when modelling data using Ontolex-Lemon because the latter requires `rdf:langString` when representing forms. This leads us to conclude that Part 4 is required in our pattern, i.e., for the inclusion of a URI shortcode in the *privateuse* portion of a language tag, which can then be mapped to a URI.

Apart from the question of how to design the language tags, a further question arises: is the granularity of our approach sufficient for the following scenarios? The language of a linguistic resource, e.g., a text or a dictionary, is written:

1. during a time span or covering a time span, e.g., a collection of 19th century legal documents or a dictionary covering several centuries such as the DEAF,
2. at different times, e.g., the *Roman de la Rose* that consists of two parts (ca.1230; ca.1275) by two authors²¹,
3. in different places or covers several places, some parts (in) region A, some parts (in) Region B.

The scenarios describe multilingual settings that require multilingual labels (a part of the RDF standard²²). Scenarios 1 and 2 can be answered with the range of Part 2 of our pattern. For scenario 3, two questions arise: how to identify (a) the language(s) of a triple subject (a lexicon, a lexical entry, etc.), and (b) the language(s) of a literal.

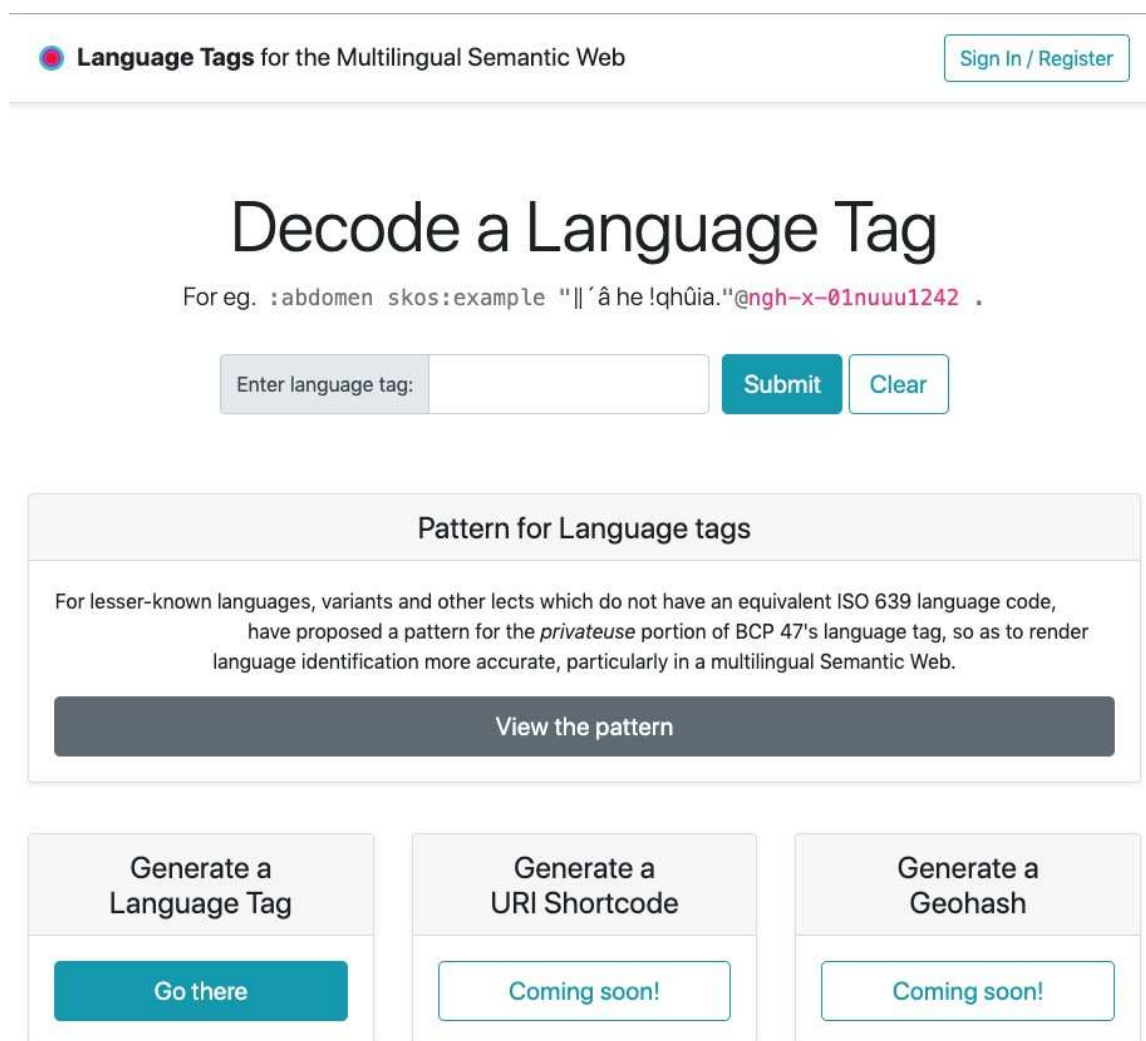
Question (a) is answerable with the property `dct:language` that has multiple values, such as `<http://example.org/language-1>` and `<http://example.org/language-2>` respectively (cp. **E0** with both ISO 639-1 and ISO 639-3 code). Question (b) is answerable with multiple literals, i.e., duplicated language-tagged literals for the same subject and predicate, with a custom language tag for each.

²¹ <http://www.deaf-page.de/bibl/bib99r.php#RosellLangl> [11-06-2019].

²² https://www.w3.org/community/bpmlod/wiki/Best_practises_-_previous_notes [12-06-2019].

6. Interface for Language Tag Generation

A user interface and REST API to both generate and decode language tags, currently in development, is to be demonstrated at eLex 2019. Language tags can be generated according to our pattern. For the decoding of language tags, the results are available in JSON, with natural language, RDF/XML and Turtle syntax to follow. Figure 3 shows the user interface. See <https://londisizwe.org/language-tags/> for more information.



Language Tags for the Multilingual Semantic Web [Sign In / Register](#)

Decode a Language Tag

For eg. :abdomen skos:example "||'â he !qhûia."@ngh-x-01nuuu1242 .

Enter language tag:

Pattern for Language tags

For lesser-known languages, variants and other lects which do not have an equivalent ISO 639 language code, have proposed a pattern for the *privateuse* portion of BCP 47's language tag, so as to render language identification more accurate, particularly in a multilingual Semantic Web.

[View the pattern](#)

Generate a Language Tag

[Go there](#)

Generate a URI Shortcode

Coming soon!

Generate a Geohash

Coming soon!

Figure 3: User interface for generating and decoding language tags.

7. Conclusions & Future Work

In this paper, we have discussed how to create language tags when modelling linguistic data as LD for languages for which ISO 639 does not provide language codes. We have focused on linguistic resources of French that are of interest for diatopic studies, and we have chosen exemplary data with a diachronic view, including Old-, Middle- and Modern French dictionaries and a Modern French linguistic atlas. For each exemplar, we have created a language tag, in line with a proposed pattern. These language tags

identify the language, its historical language stage, a subset of the language (dialect or patois) in an unambiguous way. Using a URI shortcode, the language tags can be reduced to a more user-friendly length. This, however, makes them opaque, whereas the former is more descriptive but can be long. While the use of encoded URIs affects human-readability, it remains machine-readable nonetheless.

Extension towards MoLA. In collaboration with C. Maria Keet, the authors have been working on MoLA, a Model for Language Annotation (Gillis-Webber et al., 2019). MoLA is a lightweight ontology which allows for languoids (a language family, language, dialect cluster, or lect) to be represented in RDF. Due to its expressiveness, including MoLA in the modelling of linguistic resources enables comprehensive language information to be represented. Future work is, thus, to model the languages identified in these French resources using MoLA.

Other Resources. We conclude the paper returning to linguistic desiderata: Other linguistic atlases (of the series ALFR, e.g., Lanher et al., 1979–1988 [Lorraine Romane]; Dondaine & Dondaine, 1972–1991 [Franche-Comté]) and dictionaries should be evaluated for a future conversion to LD. Valuable dictionaries comprise those covering particular patois and dialects, the comprehensive dictionary of French regionalisms (Rézeau, 2001), etc. The modelling of lexicologically rich resources of other kinds is a further task, including a lexicographer’s standard work for historic botany, the *Flore populaire de la France...* (Rolland, 1896–1914), and corpora, e.g., the *Corpus Historique du Substandard Français* (CHSF, Thun, 2011).

8. Varieties of French

Figures 4 and 5 show the designations of French varieties, the corresponding Glottocodes and ISO 639-3 codes, respectively. We define the lists of Old French varieties given by the FEW (von Wartburg (1922–: *Beiheft* p.63)) and by the DEAF as authority lists and exclude all regional varieties listed by other resources (e.g., Lexilogos) that are not covered by the FEW- or the DEAF list.

Modern French / FEW	Old French / FEW	Old French / DEAF	Glottolog (modern)	ISO 639-3 (modern)
français moderne	—	français moderne	stan1290	fra
—	ancien français	ancien français	—	fro *
—	moyen français	moyen français	mid1316	frm *
—	—	francien	—	—
pik.	apik.	picard	pica1241 **	pcd
hain.	—	hennuyer	hain1252	—
art.	—	artésien	arto1238	—

wallon	awallon.	wallon	wall1255	wln
lütt.	alütt.	liégeois	—	—
nam.	anam.	—	—	—
flandr.	aflandr.	français de la Flandre française	—	—
Lille	alill.	—	lill1247	—
champ.	achamp.	champenois	—	—
lothr.	alothr.	lorrain	lorr1242	—
norm.	anorm.	normand	norm1245	nrf
—	agn.	anglo-normand	angl1258	xno *
hbret.	—	haut-breton	gall1275	—

* Historical language stage. ** 12 sub-languages incl. ‘hain1252’, ‘arto1238’, ‘lill1247’.

Figure 4: List of French varieties, part 1 (terms in French).

Modern French / FEW	Old French / FEW	Old French / DEAF	Glottolog (modern)	ISO 639-3 (modern)
ang.	—	angevin	ange1244	—
poit.	apoit.	poitevin	poit1240	—
saint.	—	saintongeais	sant1407	—
tour.	—	tourangeau	—	—
orl.	—	orléanais	—	—
bourbonn.	abourb.	bourbonnais	bour1246	—
bourg.	abourg.	bourguignon	bour1247	—
Lyon **	—	lyonnais	lyon1243 ***	—
frcomt.	afrcomt.	franc-comtois	fran1262 ***	—
—	—	franco-italien	—	—
—	—	Nord-Est	—	—
—	—	Nord	—	—
—	—	Nord-Ouest	—	—
—	—	Ouest	—	—
—	—	Sud-Ouest	—	—
centr.	—	Centre	—	—
—	—	Est	—	—
—	—	Sud-Est	—	—
—	—	Terre Sainte	—	—
—	judfr.	Judeofrançais	—	zrp *

* Historical language stage. ** Sub Savoy. *** Sub Francoprovençal.

Figure 5: List of French varieties, part 2 (terms in French).

9. References

- ATILF – CNRS & Université de Lorraine (2015). *Dictionnaire du Moyen Français, version 2015 (DMF 2015)*. Paris. URL <http://www.atilf.fr/dmf/>. Accessed: 17-06-2019.
- Baldinger, K., Möhren, F. & Städtler, T. (1971–). *Dictionnaire étymologique de l'ancien français (DEAF)*. Québec / Tübingen / Berlin: Presses de L'Université Laval / Niemeyer / De Gruyter. DEAFél: <https://deaf-server.adw.uni-heidelberg.de/>.
- Berners-Lee, T. (2006). *Linked Data*. World Wide Web Consortium. URL <https://www.w3.org/DesignIssues/LinkedData.html>. Accessed: 17-06-2019.
- Berschin, H., Felixberger, J. & Goebel, H. (2008). *Französische Sprachgeschichte*. Hildesheim / Zürich / New York: Olms.
- Bizer, C., Heath, T. & Berners-Lee, T. (2009). Linked Data – The Story So Far. *International Journal on Semantic Web and Information Systems*, 5, pp. 1–22.
- Bouda, P. & Cysouw, M. (2012). Treating Dictionaries as a Linked-Data Corpus. In C. Chiarcos (ed.) *Linked Data in Linguistics. Representing and Connecting Language Data and Language Metadata*. Berlin/Heidelberg, Germany: Springer, pp. 15–23.
- Carroll, J. & Pan, J. (2006). XML schema datatypes in RDF and OWL: W3C Working Group Note 14 March 2006. URL <https://www.w3.org/TR/swbp-xsch-datatypes/>. Accessed: 17-06-2019.
- Chambon, J. P. (1997). Pour la localisation d'un texte de moyen français: le Mystère de Saint Sébastien. In G. Kleiber & M. Riebel (eds.) *Les formes du sens: Etudes de linguistique française, médiévale et générale offertes à Robert Martin à l'occasion de ses 60 ans*. Louvain-la-Neuve: Duculot, pp. 201–216.
- Chauveau, J. P. (2016). Régionalismes médiévaux et dialectismes contemporains en haute-Bretagne. In M. Glessgen & D. Trotter (eds.) *La régionalité lexicale du français au Moyen Âge*. Strasbourg: ÉLiPhi, pp. 131–166.
- Chiarcos, C., McCrae, J., Cimiano, P. & Fellbaum, C. (2013). Towards Open Data for Linguistics: Lexical Linked Data. In A. Oltramari, P. Vossen & L. Qin et al. (eds.) *New Trends of Research in Ontologies and Lexical Resources: Ideas, Projects, Systems*. Berlin / Heidelberg: Springer, pp. 7–25.
- Cyganiak, R., Wood, D. & Lanthaler, M. (2014). RDF 1.1. concepts and abstract syntax: W3C recommendation 25 February 2014. URL <https://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/>. Accessed: 17-06-2019.
- de Melo, G. (2015). Lexvo.org: Language-Related Information for the Linguistic Linked Data Cloud. *Semantic Web*, 6(4), pp. 393–400.
- Denizot, J. (1910). *Vocabulaire patois de Sainte-Sabine et ses environs (Côte-d'Or)*. Beaune: Imprimerie Beaunoise.
- Dondaine, C. & Dondaine, L. (1972–1991). *Atlas linguistique et ethnographique de la Franche-Comté (ALFC)*. Paris: Éd. du CNRS.

- Gally, S., Chauvin-Payan, C. & Davoine P. A. et al. (2013). GéoDialect : Exploration des outils géomatiques pour le traitement et l'analyse des données géolinguistiques. *Géolinguistique*, 14, pp. 186–208.
- Gillis-Webber, F. & Tittel, S. (2019). The Shortcomings of Language Tags for Linked Data when Modeling Lesser-Known Languages. In *Proceedings of LDK2019, Leipzig, Germany, 21-22 May 2019, OASIs, Vol. 70*. pp. 4:1–4:15.
- Gillis-Webber, F., Tittel, S. & Keet, M. (2019). A Model for Language Annotations on the Web. In B. Villazón-Terrazas & Y. Hidalgo-Delgado (eds.) *Knowledge Graphs and Semantic Web. 1st Iberoamerican Conference, KGSWC 2019, Villa Clara, Cuba, June 23-30, 2019, Proceedings*. pp. 1–16.
- Gilliéron, J. & Edmont, E. (1902–1910). *Atlas linguistique de la France*. Paris: Champion.
- Glessgen, M. & Trotter, D. (2016). *La régionalité lexicale du français au Moyen Âge*. Strasbourg: ÉLiPhi.
- Gleißgen, M. D. & Thibaut, A. (2005). La «régionalité linguistique»: essai définitoire. In M.D. Gleißgen & A. Thibaut (eds.) *La lexicographie différentielle du français et le Dictionnaire des régionalismes de France*. Presses Univ. de Strasbourg, pp. III–XVII.
- International Organization for Standardization (n.d.). Language codes – ISO 639. URL <https://www.iso.org/iso-639-language-codes.html>. Accessed: 17-02-2019.
- International Phonetic Association (2005). International Phonetic Alphabet. Tech. rep. URL <https://www.internationalphoneticassociation.org/>. Accessed: 17-02-2019.
- Lanher, J., Litaize, A. & Richard, J. (1979–1988). *Atlas linguistique et ethnographique de la Lorraine Romane (ALLR)*. Paris: Éd. du CNRS.
- Millot, C. (1905–1922 (edition 1998)). *Dictionnaire de patois de Mancey*. Tournus: Société des amis des arts et des sciences de Tournus.
- Moran, S. (2012). Using Linked Data to Create a Typological Knowledge Base. In C. Chiarcos (ed.) *Linked Data in Linguistics. Representing and Connecting Language Data and Language Metadata*. Springer, pp. 129–138.
- Möhren, F. (2016). La régionalité dans le DEAF – historique et programme. In M. Glessgen & D. Trotter (eds.) *La régionalité lexicale du français au Moyen Âge*. Strasbourg: ÉLiPhi, pp. 37–50.
- Phillips, A. & Davis, M. (2009). Tags for Identifying Languages. *BCP*, 47. URL <https://tools.ietf.org/html/bcp47>. Accessed: 17-06-2019.
- Renders, P. (2016). La régionalité lexicale du moyen français (1350–1500). In M. Glessgen & D. Trotter (eds.) *La régionalité lexicale du français au Moyen Âge*. Strasbourg: ÉLiPhi, pp. 85–96.
- Rickard, P. (1974). *A history of the French language*. London: Hutchinson University Library.
- Rolland, E. (1896–1914). *Flore populaire de la France ou histoire naturelle des plantes dans leurs rapports avec la linguistique et le folklore*. Paris: Rolland.
- Rézeau, P. (ed.) (2001). *Dictionnaire des régionalismes de France. Géographie et histoire d'un patrimoine linguistique*. Bruxelles: De Boeck.

- Rézeau, P. (2007). *Richesses du français et géographie linguistique*. Bruxelles: De Boeck & Larcier.
- Rézeau, P. (2016). La régionalité lexicale du français après 1500, à travers des régionalismes recueillis dans les correspondances de poilus. In M. Glessgen & D. Trotter (eds.) *La régionalité lexicale du français au Moyen Âge*. Strasbourg: ÉLiPhi, pp. 111–130.
- Séguy, J. (1973). Les Atlas linguistiques de la France par régions. *Langue Française*, 18, pp. 65–90.
- Thun, H. (2011). Die diachrone Erforschung der *français régionaux* auf der Grundlage des *Corpus Historique du Substandard Français*. In C. Schlaak & L. Busse (eds.) *Sprachkontakte, Sprachvariation und Sprachwandel*. Narr, pp. 359–394.
- Tittel, S. (2016). La régionalité lexicale de l'ancien français (ca.1100 – ca.1350) : Une enquête sur la base du *Dictionnaire étymologique de l'ancien français*. In M. Glessgen & D. Trotter (eds.) *La régionalité lexicale du français au Moyen Âge*. Strasbourg: ÉLiPhi, pp. 61–84.
- Tittel, S., Bermúdez-Sabel, H. & Chiarcos, C. (2018). Using RDFa to Link Text and Dictionary Data for Medieval French. In J. P. McCrae, C. Chiarcos & T. Declerck et al. (eds.) *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). 6th Workshop on Linked Data in Linguistics (LDL-2018), 12 May 2018, Miyazaki, Japan*. Paris: ELRA, pp. 30–38.
- Tittel, S. & Chiarcos, C. (2018). Historical Lexicography of Old French and Linked Open Data: Transforming the Resources of the *Dictionnaire étymologique de l'ancien français* with OntoLex-Lemon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). GLOBALEX Workshop (GLOBALEX-2018), Miyazaki, Japan, 2018*. Paris: ELRA, pp. 58–66.
- Vannini, L. & Le Crosnier, H. (2012). *Net.lang. Towards the multilingual cyberspace*. Caen: C & F Éditions.
- Varlet, M. (1896). *Dictionnaire du patois meusien*. Verdun: Société Philomathique de Verdun.
- Vasseur, G. (1998). *Dictionnaire des parlers picards du Vimeu (Somme), avec index français-picard*. Fontenay-sous-Bois: SIDES.
- von Wartburg, W. (1922–). *Französisches Etymologisches Wörterbuch (FEW)*. Bonn, Heidelberg, Leipzig/Berlin, Basel: ATILF. [Continued by O. Jänicke, C. T. Gossen, J. P. Chambon, J.-P. Chauveau, and Yan Greub].
- Wolf, H.J. (1979). *Französische Sprachgeschichte*. Heidelberg: Quelle u. Meyer.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



Challenges for the Representation of Morphology in Ontology Lexicons

Bettina Klimek¹, John P. McCrae², Julia Bosque-Gil³,
Maxim Ionov⁴, James K. Tauber⁵, Christian Chiarcos⁴

¹ Institute for Applied Informatics (InfAI), Leipzig University

² Data Science Institute, National University of Ireland Galway

³ Ontology Engineering Group, Universidad Politécnica de Madrid

⁴ Goethe-Universität Frankfurt am Main

⁵ Open Greek and Latin Project

Abstract

Recent years have experienced a growing trend in the publication of language resources as Linguistic Linked Data (LLD) to enhance their discovery, reuse and the interoperability of tools that consume language data. To this aim, the OntoLex-*lemon* model has emerged as a *de facto* standard to represent lexical data on the Web. However, traditional dictionaries contain a considerable amount of morphological information which is not straightforwardly representable as LLD within the current model. In order to fill this gap a new Morphology Module of OntoLex-*lemon* is currently being developed. This paper presents the results of this model as on-going work as well as the underlying challenges that emerged during the module development. Based on the MMoOn Core ontology, it aims to account for a wide range of morphological information, ranging from endings to derive whole paradigms to the decomposition and generation of lexical entries which is in compliance to other OntoLex-*lemon* modules and facilitates the encoding of complex morphological data in ontology lexicons.

Keywords: morphology; RDF; OntoLex-*lemon*; MmoOn; inflection; derivation

1. Introduction

Morphology is a vital and, in many languages, very sophisticated part of language, and as such it has been an important part of the work of lexicographers. In the traditional print form, morphological information is provided in brief abbreviated terms that can only be deciphered with significant knowledge of the language, however with the transformation of the dictionary to an electronic resource a re-imagining of the morphology information in a dictionary is certainly due. We base our work within the framework of the ontology-lexicon (McCrae et al., 2012; Cimiano et al., 2014) and in particular in that of the OntoLex-*lemon* model. This model has been used not only for the conversion of existing dictionaries (Khan et al., 2017; Borin et al., 2014; Bosque-Gil et al., 2015) but also for the development of new dictionaries (Gracia et al., 2017) as Linked Data (Chiarcos et al., 2013).

In this paper, we present the current modelling as well as the underlying challenges within the development of the Morphology Module for OntoLex-*lemon*, which extends the existing work by providing modelling for representing the morphology that is associated with the entries. In many cases, morphology is an important part of the language, for example in both German and Irish noun plurals are irregular and cannot be predicted from the stem alone, so many dictionaries, especially learners' dictionaries, list these irregular forms for most or all of the entries. Further, for languages such as the Romance ones, verbs may have many forms that are frequently irregularly or semi-irregularly derived, and learners' dictionaries for these languages also list many forms. However, as electronic dictionaries become of use not only to humans but also machines, it is necessary to provide all forms in a manner that can be readily processed by the latter. To this end, the Morphology Module covers not only the description of some forms of a lemma, but also allows the generation of all forms through morphological patterns, which corresponds to the idea of declensions or conjugations of an entry. Further, we base our model on the MMoOn Core ontology (Klimek, 2017), which has been designed to more generally represent morphology as a linguistic domain, and as such this module can handle a wide range of linguistic phenomena including distinctions between derivational and inflectional morphology, allomorphy, suppletion, simulfixes and transfixes among others. Moreover, this module is, as its name suggests, part of the overall model of OntoLex-*lemon* and as such can be integrated well with other parts of OntoLex-*lemon* and is consistent with its other semantic and syntactic modules.

The rest of this paper is structured as follows. In Section 2, we provide an example based illustration of the shortcomings of morphological data representation in traditional dictionaries. In Section 3 we provide background of the OntoLex-*lemon* model for readers, who are not familiar with it, which is followed by an overview of related work in Section 4. We then present the challenges of representing morphology within the OntoLex-*lemon* framework in Section 5 before presenting the current modelling state of our proposed model in Section 6. Finally we look into the further improvements that we plan for the module in Section 7, and present some conclusions in Section 8.

2. Morphological data in dictionaries and lexical databases

The treatment of morphology in dictionaries is a complex topic which is related to the lexicographic selection process (or lemma selection) (Schierholz, 2015), and the definition of the micro-structure of entries, i.e., the data model upon which the description (Hartmann, 2001) and layout (Atkins & Rundell, 2008) of each entry will be based, with different types or 'templates' being also considered, e.g. a typical noun-entry type (Abel, 2012).

Opacity, frequency and predictability of form and meaning in words were aspects that had to be considered when deciding whether a complex lexeme or compound word should be contained in a dictionary or not (De Caluwe & Taeldeman, 2003), but

dictionaries and lexicographic traditions, in general, vary substantially. For example, derivational affixes have often received main entry status, with differences from dictionary to dictionary in their description: from dictionaries that identify them just as suffixes, to dictionaries that also point to their derivational or inflectional use (Alsina & DeCesaris, 1998).

Different approaches to lexicography also play a role in these various representations of morphological data. Linguistics-oriented dictionaries, guided by a linguistic theory for morphology and its terms, contrast with function-theoretic based (or communicative) works which are focused mainly on the morphological information needs of users in specific situations (Swanepoel, 2015; Bergenholtz & Tarp, 2005).

This context leads to a heterogeneous landscape when it comes to analysing the morphological description provided in dictionaries. Most traditional dictionaries do not cover morphological information extensively: usually, the morphological description of the lexical entry is limited to the list of the word forms that allow users to identify the morphological pattern to which the entry adheres, and hence generate the paradigm by themselves. Following this, word-forms that can be formed regularly are not listed. Moreover, the description of these ‘reduced’ inflection lists is often minimal on the assumption of users being familiar with the lexicographic tradition of the object language. For example, users of a German dictionary familiar with the German language easily interpret the description ***Na** · **me** der; -ns, -n* to refer to the gender of the entry, and its genitive singular and nominative plural endings. Other dictionaries, such as The K Dictionaries Multilingual Global Series¹, provide groups of word-forms inflected for case and number, along with the ending that is displayed in the user interface, as illustrated in Example 1.1.

This is similarly the case for Ancient Greek dictionaries, where noun entries will typically list the nominative singular form, the genitive singular ending, and the article (indicating the gender). This assumes the reader is able to work out the stem by comparing the nominative form with the abbreviated genitive ending. This, in combination with the gender, is then generally enough to produce other forms of the nominal paradigm. Additional forms of the noun are generally not given in the entry unless deemed impossible or non-obvious to produce from the standard information given.

For verbs it also very common to find verbal paradigms as a reference in the appendix of dictionaries. For example, Figure 1 shows the paradigm of the verb *amar* ‘to love’ as an example of a verb that inflects according to the 1st conjugation pattern in Spanish². Even though such tables contain all forms of a lemma, the underlying morphological

¹ <https://www.lexicala.com/resources#dictionaries>

² <http://www.rae.es/diccionario-panhispanico-de-dudas/apendices/modelos-de-conjugacion-verbal#advertencias>, last accessed on 05.06.2019.

structure separating the stems from the regular and productive inflectional suffixes remains again implicit.

```

<HeadwordBlock>
  <HeadwordCtn>
    <Headword>Stipendiat</Headword> [...]
    <GrammaticalGender value="masculine" />
    <InflectionBlock>
      <InflectionCtn>
        <Inflection>Stipendiaten</Inflection>
        <Display>-en</Display>
      </InflectionCtn>
      <InflectionCtn>
        <Inflection>Stipendiaten</Inflection>
        <Display>-en</Display>
      </InflectionCtn>
    </InflectionBlock>
  </HeadwordCtn>
  <HeadwordCtn>
    <Headword>Stipendiatin</Headword> [...]
    <GrammaticalGender value="feminine" />
    <InflectionBlock>
      <InflectionCtn>
        <Inflection>Stipendiatin</Inflection>
        <Display>-</Display>
      </InflectionCtn>
      <InflectionCtn>
        <Inflection>Stipendiatinnen</Inflection>
        >
        <Display>-nen</Display>
      </InflectionCtn>
    </InflectionBlock>
  </HeadwordCtn>
  <PartOfSpeech value="noun" />
</HeadwordBlock>

```

Example 1.1: An extract of the entry *Stipendiat* ‘scholarship holder’ from the K Dictionaries Global Series German Dictionary.

1. AMAR		Verbo modelo de la 1.ª conjugación		
INDICATIVO				
TIEMPOS SIMPLES				
presente	pret. imperfecto / copretérito	pret. perfecto simple / pretérito	futuro simple / futuro	condicional simple / pospretérito
amo	amaba	amé	amaré	amaría
amas (amás)	amabas	amaste	amarás	amarías
ama	amaba	amó	amará	amaría
amamos	amábamos	amamos	amaremos	amaríamos
amáis	amabais	amasteis	amaréis	amaríais
aman	amaban	amaron	amarán	amarían
TIEMPOS COMPUESTOS				
pret. perfecto compuesto / antepresente	pret. pluscuamperfecto / antecopretérito	pret. anterior / antepretérito	futuro compuesto / antefuturo	condicional compuesto / anteupospretérito
he amado	había amado	hube amado	habré amado	habría amado
has amado	habías amado	hubiste amado	habrás amado	habrías amado
ha amado	había amado	hubo amado	habrás amado	habría amado
hemos amado	habíamos amado	hubimos amado	habrá amado	habríamos amado
habéis amado	habíais amado	hubisteis amado	habrá amado	habríais amado
han amado	habían amado	hubieron amado	habremos amado	habrían amado
			habréis amado	
			habrán amado	

Figure 1: Table of the inflectional paradigm of the verb *amar* ‘to love’ from the *Diccionario Panhispánico de Dudas* (Real Academia Española and Asociación de Academias de la Lengua Española, 2005).

From the examples just illustrated, it becomes clear that all the common approaches regarding the representation of morphological data rely highly on the implicit knowledge of the dictionary user about the language. As a consequence, morphological data varies greatly concerning their amount, their way of representation and interconnection to the relevant element they are contained in, i.e. the lemma or a form in a paradigm.

3. Overview of OntoLex-lemon

The OntoLex-lemon model³ has been under development for several years and was originally based on the combination of the three pre-existing models (LingInfo (Buitelaar et al., 2006), LexOnto (Cimiano et al., 2007), LIR (Montiel-Ponsoda et al., 2011)) that were combined into a single model (lemon) by the EU project Monnet and later extended into the OntoLex-lemon model by the Ontology Lexicon Community

³ The full specification can be consulted here: <https://www.w3.org/2016/05/ontolex/>.

Group⁴. This model was developed around five basic principles: 1) it would be an RDF model that used the Web Ontology Language (OWL) (McGuinness, Van Harmelen, et al., 2004) for its semantics; 2) it would support multilinguality and avoid language-specific assumptions that might affect the applicability of the model to other languages; 3) it would use the principle of ‘semantics by reference’ as a basic semantic model (Cimiano et al., 2013); 4) it would embrace openness in being free of any financial costs or licensing as well as allowing contributions from any interested party, and 5) relevant standards and models would be reused wherever appropriate. This led to the core model that is depicted in Figure 2, which is based around a lexical entry, composed of a number of forms and a number of senses, which can then be linked to either lexical concepts or entities in an ontology.

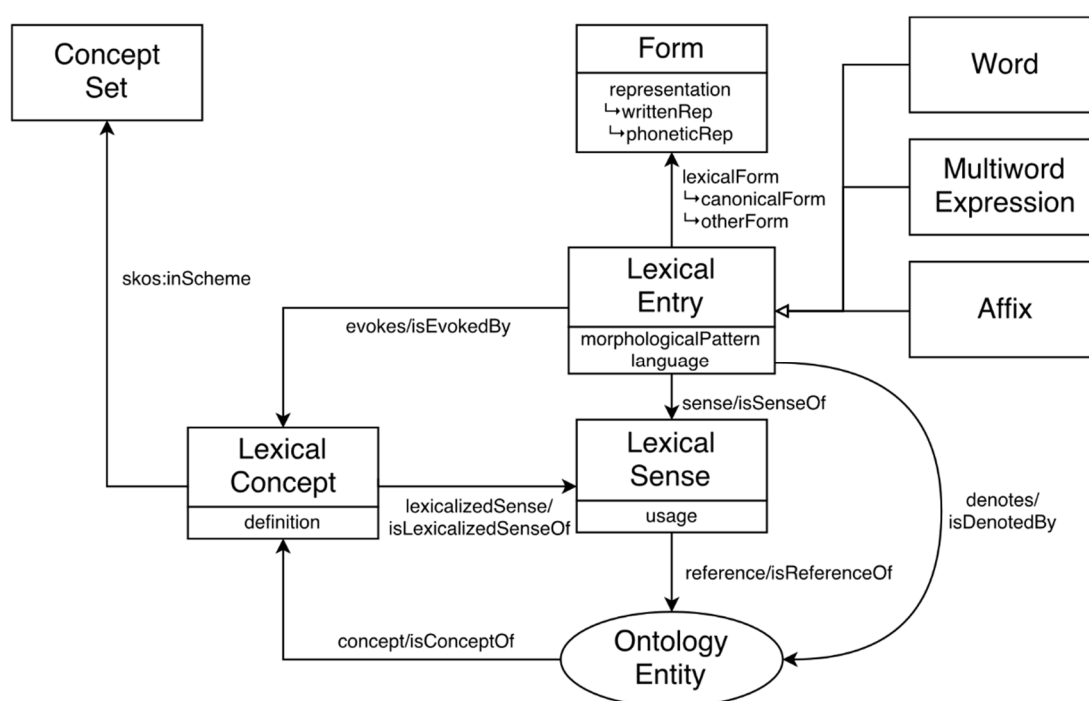


Figure 2: The core model of OntoLex-lemon.

In addition to this core, that is often also called “ontolex”, there were four further modules developed in the initial release of the model:

Syntax and Semantics (synsem) This module describes how syntactic frames may be modelled and how they can be mapped to ontology structures,

Decomposition (decomp) The decomposition of multiword expressions and compound terms is described by this module,

Variation and Translation (vartrans) Modelling of translations and other kinds of

⁴ <https://www.w3.org/community/ontolex/>

relations are provided by this module,

Linguistic Metadata (lime) This module provides metadata about the lexicon and the ontology and how this may be used to encourage interoperability between resources.

In addition, since then the group has continued to develop modules to extend the usefulness and applications of the model. One such extension, the recently released Lexicography Module (Bosque-Gil et al., 2017), has provided features for representing dictionaries in ways that are more compatible with traditional print dictionary forms. Other modules are in development, in particular this one along with a module for representing frequencies, attestations and corpus information⁵, and a module for etymological and diachronic information (Khan, 2018).

Since its development, the OntoLex-*lemon* model has been extensively used for representing a vast amount of different lexical data: In addition to traditional dictionary data mentioned in Section 1, it has been applied to lexical databases like WordNet (McCrae et al., 2014), etymological resources (Chiarcos et al., 2016; Khan, 2018), and domain-specific lexicons (Bellandi et al., 2018).

4. Related work

The emerging OntoLex-*lemon* Morphology Module described in this paper aims to enable the representation of the morphological elements and processes that are involved in the decomposition and generation of lexical data (of both lexemes and their word-forms) by overcoming the representational limitations of traditional dictionaries as outlined in Section 2 and within the technical realm and the design principles of the overall OntoLex-*lemon* model introduced in the previous section. Since the emergence of the (multilingual) Semantic Web in the early 2000s, several ontologies emerged from the lexicography, language resource and language documentation communities that already contain the modelling of morphological language data to some extent. Here we briefly describe some of these ontologies that are considered the most relevant with regard to the morphological data they allow to represent, together with an explanation to what extent they could or why they could not be reused within the OntoLex-*lemon* Morphology Module.

In the early development of the OntoLex-*lemon* model, its priorities have been on lexicalizing ontologies and knowledge bases. This was accompanied by a natural focus on lexical semantics, i.e., multilingual labels for the same concept, and, here, the original contribution of Monnet-Lemon, the predecessor of OntoLex-*lemon* has been to complement such labels with morphosyntactic information in order to facilitate context-adequate lexicalization. Morphology was only considered in the form of morphosyntax, i.e. inflectional features as well as the possibility to provide the adequate form for these.

⁵ <https://acoli-repo.github.io/ontolex-frac/>

The current OntoLex-*lemon* representation of morphological information can complement ontology concepts with morphosyntactic categories (part of speech, a property of a lexical entry), and provide different forms with different morphosyntactic features (e.g., gender, case, number, etc.) Neither derivational morphology nor morphological information beyond the specification of grammatical features was expressible with this model, and lexicalizations of the same concept with different parts of speech required independent lexical entries, without being able to represent the systematic relations on the level of form and meaning that hold between them.

OntoLex-*lemon* does not provide any vocabulary of grammatical features, instead, it endorses the reuse of the existing ontologies and vocabularies for linguistic annotations, most notably, ISocat, GOLD, OLiA, and LexInfo. ISocat, a shared repository for linguistic concepts, features and data structures, was developed as a successor of the ISO Data Category Registry (DCR), originally designed as an RDF-based knowledge graph (Ide & Romary, 2004) and is built on XML technologies and resolvable URIs (Kemps-Snijders et al., 2009). ISocat was a semistructured resource populated in a bottom-up process, so that it did not provide formal and consistent vocabulary, but its subsets became an important source of knowledge that more consolidated domain vocabularies described here drew from. GOLD, one of the first attempts in creating a linguistic ontology (Farrar & Langendoen, 2003), and OLiA (Chiarcos & Sukhareva, 2015) were designed primarily as solutions to harmonize linguistic categories and make markup schemes interoperable. In OLiA this is achieved by linking the hierarchy of abstract grammatical categories which constitutes the reference model with specific markup schemas that can vary for resources and languages.

Despite their interoperability and applicability to a vast amount of linguistic data, these ontologies are primarily focused on providing labels for the categories and lack the expressibility to represent morphosyntactic information.

LexInfo is an inventory containing various types, values and properties to describe linguistic categories (Cimiano et al., 2011). It is partially derived from ISocat and is often used to represent linguistic annotations in OntoLex-*lemon* (however, this is not a requirement). Even though it covers certain aspects of morphology, it has a focus on inflectional morphology whereas it lacks expressiveness in describing derivational morphology.

Finally, the last relevant model is the MMoOn Core ontology⁶ (Klimek et al., 2016). It is currently the only existing comprehensive domain ontology for the linguistic area of morphological language data. As such it is highly specialized and far more-fine grained than the desired modelling of the OntoLex-*lemon* Morphology Module requires. It contains, among other aspects, an extensive modelling of linguistic meanings, including derivational meanings in addition to grammatical categories. It also differentiates

⁶ <https://mmoan.org/core>

between morph and morpheme resources and comes with a set of nearly 300 morphemic glosses to provide sufficient expressivity to represent morphological data contained in Flex or Toolbox datasets. At the same time, a specification of lexical data is not provided in MMoOn Core because this ontology was envisaged to be used complementary to *OntoLex-lemon*. Therefore, there is only one existing interconnection of the two domain ontologies so far, i.e. an established subclass relation between the two classes `mmoon:LexicalEntry` and `ontolex:LexicalEntry`. A more extensive ontology alignment has been thus far only proposed from the MMoOn Core perspective (Klimek, 2017) and might be considered for future implementation. Once the *OntoLex-lemon* Morphology Module will be officially released, further alignment options might be realized. Even though the MMoOn Core ontology exceeds by far the modelling needs of the Morphology Module, it served as a modelling template since the creation of MMoOn Core was initially motivated to fill the gap of representing morphological language data in *OntoLex-lemon* that still existed back then. So far, certain types of affix classes, e.g. `mmoon:Simulfix`) as well as the two object properties `mmoon:consistsOf` and `mmoon:meaning` have been reused in the *OntoLex-lemon* module, although only in an inspirational manner. These classes and properties are defined and integrated slightly differently within the morphology module and should not be confused as long as no explicit alignment has been implemented.

From this review of relevant existing ontologies it can be concluded that the emerging *OntoLex-lemon* morphology module adheres to the Semantic Web best practice of reusing existing vocabularies. Since none of the presented ontologies sufficiently satisfies the representation needs of morphological data in particular with regard to lexical data so far, the Morphology Module will adequately fill this gap. Furthermore, as a result of the outlined reuse choices, the Morphology Module could be kept user-friendly and manageable by replacing the usually necessary modelling of grammatical categories and morphological meanings of morph resources with the recommendation to use existing vocabularies instead, and also linguistically accurate because it is influenced by the more precise MMoOn Core domain ontology.

5. Challenges in developing a Morphology Module extension

Creating a descriptive modelling foundation for representing lexical data entails several design choices that directly affect the usability of the model. This does not only hold for ontology lexicons, but also for lexicon models in general. In what follows, challenges that arose during the development of the morphology module for *OntoLexlemon* will be outlined. With the ongoing development of modules, these issues gain increasing importance and can serve as orientation points of consideration for future module extension development efforts.

5.1 Scope and coverage

Description: The first question that arises when a new ontology is being created is who should use it for what purpose? As illustrated in Section 2, morphological information is highly implicit in the landscape of traditional dictionaries. However, along with the liberation from the limits of print dictionaries came almost unlimited possibilities of lexicographic data compilation in eLexicography, which are yet again broadened by the possibilities of the Linked Data paradigm. While some lexicographers only like to digitize a printed dictionary into Linked Data using RDF, others aim at transforming their already more fine-grained lexical databases and intend to use the resulting RDF dataset to generate more lexicographic content out of it, e.g. to generate inflectional paradigms including full word-forms together with the underlying morpho-phonological formation rules.

Modelling Choice: In line with *OntoLex-lemon* model, the Morphology Module also aims at being applicable for everyone working with lexicographic content who either focuses on the transformation of traditional dictionary data into RDF or on the conversion of more structured computational lexical data. Accordingly, the scope of the module is divided into two main parts: 1) enabling the representation of elements that are involved in the decomposition of lexical entries and word-forms, and 2) enabling the representation of building patterns that are involved in the formation of lexical entries and word-forms. A fine-grained description of phonological processes that are involved in any kind of stem or word formation on the phoneme level is, however, excluded and not representable with this Morphology Module. Only the elements between the lexical entry and the morph levels will be covered.

5.2 Consistency

Description: The *ontolex* and *decomp* modules of *OntoLex-lemon* already contain various classes and properties that can be used to describe morphological data. The *ontolex:Affix* and *decomp:Component* classes for instance already exist to represent sub-word units and can be put into relation to the lexical entries in which they are contained via properties like *decomp:correspondsTo* or *decomp:subterm*. Due to the widespread usage of *OntoLex-lemon*, the development of the Morphology Module is challenged with creating the necessary missing vocabulary by taking the existing classes and properties into account, while ensuring backwards compatibility at the same time.

Modelling Choice: Due to the incremental approach of developing the module for morphology and also future *OntoLex-lemon* extensions, it is inevitable to deal with overlapping existent vocabulary. Therefore, the *OntoLex Community Group* agreed to aim for the goal of reaching consistency by reusing as much of the existent vocabulary as possible and minimize duplication that results from creating similar classes and properties. Specifically, this entails that suitable existent vocabulary can be adapted as

long as the changes made are a) only additions to domain and range restrictions of properties or b) adaptations in the `rdfs:comment` description to broaden the applicability of classes. In this way, existing vocabulary can be coherently integrated into later developed modules while simultaneously preserving already established functionalities.

5.3 Terminological ambiguity

Description: During the module development process it turned out that one of the greatest challenges is to unambiguously define the terminology that is used to label the classes and properties of the new vocabulary. As intended, the widely set scope of the Morphology Module presented in Section 5.1 attracts the use of the module for various user groups which are, however, also coming from different terminological backgrounds. The understanding and usage of linguistic concepts like *morph* or *root* diverge considerably depending on whether the user of the module is, for example, a traditional linguist, a computer linguist or a lexicographer managing data for specific languages. This entails a high risk of an inappropriate usage of the ontological vocabulary that might result in an unintentional wrong data representation the user is generally not even aware of.

Modelling Choice: While the human-readable definition of ontology elements is defined within the `rdfs:comment`, the underlying machine-processable semantics are determined by implications and restrictions for an element and its relation to other elements of the ontology. For the computational processing of the data the former is not relevant, whereas the latter is formally fixed and unambiguous. What matters is the consistent usage of the vocabulary according to the ontologically defined semantics, notwithstanding that a user would have chosen a different label for an element. Moreover, providing a definition that is interpreted in the same way by all users is almost impossible. Therefore, the `rdfs:comment` descriptions of classes and properties are discussed and refined until the highest possible consensus is reached. In addition to that, the Morphology Module specification that will be published together with the release of the module contains usage examples and recommendations that support a shared understanding to ensure the consistent application of the module vocabulary.

6. Current state of the Morphology Module

6.1 Summary of the current state

The development of the Morphology Module is an ongoing joint effort by members of the OntoLex Community Group that started in November 2018. This paper presents the intermediate results which have been reached and the state of the module as of May 2019. The documentation creation process reflecting the discussions of the scope, identified representation needs and modelling steps can be consulted on the respective

OntoLex Wiki page⁷. It contains the outcomes as well as the links to the minutes of the regular calls that have been held.

So far, half of the defined scope for the Morphology Module (cf. Section 5.1) could be modelled. In particular this includes the first scope, i.e. the representation of the decomposition of `ontolex:LexicalEntry` and `ontolex:Form` resources. An overview illustrating the resulting model structure is shown in Figure 3. The second scope of representing the automatic generation of entries and forms from morph resources is still in an early development stage and, hence, will not be addressed in detail in this paper. The model in Figure 3 displays how the Morphology Module is embedded within the existing *OntoLex-lemon* vocabulary it relates to. Classes and properties written in blue indicate the new vocabulary that is specified with the prefix `morph` with the class `morph:Morph` building the centre of the module. The two object properties `decomp:subterm` and `decomp:correspondsTo` are also represented in blue, thus, highlighting that these are vocabulary elements that will have to be adjusted by extending their ranges (as explained in Section 5.2) to arrive at an overall *OntoLex-lemon* model consistency. It has to be noted that the presented Morphology Module is not officially published yet and, therefore, not usable at this current stage. However, it can be assumed that the vocabulary elements that are described in the next Section will remain very close to their final published module specification.

6.2 New classes and properties

In order to solve the presented challenges outlined in Section 5, new classes and properties had to be developed for the Morphology Module. Altogether eleven new classes and seven object properties have been implemented into the modelling so far. In doing so, central concepts of the domain of morphological data could be reused from the *OntoLex-lemon* vocabulary, and a considerable reduction of overlap between the new and the existing vocabulary could be reached. The `ontolex:Form` class, for instance, was already appropriate to represent all forms of a lexical entry, which are crucial elements for the description of the segmentation of words. Table 1 and Table 2 present an overview of the module vocabulary with the definitions and restrictions that have been defined for all new classes and properties.

The `morph:Morph` class builds the centre of the module and is divided into six subclasses. As a result it will be possible to specify root, stem and certain affix types. The prominent affixes, i.e. prefix, suffix, infix and circumfix, are, however not part of the vocabulary because these can be reused from other ontologies such as LexInfo. The treatment and function of the `ontolex:Affix` class was highly debated for its potential re-usability. Since this class is a subclass of `ontolex:LexicalEntry` it cannot be used to represent bound morphs that are inflectional, because those are usually not described

⁷ <https://www.w3.org/community/ontolex/wiki/Morphology>

as headwords in lexical databases or dictionaries. In order to avoid uncertainty within the classification of inflectional and derivational affixes, the `morph:AffixMorph` class has been created. Affixes that should be represented as lexical entries can be described with `ontolex:Affix`, whereas those that cannot should be described with the `morph:AffixMorph` class, regardless of their derivational or inflectional nature. Moreover, an explicit declaration for these two morphological functions has been enabled by providing the object property `morph:hasMorphStatus` and the class `morph:MorphValue` that already contains the two individuals `morph:inflectional` and `morph:derivational` ready for use.

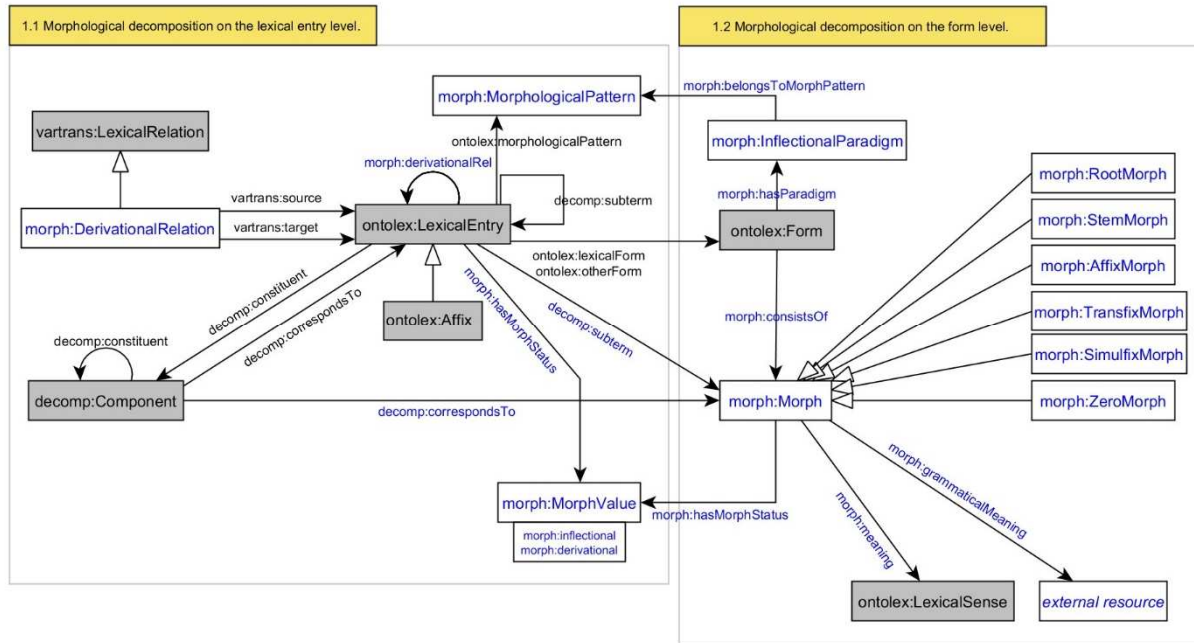


Figure 3: Current proposal of the Ontolex-*lemon* morphology module.

Since the derivational morphs of a derived lexical entry are now explicitly representable within the Morphology Module, a possibility to state that one derived lexical entry is derived from another lexical entry should be provided. This has been achieved by creating the class `morph:DerivationalRelation` that is defined as a subclass of `vartrans:LexicalRelation`. Therefore, it inherits the same domain and range restrictions which mean it can represent the direction of the derivational relation between two lexical entries, i.e. one can explicate that one derived lexical entry is derived by a specific derivational relation from another lexical entry. Furthermore, more generically all lexical entries that can be created through a derivational relation from another lexical entry can be expressed by using the object property `morph:derivationalRel`. Examples illustrating the use of this class and this property will be provided in Section 6.3.1.

Class Name	Definition	Class Relation
Morph	A morph is a concrete primitive element of morphological analysis.	owl:disjointWith ontolex:LexicalEntry
RootMorph	A morph that constitutes the semantic nucleus of a stem. It cannot be further segmented and is often not specified for a part of speech.	rdfs:subclassOf morph:Morph
StemMorph	The stem is the morph to which inflectional marking applies.	rdfs:subclassOf morph:Morph
AffixMorph	An affix is a bound segmental morph.	rdfs:subclassOf morph:Morph
TransfixMorph	A transfix is a discontinuous affix.	rdfs:subclassOf morph:Morph
SimulfixMorph	A simulfix is a bound morph that entails a change or replacement of vowels or consonants (usually vowels) which changes the meaning of a word, e.g. <i>eat</i> in past tense becomes <i>ate</i> .	rdfs:subclassOf morph:Morph
ZeroMorph	A morph that that corresponds to no overt form, i.e. orthographic or phonetic representation.	rdfs:subclassOf morph:Morph
MorphValue	The value of a morph states the relationship that holds between the morph and the forms or lexical entries in which it can occur.	class instances: morph:inflectional morph:derivational
DerivationalRelation	A 'derivational relation' is a lexical relation that relates two lexical entries by means of a derivational affix.	rdfs:subclassOf vartrans:LexicalRelation
MorphologicalPattern	The morphological pattern states the inflectional, derivational or compositional building pattern that applies to a lexical entry.	none
InflectionalParadigm	A structured set of inflected forms according to specific grammatical parameters.	none

Table 1: Overview of new classes of the Morphology Module.

With the foresight to enable also the automatic generation of `ontolex:LexicalEntry` resources from given `morph:Morph` and `ontolex:Affix` resources, the necessary conceptual frame has been modelled already. Figure 3 shows that the existing `ontolex:morphologicalPattern` object property was an initial proposal but remained under specified due to the non-existent Morphology Module at the point of its creation. This lack of expressivity has been now resolved by creating the two classes `morph:MorphologicalPattern` and `morph:InflectionalParadigm` which interrelate

ontolex:LexicalEntry and ontolex:Form within the graph structure of the module via the two established object properties morph:hasParadigm and morph:belongsToMorphPattern. Even though the specific usage of this part of the module is not sufficiently attested yet, the example for it provided in Section 6.3 illustrates the intended utilization.

As a central component of the morphological data domain the representation of the meaning of morph:Morph resources had to be modelled as well. Therefore, the two object properties morph:meaning and morph:grammaticalMeaning have been implemented in the module. The underlying concepts of morph:StemMorph and morph:RootMorph resources can be expressed by the former property by pointing to a ontolex:LexicalSense resource and the grammatical categories that are encoded in resources that represent grammatical morphs, usually bound affixes, can be expressed by pointing to an external resource. As already mentioned, the creation of an extensive modelling of possible linguistic categories has been considered to be out of scope for this module, and it is recommended to reuse existing vocabulary elements, e.g. from LexInfo, instead. The possible lack of a grammatical category in any existing ontology can be then compensated by using the morph:grammaticalMeaning property alternatively together with a newly created vocabulary.

Property Name	Definition	Restrictions
derivationalRel	The property relates two lexical entries that stand in some derivational relation.	domain: ontolex:LexicalEntry ontolex:LexicalEntry
consistsOf	This property states into which Morph resources a Form resource can be segmented.	domain: ontolex:Form morph:Morph
hasMorphStatus	The property states whether a morphological element functions as inflectional or derivational.	domain: morph:Morph, ontolex:Affix morph:MorphValue
hasParadigm	This property assigns a form to an inflectional paradigm.	domain: ontolex:Form morph:InflectionalParadigm
belongsToMorphPattern	This property assigns an inflectional pattern of a form as belonging to a morphological pattern of a lexical entry.	domain: morph:InflectionalParadigm morph:MorphologicalPattern
meaning	This property assigns a lexical sense to a morph resource.	domain: morph:Morph ontolex:LexicalSense
grammaticalMeaning	This property assigns a grammatical meaning to a morph resource.	domain: morph:Morph

Table 2: Overview of new object properties of the Morphology Module.

Finally, a relation was needed that states that an `ontolex:Form` resource consists of `morph:Morph` resources analogously to the `ontolex:constituent` object property that interrelates `ontolex:LexicalEntry` resources and `decomp:Component` resources. This relation manifests itself in the object property `morph:consistsOf` which is used to identify the segmentable morphs of inflected words, whereas `ontolex:constituent` can identify the lexical parts of derived or compounded words. By further extending the range of `ontolex:correspondsTo` and `ontolex:subterm` for the class `morph:Morph` it is even possible to identify inflectional affixes within complex lexical entries. This is a particularly useful functionality of the morphology module for many languages that involve the expression of an inflectional morph in the process of word-formation. German nominal compounds, for example, can consist of some linking morph that can be identified as a case marking morph (or depending on the underlying linguistic theory as a zero morph), e.g. as in *Haushalt-s-kasse*, ‘household-GEN-budget’.

6.3 Representing morphological decomposition

In what follows the usage of the introduced vocabulary of the Morphology Module will be illustrated by the example displayed in Figure 4. It shows the graph modelling evolving around the English noun *speaker*, including all the properties, classes and instances that are involved. For better understandability the graph is reduced to the representation of only one derived lexical entry, i.e. the adjective *speakerless* and only two word-forms of *speaker*, assuming that there are more. All boxes highlighted in yellow represent the new classes of the Morphology Module vocabulary.

6.3.1 On the lexical entry level

Looking at the resource `:lex_speaker_n` as the subject of this graph clarifies which morphological information can be explicated by creating the following statements:

- 1) It consists of two constituents which are `decomp:Component` resources which again can be said to correspond to another `ontolex:LexicalEntry` and a `morph:AffixMorph` resource, i.e. the verb `:lex_speak_v` and the derivational suffix `:suffix_er`. This suffix has been specified with the value `morph:derivational` and the `ontolex:LexicalSense` `:agentNominalizer`. This modelling indicates that in this example dataset this derivational suffix *-er* is explicitly not a lexical entry but could, however, be easily turned into one by changing its type assertion to `ontolex:Affix`.
- 2) It can be created with the morphological pattern `:pattern_CommonNouns`. As mentioned already, this is technically not implemented yet but it is intended to use the two `decomp:Component` resources `:component_speak` and `:component_er` for this purpose.

- 3) It can be linked to other lexical entries by using the `morph:derivationalRel` property in order to state which other derived words can be derived from `:lex_speaker_n`. This is, however, only a very generic statement but one that is often found in lexical or dictionary data.

Finally, the statement in 3) can be specified in a fourth statement by turning `:lex_speaker_n` into an object of a statement that describes it as the target of the derivational relation `:derivRel_speaker_AgentNoun`. While the property in statement 3) just states that there is some derivational relation between two `ontolex:LexicalEntry` resources, triples with a `morph:DerivationalRelation` instance in the subject position explicitly interlink the source lexical entry and the target lexical entry for which a unique derivational relation holds.

6.3.2 On the form level

The interconnection between lexical entries and the forms that can be built from them has been already established within *OntoLex-lemon* with the `ontolex:otherForm` property and has been, therefore, used in this example accordingly to relate the two forms `:form_speakers1` and `:form_speakers2` to the lexical entry `:lex_speaker_n`.

Considering these two instances as the subjects when consulting Figure 4 makes it possible to create the following statements about them:

- 1) They are both specified to belong to the inflectional paradigm `:paradigm_NounInflection`. This paradigm defines the grammatical form variants of the `ontolex:Form` resources, i.e. case and number, and is itself assigned to the overall building pattern `:pattern_CommonNouns` for `ontolex:LexicalEntry` resources that are nouns like `:lex_speaker_n`.
- 2) They are both segmentable into `morph:Morph` resources that are stated with the `morph:consistsOf` property. As it is clear from Figure 4, they both share the same `morph:StemMorph` resource but consist of two different `morph:SuffixMorph` resources.

In addition to that, the three morphs `:stem_speaker_n`, `:suffix_s1` and `:suffix_s2` can be further specified for their meanings by pointing to `ontolex:LexicalSense` instances and grammatical values for the linguistic category case reused from the *LexInfo* vocabulary. It is essentially due to this enabled decomposition chain that makes it possible to not only identify, specify and interrelate all meaningful sub-word units but also the lexical entries and forms contained in lexical data, that all these elements can be disambiguated and described within a dataset modelled with the Morphology Module and *OntoLex-lemon*.

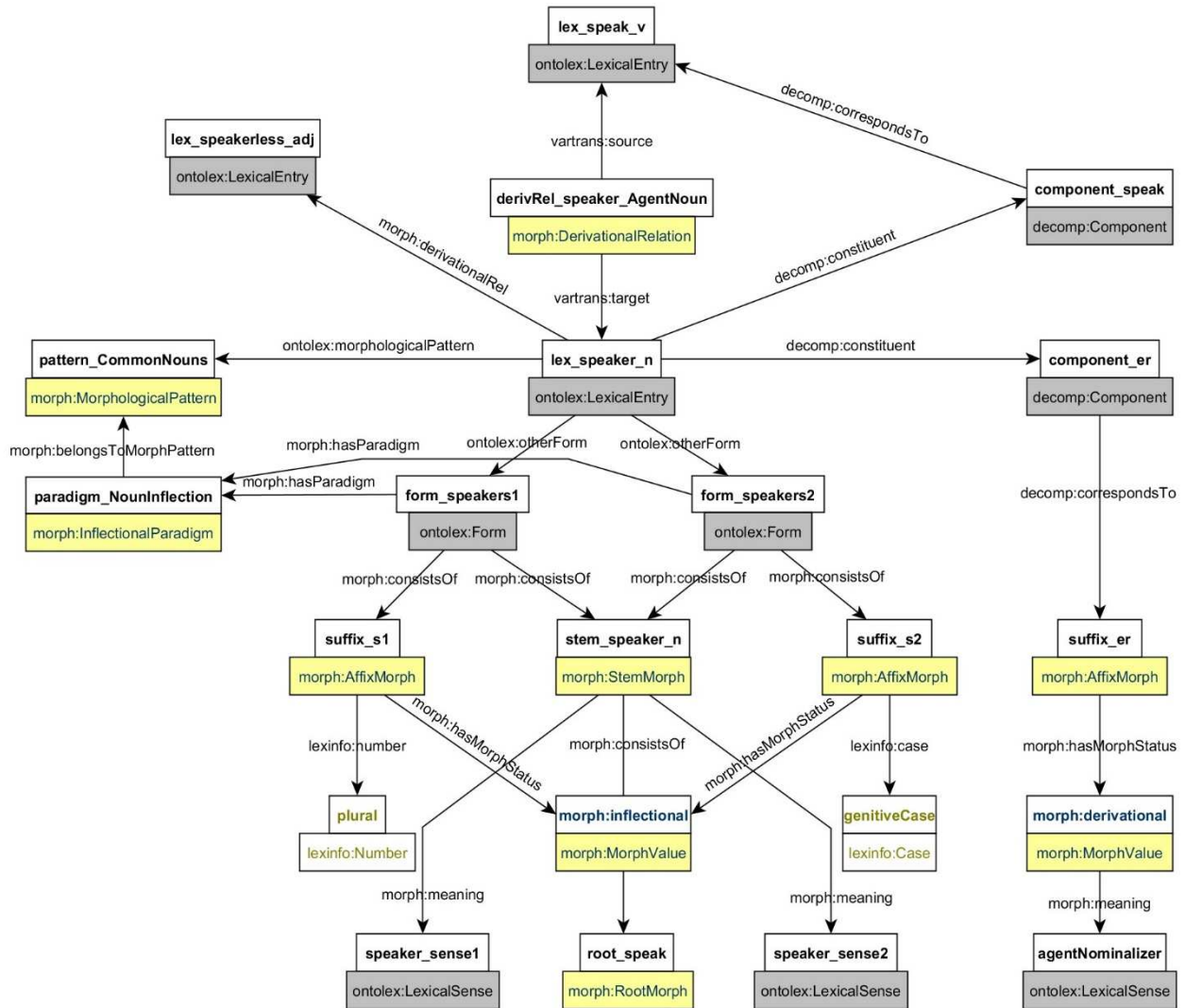


Figure 4: Graph representation for the example entry :lex_speaker_n.

7. Future work

Even though the modelling outcomes presented here have been largely agreed upon, several issues remain open for future work. Due to the various linguistic backgrounds of the OntoLex Community Group members some desired implementation options have been raised that might be still realized and included within the final Morphology Module specification. The following three features have been proposed for additional realization and are still under discussion:

- 1) **Morphemic glosses:** Since interlinear glossed text language data is an emerging source of lexical data that can be also represented in RDF, interest has been indicated to include the representation of morphemic glosses. So far it has been discussed if a modelling of glosses would exceed the scope of the Morphology Module, while the option to provide a shallow modelling with an

alignment to the MMoOn Core vocabulary that already provides a representation of glosses is also considered.

- 2) **Ordering:** For some highly polysynthetic and morphology-rich languages it is desirable to have a more precise representation of the internal morphological structure of lexical entries and forms. Therefore, it has been decided that a more expressive possibility for representing the position and ordering of morphs should be implemented to be available next to the currently used but very inexpressive `rdflist` object property. Proposals for that have been already made, but no agreement has been reached yet.
- 3) **Multiple segmentations:** Taking into account that a lexical dataset created based on the Morphology Module could be also applied in the context of computational linguistics, the processability of this data for machines might require the representation of more than one possible segmentation strategy. Allowing for the explication of that would be also interesting for linguists who want to document and analyse competing segmentations of words in their research.

In addition to these yet unrealized features it is necessary to focus on the refinement of the definitions of the newly created vocabulary elements. The exchanges within the community group have revealed that some of the presented `rdflist:comment` information is not precise enough and might lead to misunderstandings. In order to avoid misunderstandings in the usage of the vocabulary, time and attention will be invested again to resolve currently ambiguous or unclear definitions.

Furthermore, the second part of the Morphology Module that will enable the generation of forms with existing productive morphs in a dataset is also a part of the future work. However, the modelling is envisaged to produce lexical entries and forms based on patterns and paradigms, including also discontinuous morphs like transfixes and infixes. As it turned out in previous discussions such a formal representation is not trivial to model, especially with regard to the aim to be language-independently applicable.

8. Conclusion

To summarize, the current state of the Ontolex-*lemon* Morphology Module has been presented. The created vocabulary has been introduced and its usage illustrated. From that it becomes clear that the new module overcomes the limitations of the current representation of morphological data contained in traditional dictionaries by enabling the explication of formerly implicit information. With the Morphology Module modelled so far it is possible to represent the decomposition of lexical entries and forms with regard to both their derivational and inflectional morphs and underlying building patterns.

Furthermore, the challenges that arose from integrating the module into the existing

Ontolex-*lemon* model have been explained and design choices have been supported. It has been also shown that the module applies to existing Semantic Web standards by reusing relevant existing ontologies within its framework.

The remaining open issues have been presented and will be addressed in future work in order to arrive at the release of the final Morphology Module specification.

9. Acknowledgements

John McCrae is supported in part by a research grant from Science Foundation Ireland, cofunded by the European Regional Development Fund, for the Insight Centre under Grant Number SFI/12/RC/2289, as well as by the EU H2020 programme under grant agreements 731015 (ELEXIS - European Lexical Infrastructure) and 825182 (Prêt-à-LLOD).

Julia Bosque-Gil is supported by the Spanish Ministry of Education, Culture and Sports through the Formación del Profesorado Universitario (FPU) program.

Maxim Ionov and Christian Chiarcos are supported by the German Ministry for Education and Research (BMBF) through a project Linked Open Dictionaries (LiODi, 2015-2020) as a part of an Early Career Research Group on eHumanities.

10. References

- Abel, A. (2012). Dictionary Writing Systems and Beyond. In S. Granger and M. Paquot (eds.) *Electronic Lexicography*. Oxford: Oxford University Press. Chap. 5, pp. 86–106.
- Alsina, V. & DeCesaris, J. (1998). *Morphological structure and lexicographic definitions: The case of -ful and -like*.
- Atkins, B. T. S. & Rundell, M. (2008). *The Oxford guide to practical lexicography*. Oxford University Press.
- Bellandi, A., Giovannetti, E. & Weingart, A. (2018). Multilingual and Multiword Phenomena in a lemon Old Occitan Medico-Botanical Lexicon. *Information* 9.3, p. 52.
- Bergenholtz, H. & Tarp, S. (2005). Dictionaries and inflectional morphology. In *Encyclopedia of Language and Linguistics*. Pergamon Press, pp. 577–580.
- Borin, L. et al. (2014). Representing Swedish Lexical Resources in RDF with lemon. In: *International Semantic Web Conference (Posters & Demos)*. Citeseer, pp. 329–332.
- Bosque-Gil, J., Gracia, J. & Montiel-Ponsoda, E. (2017). Towards a module for lexicography in OntoLex. In: *DICTIONARY News* 7.
- Bosque-Gil, J., Gracia, J., Aguado-de-Cea, G. et al. (2015). Applying the ontalex model to a multilingual terminological resource. In *European Semantic Web Conference*. Springer, pp. 283–294.

- Buitelaar, P. et al. (2006). LingInfo: Design and applications of a model for the integration of linguistic information in ontologies. In *Proceedings of the OntoLex Workshop at LREC*.
- Chiarcos, C. & Sukhareva, M. (2015). Olia – ontologies of linguistic annotation. *Semantic Web* 6.4, pp. 379–386.
- Chiarcos, C., Abromeit, F. et al. (2016). Etymology Meets Linked Data. A Case Study In Turkic. In *Digital Humanities 2016, DH 2016, Conference Abstracts*. Krakow, Poland: Alliance of Digital Humanities Organizations (ADHO), pp. 458– 460. ISBN: 978-83-942760-3-4.
- Chiarcos, C., McCrae, J. et al. (2013). Towards open data for linguistics: Linguistic linked data. In *New Trends of Research in Ontologies and Lexical Resources*. Springer, pp. 7–25.
- Cimiano, P., McCrae, J. et al. (2013). “On the role of senses in the ontology lexicon”. In *New trends of research in ontologies and Lexical resources*. Springer, pp. 43–62.
- Cimiano, P., McCrae, J. & Buitelaar, P. (2014). *Lexicon Model for Ontologies: Community Report*. W3C Community Group Final Report.
- Cimiano, P., Buitelaar, P. et al. (2011). LexInfo: A declarative model for the lexicon-ontology interface. *Web Semantics: Science, Services and Agents on the World Wide Web* 9.1, pp. 29–51.
- Cimiano, P., Haase, P. et al. (2007). “LexOnto: A model for ontology lexicons for ontology-based NLP”. In *Proceedings of the OntoLex07 Workshop held in conjunction with ISWC’07*.
- De Caluwe, J. & Taeldeman, J. (2003). 2.5 Morphology in dictionaries. In P. van Sterkenburg (ed.) *A Practical Guide to Lexicography*. Vol. 6. John Benjamins Publishing, pp. 114–126.
- Farrar, S. & Langendoen, D. T. (2003). A linguistic ontology for the semantic web. *GLOT international* 7.3, pp. 97–100.
- Gracia, J., Kernerman, I. & Bosque-Gil, J. (2017). Toward linked data-native dictionaries. In I. Kosem et al. (eds.) *Proceedings of eLex 2017, Leiden, Netherlands*. Brno: Lexical Computing Ltd.
- Hartmann, R. R. K. (2001). *Teaching and researching lexicography*. Routledge.
- Ide, N. & Romary, L. (2004). A registry of standard data categories for linguistic annotation. In *4th International Conference on Language Resources and Evaluation-LREC’04*, pp. 135–138.
- Kemps-Snijders, M. et al. (2009). ISOcat: Remodeling metadata for language resources. *International Journal of Metadata, Semantics and Ontologies (IJMSO)* 4.4, pp. 261–276.
- Khan, F. (2018). Towards the Representation of Etymological and Diachronic Lexical Data on the Semantic Web. In J. P. McCrae et al. (eds.) *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA). isbn: 979-10-95546-19-1.

- Khan, F. et al. (2017). The Challenges of Converting Legacy Lexical Resources to Linked Open Data using Ontolex-Lemon: The Case of the Intermediate Liddell-Scott Lexicon. In: *LDK Workshops*, pp. 43–50.
- Klimek, B. et al. (2016). Creating Linked Data Morphological Language Resources with MMoOn - The Hebrew Morpheme Inventory. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.
- Klimek, B. (2017). Proposing an OntoLex - MMoOn Alignment: Towards an Interconnection of two Linguistic Domain Models. In *Proceedings of the LDK workshops: OntoLex, TIAD and Challenges for Wordnets*.
- McCrae, J., Fellbaum, C. & Cimiano, P. (2014). Publishing and Linking WordNet using lemon and RDF. In *Proceedings of the 3rd Workshop on Linked Data in Linguistics*.
- McCrae, J., Aguado-de-Cea, G. et al. (2012). Interchanging lexical resources on the Semantic Web. *Language Resources and Evaluation* 46.6, pp. 701–709.
- McGuinness, D. L., Van Harmelen, F. et al. (2004). *OWL: Web Ontology Language overview*. W3C recommendation.
- Montiel-Ponsoda, E. et al. (2011). Enriching ontologies with multilingual information. *Natural language engineering* 17.3, pp. 283–309.
- Real Academia Española and Asociación de Academias de la Lengua Española (2005). *Diccionario panhispánico de dudas*. Santillana Ediciones Generales.
- Schierholz, S. J. (2015). Methods in Lexicography and Dictionary Research. *Lexikos* 25, pp. 323–352.
- Swanepoel, P. H. (2015). The design of morphological/linguistic data in L1 and L2 monolingual, explanatory dictionaries: a functional and/or linguistic approach? *Lexikos* 25, pp. 353–386.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



Proto-Indo-European Lexicon and the Next Generation of Smart Etymological Dictionaries: The Technical Issues of the Preparation

Jouna Pyysalo¹, Fedu Kotiranta², Aleksi Sahala¹, Mans Hulden³

¹ University of Helsinki, Faculty of Arts, PL 24, 00014 Helsingin yliopisto

² Independent

³ University of Colorado Boulder, Department of linguistics, Hellems 290,
295 UCB Boulder, CO 80309

E-mail: jouna.pyysalo@helsinki.fi, fedu@mediamoguli.fi, aleksi.sahala@helsinki.fi,
mans.hulden@gmail.com

Abstract

Proto-Indo-European Lexicon (PIELex) is the generative etymological dictionary of Indo-European (IE) languages at <http://pielexicon.hum.helsinki.fi>. It is the first dictionary in the world capable of mechanically generating its data entries, i.e. the lexical stems of more than 120 of the most archaic IE languages. In addition, in order to solve the reverse process work has already begun on the problem of the mechanical generation of Proto-Indo-European (PIE) from the IE data,. The plan of the project as a whole is to run PIE Lexicon using an operating system (OS), a computer, under which the dictionary and its data are exclusively governed by smart features ranging from semantics to morphology, and the very root structure of Proto-Indo-European itself.

In principle PIE Lexicon is compatible with all digitized etymological dictionaries of IE languages, and as the operating system is scientifically neutral, material of any language or language family can be implemented onto the platform. By outlining the key features of the future coding plan we hope to offer ideas, assistance and support for other enterprises in the field of electronic lexicography.

Keywords: Indo-European linguistics; Proto-Indo-European; electronic lexicography; finite-state technology; historical linguistics

1. General introduction to PIE Lexicon

An etymological dictionary deals with at least two genetically related languages, and is therefore smart by default when compared to dictionaries of a single language. The Indo-European (IE) language family is one of the largest in the world, comprising some 400 languages. This naturally increases the complexity at the outset, as the preserved inherited data appear in mutually incompatible native writing systems. This problem is solved by means of the comparative method of reconstruction, a procedure that allows arranging etymologically related items into correspondence sets and projecting them back into the unitary phoneme system of a single language, Proto-Indo-European.

Furthermore, the IE languages are usually attested in several successive chronological phases. This entails additional complex requirements, the most important of them being that since the older the language is, the fewer changes it has undergone, it is necessary to start the reconstruction with the oldest form of every language in order to optimize the output. The full addition of the later layers becomes possible once these preconditions have been met.

Initially PIE Lexicon will be dealing with perhaps some 150-200 languages, mostly representing the oldest or a middle period in the written history of the languages, but also already including modern ones when the language is attested only in two periods such as, for instance, Lithuanian and Russian.

The etymological entries of PIE Lexicon, an example of which is shown in Figure 1, are of the following form:

PIE √hai- (vb.) ‘glänzen, brennen’			(IEW 11-12 *ai-)
√hai-			(HEG A:3)
PIE *h ₂ ai-	Pal. haa-	(vb.) ‘heiß, warm sein’	(DPal. 53)
√hain-			
PIE *h ₂ ai-	Pal. haan-	(pt.n.) ‘warm, heiß’	(DPal. 53)
PIE *h ₂ ai-	gAv. ayān-	(n.) ‘Tag’	(AIWb. 157)
√hair-			(IEW 12)
PIE *h ₂ ai-	gAv. ayar-	(n.) ‘Tag’	(AIWb. 157)
PIE *h ₂ ai-	Arm. aire-	(vb.) ‘verbrennen, anzünden’	(ArmGr. 1:418-9)
PIE *h ₂ ai-ino-	Hitt. ḫir-ina-	(UDUNm.) ‘Schmelzofen’	(HEG H:237)
PIE *h ₂ ai-h ₂ airos-	LAv. uz-īrah-	(n.) ‘Nachmittag’	(AIWb. 410)

Figure 1: An etymological entry of PIE Lexicon

The topmost horizontal line (in bold) starting with PIE √hai- (vb.) ‘glänzen, brennen’ and ending with (IEW 11-12 *ai-) represents a Proto-Indo-European root with a reference to earlier research. The root and its extensions (PIE √h₂ai-, √h₂ain- √h₂air-) are morphologically arranged as nodes of the root.

The PIE Lexicon data entries, consisting of a PIE reconstruction (e.g. PIE *h₂ai-) and the respective IE stem, (e.g. Pal. haa-), the morphological classifier of the IE stem ‘(vb.)’, translation (‘heiß, warm sein’),¹ and the reference ‘(DPal. 43)’ are arranged under the nodes from which they were originally derived.

¹ Note that in the initial version of PIE Lexicon the translations are those provided in the quoted source (usually a dictionary). In addition to this, future versions of PIE Lexicon will provide translations in several main languages, initially at least German and English.

2. Mechanical generation of the Indo-European data from PIE

In traditional (non-digital) etymological dictionaries the PIE reconstructions and the proto-phoneme system are not necessarily explicit. Furthermore, the sound laws leading from PIE to the IE languages are not always evident, and sometimes they are even inconsistent. In short, the entire traditional reconstruction is more or less intuitive, to a degree necessitating scholars to take leaps of faith instead of allowing them to rely on robust proofs by digitized sound laws.

In contrast to the traditional etymology, PIE Lexicon uses an explicitly defined PIE proto-phoneme inventory shown in Figure 2:

*o	*e	*a	*h	*i	*k	*l	*m	*n	*p	*r	*s	*t	*u
*ō	*ē	*ā	*ḥ	*ī	*g	*ļ	*ṁ	*ṇ	*b	*ṛ	*z	*d	*ū

Figure 2: The PIE phoneme inventory of PIE Lexicon

In the reconstruction these and only these phonemes are allowed, which blocks the use of ad hoc-phonemes.² The fact that the set is sufficient to reconstruct the IE forms proves the completeness of the PIE phoneme inventory.³

The most archaic IE sound laws, revised in Pyysalo (2013) have been digitized with the foma finite-state-compiler developed by Mans Hulden (2009).⁴ In practice this means that the non-formal sound laws used by the rest of the field have been replaced with their foma counterparts, 800 unique sound laws having been coded at this point. For illustration's sake, the loss of PIE *ḥ/h as a segmental phoneme is coded with the following two rules:

define Rḥ>0 ḥ -> 0 || .#. | \Stop _ ; define Rh>0 h -> 0 || .#. | \Stop _ ;

In order to facilitate the mechanical generation of the IE stems the individual sound laws coded in foma have been arranged in chronological order for each language, forming the sound law system of that language in digitized form. These sound law (foma) scripts can in turn be used to mechanically generate the actual forms of the language from their respective PIE reconstructions. The sound law scripts, as far as coded, can be found in the control bar at the bottom of the PIE Lexicon site. By

² For the revised PIE phoneme inventory used in PIE Lexicon, a further revision of Szemerényi (1967), see Pyysalo (2013).

³ For the completeness (i.e. sufficiency in the generation of the IE data) of the phoneme inventory, see Pyysalo, Sahala and Hulden (2018).

⁴ For the latest version of *foma*, see <https://code.google.com/archive/p/foma/>.

clicking ‘Select rule set’, choosing one (e.g. gAv.) and clicking ‘Show rules’, the respective sound law script is opened:

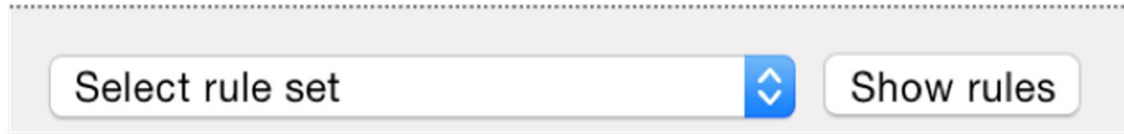


Figure 3: The control bar access to PIE Lexicon sound law scripts

By now some 120 of the most archaic IE languages have been provided with a sound law script in PIE Lexicon, and new scripts are constantly added as new languages emerge when new data is published. The sound laws provably form a consistent system and generate the IE data with an accuracy rate exceeding 99% (see Pyysalo, Hulden & Sahala 2018), strongly suggesting that the system is valid, i.e. sound and complete.

Due to the availability of the sound law scripts the PIE Lexicon operating system mechanically generates the IE stems (output) from their PIE reconstructions (input). PIE Lexicon editors, users, and visitors can explicitly verify the mechanical generation of the data by clicking a reconstruction (in blue). This is a command for the code reader to execute the foma script and create an explicit foma proof chain consisting of successive, explicitly stated sound laws leading from the PIE reconstruction to the respective IE stem, as shown in Figure 4:

PIE * <i>h₂ai-</i>	Pal. <i>ḥaa-</i>	(vb.) ‘heiß, warm sein’	(HEG A:3) (DPal. 53)
1. PIE * <i>h₂ai-</i>	PIE * <i>a</i> → Ø	Loss of * <i>a</i>	Ra>0 → * <i>hojo-</i>
2. <i>hojo-</i>	PIE * <i>o</i> → <i>a</i>	Change of * <i>o</i> into <i>a</i>	Ro>a → * <i>haja-</i>
3. <i>haja-</i>	* <i>aia</i> → <i>aa</i>	Loss of * <i>i</i> between * <i>a</i> and * <i>a</i>	Raja>aa → * <i>haa-</i>
4. <i>haa-</i>	PIE * <i>h₂/h̥</i> → Hitt. <i>ḥ</i>	Orthographic change of * <i>h₂/h̥</i> into <i>ḥ</i>	RH>ḥ → Pal. <i>ḥaa-</i>

Figure 4: An example of a foma proof chain in PIE Lexicon

When the output form has been generated, an additional function of the operating system (OS) compares the output to the actual stem form, and if these match, the letters of the attested form are shown in black as in the previous screenshot. If, on the other hand, any phoneme is erroneously generated, the error is shown in red in the attested form (Figure 5).

PIE * <i>ha₂pahont-</i>	LAv. <i>āfant-</i>	(a.) ‘wasserreich’	(AIWb. 330)
1. PIE * <i>ha₂pahont-</i>	PIE * <i>aē</i> → <i>aā</i>	Pyysalo’s rule for PIE * <i>aē</i>	Raē>aā → * <i>hāpahont-</i>
2. <i>hāpahont-</i>	PIE * <i>a</i> → Ø	Loss of * <i>a</i>	Ra>0 → * <i>hāphont-</i>
3. <i>hāphont-</i>	PIE * <i>o</i> → <i>a</i>	Change of * <i>o</i> into <i>a</i>	Ro>a → * <i>hāphant-</i>
4. <i>hāphant-</i>	PIE * <i>h</i> → Ø	Loss of segmental * <i>h</i>	Rh>0 → * <i>āphant-</i>
5. <i>āphant-</i>	* <i>ph</i> → <i>f</i>	Affricativization of * <i>ph</i>	Rph>f → LAv. <i>āfant-</i>

Figure 5: An example of an error (in red) in foma proof chain

All errors have been collected on a separate ‘mismatch’ page at the address <http://pielexicon.hum.helsinki.fi/?alpha=ALL&view=mismatch>. Although about half of the currently listed errors are typos or result from a necessary rule have not yet been coded, there are some 200 errors forming a dozen (or so) open research (sound law) problems to be solved.

Finally, and as particularly relevant to lexicography, the capability of the operating system to generate the Indo-European languages from the PIE phoneme inventory reduces the some 150 IE languages which are to be treated into a single, uniform language to manage, an advantage readily understood by anyone familiar with the complexities of lexicography in an environment requiring the treatment of a relatively large set of languages.

3. The automatic generation of PIE on the basis of Indo-European data

The second most challenging problem of historical linguistics in language technology after the automatic generation of IE data from PIE discussed above involves the mechanical reconstruction of the proto-language (here: PIE) and the definition of etymologies based on the attested data (here: IE). With regard to this problem there are two main solutions available, the original (traditional) and the recently emerged digital one. These ultimately represent the same process, that of reversing the order of the historical sound changes that have taken place during the development of a language and, based upon this, engineering a decision method allowing for the identification of originally identical Indo-European forms and their etymologies.

The traditional decision method of Indo-European etymology was originally outlined by August Schleicher. In Schleicher’s (1852b: iv-v) words, quoted here in Koerner’s (1982: 24) translation:

“When comparing the linguistic forms of two related languages, I firstly try to trace the forms to be compared back to their probable base forms, i.e., that structure [gestalt] which they must have [had], excepting phonetic laws [lautgesetze] which became effective at a later time, or at least I try to establish identical phonetic situations in historical terms for both of them.”

In modern terminology the identification of a PIE prototype and its reconstruction is based on creating a disjunction of possible PIE prototypes of an Indo-European morpheme. This disjunction, in turn, is compared to the similar disjunctions of other Indo-European languages, and if a formal match that is also semantically acceptable is found between two disjunctions, then an etymology (and a reconstruction) has been found.

This procedure is a decision method in a mathematical sense, i.e. it leads to the solution

if sufficient data have been preserved. For this reason the comparative method has proven its worth in allowing scholars to reconstruct the proto-forms of the discovered correspondences, simultaneously settling their etymologies.

The attempts to mechanize the reconstruction (here: PIE) have been unsuccessful up to this day, and have by now been largely abandoned and replaced by AI-based attempts to identify the processes involved (see Sims-Williams 2018). In the case of the Indo-European languages, the reason for the failure does not lie in the decision method or in its digitized formulation, the latter equally functional as the former, but in an imperfect set of sound laws leading from IE to PIE. If this (or any similar) set does not actually represent a consistent system of historical sound laws, then the system does not yield correct reconstructions, because the decision method essentially consists of reversing the sound laws, allowing the back-projection of the PIE prototypes mentioned by Schleicher. This can be seen from the digitized version of the method, consisting in essence of the following steps:

- a) The order of the sound law (foma) scripts is reversed so that the first rules become the last ones and the last ones become the first.
- b) In addition, the direction of the individual sound laws of the scripts, basically implications of the form ‘if X, then Y’, is also reversed, i.e. each rule $X \rightarrow Y$ is turned into $Y \rightarrow X$.

This reversing of the sound law scripts makes it possible to generate digital counterparts of Schleicher’s disjunctions, except for the fact that the code reader lacks the common sense applied in the intuitive use of the method. Without this the code reader generates infinite chains of phonemes, especially lost ones. In order to eliminate the problems related to this it is necessary to add morphophonological constraints to the code that exclude impossible prototypes such as †hhhhhhhhhhhep-.

Once the morphophonological constraints have been added to the reversed sound law scripts, their output is in essence identical with the intuitively used decision method, i.e. the algorithm generates disjunctions of possible PIE prototypes for the IE forms used as input. At this point it is possible to code and implement the decision method function, basically an intersection seeking identities from the terms of each two PIE disjunctions. If a common denominator is identified by the function then a PIE reconstruction has been defined and an etymology has been found, if the equation satisfies the semantic criteria.

With the decision method function coded, also the intuitive comparison, done manually until now, has been explicated and may be used in automatically reconstructing PIE prototypes, testing the hitherto suggested etymologies as well as finding new ones, discovered by a computer for the first time in the history of the field.

4. On the digitalization of other key features of PIE Lexicon

The core idea of PIE Lexicon, illustrated above with mechanized generation of IE data and the PIE reconstructions, is to digitize every possible feature and aspect of the linguistic data. This will ultimately result in an etymological dictionary exclusively containing smart or digitized features. In order to illustrate this in further detail several other key features to be digitized will be outlined in this paragraph.

Initially the focus of PIE Lexicon is placed on etymology and therefore we do not aim at full coverage of the entire IE data like the dictionaries of individual IE languages. This partial display of the material is compensated for with active links attaching the IE data entries of PIE Lexicon to other electronic dictionaries available on the internet. This automatic linking has already begun in a manner illustrated by the screenshot below, where the blue in ‘(Poucha 22)’ indicates an active link leading to the respective entry in another electronic dictionary:⁵

$\sqrt{\text{hai-}}$			
PIE * haōj-	TochA. āy-	(m.) ‘os’	(Pyysalo)
PIE * haōj-	TochB. āy-	(sb.) ‘Bein: bone’	(Poucha 22)
			(DTochB. 45-6)

Figure 6: An automatic external hyperlink in PIE Lexicon

This exploitation of language resources allows the PIE Lexicon users to verify the entries and, something of equal importance, reach comprehensive internal data and description of the IE entries.

Automated customization of the dictionary to the users’ needs and characteristics is already provided in a preliminary form in the search function located in the control bar at the bottom of the site:

Figure 7: The PIE Lexicon search engine window

Initially the search function is referential, only allowing the user to search for a single, untagged item, but this function will be upgraded into a full-scope advanced search with any number of search variables of all categories to exactly define any data segments needed by scholars in their work.

⁵ For the actual entry in CEToM, see <https://www.univie.ac.at/tocharian/?āy>.

The rightmost (optional) column is reserved for the attested forms of the IE stems and their grammatical analysis, as shown in Figure 8:

Hitt. ḫaa-	(vb.) 'vertrauen, jemandem etwas glauben'	(HHand. 34)	(Hitt. ḫa-a, ḫa-a-mi [1sg], ḫa-a-ši, ḫa-a-ir)
------------	---	-------------	---

Figure 8: A PIE Lexicon data entry line including attested forms

Once the priority coding tasks have been established, a key NPL tool, the automatic grammatical analysis of the attested forms, will be implemented in this section. In addition, the attested forms and their exact locus, possibly in the context of the original text, will be added to each form, if not already present.

Until this point the data of the pilot versions of PIE Lexicon have been limited to correspondence sets containing at least one of the best preserved Old Anatolian languages: Hittite, Palaic, Cuneiform Luwian, or Hieroglyphic Luwian. These languages have uniquely preserved the PIE 'laryngeal' (i.e. glottal fricative PIE *h) as such, giving them priority in the reconstruction of PIE ever since their discovery. In the next coding phase, however, such limitations no longer apply, and inherited data of all languages will be used equally to compile the first complete initial PIE *u/ṽ, comprising the main bulk of the entire most archaic data starting with this initial. As this data segment, the first of the total of eleven main entries,⁶ will be about a thousand pages long, its publication will turn PIE Lexicon into a big data program proper and, equally importantly, the stable, largely permanent initial display of the data will allow scholars of IE linguistics as well as other fields to begin the study of the data in earnest.

As the entry PIE *u/ṽ is representative in terms of the preserved material, its publication will make possible especially the study of the morphology, the original structure, formation and the origin of Proto-Indo-European. This is facilitated by the fact that the reconstructions contain the information of the respective IE correspondence sets in compressed form, i.e. this single, unified language can be taken as the primary object of the study instead of the earlier material divided into some 150 distinct languages. This study has already been anticipated in the control bar at the bottom of the site (Figure 9).

⁶ The PIE phoneme inventory (see §2.2.1) comprises of fourteen items, each with two varieties in columns. Of these fourteen phonemes the three leftmost are vowels, which occur as independent roots in only a few cases, to be dealt with the introduction in a small separate work. Due to this the dictionary proper splits into eleven main entries corresponding to the remaining consonantal phonemes of the inventory.

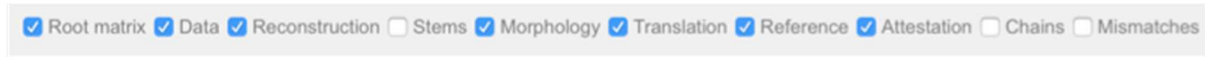


Figure 9: The PIE Lexicon data selection control bar

When the ‘Stems’ option is deactivated in the manner shown in the screenshot above, the IE forms are not shown and the translations apply to the PIE reconstructions instead. A description of a single language, PIE, gives the following results:

PIE √ <i>hai-</i> (vb.) ‘glänzen, brennen’		(IEW 11-12 * <i>ai-</i>)	(Pyysalo)
√ <i>hai-</i>		(HEG A:3)	(Pyysalo)
PIE * <i>haojo-</i>	(vb.) ‘heiß, warm sein’	(DPal. 53)	(Pal. <i>ha-a-ri</i> [3sg], <i>ha-a-an-ta</i> , <i>ha-an-ta</i>)
√ <i>hain-</i>			
PIE * <i>haojon-</i>	(pt.n.) ‘warm, heiß’	(DPal. 53)	(Pal. <i>ha-a-an</i> [-?])
PIE * <i>haojōn-</i>	(n.) ‘Tag’	(AIWb. 157)	(LAv. <i>ayq</i> [sgL], gAv. <i>ayan</i> [sgG])
√ <i>hair-</i>			
PIE * <i>haejor-</i>	(n.) ‘Tag’	(AIWb. 157)	(IEW 12) (SPIE 217) (gAv. <i>ayarš</i> [sgNA])
PIE * <i>haeire-</i>	(vb.) ‘verbrennen, anzünden’	(ArmGr. 1:418-9)	(Arm. <i>airel</i> [inf.])
PIE * <i>hair-ino-</i>	(UDUNm.) ‘Schmelzofen’	(HEG H:237)	(Hitt. <i>hi-ri-na-aš</i>)
PIE * <i>háiros-</i>	(n.) ‘Nachmittag’	(AIWb. 410)	

Figure 10: The PIE Lexicon in the PIE mode without IE languages

With this simple device the IE data has turned into PIE data, and the further digitalization of these structures, including simplifications, enables us to digitally define and manage the entire word formation of Proto-Indo-European.

Similarly, by releasing all buttons in the control bar except ‘Root matrix’, the root structure of PIE becomes directly observable, as shown in Figure 11:

PIE √ <i>hai-</i> (vb.) ‘glänzen, brennen’	(IEW 11-12 * <i>ai-</i>)	(Pyysalo)
√ <i>hai-</i>	(HEG A:3)	(Pyysalo)
√ <i>hain-</i>		
√ <i>hair-</i>	(IEW 12)	(SPIE 217)

Figure 11: PIE Lexicon in PIE root and extension mode

As soon as the first representative data set becomes available, these and other similar devices will facilitate the study and mechanization of the proto-language PIE in an exact manner, similar to how the Indo-European languages themselves have already been mechanized in PIE Lexicon.

The complete data entries enable the coding and digital management of the semantics of Proto-Indo-European. This observation is based on the fact that every PIE morpheme is associated with the meaning conveyed by its IE counterpart, which associates the morpheme with a specific morphological category (e.g. verb or adjective). Under these circumstances it is possible to define the semantic fields of the PIE roots. Each of these contains a number of IE stems (e.g. verbs and nouns) with meanings, the combination of which defines the semantic field of the root in question. Once these meanings have been defined and coded for the individual PIE roots, it becomes possible to compare multiple PIE roots having similar semantic fields. This will provide a warning of potential errors in the classification of the data if a parallel for the meaning of a semantic field is absent in other roots with otherwise identical semantic fields. Reversely, forms that have hitherto failed to be connected to any root can be attached to one, if a semantic parallel is available in the semantic field of another, morphologically different root. As a whole this means that the relatively complex and abstract study of meaning in Proto-Indo-European can be established in a strictly scientific environment.

Initially PIE Lexicon uses IE stems, supported by some attested forms, as its data entries. Naturally this restriction is artificial, and PIE Lexicon can be expanded to contain all the attested data and the related scientific discussion so far. Achieving this is not problematic, because a separate article page can be simply opened for each data entry, allowing the editors and contributors to compile an article containing the full attested data, the related scientific discussion so far, and other relevant observations.

5. Summary

As a whole the underlying plan of PIE Lexicon is to digitize (and turn smart) all of its features, ranging from reconstruction to semantics and its data. In other words, the long-term aim is to critically summarize two centuries of Indo-European linguistics as a whole into a single file, ultimately containing every piece of data or material bearing relevance to it, and offer it to scholars and others interested. While this task is too ambitious to be achieved by a single person or even a single team, the PIE project is built upon the chassis of natural science and is thus open-ended. This allows new administrators and teams to take over the management and continuation of the project in future decades, possibly even centuries, during which corrections, improvements, supplementations, and extensions to the original can be executed when needed until all problems of the field, including the new high-level ones emerging during the process, have been solved.

As specifically related to the content, the project is initially designed to optimize the digital treatment, analysis and presentation of the primary material, the Indo-European languages themselves. However, as soon as the basic problems involved are satisfactorily managed, the aim is to increasingly shift the focus to the digitized study of Proto-Indo-European, the inductive equivalent of the Indo-European languages. This will take the

field far beyond the scope of traditional Indo-European linguistics, resulting not only in the triumph of the electronic Neogrammarians mentioned by Sims-Williams (2018), but also of electronic lexicography as a whole in the 21st century.

In order to reach such ambitious goals the importance of electronic lexicography cannot be exaggerated: As an empirical science Indo-European linguistics is exclusively data-based. Accordingly, the more advanced and smarter electronic dictionaries of the field get, the more advantages result for science. In addition, the cooperation of electronic dictionaries will play a vital role in future science: Not only the active links, guiding the users to other sites and thus promoting these, but more abstract sharing of data, e.g. in the forms of etymologies, improves the content of electronic dictionaries..

6. References

- Hulden, M. (2009). *Finite-State Machine Construction Methods and Algorithms for Phonology and Morphology*. PhD Dissertation, University of Arizona.
- Koerner, K. (1982). The Schleicherian Paradigm in Linguistics. *General Linguistics* 22: 1-39.
- Pyysalo, J. (2013). *System PIE: The Primary Phoneme Inventory and Sound Law System for Proto-Indo-European*. PhD Dissertation, University of Helsinki. Publications of the Institute for Asian and African Studies 15. Helsinki: Unigrafia Oy. <https://helda.helsinki.fi/handle/10138/41760>
- Pyysalo, J., Sahala, A. & Hulden, M. (2018). Verifying the Consistency of the Digitized Indo-European Sound Law System Generating the Data of the 120 Most Archaic Languages from Proto-Indo-European. In Eetu Mäkelä, Mikko Tolonen, and Jouni Tuominen (eds) *Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference Helsinki, Finland, March 7-9, 2018*. Helsinki, University of Helsinki. <http://ceur-ws.org/Vol-2084/paper7.pdf>
- Schleicher, A. (1852b). *Die Formenlehre der kirchenslawischen Sprache, erklärend und vergleichend dargestellt*. Bonn: H.B. König.
- Sims-Williams, P. (2018). Mechanising historical Phonology. *Transactions of the Philological Society* 116, pp. 555-573.
- Szemerényi, O. (1967). The new look of Indo-European reconstruction and typology. *Phonetica* 17, pp. 65-99.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



Converting and Structuring a Digital Historical Dictionary of Italian: A Case Study

Eva Sassolini¹, Anas Fahad Khan¹, Marco Biffi^{2,3},

Monica Monachini¹, Simonetta Montemagni¹

¹ Istituto di Linguistica Computazionale “A. Zampolli” - CNR (Pisa, Italy)

² Accademia della Crusca (Firenze, Italy)

³ Università degli Studi di Firenze (Italy)

E-mail: {eva.sassolini, fahad.khan, monica.monachini, simonetta.montemagni}@ilc.cnr.it,
marco.biffi@unifi.it

Abstract

The paper describes ongoing work on the digitization of an authoritative historical Italian dictionary, namely *Il Grande Dizionario della Lingua Italiana* (GDLI), with a specific view to creating the prerequisites for advanced human-oriented querying. After discussing the general approach taken to extract and structure the GDLI contents, in the paper we report the encouraging results of a case study carried out against two volumes which have been selected for the different conversion issues raised. Dictionary content extraction and structuring is being carried out through an iterative process based on hand coded patterns: starting from the recognition of the entry headword, a series of truth conditions are tested which allow the building and progressive structuring, in successive steps, of the whole lexical entry. We also started to design the representation of extracted and structured entries in a standard format, encoded in TEI. An outline of an example entry is also provided and illustrated in order to show what the end result will look like.

Keywords: historical dictionaries; automatic acquisition; TEI representation

1. Introduction

The digitization of historical dictionaries represents a growing convergence between lexicographers, computational linguists and digital humanists.

Research in the area dates back to the origins of computational lexicography, and has proceeded along two main lines. Since the 1980s, pioneering studies have been carried into the transformation of Machine Readable Dictionaries (MRDs) into Computational Lexicons, mainly for use in machine-oriented applications. This strategy was proposed as a way to tackle the so-called “lexical bottleneck” caused by the lack of large-scale lexical resources, indispensable for the success of realistic applications in the field of Natural Language Processing (NLP), involving e.g. syntactic parsing, word sense disambiguation, speech synthesis, information extraction, etc. Such information was acquired by exploiting the lexical entry structure of dictionaries as well as through the automatic analysis of natural language definitions: a large literature exists on this

subject, from Amsler (1981) to Calzolari (1984), Boguraev and Briscoe (1989), Montemagni and Vanderwende (1992), to mention only a few. By the mid-1990s this line of research started to go into decline as it was concluded that MRDs could not be usefully exploited for NLP applications, especially when compared with other knowledge sources such as corpora (Ide & Veronis, 1993).

Together with the acquisition of lexical knowledge from MRDs, another important issue to be tackled concerns the identification of the optimal structure, organization and representation of the resulting computational lexicons. Since the 1990s, research has started to focus on the definition of lexical representation standards, which eventually led to the definition of i) the “Lexical Markup Framework” (LMF; Francopoulo, 2013), a framework for publishing computational lexicons that today is also an ISO standard (ISO-24613:2008), and ii) Ontolex-Lemon¹ which is a *de facto* standard for publishing lexicons as linked data. In addition, the Text Encoding Initiative (TEI)² is now very popular for representing digital editions of lexicographic resources in XML.

Although these lines of research were focused on the development of computational lexicons mainly designed for use within Natural Language Processing applications, methods and techniques developed for extracting, structuring and representing machine-oriented dictionaries still have a potential role to play in lexicographic tasks for dictionary publishers and lexicographers, i.e. for the design and construction of human-oriented resources. As pointed out by Granger (2012), the line between machine- vs human-oriented lexical resources is progressively narrowing, thus making the synergy between these two areas of research ever more interesting.

Over the last few years, e-lexicography research has moved towards the design and construction of human-oriented online dictionaries which allow for efficient access by multiple users and which can also be easily integrated with other lexical resources and corpora (Krek, 2019). In Italy, the *Accademia della Crusca*,³ an institution regarded as the pre-eminent authority in the study of Italian language, is moving in this direction thanks to its work on the design and construction of a dictionary of the post-Unification⁴ Italian language.

The current paper reports on preliminary results of a collaboration between the *Accademia della Crusca* and the *Istituto di Linguistica Computazionale* of the Italian National Research Council (ILC-CNR) with the aim of extracting the contents of the

¹ <https://www.w3.org/2016/05/ontolex/>

² <https://tei-c.org/guidelines/P5/>

³ <http://www.accademiadellacrusca.it/en/pagina-d-entrata>

⁴ The process of Italian unification took place in the 19th century; it began in 1815 with the Congress of Vienna and was completed in 1871 when Rome became the capital of the Kingdom of Italy: during this period the different states of the Italian peninsula were unified into the single state of the Kingdom of Italy.

Grande Dizionario della Lingua Italiana (‘Great Dictionary of Italian Language’, henceforth GDLI) in order to convert them into structured digital data for human use and to integrate them with other language resources, both dictionaries and corpora. This collaboration is being carried out within the framework of a national project strategic for the *Accademia della Crusca* and which aims at the construction of a *Dynamic Vocabulary of Modern Italian* (‘Vocabolario dinamico dell’italiano moderno’, in short VoDIM)⁵, within which GDLI plays a central role. A prototype digital version of GDLI, recently released by *Accademia della Crusca*, represents the starting point of the case study presented in this paper.

This case study presents itself as a challenging test bed at different levels, in particular: the extraction and structuring of the contents of the dictionary, starting from methods and techniques developed over the years for acquiring lexical knowledge from digital dictionaries; the design of a lexical representation model for the extracted and structured entries of such a complex historical digital dictionary in a standard format, encoded in TEI, with a specific view to enabling interoperability, comparability and further ease of exploitation. In what follows, the results achieved so far are presented, together with the current directions of research. After a short introduction to the GDLI dictionary and its main features (Section 2), Section 3 illustrates the general strategy adopted for extracting and structuring the dictionary contents from the OCRred version of the dictionary, the challenges to be tackled, the solutions adopted and a preliminary evaluation of results achieved so far. The final section of the paper (4) discusses the issues which are being addressed to convert the extracted contents in a standardized lexical representation format and shows how the end result will look.

2. The dictionary

The *Grande Dizionario della Lingua Italiana*, conceived by Salvatore Battaglia and released periodically in successive volumes between 1961 and 2002, is the most important historical dictionary of Italian ever published and covers the entire chronological period of the language, from its origins in the XIII century to the present day. The dictionary was published under the aegis of UTET *Grandi Opere* and maintains the legacy of a great publishing tradition: the UTET publishing house is, in fact, the oldest in Italy, having been founded in 1791. GDLI consists of 22,700 pages divided into 21 volumes, containing 183,594 entries. Word usage is documented through

⁵ The main goal of the VoDIM project is the construction of a vocabulary of post-unitary Italian that gathers together the national linguistic heritage of the official language of the State from 1861 to the present day. It was funded through two Research Projects of National Relevance (PRIN), in 2012 (‘Corpus di riferimento per un Nuovo vocabolario dell’italiano moderno e contemporaneo’), and in 2015 (‘Vocabolario dinamico dell’italiano post-unitario’). Numerous Italian universities and research centres are involved in the project: Piemonte Orientale, Milano, Genova, Firenze, Viterbo, Napoli, Catania, ITTIG-CNR (first phase only) and Università di Torino (second phase only). The *Accademia della Crusca* has collaborated in both projects as an external partner, the post-unitary Italian dictionary being one of its strategic activities.

14,061 citations by 6,077 authors: authors and works are indicated in the lexical entry with abbreviations, which are gathered in a separate volume with the index to authors and quotations (*Indice degli autori citati*). The dictionary also includes update volumes, published in 2004 and 2009, which document most recent and innovative uses of language.

The dictionary offers valuable information on the first attestations of words, on their variants (ranging e.g. from formal to diachronic or diatopic kinds), on the authors who quote them, and on their etymologies. The potential advantages of the digitization of such a monumental dictionary have always been clear to scholars who would have liked the same search functionalities for GDLI as those offered by the electronic version of the five editions of the *Vocabolario degli Accademici della Crusca* (1612, 1623, 1691, 1729-1738, 1863-1923) which can be accessed from the web site *Lessicografia della Crusca in Rete*.⁶ The digitization and structuring of the GDLI text, by explicitly marking “macro-contexts” (e.g. lemmas, definitions, examples) as well as “micro-contexts” (e.g. foreign words, proverbs, idioms, etc.), would allow for more refined and in-depth search functionalities, permitting scholars to navigate through a rich and representative diachronic corpus of the Italian language (Biffi, 2018). This becomes even more crucial if we consider that from a careful analysis of the rich historical corpus of citations of GDLI it turned out that there are words occurring in it which were not selected as lemma entries.

Taking this idea as a starting point, the *Accademia della Crusca* signed an agreement with UTET *Grandi Opere* in September 2017 which led to the latter making the electronic version of the dictionary available for digitization and online publication. In May 2019, a prototype digital version of GDLI was released via the *Accademia della Crusca* “Digital Shelves”⁷, which can be accessed and queried with basic full text functionalities. This version was acquired through optical character recognition (OCR) carried out with the FineReader application operating against the dictionary PDF files made available by UTET. Up till recently the process of text correction was limited to correcting page boundaries to avoid the erroneous splitting of words and entries. However, the manual correction of the text is now proceeding, including the correction of words in Greek. In parallel, a semi-automatic approach to text correction and structuring is being developed: the case study presented in this paper presents the general approach and the first steps taken in this direction so far. The OCR output used for the GDLI digital prototype represents the starting point of this case study.

⁶ www.lessicografia.it

⁷ <http://www.gdli.it/>

3. Extraction and structuring of dictionary contents

3.1 General approach

The process of extracting and structuring dictionary contents and converting them into TEI XML has been organized into several iterative steps, each with the function of progressively refining and organizing the dictionary structure previously identified. The iterative approach we follow consists of a series of successive refinement phases which, starting from the identification of the lemma vs the body of the lexical entry, aim to further refine this segmentation by recognizing, around this nucleus, the other fields/parts of the lexical entry. Each field requires specific strategies to identify its distinguishing features. Constraints are set incrementally, leading to an increasingly granular recognition of distinct sections/fields of the entry structure.

The final aim of the work is to structure the entire dictionary entry, but the problems due to the non-standard format do not currently allow us to make a precise estimation as to how long it will take to reach the goal. This is a long process, full of unknowns, in terms of both extraction strategies and the quality of the results. We have made a long-term work plan, that consists of milestones to be achieved progressively: 1) recognition of the headword; 2) identification of all fields of the main lemma; 3) number of main senses; 4) number of nested senses; 5) fields of every main sense; 6) fields of each nested sense 7) mapping to the standardized TEI format. To optimize the time required to complete the overall work, we decided to work on several objectives in parallel. In the case of milestone 7) it is in fact a matter of defining a final structure and format that can be implemented parallel to the extraction work. In this paper we describe the extraction work foreseen in 1) and 2) above (this section) and the mapping in TEI (Section 4).

This iterative approach to entry structure recognition was also designed to reduce the number of unavoidable errors, thanks to the semi-automatic correction of extracted and structured contents to be used as input for the further processing stages. For this reason, in parallel with the content parsing strategies, we have defined methods to facilitate manual data review and correction. At the present time we have not defined a final protocol for the treatment of cases like this, but we wanted to propose our approach as a case study for similar situations anyway, that is in situations where it is not possible to use consolidated or experimental tools and or procedures already known in the literature, and the data has a significant amount of errors. In fact, in these cases we cannot define only the extraction procedures, but at the same time we have to implement strategies to support the correction and an efficient system of revision and subsequent realignment of the extracted data.

3.2 Input data

The richly detailed resource described above poses numerous challenges for the extraction and structuring of dictionary contents which are carried out against an

OCRed version of the dictionary. As pointed out in Section 2, OCR was carried out with the conventional FineReader application operating against the PDF files made available by the publisher. Although desirable, due to time and resource constraints it was not possible to improve OCR accuracy through pre- and/or post-processing techniques on the output of a single or multiple OCR engines, as currently proposed in the literature on novel approaches for OCR accuracy enhancing.

The original text in paper format shows some stylistic features and layout choices that make OCR extremely complicated, and we had to deal with the problems which resulted. The published edition which was used adopted a subdivision of the page into 3 columns, used a non-white paper colour, as well as a very small typographic font and an equally small interline one. With a work covering a time span of 40 years, it was unavoidable that there have been changes and adjustments (even minor) which have been introduced over time to the structuring of entries and the reference corpus of GDLI. Although the basic entry structure remained constant through time there have been slight changes in its internal organization, even just at the level of layout, as exemplified in Figure 1 which reports OCRed text samples from different volumes. For this reason, this case study has been carried out on two different GDLI volumes (namely, I and XII), which were selected for the different challenges and parsing problems posed by the OCR results.

All these features made the acquisition via OCR subject to errors of various types, which prevented the possibility of using already available state of the art automatic parsing tools.

Vol.	OCR text
01	<p>Ammannimento, sm. Allestimento. <i>Fra Giordano [Crusca]:</i> Facevano per la guerra gli ammannimenti necessari. <i>Soderini</i>, I-215: Così fatti e simili ed altri deono essere gli ammannimenti che s'hanno ad avere in preparamento per potere a dilungo fabbricare. <i>Salvini</i>, 30-2-158: Le ore gloriosamente spendete nell'ammannimento delle nuove voci, e nella correzione, che molto importa., del Vocabolario di nostra lingua. = Deriv. da <i>ammannire</i>.</p>
03	<p>Capocameriere, sm. (plur. <i>capicamerieri</i>). In un albergo, in un ristorante, il primo cameriere, quello da cui dipendono tutti gli altri. <i>Moretti</i>, 17-296: Vuoi sapere che cosa si fa laggiù? Si fa : il cameriere e il capo-cameriere, il fornitore di viveri ai piroscafi italiani, il padrone di trattoria. = Comp. da <i>capo</i> e <i>cameriere</i> (v.).</p>
09	<p>Iniare, intr. con la particella pronom. (m'inio). Ant. Diventare simile, identificarsi. <i>Dante</i>, <i>Par.</i>, 33-44: Indi a l'eterno lume s'addrizzaro, / nel qual non si dee credere che s'ini / per creatura l'occhio tanto chiaro. <i>Ottimo</i>, III-729: 4 Nel qual non si de' credere ec. : cioè, si come più volte è detto, occhio creato non può iniarsi al fondo della divinitade. 4 Inii si è verbo informativo, ed è tanto a dire, come diventare simile di quella cosa ch'è considerata. = Denom. dal pronom. io col pref. in- con valore illativo.</p>
11	<p>Nauseante (part. pres. di <i>nauseare</i>), agg. Che provoca nausea, disgusto, voltastomaco; disgustoso, nauseabondo. <i>O. Targioni Pozzetti</i>, I-213: Linneo li aveva divisi e classati [gli odori] in... ambrosiaci..., fragranti..., tetri o virosi..., nauseanti. <i>Massaia</i>, Vili-163: Riacquistate, con l'aiuto di quella putrida è nauseante acqua, alquanto le forze, si continuo il cammino per l'arido deserto. <i>Tarchetti</i>, 6-II-639: Permetti, bevo un bicchiere di decotto di gramigna, serrandomi prima delicatamente la punta del naso tra il pollice e l'indice della mano sinistra. Dio, che roba nauseante, è un beverage da cavallo. <i>Svevo</i>, 3-569: Restai tranquillo a quel posto fumando quelle sigarette nauseanti. <i>Stuparich</i>, I-333: Avverto intorno un puzzo di pesce rancido, nauseante.</p>
	<p>Arricavo, sm. Marin. Estremità di un cavo fissata all'oggetto che deve essere alzato o trasportato; dormiente. <i>Arriicare</i>, tr. e rifl. Ant. <i>Arriichire</i>. <i>Rugieri d'Amici</i>, I-20: Amor m'à sì arricato / in tutto 'l meo volere, / e dato m'à a tenere / più ricca gioia mai non fue visato. <i>Iacopone</i>, 18-20: O taupino, a cui aduni? A arricar li toi garzuni? / Da ch'èi morto, i gran bocconi se fo del tuo guadagnato. <i>Idem</i>, 26-34: Frate, non m'esser sì avaro, / ca molto caro me costi per volerte arriicare. = Deriv. da <i>ricco</i> (v.).</p>
	<p>Certatōre, sm. Ant. e letter. Combattente. - Anche al figur. <i>Alberti</i>, 267: Mai mi lascio stare in ozio, fugo il sonno, né giaccio se non vinto dalla strachezza, che sozza cosa mi pare senza ripugnare cadere e giacere vinto, o come molti prima aversi vinti che certatori. = Deriv. da <i>certare</i>.</p>
	<p>Ludro, agg. e sm. Dial. Mascalzone, birbante, imbroglione, canaglia; persona avida, ingorda, insaziabile.</p>
	<p>Motosilurante, sf. Marin. Milit. Piccola unità navale da guerra, leggera e assai veloce, munita di motore a propulsione endotermica, per lo più a scoppio o Diesel o, anche, a turbina, armata con siluri e con qualche cannone di piccolo calibro a cadenza di tiro assai elevata, impiegata per attacchi di sorpresa in acque ristrette. <i>Migliorini</i> [s. v.]: 'Motosilurante', nome masch. o femm.: leggera imbarcazione a motore per il lancio dei siluri. Lo stesso che 'mas'. = Comp. da <i>motore</i> e <i>silurante</i> (v.).</p>

Figure 1: OCRed text samples from different volumes in Word format.

The input of the extraction and structuring work is represented by more than 23,000 pages of dictionary text, provided in a (non-standard) Word format and organized into 21 volumes preserving the same subdivision as the GDLI paper format. Since the resulting Word files are very heavy and difficult to manage, we tried to convert these to other formats (XML and TXT). It turned out that for the lemma extraction we had substantially the same problems as with the Word format, but errors in other parts of the structure made the extraction procedure more complex. Although lighter to handle and more readable, the TXT format extracted from the Word format left out important information pertaining to format and style, which is often crucial in the discrimination between, e.g. a lemma and a simple paragraph beginning (see below).

3.3 Segmentation strategy

The first phase of the work concerned the segmentation of the Word format (“.doc”) file of each individual volume into portions of no more than 50-60 pages, each of which was saved in a separate file, and analysed in succession by the parsing program. The entire process required the use of numerous software libraries capable of parsing the Word format and identifying the peculiarities of the structural and formatting characteristics of the text. The segmentation procedure was performed manually to avoid the inappropriate cutting up of individual entries across pages. At this stage and with unavoidably noisy input, a fully automatic system would have not produced a sufficiently good result when applied to dictionary texts in which lexical entries are typically organized in relatively long enumerations of nested senses each of which also contains related quotations.

The second step consisted in the segmentation of individual pages recognized at the previous step into lexical entries, whose boundaries were explicitly marked. For each identified lexical entry, the headword (or lemma) and a text area corresponding to the body of the entire entry is recognized. The segmentation procedure proceeds with the identification of the other entry fields, according to similar methods used for the headword.

These further steps include the iterative segmentation of the body of the lexical entry into different blocks with grammatical information (including the indication of possible variants, e.g. orthographic, diatopic, diachronic, etc.), main senses, sense attestations and examples, other numbered sub-senses with examples (if any), and etymology. Each main sense block is in its turn articulated into different sections within which quotations play a central role: to quote Beltrami and Fornara (2004), “the veritable fulcrum of the dictionary is the massive presence of text quotations from authors”. These quotations cover a wide variety of language use, from everyday and literary language, dialectal and regional languages, to technical and scientific language, specialized languages, neologisms and foreign words.

The results of this further segmentation, which are currently being analysed in detail, are strongly influenced by the success of the lemma extraction phase. However, the type of recognition errors generated by the extraction system has also been analysed on each individual structural feature of the dictionary: lemma, spelling variants, grammatical category, usage codes, definition, etymology, main senses and additional senses (nested). Each of the fields shows errors of various types, ranging from errors in the segmentation of paragraphs, to those in the rendering of punctuation marks, to spelling errors, to the failure to identify the structural elements that define the different sections of the dictionary entry (bullet points, indentation, font size etc.). Figure 2 exemplifies some OCR errors negatively impacting on the further recognition process.

N.	Paper text	OCR output
1	Amminoazobenzene (<i>aminoazobenzene</i>), sm. Chim. Composto organico classificato tra i coloranti azoici conosciuto anche col nome di <i>giallo d'anilina</i> : cristalli gialli che si sciolgono in alcole ed etere, assai meno in acqua (usato nella colorazione di prodotti alimentari e per preparare altri coloranti).	Am mi no a z ob e nz è ne (<i>aminoazobenzene</i>), sm. Chim. Composto organico classificato tra i coloranti azoici conosciuto anche col nome di <i>giallo d'anilina</i> : cristalli gialli che si sciolgono in alcole ed etere, assai meno in acqua (usato nella colorazione di prodotti alimentari e per preparare altri coloranti).
2	Assolare ¹ , tr. (<i>assòlo</i>). Disus. Rendere solo. - <i>Assolare una carta</i> : tenere scompagnata, nel gioco, una carta di un dato segno. = Deriv. da <i>solo</i> (v.). Assolare ² , tr. (<i>assòlo</i>). Esporre al sole; rendere soleggiato. = Deriv. da <i>sole</i> (v.). Assolare ³ (<i>assuolare</i>), tr. (<i>assòlo</i> o <i>assuòlo</i>). Disporre a strati. = Deriv. da <i>suolo</i> (v.).	Assolare ¹ , tr. (<i>assòlo</i>). Disus. Rendere solo. - <i>Assolare una carta</i> : tenere scompagnata, nel gioco, una carta di un dato segno. = Deriv. da <i>solo</i> (v.). Assolare ² , tr. (<i>assòlo</i>). Esporre al sole; rendere soleggiato. = Deriv. da <i>sole</i> (v.). Assolare ³ (<i>assuolare</i>) tr. (<i>assòlo</i> o <i>assuòlo</i>). Disporre a strati. = Deriv. da <i>suplo</i> (v.).
3	Ammacchiare ¹ , rifl. (<i>m'ammacchio, t'ammacchi</i>). Raro. Nascondersi nella macchia. B. Davanzali, I-136: Floro s'ammacchiò: vedendo poi presi i passi dell'uscita, s'uccise.	Ammacchiare ¹ rifl. (<i>m'ammacchio, Vammacchi</i>). Raro. Nascondersi nella macchia. B. Davanzali, I-136: Floro s'ammacchiò: vedendo poi presi i passi dell'uscita, s'uccise.
4	Attendista , agg. e sm. e f. (plur. m. -i). Neol. Chi evita di prendere posizione (e resta in attesa degli avvenimenti, riservandosi di decidere secondo il loro svolgersi). = Fr. <i>attentiste</i> (1941), da <i>attendre</i> 'attendere'. Attenditore , agg. e sm. (femm. -trice). Ant. Che attende, aspetta.	=Deriv. da <i>attendere</i> . Attendista , agg. e sm. e f. (plur. m. -i). Neol. Chi evita di prendere posizione (e resta in attesa degli avvenimenti, riservandosi di decidere secondo il loro svolgersi). =Fr. <i>attentiste</i> (1941), da <i>attendre</i> *attendere. Attenditore , agg. e sm. (femm. -trice). Ant. Che attende, aspetta.

Figure 2: Examples of blocking OCR errors.

These errors often block the correct segmentation of the internal structure of the entry, especially for what concerns the extraction of senses and sub-senses. The frequent co-presence of more than one error within the same entry makes the recognition of the internal structure a more challenging problem.

3.4 Main error types

The main types of errors concern the OCR format, and they impose an unavoidable conditioning on the quality of the extraction phase. Other errors, introduced by the parsing phase, could be added to these. A bad interpretation of the structure of the entry during the OCR process will obviously mislead the system, invalidating the extraction both of the lemma and other fields. We tried to organize the variety of anomalous phenomena encountered so far into six main error types, listed below:

1. “omission”, occurring when parts of the lexical entry (including substrings of characters) have been omitted;
2. “illegal merger”, occurring when different fields within a lexical entry or two lexical entries are wrongly merged (see example n. 4 in Figure 2);
3. “illegal disjunction”, corresponding to wrongly segmented words: e.g. ‘Ab borrire e deriv.’ for ‘Aborrire e deriv.’; ‘A c cespugli are’ for ‘Accespugliare’; ‘Acetilèni co’ for ‘Acetilenico’; ‘Acòre e a còro’ for ‘Acòre e acòro’; etc.;
4. “incorrect graphemes”, corresponding to wrongly interpreted sequences of graphemes of the same length: e.g. ‘sl’ for ‘sì’, ‘ero’ for ‘cro’, ‘cto’ for ‘chi’; ‘ln’ for ‘in’ or ‘li’ or ‘li’, etc.;
5. “exchange of graphemes”, corresponding to wrongly interpreted sequences of graphemes of different length (i.e. expansion or contraction): e.g. ‘lite’ for ‘nte’; ‘til’ for ‘rell’; ‘fif’, ‘flf’ or ‘tif’ for ‘ff’; ‘dd’ for ‘cìcl’; ‘g’ for ‘ci’, etc.;
6. “missing bullet points”, which are mainly concerned with the recognition of senses as exemplified in Figure 3, where the OCRred text on the right lacks sense numbers.

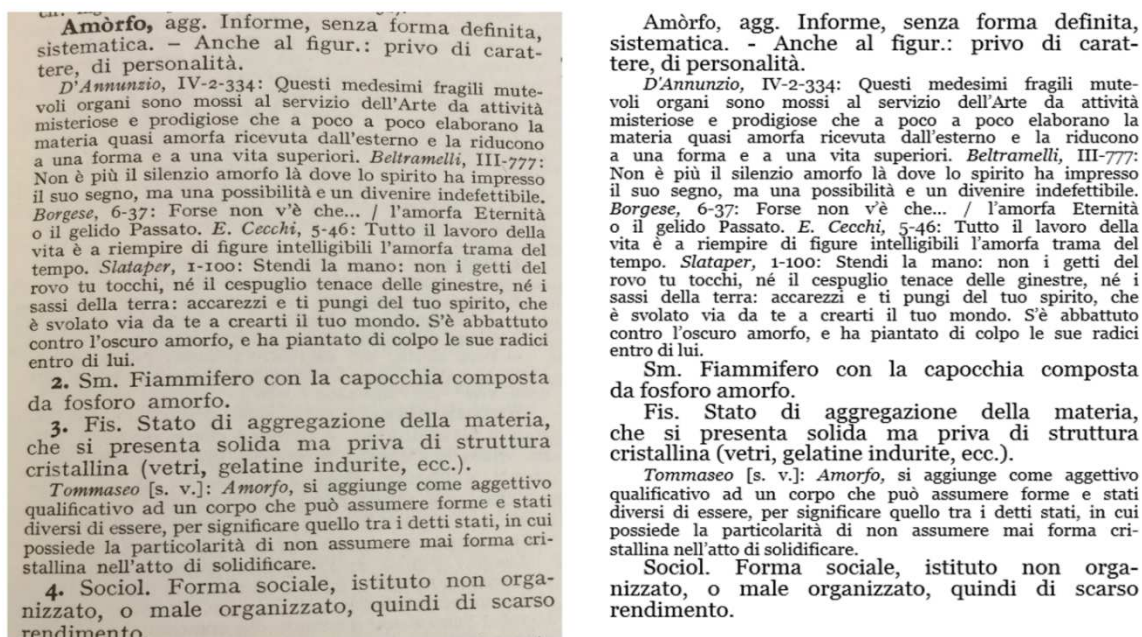


Figure 3: Bullet points in printed vs Word formats.

In Figure 4 below is a graph showing the percentage distribution of the six error types in the two volumes of the dictionary which were selected for this case study:

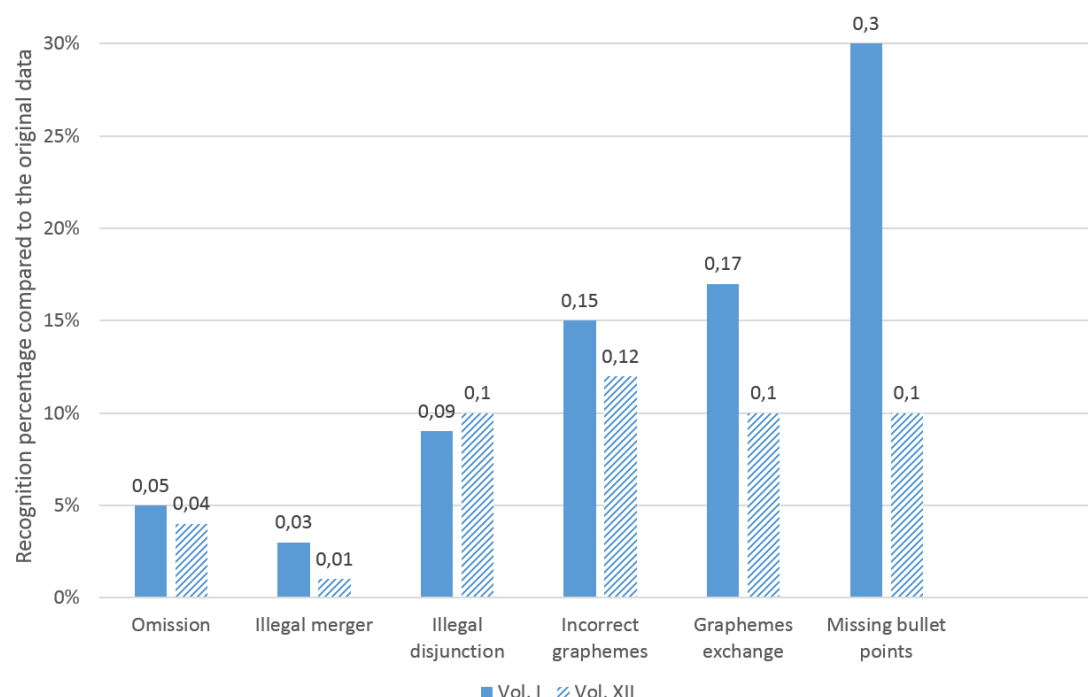


Figure 4: Distribution of error types between volumes I vs XII

It can be noted that the distribution of some types of errors differs significantly between volumes, suggesting a discrepancy of quality of OCR across them: this is the case, for instance, with “missing bullet points” and “exchange of graphemes”. It is possible that the long phase of preparation of the work influenced the differences between the volumes: the quality of the print, the colours of the paper and the ink, etc. As we have already said, we have noticed differences across volumes which already visually explain the differences in the performance of OCR procedures. It is likely that the conservation status of the volumes from which the OCR was made also comes into play, and it is not certain that all the volumes were in the same condition.

3.5 Lemma extraction

The approach to the extraction and structuring of GDLI contents is illustrated here with respect to the first segmentation step, mainly aimed at lemma extraction. Due to the complexities sketched above, we decided, at least initially, to follow an approach based on pattern matching. The patterns we work with cover a wide range of characteristics ranging from the layout of the page to structural information relating to the different parts of the lexical entry. They also relate to linguistic aspects regarding the format and spelling of the lemma as well as lexicographic ordering, with respect to the lemmas that precede and follow an entry. The patterns have been defined manually

and start from the recognition of the lexical entry and its headword (lemma). The extraction phase is then determined by the identification of the characteristics listed above and the testing of different truth conditions that, placed in combination with each other, confirm, to a reasonable approximation, the beginning of the entry of the dictionary and its end.

The recognition phase of the lemma is integrated with strategies supporting the correction of incompletely or erroneously extracted lemmas.

Whenever the lemma cannot be recognized with certainty, a check on the number of conditions satisfied is activated: a lower number of verified conditions causes the positive matching of entries that are often erroneous. Based on experiments, two different thresholds have been defined: cases that verify 2/3 of the conditions for the correct recognition of the headword are reported as requiring a manual verification; those that reach 3/4 of the conditions, already acquired as headwords, are suggested for manual control, although with a lower priority assigned. These cases are recorded within a report file which is generated together with the outcome of the parsing phase. In this report file, the “candidate” lemma is written, followed by page indication and listing of conditions which have not been verified.

Even when the lemma is correctly segmented, there may be spelling errors. We analysed these cases in order to find a suitable reporting method. Starting from a cost/benefit evaluation, we studied different techniques to identify and report this type of error. One technique consists of applying lexicographic sorting criteria to the lists of lemmas extracted automatically. The comparison of the natural sequence of the headwords found in the pages, with the same lexicographically ordered list, brings out the differences in the cases of spelling errors. We have decided to turn this evidence into a correction support report. In particular, parallel to the parsing, the extraction system, for each volume analysed, produces a file containing the list of all the headwords extracted, ordered lexicographically and followed by the page number where each headword was found. In this way the misalignment between the page sequence and the ordering of the headwords is evident and provides concrete help to the manual correction phase. Another technique to test the correctness of the acquired lemma consists of looking up the acquired candidate lemma string in other reference lexical resources, historical dictionaries (for example, the *Tesoro della Lingua Italiana delle Origini* or TLIO)⁸ as well as wide coverage contemporary dictionaries including historical lexical variants. Those entries for which no corresponding lemma has been found are reported for manual checking.

⁸ <http://tlio.ovl.cnr.it/TLIO/>

Error types	Fields	Adopted solutions	Examples
Orthographic (type n.1, 2, 3)	lexical entry	Ref. in the lemmas report	<div>for Affoltito</div> <p>A Abiti to (part. pass. di <i>affoltire</i>), agg. Folto; gremito. <i>Viani</i>, 10-189: Quando uno spiritato urlava sulla piazza della chiesa, subito dopo la messa di mezzogiorno affoltita di cavalieri: - Cavaliere! - l'unico che si voltava era il cavaliere Grotta. <i>Affondamento</i>, sm. L'affondare; l'andare a fondo.</p>
Punctuation	lexical entry	Amendment strategies embedded in the parser	<p>Accampionare «» r. (<i>accampiótto</i>). Disus. Ammin. Registrare nel censimento comunale, a scopi fiscali. <i>Fil. Ugolini</i>, 5: <i>Accampionare</i> è da fuggirsi insieme con <i>campionare</i>: dirai meglio 'porre a campione'. <i>Arla</i>, 8: <i>Accampionare</i>, registrare o notare su' registri pubblici, che si addimandano <i>campioni</i>, beni stabili per sottoporli al pagamento delle tasse. I lustrini la scomunicano, ma è di uso, e ben si attaglia alla cosa.</p>
Omissions	lexical entry	NA	<p>niaccio², sm. Marin. Agghiaccio. AGGIACCIO che pare la forma più antica rispetto ad <i>agghiaccio</i> (<i>Dizionario di Marina</i>, 11: <i>Agghiaccio</i> oggi in luogo di <i>agghiaccio</i>). <div>for Agghiaccio²</div> </p>
Lemma not found at the paragraph beginning	etym.	Event reported in the error report	(see Tab. 2. n.4)
Incorrect sequence of characters	lexical entry	Ref. in the lemmas report	<p>Afilosòfico»228»1? Afiòssatóre»223»1? Aflferratóio» 203»1? Aflferratóre» 203»1? Aflfettibilità» 204»1? Aflfrenare» 225»1? Aflfrettatóre»226»1? Aflfrettóso»226»1? Aflfricógno»226»0? Aflfricógnolo»226»0? Aflfrigolito» 226»1? Aflfrontatura 227 1 Aflreddato» 225»1?</p>

Table 1: Typical errors in lemma recognition

3.5.1 Specific error types

As far as lemma recognition is concerned, the largest number of errors found is distributed among error types 3), 4) and 5) listed in Section 3.4, namely “illegal disjunction”, “incorrect graphemes” and “exchange of graphemes”. Since these three error types have a greater impact on content extraction and structuring, it is on them that we have focused our strategies of manual correction support. Table 1 shows how these error types impact on the recognition of the lemma and the related strategies adopted to support the manual correction.

As for the “omission” type, besides manual correction, we have not found a solution at the moment. There are also possible errors when a string of characters corresponding to the true lemma is incorrectly interpreted by OCR, such that it overlaps with a previously recognized lemma.

3.5.2 Preliminary results

At the end of the acquisition experiments carried out against volumes I and XII, the results obtained for what concerns lemma extraction are promising, with an over 94% success rate, as shown in the pie chart in Figure 5. Lemmas are correctly extracted and identified in 75% of the cases; 15% of correctly acquired lemmas contain an OCR error, and 6% of them contain spelling errors (originating, for example, in the overlap with lemmas already extracted). This result, however, cannot be seen as exhaustive, because the amount of entries analysed, set against the total number contained in the GDLI, is around 10%.

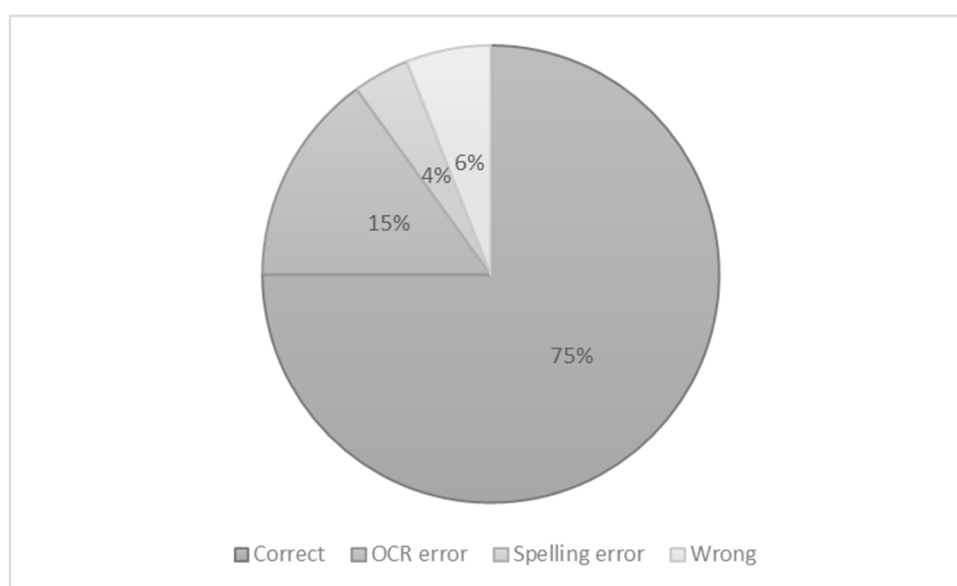


Figure 5: Lemma acquisition and identification results

4. TEI Mapping

4.1 Introduction

Although as regards the current state of progress of the work described in this paper we are still not in a position to discuss the technical details of the final conversion of the original source files into a standardized format for lexical resources such as TEI, we can show what it is we are aiming for and what the end result will look like. As pointed out above, we decided to work on both extraction and representation objectives in parallel: the reasons underlying this choice range from the optimization of the time required to complete the overall work to the fact that the adopted lexical representation model can influence, at least to some extent, the structuring of extracted contents.

In the following subsections we will describe the importance of using a specialized standard to encode the information in a resource such as the GDLI, as well as explain why we chose the TEI guidelines, and we will present an example entry from the GDLI and describe what a TEI encoding of the entry looks like.

4.2 Background on standards for lexical resources

The importance of the role of standards in the modelling, creation, and publication of computational lexical resources has gained increasing recognition in recent years. This is thanks not only to more general initiatives relating to the FAIR data principles (Wilkinson et al., 2016) but also to a growing appreciation of the critical worth of well-made lexical resources to much work in Computational Linguistics and Digital Humanities. There are several reasons why standards play such an important role in the specific case of lexical resources. For one thing the existence of lexical standards facilitates the harmonization of the different linguistic and metadata categories used in such resources, and is an important prerequisite to ensuring the interoperability of lexical datasets. Standards also allow resources to be re-used more easily and in various different contexts and tasks, and this is especially important in NLP where one single resource, such as WordNet, can be used in numerous different kinds of task. It is also more likely that, at least for the most popular and well known standards, there already exists software for creating, maintaining and publishing resources that adhere to the standards in question. Finally, in many cases these standards represent a community endorsed solution to those problems that are likely to arise when encoding various different types of lexical information.

When it comes to encoding lexical and, more specifically, lexicographic resources, there are a number of different relevant standards which should be taken into consideration, and in some cases a choice needs to be made between two or more competing standards encoding the same kinds of information. In our case it was important to choose a standard that was as widely used as possible and made use of common formats but that was also sufficiently expressive for our modelling needs. We wanted to annotate both those aspects of the resource pertaining to the source dictionary’s status as a printed text, as well as to its conceptual, bibliographic and linguistic content: that is, we wanted a model that would allow us to annotate things like bibliographic citations, quotes, as well as lexical entries, senses, and etymologies. For these reasons and others we decided to choose the Text Encoding Initiative (TEI) guidelines, and especially the chapter on encoding dictionaries, as our main standard in encoding the GDLI.

In the next subsection we will look at two GDLI entries encoded in TEI to show what the end result will look like.

4.3 Example entry

In order to show what the end result of the process described in this paper will look like, as well as to highlight some of the most typical features of GDLI lexical entries and how the TEI guidelines allow us to encode these features, we present an example entry from the GDLI. The entry in question concerns the adjective *padronale*, which has the primary sense of ‘pertaining to or deriving from the condition of being a boss

or master' and derives from the noun *padrone* 'boss, master'. The entry for *padronale* has four different senses, each of which is further subdivided into more specific sub-senses and each of which is provided with a list of citations from the corpus of historical Italian texts referred to by the GDLI. For reasons of space we will only discuss the first sense, which we show as Figure 6 (the page containing the full entry can be found here: <http://www.gdli.it/JPG/GDLI12/00000348.jpg>).

Padronale (*patronale*), agg. Che si riferisce o che deriva dalla condizione di padrone, dalle sue prerogative giuridiche di proprietario o dalla sua autorità nei confronti dei dipendenti, dei familiari o di altre persone; che denota, talora in modo ostentato, tale condizione; commesso, esercitato da un padrone; da padrone.

De Luca, 1-1-50: Quell'entrate e robbe, le quali abbiano annessa qualche giurisdizione o preminenza padronale, come per esempio sono li molini e forni. *Foscolo*, XVIII-253: Permetterò a Pietro d'incamminarsi tanto che dura la buona stagione; e se non altro sono consolato ch'egli non si dorrà mai giustamente di me, perché l'ho sempre trattato con volto padronale, ma con cuore fraterno. *Franzoni*, 13: I facili incrociamenti, le violenze patronali, le vergogne sifilitiche... ne hanno deturpato [degli arabi] il tipo fisico. *D'Annunzio*, IV-2-205: Don Giovanni Ussorio, presente sempre, aveva delle arie padronali. *Borghese*, 1-64: Egli passeggiava velocemente facendo cantare gli sproni..., avviato verso un'indignazione metà fredda e metà calda, donde desumeva chi sa che autorità maritale o padronale sulla donna di cui presentiva l'avvicinarsi. *Piovene*, 6-153: Era gentile per diplomazia padronale e per naturale indulgenza della persona superiore con la gente bassa.

– Per simil. Spavaldo, privo di ritegno.

Baldini, I-6526: Uscendo il treno dai monti in corsa verso il mare, il fischio della locomotiva righerebbe l'aria con quella padronale allegria della quale qui si sente propriamente la mancanza.

– Che spetta al proprietario di un podere, dominicale.

Tommaso [s. v.]: 'Parte padronale': quella che in Toscana e altrove 'domenicale', la parte della rendita appartiene al padrone del fondo, a distinguerla da quel che ne viene al colono. *Einaudi*, 2-280: È vero il vecchio adagio del mezzadro il quale: « signor padrone – dice – venga a dividere la sua metà », ed il quarto padronale non basta a pagare le imposte.

Figure 6: Sense 1 of the *padronale* GDLI entry.

Here the nesting structure of the first sense is implicit in the sense that the sub-senses are not given identifiers (the other sub-senses of the entry are given numbers) but can be identified by the tab space and the dash. The first sense has a main sense (that starts after the part of speech information), and two more specific sub-senses.

The entry (seen at the top level with sense nodes unexpanded) is shown in Figure 7.

```
<entry>
  <form type="lemma">
    <orth>Padronale</orth>
  </form>
  <form type="variant">
    <orth>patronale</orth>
  </form>
  <gramGrp>
    <pos>agg.</pos>
  </gramGrp>
  <etym>= Deriv. da <mentioned xml:lang="it">padrone</mentioned>.</etym>
  <sense level="1" n="1"> [47 lines]
  <sense level="1" n="2"> [32 lines]
  <sense level="1" n="3"> [5 lines]
  <sense level="1" n="4"> [13 lines]
  <sense level="1" n="5"> [5 lines]
  <sense level="1" n="6"> [5 lines]
</entry>
```

Figure 7: TEI representation of the *padronale* GDLI entry with sense nodes unexpanded.

Here we have annotated the fact that the entry has the lemma *Padronale* using the TEI `<form>` element, specifying its type attribute as “lemma”, as well as the alternative form *patronale*. We have also annotated its part of speech using the `<gramGrp>` and `<pos>` elements, and represented the fact that the word is derived from another word using the `<etym>` element. Next we represent the fact that the entry has six senses (at the first level of nesting) using the `<sense>` element and the attributes `@level` and `@n`.

In Figure 8, we show the structure of the first sense and its two sub-senses (with the `<cit>` node unexpanded). All three senses have their definitions marked out using the `<def>` element, with each citation annotated using the `<cit>` element.

```
<sense level="1" n="1">
  <def>Che si riferisce o che deriva dalla condizione di padrone,
    dalle sue prerogative giuridiche di proprietario o dalla sua autorità nei confronti dei dipendenti,
    dei familiari o di altre persone; che denota, talora in modo ostentato, tale condizione;
    commesso, esercitato da un padrone; da padrone.</def>
  <cit> [2 lines]
  <cit> [3 lines]
  <cit> [3 lines]
  <cit> [2 lines]
  <cit> [4 lines]
  <cit> [3 lines]
  <sense level="2" n="1">
    <def>Per simil. Spavaldo, privo di ritegno.</def>
    <cit> [2 lines]
  </sense>
  <sense level="2" n="2">
    <def>Che spetta al proprietario di un podere, dominicale.</def>
    <cit> [2 lines]
    <cit> [2 lines]
  </sense>
</sense>
```

Figure 8: TEI representation of sense 1 of the *padronale* GDLI entry.

Finally, in Figure 9, we expand the first two citations of the first sense.

```
<sense level="1" n="1">
  <def>Che si riferisce o che deriva dalla condizione di padrone,
    dalle sue prerogative giuridiche di proprietario o dalla sua autorità nei confronti dei dipendenti,
    dei familiari o di altre persone; che denota, talora in modo ostentato, tale condizione;
    commesso, esercitato da un padrone; da padrone.</def>
  <cit>
    <bibl>De Luca, 1-1-50:</bibl><quote> Quell'entrare e robbe,
      le quali abbiano annessa qualche giurisdizione o preminenza padronale,
      come per esempio sono li molini e forni.</quote>
    </cit>
  <cit>
    <bibl>Foscolo, XVIII-253:</bibl>
    <quote>Permetterò a Pietro d'incamminarsi tanto che dura la buona stagione;
      e se non altro sono consolato ch'egli non si dorrà mai giustamente di me,
      perché l'ho sempre trattato con volto padronale, ma con cuore fraterno.</quote>
  </cit>
  <cit> [3 lines]
  <cit> [2 lines]
  <cit> [4 lines]
  <cit> [3 lines]
```

Figure 9: TEI representation of citations in the *padronale* GDLI entry.

The first citation is from Giovanni Battista De Luca, the noted 17th century jurist and cardinal, and the second citation is taken from the works of Ugo Foscolo, the well-known 19th century Italian poet and political exile. In future work we are planning to add links to virtual authority files for the authors cited in the GDLI in the TEI-XML encoding itself.

From this brief description of the (manual) encoding of a single entry we hope it is clear how important such a conversion of the original resource is for rendering the linguistic, historical and cultural information inside the dictionary more machine actionable and more amenable to querying by human users.

5. Conclusion

In this paper, we presented the preliminary and encouraging results of a case study carried out to define the strategy to be adopted to extract and structure the contents of the most important historical dictionary of Italian, *Il Grande Dizionario della Lingua Italiana*, with a specific view to creating the prerequisites for advanced human-oriented querying, which allows for multiple and efficient access, can be integrated with other lexical resources and corpora, can be customized to meet specific user needs, etc. Dictionary content extraction and structuring is being carried out through an iterative process based on hand coded patterns: starting from the recognition of the entry headword, a series of truth conditions are tested which allow the building and progressive structuring, in successive steps, of the whole lexical entry. We also started to design the representation of extracted and structured entries in a standard format, encoded in TEI. After discussing the general approach taken, in the paper we focused on the early stages of the conversion of the dictionary contents into structured digital

data, with particular attention to supporting the semi-automatic correction of errors mainly originating in the OCRed parsed text.

The complex situation of the digitized version of the GDLI dictionary described in the previous sections, characterized by slightly different entry formatting and/or structuring conventions across volumes and the presence of OCR errors, led us to opt, at least for this first explorative phase, for a pattern-based approach. We are aware of the limits of this approach, i.e. the costly manual elaboration of complex patterns based on observing the organisation of the lexical information in dictionary entries, but at this stage this turned out to be the only viable approach. We are currently evaluating whether, once an appropriate quantity of dictionary entries from consistent GDLI portions has been reconstructed and corrected, a machine learning approach, such as that used by GROBID-Dictionaries (Khemakhem et al., 2017), could be usefully exploited for completing this work. The iterative approach to extraction and structuring of GDLI lexical entries proposed here creates the prerequisites for the creation of cascading extraction models which represent one of the main features of the GROBID-Dictionaries strategy for structuring digitized dictionaries.

For what concerns the GDLI representation, we are planning to evaluate whether and to what extent the representation model which is being developed within the European ELEXIS project (“European Lexicographic Infrastructure”, Krek et al., 2018) aiming to establish a pan-European infrastructure for lexicography could effectively be used to represent such a complex historical digital dictionary, with a specific view to enabling efficient access to high quality lexicographic data.

6. Acknowledgments

The authors have been partly supported by the EU H2020 programme under grant agreement 731015 (ELEXIS – European Lexicographic Infrastructure).

7. References

- Amsler, M. A. (1981). A taxonomy for English nouns and verbs. In *Proceedings of the 19th Annual Meeting of the ACL*, pp. 133-138.
- Beltrami, P. G. & Fornara, S. (2004). Italian historical dictionaries: from the Accademia della Crusca to the web. *International Journal of Lexicography*, 17(4), pp. 357-384.
- Biffi, M. (2018). Strumenti informatico-linguistici per la realizzazione di un dizionario dell’italiano post-unitario. In D. Fioredistella et al. (eds.) *Proceedings of the 14th International Conference on Statistical Analysis of Textual Data (JADT '18)*, Roma, Universitalia, vol. 1, pp. 99-107.
- Boguraev, B. & Briscoe T. (eds.) (1989). *Computational Lexicography for Natural Language Processing*. Longman.
- Calzolari, N. (1984). Detecting Patterns in a Lexical Database. In *Proceedings of the*

- 10th International Conference on Computational Linguistics, Stanford, California, pp. 170-173.
- Francopoulo, G. (ed.) (2013). *LMF Lexical Markup Framework*. John Wiley & Sons
- Grande Dizionario della lingua italiana*, Opera Diretta da Salvatore Battaglia, Torino, UTET, 1961-2002.
- Granger, S. (2012). Electronic lexicography: From challenge to opportunity. In S. Granger, M. Paquot (eds.) *Electronic Lexicography*, Oxford University Press, pp.1-11.
- Ide, N. & Veronis, J. (1993). Extracting knowledge bases from machine-readable dictionaries: Have we wasted our time? *Knowledge Bases & Knowledge Structures*, 93, Tokyo.
- Khemakhem, M., Foppiano, L. & Romary, L. (2017). Automatic extraction of TEI structures in digitized lexical resources using conditional random fields. In I. Kosem et al. (eds.) *Proceedings of eLex 2017*, September 2017, Leiden, Netherlands. Brno, Lexical Computing.
- Krek, S. (2019). Natural Language Processing and Automatic Knowledge Extraction for Lexicography. *International Journal of Lexicography*, 32(2), pp. 115-118.
- Krek, S., Kosem, I., McCrae, J. P., Navigli, R., Pedersen, B., Tiberius, C. & Wissik, T. (2018). European Lexicographic Infrastructure (ELEXIS). In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (eds.) *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*. Ljubljana: Ljubljana University Press, pp. 881–891.
- Montemagni, S. & Vanderwende, L. (1992). Structural Patterns versus String Patterns for Extracting Semantic Information from Dictionaries. In *Proceedings of COLING-1992*, Nantes, France, pp. 546-552.
- TEI Consortium, Eds. “9. Dictionaries.” TEI P5: Guidelines for Electronic Text Encoding and Interchange. [3.5.0]. [Last modified 29th January 2019]. TEI Consortium. <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/DS.html#DSFLT> (17 June 2019)
- Wilkinson, M. D. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 160018. doi:10.1038/sdata.2016.18.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



Challenges and Difficulties in the Development of Dicionário Olímpico (2016)

Rove Chishman, Aline Nardes dos Santos, Bruna da Silva,
Larissa Brangel

Unisinos University, São Leopoldo, Brazil

E-mail: rove@unisinos.br, aline.nardes@gmail.com, broonamoraes@gmail.com,
larissabrangel@gmail.com

Abstract

This paper discusses some theoretical and practical implications arising from the development of the Dicionário Olímpico (2016), created by the SemanTec (Semantics & Technology) research group. The Dicionário Olímpico (available at <http://www.dicionarioolimpico.com.br/>) is a bilingual lexicographic resource (Portuguese-English) which describes the lexicon of 40 Olympic sports. The dictionary is based on the theoretical-methodological framework of Frame Semantics, developed by Charles J. Fillmore. The paper brings some background to the Dicionário Olímpico's methodological approach. In addition, it describes the lexicographical structure of the resource and the way frame-semantic features were incorporated and adapted in this context. Finally, it explores two kinds of challenges faced by the project: the identification and description of semantic frames, and the design of a template for frame definitions. These stages of development have included some adaptations of frame-semantic concepts with the purpose of building a user-friendly, frame-based dictionary. Such challenges have enriched the lexicographic work and impacted subsequent projects that are yet to be developed by the authors.

Keywords: Frame Semantics; Frame-based dictionary; Dicionário Olímpico.

1. Introduction

The contributions of Frame Semantics (Fillmore, 1982, 1985) to lexicography have been widely addressed since Fillmore's first research works within the context of FrameNet Berkeley, the first frame-based lexicographical database ever published (<https://framenet.icsi.berkeley.edu/>). For example, Atkins, Rundell and Sato (2003) and Atkins, Fillmore and Johnson (2003) approached the contributions of FrameNet to practical lexicography, especially in the process of managing and manipulating corpus data to extract lexicographically relevant information. In this regard, Fillmore and Atkins (1992:75) explored the idea of building an online frame-based lexicographical resource: "In such a dictionary [...], individual word senses, relationships among the senses of polysemous words, and relationships between (senses of) semantically related words will be linked with the cognitive structures (or 'frames'), knowledge of which is presupposed for the concepts encoded by the words."

More recently, advances towards a richer convergence between Frame Semantics and dictionary writing have increased. Specifically, we highlight the works by Ostermann (2012, 2016) concerning Cognitive Lexicography and the improvement of dictionary sections by the inclusion of information based on cognitive theories. In this sense,

practical lexicography imposes many challenges when it comes to articulating cognitive-linguistic theories such as Frame Semantics with dictionary-making processes, since “The craft of lexicography demands not only the ability to collect data, [...] we need to set out these facts in an intelligible and orderly way.” (Atkins, 2002: 171).

This paper aims at discussing some of these challenges within the context of development of the *Dicionário Olímpico* (DO) (<http://www.dicionarioolimpico.com.br/>), a bilingual lexicographic resource (Portuguese-English) which describes the lexicon of 40 Olympic sports. The dictionary is based on the theoretical-methodological framework of Frame Semantics (Fillmore, 1982, 1985). More specifically, the paper approaches some of the challenges faced by the developers during the process of compilation of the *Dicionário Olímpico*, considering that such challenges have enriched the lexicographic work and impacted on subsequent projects that are yet to be developed by the authors. The rest of the paper is structured as follows. Section 2 provides some background to the development of the DO, including its methodological approach. Section 3 describes the lexicographical structure of the dictionary and the way frame-semantic features were incorporated and adapted in this context. Section 4 focuses on two kinds of challenges faced by the project: identification and description of semantic frames (section 4.1), and the ongoing design of a template for frame glosses (section 4.2).

2. Background to the *Dicionário Olímpico*

The *Dicionário Olímpico* is a Brazilian bilingual dictionary of Olympic sports developed within the context of the 2016 Olympic Games. It is the result of a broader academic project whose purpose was to study the potential convergence between Frame Semantics and lexicography for the purpose of describing the lexicon of sports. Two years earlier, the research group responsible for building this resource had already launched a frame-based football dictionary called *Dicionário Field* (<http://dicionariofield.com.br/>), a trilingual resource (in English, Spanish, and Portuguese) structured by semantic frames. During this first lexicographical project, among other results, the group explored the relevance of Frame Semantics for lexicographical practice, not only in terms of enhancing the process of collecting lexicographically relevant information (Chishman et al., 2015), but also with regard to making a dictionary more contextualized by duplicating its macrostructure and enabling users to look up words, frames and different evokers of the same scenario (Santos & Chishman, 2015).

Although some frame-semantic assumptions are adapted in these projects (see Section 3.1 for more details), it is important to approach the theory’s core concepts that underlie the building of the *Dicionário Field* and the *Dicionário Olímpico*, enriching their content and access structures. According to Fillmore and Baker (2010: 237), Frame Semantics assumes that “[...] the meaning dimension is expressed in terms of the cognitive structures (frames) that shape speakers’ understanding of linguistic expressions.” For example, in football, a word such as *assist* can only be understood if a speaker recognizes the cognitive structure it evokes, which is constituted of encyclopaedic and sociocultural information:

in football, a player *assists* a scorer of a goal when he passes him the ball. Therefore, to *assist* means to supply a specific (and decisive) kind of pass in football – hence this word evokes the Pass frame. As Fillmore (1985: 229) states, “Frame semantics allows the possibility that speakers can have full knowledge of the meaning of a given word in a domain [...]”. In other words, understanding a word (or, in frame-semantic terms, understanding a lexical unit) implies recognizing the frame it evokes.

The challenge of describing the language of sports through semantic frames became bigger with the development of the *Dicionário Olímpico*. Firstly, while *Field* is a football dictionary, *DO* describes 40 Olympic sports. Secondly, the corpus compilation imposed other difficulties: to build a corpora for the basis of *Field*’s lexicographical work, the editors selected match reports from football websites, which is a pervasive text genre both in Brazilian Portuguese and in English (more specifically, those on British websites). However, in the context of the *Dicionário Olímpico*, only a few Olympic sports, such as volleyball and basketball, are as popular as football in Brazil; thus match reports could not be used as the main sources to build all corpora. In case of less popular games, sometimes the only reliable written documents available concerned the rules of these sports.

Therefore, in order to broaden the range of text genres for corpus compilation purposes, the following procedures were adopted: transcription of match videos available online; compilation of documents such as sports rules and other official materials; and a qualitative study of sports-related videos and other multimodal materials whose content was not processable by a corpus tool, nor worth transcribing – since it is a very time-consuming task. Indeed, these multimodal sources provided supporting information and were used as a reference material for comparing and complementing the study corpora.

The *Dicionário Olímpico*’s corpora were processed and managed through Sketch Engine. This tool is renowned for its relevance for dictionary writing, especially due to the word sketches it provides, which “combine information of two types: grammatical relations in the corpus, and statistically significant frequencies of co-occurrence” (Atkins et al., 2003: 336). As described by Chishman et al (2017), after planning the macro- and microstructure of the dictionary, the development of the *Dicionário Olímpico* included the following stages: (i) study of sports and systematization of their main characteristics; (ii) corpus design and compilation, including documents such as sports rules and match reports, if available; (iii) gathering of multimodal supporting material, especially in case of little-known sports; (iv) creation of conceptual maps regarding the respective domains, which were based on the previously collected written corpora and multimodal sources; (v) description of semantic frames, based on the previously designed and discussed conceptual maps; (vi) corpus extraction of possible lexical frame evokers and their equivalents; (vii) writing and collective revision of the entries by the editors, with the assistance of sports experts (for example, coaches and former players); (viii) building of the entries on the dictionary website database. All these stages have brought many challenges that have been, or are yet to be, discussed.

3. The Dicionário Olímpico: lexicographic structure

As we saw earlier, since the Dicionário Olímpico (DO) was developed from the theoretical-methodological framework of Frames Semantics, many aspects of the lexicographic structure of this tool were based on FrameNet's lexicographic structure. However, there is only a slight degree of similarity between these two tools, since the target audiences also differ.

At this point, it is relevant to mention that the target audience consists of people who relate directly to the Olympic modalities, such as students, athletes and other sports professionals; and also includes users whose relationship with Olympic sports is indirect, such as translators and people interested in this topic. Above all, the DO audience includes people who do not necessarily have any extensive knowledge of linguistics' or lexicography's theoretical concepts.

With this in mind, in this section, we describe the lexicographic structure of the DO: how to access data and the levels of the dictionary. In addition, we discuss how the notion of frame has been incorporated into the project, emphasizing the centrality of the intended audience in the process of definitions regarding the content and form of the dictionary.

3.1 Access to data

Considering that the DO is composed of 40 dictionaries, each one corresponding to one of the sports that comprise the framework of the Summer Olympics, the resource's homepage enables users to select a specific Olympic modality (from the respective icons) or the search for a word, scenario, or modality (from the search box), as shown in Figure 1.

At this point, it is necessary to approach the first adaptation that was necessary in the development of the DO. In the context of FrameNet, the terminological concepts 'frame' and 'lexical unit' are used. This is due to the fact that the target audience comprises predominantly linguistics researchers, teachers, and students, i.e., people who are familiar with these theoretical concepts.

On the other hand, the potential audience of the Dicionário Olímpico is composed of non-specialists. For this target audience, the use of theoretical concepts could lead to a communication failure. With that in mind, the SemanTec research group adopted words that sound more familiar to the user. Thus, the word 'frame' was replaced by 'scenario', and the expression 'lexical unit' was replaced by 'word' in the structure of the dictionary.

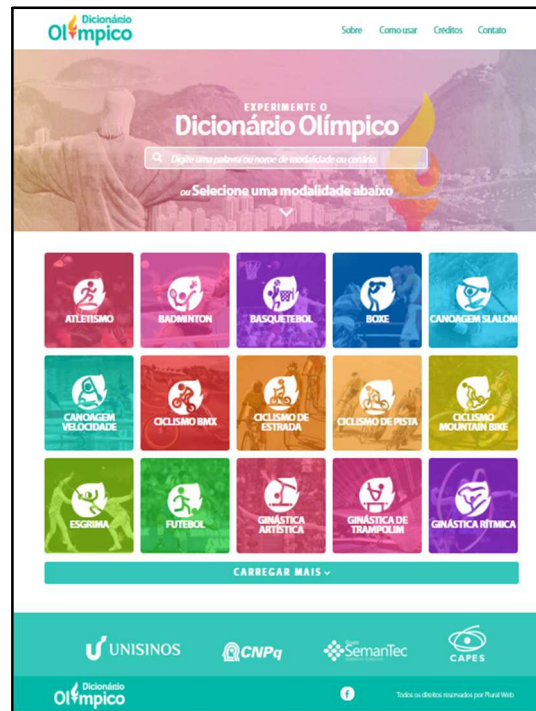


Figure 1: Dicionário Olímpico's homepage

3.2 Access levels of the Dicionário Olímpico

When selecting one of the forms of access, users are directed to one of the three levels of the DO: the modality level, the scenario level, or the word level. Each of them is presented in the following subsections.

3.2.1 First level: Olympic modality

When selecting one of the sports on the homepage, users are directed to a page containing this set of information: gloss (supergloss), conceptual map, scenario list, word list, trivia section, related sports, and image, as shown in Figure 2.

The most significant differences between the Dicionário Olímpico and FrameNet are at this level. While FrameNet describes general frames, the Olympic Dictionary describes the frames of Olympic sports, which are called, in this context, superframes. Thus, each frame of the Dicionário Olímpico corresponds to an Olympic modality, and not to the Olympic domain as a whole. In contrast, FrameNet does not group frames by domains.

For this reason, this level presents elements that do not exist in FrameNet, such as conceptual maps for each Olympic modality and the trivia section, which are a result of decisions made during the Dicionário Olímpico development process. The reasons for these decisions are explained in the next sections.

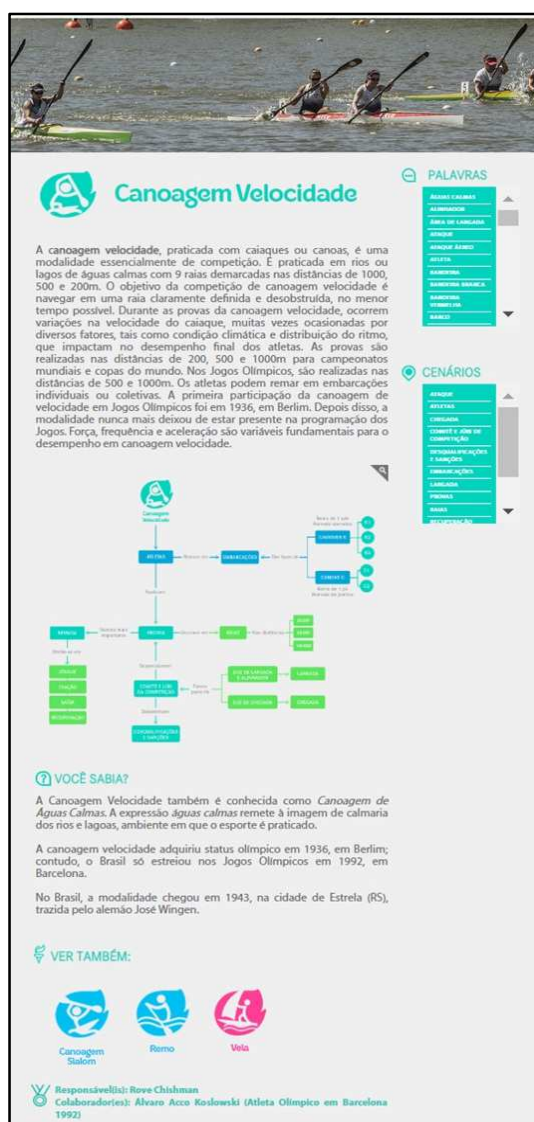


Figure 2: Level of the Olympic modality or superframe

3.2.2 Second level: scenario

In terms of content, the level of the scenario resembles the modality level. The elements that constitute it are: gloss, list of words, related frames, image, and conceptual map, as shown in Figure 3.

Elements that were based on FrameNet's structure, such as gloss, word list, and relations between scenarios, have undergone some modifications. Regarding the relations between scenarios, it is worth mentioning that initially the editors intended to use the set of frame relations created by FrameNet: inheritance, perspective, use, subframe and precedence. However, on submitting the dictionary content to the experts' inspection, the research group received negative feedback. According to these professionals, these relations were obscure; they were not user-friendly. For this reason,

FrameNet relations were not used, and those responsible for each Olympic sport were in charge of identifying the types of relations that could be established between the frames, based on the study of each discipline.

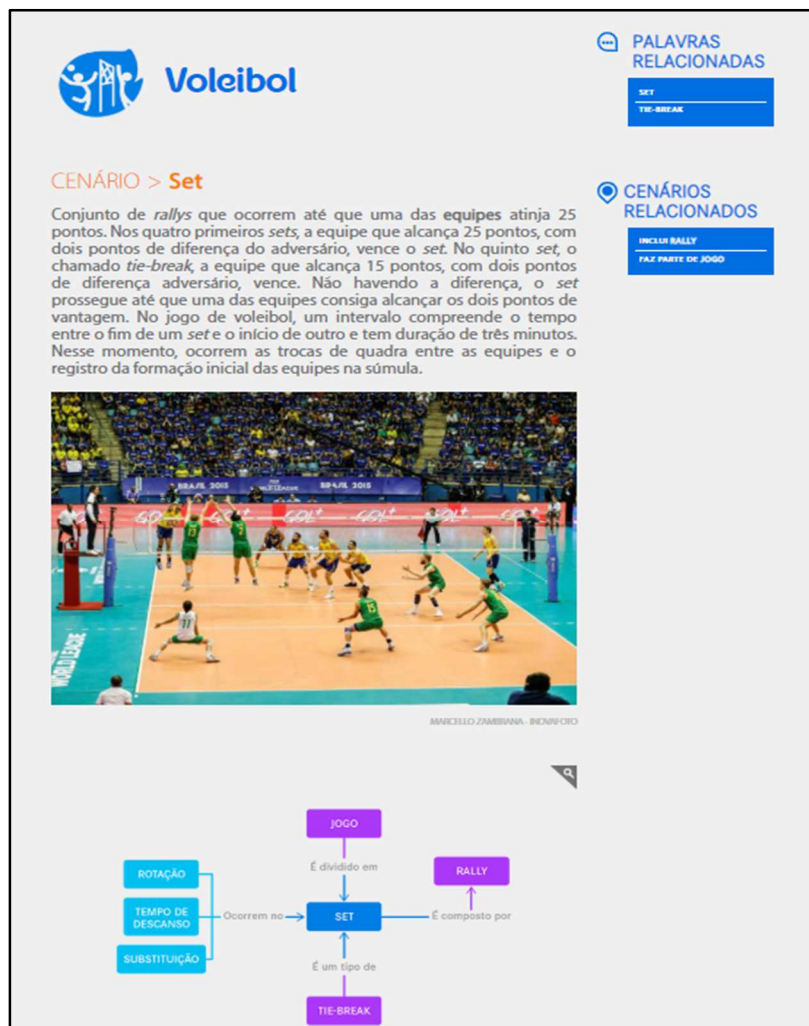


Figure 3: Scenario level.

The following figure presents the relations for the basketball frame called Basket: basket depends on Shot; generates Throw-in; uses Team; and uses Court. Other types of relations used in this context were ‘part of’, such as in the badminton frame Equipment (Equipment is part of Court); ‘to control’, as in the beach volleyball frame Refereeing (Refereeing controls the Match); and ‘to execute’, as in the tennis frame Tennis Players (Tennis Players execute Shot).

CENÁRIO > Cesta

O principal objetivo de uma equipe em um jogo de basquete é obter pontos ao fazer com que a **bola** passe por dentro do **aro** localizado na quadra de ataque. A conversão de um lance livre vale 1 ponto. Os arremessos convertidos durante as jogadas normais de jogo – as chamadas cestas de campo – podem valer 2 pontos, quando originadas de arremessos convertidos de dentro da linha de três pontos, ou 3 pontos, quando provenientes de arremessos realizados de fora da linha de três pontos.

📍

CENÁRIOS RELACIONADOS

DEPENDE DE ARREMESSO

GERA REPOSIÇÃO DE BOLA

USA EQUIPE

USA QUADRA

Figure 4: Relations between scenarios

Furthermore, the glosses of the Dicionário Olímpico, an element that is discussed in more detail in the next section, do not follow the structure of FrameNet’s standard glosses, which are built through the following steps: (i) characterizing the frame; and (ii) describing and naming frame elements (Fillmore & Baker, 2009).

Revenge

[Lexical Unit Index](#)

Definition:

This frame concerns the infliction of punishment in return for a wrong suffered. An **Avenger** performs a **Punishment** on a **Offender** as a consequence of an earlier action by the **Offender**, the **Injury**. The **Avenger** inflicting the **Punishment** need not be the same as the **Injured party** who suffered the **Injury**, but the **Avenger** does have to share the judgment that the **Offender**’s action was wrong. The judgment that the **Offender** had inflicted an **Injury** is made without regard to the law.

Figure 5: FrameNet gloss model

In Dicionário Olímpico, it was considered that these elements would not receive the prominence they have in FrameNet. Instead, glosses – both glosses (scenarios) and superglosses (Olympic modalities) – feature prominent words that are not necessarily frame elements, but can be viewed as keywords that are necessary to understand the respective frame.

CENÁRIO > Resultado

O resultado de um **jogo** de badminton pode ser parcial ou final. O **resultado parcial** envolve o número de **pontos** marcados pelos jogadores durante um **set**. O **resultado final** envolve o número de **sets** vencidos por cada jogador ou equipe ao final do jogo. Para vencer um set, um jogador ou dupla precisa conquistar marcar 21 pontos, com dois pontos de diferença do adversário. Para vencer um jogo, os jogadores precisam conquistar o **melhor resultado de 3 sets**.

📍

CENÁRIOS RELACIONADOS

DEPENDE DE ARREMESSO

GERA REPOSIÇÃO DE BOLA

USA EQUIPE

USA QUADRA

Figure 6: Gloss model of the Dicionário Olímpico



Figure 7: Dicionário Olímpico supergloss model

The structure of the modality and the scenario levels resemble each other, according to their nature. In the context of the DO, modalities are considered more comprehensive frames (superframes), and for that reason they should be described in a similar way to how are described.

An element that integrates the levels of the Olympic modality and the scenario is the conceptual map. Initially used only as a methodological strategy for the organization of information about modalities, the conceptual maps were later included in the access structure because they include, albeit implicitly, some notions underlying Frame Semantics.

The task of connecting frames and frame elements refers to the notion of frames as sets of related concepts, in such a way that to understand one of them it is necessary to understand the system as a whole (Fillmore, 1982). Thus, by locating a frame or frame element on the conceptual map, users identify the role that such unit plays within the system. In addition, the way these relations between concepts are presented refers to FrameNet's frame-to-frame relations. From this information, users identify the ways in which, for example, one frame contributes to a preceding one or how a frame integrates a larger one (subframe).

Finally, the inclusion of images (modality level and scenario level), the trivia section and "see also" (both at the modality level) aim at meeting the encyclopaedic character of the dictionary. Images, for example, play a role as frame evokers. The "see also" section, in turn, highlights the similarities between sports whose structures share the same bases (for example, rhythmic and artistic gymnastics).

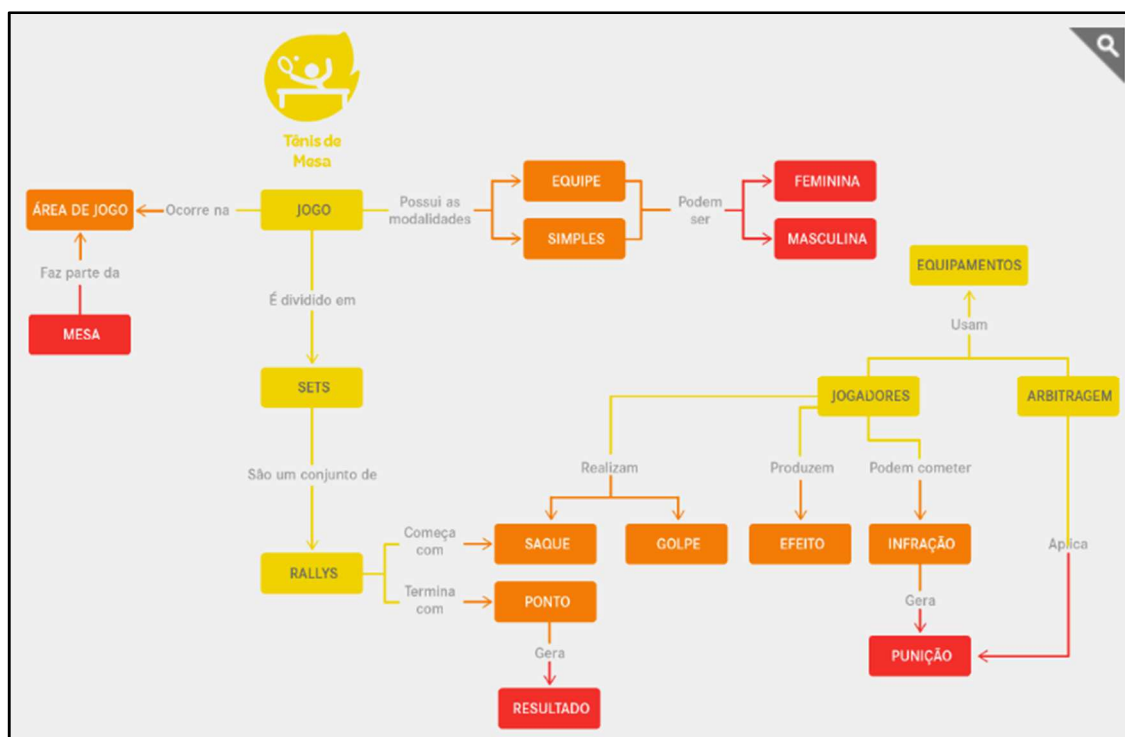



Figure 8: Conceptual map of table tennis

3.2.3 Third level: word

The third and last level of the DO presents information related to the words of the Olympic modalities. From this level, users have access to the grammatical classification of the word, the scenario which the word searched evokes, the English equivalent, an example and a list of other words that integrate the correspondent scenario. Notes are presented in some cases, for the purpose of providing more specific information about a word. In addition, variants are presented when the same phenomenon can be named in two or more different ways.



Voleibol

PALAVRA > saque *sm.*

CENÁRIO: Saque

VARIANTE: serviço

INGLÊS: service

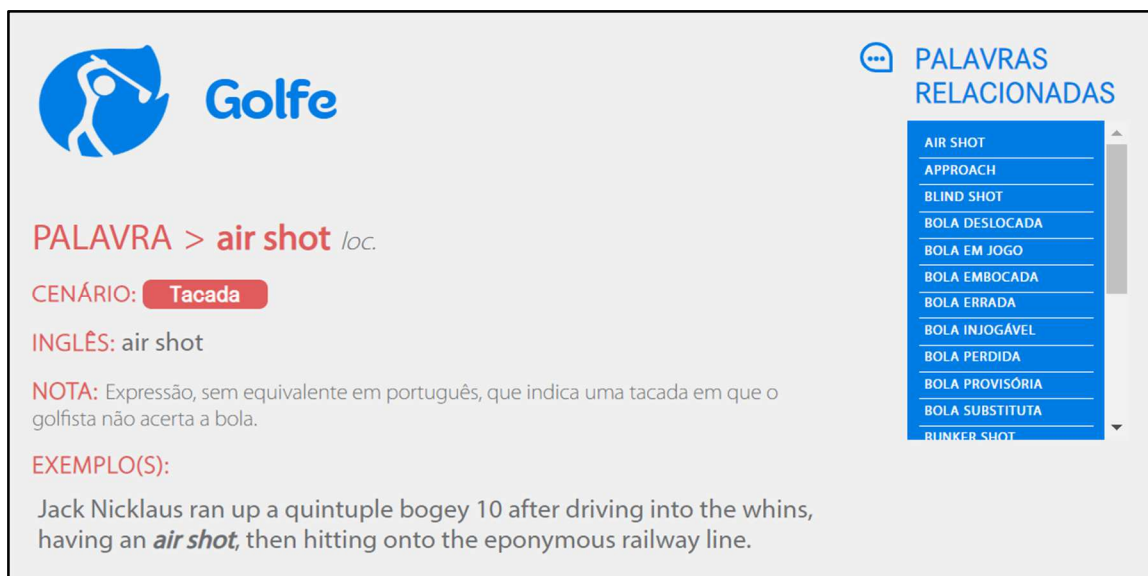
EXEMPLO(S):

The Brazilian *service* was more effective on the first set.

PALAVRAS RELACIONADAS

- AUTORIZAÇÃO DO SAQUE
- AUTORIZAR O SAQUE
- BOLA EM JOGO
- CAIXINHA
- FORÇAR O SAQUE
- SACAR
- SAQUE
- SAQUE FLUTUANTE
- SAQUE POR BAIXO
- SAQUE POR CIMA
- SAQUE VIAGEM

Figure 9: Level of word 1



The screenshot shows a dictionary interface for the word 'air shot' in the context of golf. At the top left is a logo with a golfer silhouette and the word 'Golfe'. The main entry for 'PALAVRA > air shot' includes its locutionary form, a scenario ('Tacada'), its English translation, a note explaining its meaning in Portuguese, and an example sentence. On the right, a sidebar titled 'PALAVRAS RELACIONADAS' lists related golf terms.

Golfe

PALAVRA > air shot *loc.*

CENÁRIO: Tacada

INGLÊS: air shot

NOTA: Expressão, sem equivalente em português, que indica uma tacada em que o golfista não acerta a bola.

EXEMPLO(S):

Jack Nicklaus ran up a quintuple bogey 10 after driving into the whins, having an *air shot*, then hitting onto the eponymous railway line.

PALAVRAS RELACIONADAS

- AIR SHOT
- APPROACH
- BLIND SHOT
- BOLA DESLOCADA
- BOLA EM JOGO
- BOLA EMBOCADA
- BOLA ERRADA
- BOLA INJOGÁVEL
- BOLA PERDIDA
- BOLA PROVISÓRIA
- BOLA SUBSTITUTA
- BUNKER SHOT

Figure 10: Level of word 2

In comparison to FrameNet, the DO's lexical unit entries do not include features such as semantic type, frame elements and their syntactic realizations, and valence patterns. The editors considered that such information could represent an overly theoretical level, considering the intended audience for the dictionary. Other kinds of information that was suppressed concerned the definition of lexical units. However, the notes on the DO have a similar function to FrameNet's definitions.

Variants and translation equivalents were also proposed. Regarding variants, it is worth noting that their use was quite broad in the dictionary. This was due both to the regional differences in Brazil and to the fact that Olympic sports that are not widespread in the country present many words in English which have not yet been fully adopted in Brazilian Portuguese.

In this section, we presented the lexicographic structure of DO, showing the similarities and differences that this tool presents in comparison to FrameNet. We intended to highlight the reasons that led to adaptations of some of the FrameNet's features and to the inclusion of new elements in the dictionary. In the next section, we address some of the key challenges faced in the process of developing the DO and discuss how we dealt with such difficulties.

4. Challenges in the development of the Dicionário Olímpico

As demonstrated in the previous sections, the lexicographical structure of the DO provides the user with all the modalities of the Olympic sports in the form of superframes. Inside each superframe (that can be accessed by a hyperlink), the user can find a set of information about the sports, such as images, conceptual maps, lists of words, lists of frames and glosses. This section presents a brief overview about the challenges and difficulties faced by the SemanTec group during the description of frames: identification and description of semantic frames (section 4.1), and writing of

the glosses (section 4.2). Moreover, these sections discuss how such difficulties have been circumvented in the compilation of *Dicionário Paralímpico*, a dictionary of the Paralympic sports that is currently under development.

4.1 Identification and description of semantic frames

In a frame-based dictionary, all structural elements are somehow subordinate to the set of frames described. Therefore, among the tasks of compiling a dictionary of this nature, the frame definition step occupies a central position, since it is the stage from which the dictionary begins to be constructed.

As we saw in the previous section, the DO compilation process comprised a series of adaptations of FrameNet's lexicographic model. In this regard, one of the stages that was not based on FrameNet's methodology was the step of identification of frames.

Regarding FrameNet, Fillmore and Baker (2009: 320) state that "The method of inquiry is to find groups of words whose frame structures can be described together, by virtue of their shared common schematic backgrounds." However, the method of identifying frames used by FrameNet compilers is not explicit. In describing the process of lexical analysis of the platform, for example, the authors begin the process of frame identification with a step related to the characterization of the frame.

It is important to highlight that not even the frames already described by FrameNet could be used as a starting point to describe Olympic sports' scenarios, since FrameNet does not describe frames of more specific domains. In addition, establishing the frameset of a general language and describing frames from a sports domain are not equivalent activities. With this in mind, the identification of the Olympic frames started from scratch and can therefore be considered one of the most challenging tasks performed during the development of the DO.

Therefore, the SemanTec research group outlined a methodology for identifying the frames based on the conceptual mapping of the Olympic modalities. A similar procedure was used in the development of the Field dictionary. However, it was in the context of the *Dicionário Olímpico* that the use of this methodology acquired more definite contours.

This procedure was constituted of two steps: elaboration of the general map and design of the map of the frames. In the first stage, the editors in charge of the description of each Olympic modality elaborated a more comprehensive conceptual map, describing the sports with a high level of detail. In order to do so, the group studied the support materials, mentioned in the beginning of this paper, from which detailed information of the sports, including terms, expressions and specific concepts, were extracted. At this stage, conceptual maps eventually incorporated the organizational structure of sports manuals, since many titles and sections of these materials were converted into central map nodes.

In the second stage, from the more general map, it was possible to design a conceptual map containing only the Olympic sport's frames. In this process of refinement, the objective was to build the final conceptual map of each Olympic modality and to establish a definitive list of frames. The main methodological procedure for this step was the systematization of the list of lexical units, in order to divide them into groups of words that together evoked the frames of each sport. As a final step, the material was sent to an expert.

In view of the innovation represented by the use of conceptual maps in a frame-based dictionary, and considering the lack of methodological support for the elaboration of these maps, the strategies described above represent a first step towards dealing with challenges of this nature. Currently, in the process of compiling the *Dicionário Paralímpico*, the group has been discussing new forms of frame identification, in order to improve this method and to evaluate the most efficient methodological procedures.

4.2 The glosses

First of all, it is necessary to define what we understand by gloss in the context of the DO. In semasiological dictionaries, a gloss is usually regarded as “a paraphrase or synonym used within a dictionary entry to provide an explanation of the sense of a word or phrase related to the headword” (Hartmann & James, 2002, s.v. gloss). This is not, however, an applicable definition to the glosses of the *Dicionário Olímpico*, which are, in fact, brief texts located in specific sections of the dictionary with the purpose of providing the user with information about the Olympic sports. Once the glosses of the DO comprise information classified as “encyclopaedic”, they are more closely related to an encyclopaedic definition: “a definition which reflects encyclopaedic knowledge (about facts) rather than linguistic knowledge (about words)” (Hartmann & James, 2002, s.v. encyclopaedic definition). The term “definition”, however, is still often related to the brief explanations found in the entries of semasiologic monolingual dictionaries, and this is the reason why we gave a proper nomenclature to the textual information about the sports in DO: gloss.

The glosses of DO are located in two specific parts of the dictionary: in modality entries and in frame entries. In the modality entries, the glosses provide the user with a set of information about a specific Olympic sport, helping them to know the main facts and features about the sport. This part of the dictionary is called superframe (as referred in Section 3.2.2), and this kind of gloss is called supergloss. In the frame entries, glosses intend to describe the frames of each sport, helping users to understand some specifics of each Olympic modality, such as the equipment used and the rules of the games. The figures below present the supergloss of artistic gymnastics and the gloss of one of its frames:



Figure 11: The supergloss of artistic gymnastics¹

¹ Translation: “Artistic gymnastics is a sport of formal precision in which gymnasts must present a routine composed of acrobatic and gymnastic elements in one of the apparatuses of the competition. In women’s competition, gymnasts perform on four events: uneven bars, vault, floor and balance beam. In men’s competition, gymnasts compete on vault and floor too, and also on the still rings, the horizontal bar, the parallel bars, and the pommel horse. A jury composed of 8 judges evaluate the gymnasts according to the level of difficulty of the routine (based on the value of the elements that make up each routine, established by the punctuation code) and also according to its execution (according to the quality and technical accuracy of the movements performed by the gymnasts). When performing the routine, gymnasts can make mistakes, which lead to score deduction, or perform combined movements or highly difficult movements, which lead to bonus points. These items determinate the gymnast routine score. The routines usually present an entry, a way of starting the presentation and contacting the apparatus, the execution of the elements of the routine, and dismount, the ending of the routine and the termination of contact with the apparatus. Different kinds of elements are performed by the gymnasts in acrobatics. Somersaults, pirouettes, dance jumps and supports are some of the elements which, performed in sequence, make up the routine. Present in Olympics since the first edition of the Modern Games, in Athens in 1886, artistic gymnastics competitions consist of individual all-around, team, or individual events. In the finals, only 8 gymnasts or teams that get the best scores in the qualifying round compete.”

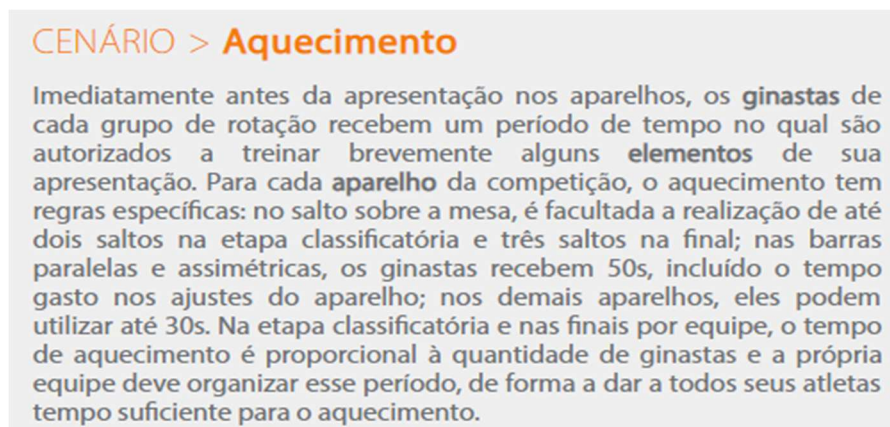


Figure 12: The gloss of one of artistic gymnastics frames²

The examples above show that the glosses of DO are strongly characterized by the use of encyclopaedic information and by their extended size, which are two important features that distinguish the DO glosses from the lexicographical definitions usually found in semasiologic, monolingual dictionaries. The first feature (the encyclopaedic information) brought to the DO compilation one of the biggest challenges faced by the SemanTec group during the writing of the glosses, impacting also on the second feature (the size of the glosses).

The next paragraphs approach this experience of writing the glosses. First and foremost, it is important to highlight that the distinction between linguistic knowledge and encyclopaedic knowledge has pervaded debates in Linguistics for a long time. A very important contribution from Cognitive Semantics to this discussion is the intensification of the idea that it is not always possible to make a rigid distinction between knowledge of language and knowledge of facts (see Evans & Green, 2006: 160-162; Riemer, 2010: 100-105; Geeraerts, 2010:222-224). As Riemer (2010: 104) postulates, “we know a variety of things about words and their denotation, and the greater the likelihood that a particular piece of this knowledge is shared between speaker and hearer, the greater the likelihood that it will determine the word’s linguistic properties”.

One of the consequences of this discussion to lexicography concerns the lexicographical definition, and, in particular, the content of definitions: how can lexicographers choose the best encyclopaedic information to define lexical items? If, on the one hand, “linguistic” information seems easier to be identified and chosen for the writing of definitions, on the other hand, encyclopaedic information corresponds to a larger

² Translation: “Immediately before the performance on the apparatus, the gymnasts in each rotation group are given a period of time in which they are allowed to briefly train some elements of their presentation. For each apparatus of the competition, the warm-ups have specific rules: for the vault, it is allowed to perform up to two vaults in the qualifying stage and three vaults in the final; for parallel and asymmetric bars, gymnasts have 50 seconds, including time spent adjusting the apparatus; for other apparatus, they can use up to 30 seconds. In the qualifying rounds and finals per team, the warm-up time is proportional to the number of gymnasts, and the teams must organize themselves in order to give all their athletes enough time to warm up.”

amount of information, once it consists of our “knowledge of the world” (Matthews, 2007, s.v. encyclopaedic knowledge). This knowledge of the world represents an immeasurable amount of information; and it would be obviously impossible to allocate all the encyclopaedic information about a word in a single dictionary entry. Thus, when a lexicographer proposes to create encyclopaedic definitions for dictionary entries – whether brief definitions of printed semasiological dictionaries or longer definitions, such as the definitions of DO – this lexicographer will always face the challenge of choosing the most appropriate information to describe lexical units.

Let us take the example of football. Which information would be indispensable to describe its meaning? The fact that it is a sport in which players use their feet? That the objective is scoring goals? That the teams have supporters? That the games take place at stadiums? That the match is played by two teams? That the teams are composed by eleven players? That each match is divided into a first and a second half of 45 minutes each? That between the first and the second half there is a break of 30 minutes? That there is an official football World Cup? That Pelé is considered the king of football? We emphasize that we are not even trying to separate linguistic from encyclopaedic knowledge – we are just trying to list what is essential in the definition of the word *football*.

In the context of the DO, without having a methodology that could guide the lexicographers to choose the most adequate information for the description of the sports, each editor found their own way to describe the sports they were responsible for. They used especially their linguistics intuition based on the studies previously developed about the sports. At the end, the editors compiled a group of glosses which could meet the demands of the DO users, although there were significant differences between them, especially because of the size and kind of information presented. The two examples below demonstrate this:

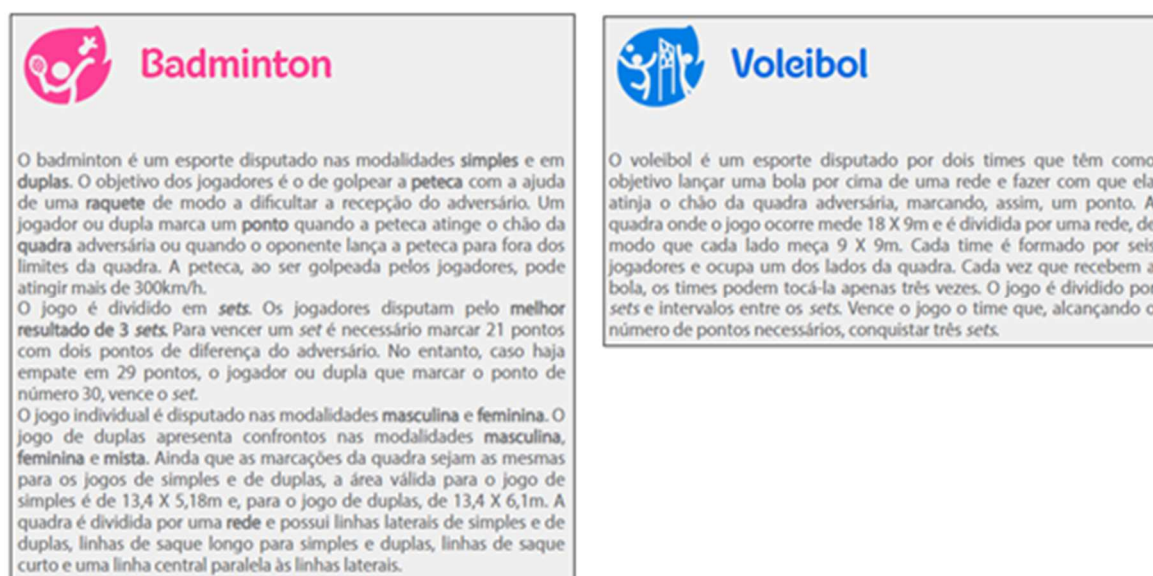


Figure 13: Size difference between two glosses

The images above present the size difference between two superglosses. Once they define different sports, it would be expected the two could diverge from each other with regard to length, especially because some sports may require specific explanations about specific features, while other sports may not. Even so, we believe that it would be possible to create a methodology for the writing of the superglosses, presenting them in a more standardized form, especially in terms of size and content.

We have put this into practice during the compilation of the *Dicionário Paraolímpico*, the most recent dictionary produced by the *SemanTec* group that is currently under construction. *Dicionário Paraolímpico* will present the same lexicographic structure as the *DO* and will also have *Frame Semantics* as a guideline. This dictionary has benefited from all the expertise acquired by the group during the compilation of the *DO*, which has been helping the group to reflect on new strategies to solve some challenges such as the writing of the glosses.

The methodology for the writing of the glosses of *Dicionário Paraolímpico* is currently under development. It proposes to split the gloss into two parts. The first part is intended to have the form of an intensional definition, which enumerates a set of important features of the Paralympic sports. To construct this part, we base our work on a study that proposes a classification of sports (Gonzalez, 2004). In this study, Gonzalez (2004) classifies the sports based on four parameters that he calls “relation to the opponent”, “relation to the objective”, “relation to the partner” and “relation to the environment”. Considering this division, the first part of the gloss will present the information below (we added one more parameter, the “objective”):

- 1) Kind of sport (relation to the objective): translation / fight / field and bat or court / split court or wall / by demarcation / aesthetic or technical combinatory / precision or target
- 2) Relation to the partner: individual / collective
- 3) Objective
- 4) Relation to the opponent: interaction with the opponent or direct opposition to the opponent / no interaction and no direct opposition
- 5) Relation to the environment: stability / no stability

The second part of the gloss describes some specifics of the Paralympic competition, opposing, if possible, the Paralympic sport to its Olympic counterpart. In this part of the gloss, we intend to include an extensional format of definition that provides the user with encyclopaedic information about the Paralympic sports. Putting this methodology into practice, we have developed the following template to guide the writing of Paralympic sports' glosses:

PART 1: ____ (name of the sport) ____ é um esporte de/do tipo ____ (1) ____ disputado/que pode ser disputado ____ (2) ____ cujo objetivo é ____ (3) ____ [descrição da sequência do ato esportivo]. No(a) ____ (name of the sport) ____, a relação com o adversário ocorre de maneira ____ (4) ____ através de [descrever relação entre atletas no ato da competição]. O ____ (name of the sport) ____ é praticado em [descrever ambiente], ambiente que oferece/não oferece ____ 5 ____ para o atleta.

PART 2: specificities of Paralympic competition and the differences between the Paralympic sport and its Olympic counterpart.

The following gloss is an example of application of this template to a Paralympic sport – football 5-a-side:

O futebol de 5 é um esporte de quadra disputado coletivamente cujo objetivo de cada equipe é marcar gols na área adversária. No futebol de 5, a relação com o adversário ocorre por oposição direta através de disputas de bola, dribles, passes e chutes a gol. O futebol de 5 é geralmente praticado em quadras adaptadas de futebol de salão, podendo também acontecer em campos de grama sintética, ambientes que oferecem estabilidade para o atleta. Em relação a sua contraparte olímpica, o futebol de cinco diferencia-se por ser disputado por atletas cegos, que utilizam vendas nos olhos para garantir condições iguais a todos os participantes. A bola da partida possui guizos internos para que os jogadores possam localizá-la e a quadra possui bandas junto às linhas laterais, para evitar que a bola saia. Durante a partida, existe um guia, que recebe o nome de chamador, que fica atrás do gol para orientar os jogadores em relação ao seu posicionamento em campo e chutes a gol. As partidas de futebol de 5 acontecem de maneira silenciosa; a torcida só tem permissão de se manifestar quando acontecem gols.³

³ Translation: “5-a-side football is an indoor sport played collectively whose purpose is to score goals in the opposing area. In a 5-a-side football match, the relationship of opponents occurs by direct opposition through ball disputes, dribbling, passes and goal shots. 5-a-side football is usually played on courts adapted from indoor soccer and may also take place on synthetic grass, places that offer stability for the athletes. Differently from its Olympic counterpart, 5-a-side football is played by blind athletes who use blindfolds to ensure equal conditions to all participants. The ball has bells inside to aid the players in their movements, and the court has bands along the lines to prevent the ball from coming out. During the match there is a guide, who receives the name of caller and stands behind the goal to guide the players’ positions and shots. 5-a-side football has quiet matches; the public is allowed to cheer only when a goal is scored.”

5. Final considerations

This paper presented an overview of the challenges and difficulties faced in the development of the Dicionário Olímpico. In the previous pages, we presented some problems we faced during our work and how we dealt with some of these issues. As many studies have shown, Cognitive Linguistics and Frame Semantics have proved to be important theoretical frameworks for lexicography (especially for online dictionaries). Considering this potential, one of the biggest challenges of Cognitive Lexicography is to build its own methods to convert the principles of a cognitive theory of language into tools for dictionary making. The lexicographic products presented in this paper integrate this enterprise.

6. Acknowledgements

This study was financed by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) – Finance Code 001 –; the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq); and the Fundação de Amparo à pesquisa do Estado do RS (FAPERGS).

7. References

- Atkins, S. (2002) Then and now: competence and performance in 35 years of lexicography. In A. Braasch, & C. Povlsen (eds.) *Proceedings of the Tenth EURALEX International Congress*. Copenhagen: Center for Sprogteknologi, pp. 247-272.
- Atkins, S., Fillmore, C. J. & Johnson, C. R. (2003). Fuzzy SF: Lexicographic Relevance: Selecting Information from Corpus Evidence. *International Journal of Lexicography*, 16(3), pp. 251–281.
- Atkins, S., Rundell, M. & Sato, H. (2003). The contribution of FrameNet to Practical Lexicography. *International Journal of Lexicography*, 16(3), pp. 333–357.
- Chishman, R. et al. (2015). The relevance of the Sketch Engine software to build Field - Football Expressions Dictionary. *Revista de Estudos da Linguagem*, 23, pp. 769-796.
- Chishman, R. et al. (2017) Dicionário Olímpico: a semântica de frames encontra a lexicografia eletrônica. In M. J. B. Finatto et al (eds.) *Linguística de Corpus: Perspectivas*. Porto Alegre: Instituto de Letras - UFRGS, pp. 265-298.
- Evans, V. & Green, M. (2006). *Cognitive Linguistics: an introduction*. Edinburgh: Edinburgh University Press.
- Fillmore, C. J. (1982). Frame Semantics. In The Linguistics Society of Korea (ed.) *Linguistics in the Morning Calm*. Seoul: Hansinh Publishing Co., pp. 111–137.
- Fillmore, C. J. (1985). Frames and the semantics of understanding. *Quaderni di Semantica*, 6(2), pp. 222-254.
- Fillmore, C. J. & Atkins, S. (1992). Toward a Frame-based Lexicon: The Semantics of RISK and its Neighbors. In A. Lehrer & E. Kittay (eds.) *Frames, Fields and Contrasts: New Essays in Semantic and Lexical Organization*. Hillsdale: Erlbaum,

- pp. 75-102.
- Fillmore, C. J. & Baker, C. (2009). A frames approach to semantic analysis. In B. Heine & H. Narrog (eds.) *The Oxford Handbook of Linguistic Analysis*. New York: Oxford University Press, pp. 313-339.
- Geeraerts, D. (2010). *Theories of Lexical Semantics*. New York: Oxford University Press.
- Gonzalez, F. J. (2004). Sistema de classificação de esportes com base nos critérios: cooperação, interação com o adversário, ambiente, desempenho comparado e objetivos táticos da ação. *Lecturas: Educación Física y Deportes*, 10(71).
- Hartmann, R. R. K. & James, G. (2002). *Dictionary of lexicography*. 2.ed. London/New York: Routledge.
- Matthews, P. H. (2007). *Oxford Concise Dictionary of Linguistics*. 2nd ed. Oxford: Oxford University Press.
- Ostermann, C. (2012). Cognitive lexicography of emotion terms. In R. V. Fjeld. & J. M. Torjusén (eds) *Proceedings of the 15th EURALEX International Congress*. Oslo: Department of Linguistics and Scandinavian Studies/University of Oslo, pp. 493-501.
- Ostermann, C. (2016). *Cognitive Lexicography: a new approach to lexicography making use of Cognitive Semantics*. Berlin; New York: Mouton de Gruyter.
- Riemer, N. (2010). *Introducing Semantics*. New York: Cambridge University Press.
- Santos, A. & Chishman, R. (2015). O papel da Semântica de Frames na construção de um recurso dicionarístico: a organização lexicográfica do FIELD - Dicionário de Expressões do Futebol. *Revista da ABRALIN*, 14, pp. 433-468.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



The ELEXIS Interface for Interoperable Lexical Resources

John P. McCrae¹, Carole Tiberius², Anas Fahad Khan³,

Ilan Kernerman⁴, Thierry Declerck^{5,7}, Simon Krek⁶,

Monica Monachini³ and Sina Ahmadi¹

¹ Data Science Institute, National University of Ireland Galway

² Instituut voor de Nederlandse Taal

³ CNR- Istituto di Linguistica Computazionale «A. Zampolli»

⁴ K Dictionaries

⁵ Austrian Centre for Digital Humanities, Austrian Academy of Sciences

⁶ Jožef Stefan Institute/University of Ljubljana

⁷ DFKI GmbH, Multilinguality and Language Technology Lab

Abstract

ELEXIS is a project that aims to create a European network of lexical resources, and one of the key challenges for this is the development of an interoperable interface for different lexical resources so that further tools may improve the data. This paper describes this interface and in particular describes the five methods of entrance into the infrastructure, through retrodigitization, by conversion to TEI-Lex0, by the TEI-Lex0 format, by the OntoLex format or through the REST interface described in this paper. The interface has the role of allowing dictionaries to be ingested into the ELEXIS system, so that they can be linked to each other, used by NLP tools and made available through tools to Sketch Engine and Lexonomy. Most importantly, these dictionaries will all be linked to each other through the Dictionary Matrix, a collection of linked dictionaries that will be created by the project. There are five principal ways that a dictionary maybe entered into the Matrix Dictionary: either through retrodigitization; by conversion to TEI Lex-0 by means of the forthcoming ELEXIS conversion tool; by directly providing TEI Lex-0 data; by providing data in a compatible format (including OntoLex); or by implementing the REST interface described in this paper.

Keywords: lexicography; linked data; infrastructure; ELEXIS; REST; RDF; TEI; JSON

1. Introduction

ELEXIS is a Horizon 2020 infrastructure project dedicated to lexicography. This new infrastructure will (1) enable efficient access to high quality lexicographic data, and (2) bridge the gap between more advanced and less-resourced scholarly communities working on lexicographic resources. In most European countries, elaborate efforts are put into the development of lexicographic resources describing the language(s) of the community. Although confronted with similar problems relating to technologies for producing and making these resources available, cooperation on a larger European scale

has long been limited. Consequently, the lexicographic landscape in Europe is rather heterogeneous. Firstly, it is characterized by stand-alone lexicographic resources, which are typically encoded in incompatible data structures due to the isolation of efforts, prohibiting reuse of this valuable data in other fields. Secondly, there is a significant variation in the level of expertise and resources available to lexicographers across Europe. Within ELEXIS, strategies, tools and standards are under development for extracting, structuring and linking lexicographic resources to unlock their full potential for Linked Open Data, NLP and the Semantic Web, as well as in the context of digital humanities. In a virtuous cycle of cross-disciplinary exchange of knowledge and data, a higher level of language description and text processing will be achieved. By harmonizing and integrating lexicographic data into the Linked Open Data cloud, ELEXIS will make this data available to AI and NLP for semantic processing of unstructured data, considerably enhancing applications such as machine translation, machine reading and intelligent digital assistance thanks to the ability to scale to wide coverage in multiple languages. This, in turn, will enable the development of improved tools for the production of structured proto-lexicographic data in an automated process, using machine learning, data mining and information extraction techniques, where the extracted data can be used as a starting point for further processing either in the traditional lexicographic process or through crowdsourcing platforms.

In the context of the ELEXIS project it has been necessary to develop an interface that allows all different kinds of dictionary data to be included in the infrastructure. As such, the ELEXIS interface is a set of common protocols which take the form of a REST API and which allows dictionaries and lexicographic resources to be accessed through a common interface and in a uniform manner. The REST interface will allow users who wish to query a given endpoint to get back the metadata of the different lexicographic resources accessible from that endpoint, as well as to query individual dictionaries with the possibility of getting back lexical entries in either JSON-LD, OntoLex or TEI Lex-0 (at least one of which must be implemented), these comprise the formats for interoperability of the ELEXIS project. The data model ensures that key elements of the dictionary data are referred to in a uniform manner, and as a particular example of this we require that all the part of speech values are mapped to the Universal Dependencies (UD) part of speech tagset (Petrov et al., 2012; Nivre et al., 2016).

In this paper, we describe this interface and its usage as a tool for getting dictionary data into the ELEXIS infrastructure, so that they can be linked to each other, used by NLP tools and made available through tools to Sketch Engine (Kilgarrieff et al., 2014) and Lexonomy (Měchura, 2017). Most importantly these dictionaries will all be linked to each other as part of the **Dictionary Matrix**, a collection of linked dictionaries that will be created by the project. There are five principal ways that a dictionary may be entered into the Matrix Dictionary: either through retrodigitization; by conversion to TEI Lex-0 by means of the forthcoming ELEXIS conversion tool; by directly providing TEI Lex-0 data; by providing data in a compatible format (including

OntoLex, Cimiano et al., 2014); or by implementing the REST interface¹ described in this paper.

2. The REST interface

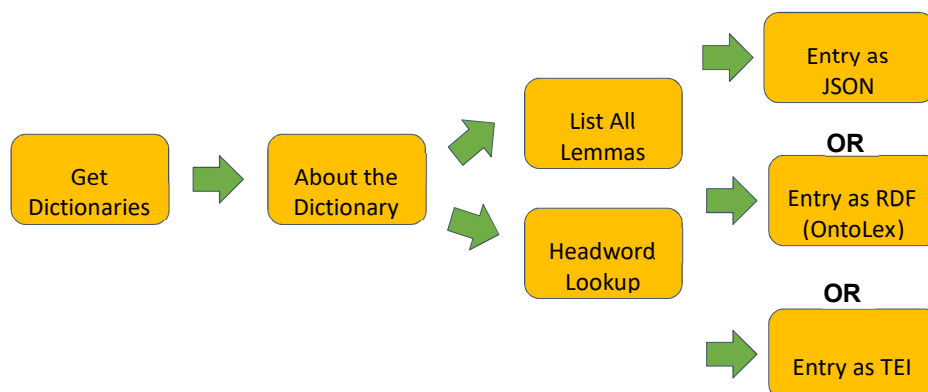


Figure 1: The access protocol for the REST interface

The goal of the REST interface (depicted in Figure 1) is to provide access to the dictionary for the Dictionary Matrix. To this extent it provides a number of basic tools to provide indexing and search over the dictionary interface. As the interface is intended to be implemented with very little effort for the contributors to the ELEXIS network there is a focus on making minimal and simple queries, as such the interface only documents very basic usage. More sophisticated usage can be provided by either custom extensions or by downloading all the data and querying it offline. The first query is to show the set of dictionaries that are available at a particular endpoint, which is done with the following call:

Method Name:	/dictionaries
Parameters:	<i>None</i>
Returns:	A list of dictionary IDs
Example Request:	http://www.example.com/dictionaries
Example Response:	<pre>{ "dictionaries": ["dict1", "dict2"] }</pre>

¹ <http://elexis-eu.github.io/elexis-rest/elexis.html>

The next call in the interface is normally to retrieve the metadata about this dictionary that is necessary to show the dictionary in the dictionary interface. We require a small number of custom parameters that are especially helpful to the ELEXIS interface, including information about the release level, which is whether the data is public, limited to signed-in academic users or private, as well as information about the genre of the dictionary and languages. For genres, we use the previous categorization at the EU dictionary portal, which is as follows:

- **General dictionaries** are dictionaries that document contemporary vocabulary and are intended for everyday reference by native and fluent speakers.
- **Learners' dictionaries** are intended for people who are learning the language as a second language.
- **Etymological dictionaries** are dictionaries that explain the origins of words.
- **Dictionaries on special topics** are dictionaries that focus on a specific subset of the vocabulary (such as new words or phrasal verbs) or which focus on a specific dialect or variant of the language.
- **Historical dictionaries** are dictionaries that document previous historical states of the language.
- **Spelling dictionaries** are dictionaries which codify the correct spelling and other aspects of the orthography of words.
- **Terminological dictionaries** describe the vocabulary of specialized domains such as biology, mathematics or economics.

For languages, we consider that the dictionary has a single language for its headwords, but that the definitions may be in different languages. As such, a bidirectional, bilingual dictionary is split into two 'dictionaries' based on the direction in which we are querying. In addition, there are over 40 other metadata properties, mostly derived from Dublin Core, which may be included in the metadata, although these have no functional role and are merely reproduced for the user at the dictionary portal.

Method Name:	/about
Parameters:	The dictionary ID
Returns:	An object describing the dictionary
Example Request:	http://www.example.com/about/example-dictionary

Example Response:	<pre>{ "release": "PUBLIC", "sourceLanguage": "en", "targetLanguage": ["en", "de"], "genre": ["gen"], "license": "https://creativecommons.org/licenses/by/4.0/", "title": "The Human-Readable Name of this resource", "creator": [{ "name": "Institute of This Resource", "email": "contact@institute.com" }], "publisher": [{ "name": "Publishing Company" }] }</pre>
--------------------------	--

The next issue is obtaining individual entries from the dictionary, in which two principle modes are planned: firstly, retrieval of all entries in the dictionary in order and, secondly, search by lemma. Entries in the dictionary are defined by their lemma, their part-of-speech values and the formats that they are available in. For part-of-speech we use the universal dependencies categories as this provides a broad but good categorization of part-of-speech values, and these values have already been documented and tested in a wide range of languages². As such, we believe that these categories are a good general purpose categorization of part-of-speech values. The full list is given below.

adjective	interjection	punctuation
adposition	(common) noun	subordinating conjunction
adverb	numeral	symbol
auxiliary	particle	verb
coordinating conjunction	pronoun	other
determiner	proper noun	

The querying of entries in the order they appear in the dictionary is limited only by the offset and limit that states how many entries into the dictionary to read and how many to return:

² See <https://universaldependencies.org/u/pos/> for more details.

Method Name:	/list/ <i>dictionary</i>
Parameters:	A limit and an offset
Returns:	A list of lexical entry descriptions
Example Request:	http://www.example.com/list/example-dictionary?limit=2
Example Response:	[<pre> { "release": "PUBLIC", "lemma": "work", "language": "en", "id": "work-n", "partOfSpeech": ["NOUN"], "formats": ["tei"] }, { "release": "PUBLIC", "lemma": "work", "language": "en", "id": "work-v", "partOfSpeech": ["VERB"], "formats": ["tei"] }]</pre>

The lemma lookup requires specifying a lemma, as well as an offset and limit and a flag to say if the query should also look for inflected forms that match this lemma.

Method Name:	/lemma/ <i>dictionary</i> / <i>query</i>
Parameters:	A limit and an offset and flag to state if the entry should be inflected
Returns:	A list of lexical entry descriptions
Example Request:	http://www.example.com/lemma/example-dictionary/works?inflected
Example Response:	<i>As previous</i>

The final part of the API is to return the relevant documents in one of the interoperability formats. The interface can be used to access each of the three formats with a URL such as below. It is up to the implementer to decide which of the three (or all three) to implement.

- <http://www.example.com/json/dictionary/lemma>
- <http://www.example.com/ontolex/dictionary/lemma>
- <http://www.example.com/tei/dictionary/lemma>

It should be noted that this interface does not see any modification of the content of the dictionaries, and by participating in the infrastructure content providers allow the ELEXIS infrastructure to provide links and to make public the list of lemmas through the dictionary portal.

2.1 Design considerations

In general, the interface is designed to be lightweight and easy to implement so that many different dictionary providers can contribute their data to the ELEXIS infrastructure. The interface provides only very simple query methods that should be easy to implement with high performance in the database of the third party who is already responsible for ingesting the data into their infrastructure. It also follows that implementations will need to provide their own mapping of their data into one of the formats provided in the next section and in particular find a mechanism for mapping their part-of-speech categories to the universal dependency list. More sophisticated alignment of properties of lexical entries, e.g., domain or region labels, grammatical information, is not covered from this interface as there is little demand and these properties are generally not well-aligned across resources. While the categories presented in universal dependencies are very broad, they are used primarily for indexing and the entries in the formats below can provide very specific part-of-speech categories to be shown to the user.

3. Formats for interoperability

3.1 JSON

The JSON format is provided for the convenience of those who do not have their data already in TEI Lex-0 or OntoLex, and wish to develop an implementation without reference to other standards. This format is a highly reduced version of OntoLex and as such does not capture all the elements that may be present in a dictionary, nor does it preserve the format of the original dictionary. In fact, the JSON document is a version of the OntoLex model using the JSON-LD model. The JSON object returned should have the following fields:

@context	This should have the fixed value https://elexis-eu.github.io/elexis-rest/context.json
@id	Should be the same as the request ID
@type	One of “LexicalEntry” or more specifically “Word”, “MultiWordExpression” or “Affix”
canonicalForm	A JSON object with two fields: <ul style="list-style-type: none"> writtenRep: The lemma goes here phoneticRep: A pronunciation guide (if any)
partOfSpeech	One of the Universal Dependency values
otherForm	An array of objects with two fields: <ul style="list-style-type: none"> writtenRep: The form goes here phoneticRep: A pronunciation guide (if any)
morphologicalPattern	A morphological class if relevant
senses	An array of objects with the following fields: <ul style="list-style-type: none"> definition: A definition of the sense reference: A URL pointing to an external definition of the entry
etymology	A string giving the etymology of the entry
usage	Notes about the usage of the entry

```
{
  "@context": "https://elexis-eu.github.io/elexis-rest/context.json",
  "@type": "Word",
  "@id": "work-n",
  "canonicalForm": { "writtenRep": "work" },
  "partOfSpeech":
  "commonNoun", "senses": [{
    "definition": "a product produced or accomplished through the effort or activity or
      agency of a person or thing",
    "reference": "http://ili.globalwordnet.org/ili/i61245"
  },{
    "definition": "(physics) a manifestation of energy; the transfer of energy from one
      physical system to another expressed as the product of a force and the distance
      through which it moves a body in the direction of that force;", "reference":
    "http://ili.globalwordnet.org/ili/i97775" }]
}]
```

Figure 2: Code example based on <http://wordnet-rdf.princeton.edu/lemma/work>. NB “commonNoun” is used in the JSON schema for the UD class ‘(common) noun’.

3.2 OntoLex

The OntoLex-Lemon model was developed by the OntoLex Community Group (Cimiano et al., 2016, see also <https://www.w3.org/2016/05/ontolex/> for the Final Community Group Report) based on previous models, in particular the *lemon* model (McCrae et al., 2012; McCrae et al., 2011). This model provides a general framework for the representation of lexical information relative to ontologies, as well as providing for the general modelling of lexical graphs in terms of senses and concepts, in a model that is inspired by the Princeton WordNet model (Fellbaum, 1998). The OntoLex-Lemon model is based on the Resource Description Framework (Lassila & Swick, 1999), and is divided into five modules, with two more in development

- **OntoLex Core:** This describes the key elements of the lexicon, e.g., the lexical entry and its forms, the lexical sense and its associated lexical concept and the reference to the ontology.
- **Syntax and Semantics:** This module describes how the syntactic frames of an entry can be described and how they can be mapped onto the formal semantics in the ontology.
- **Decomposition:** The decomposition module is concerned with how lexical entries can be decomposed into sub-entries, for example in multi-word expressions.
- **Variation and Translation:** Variation (and specifically translation) represents relations between words and in this model such relations can be across entries, part-of-speech and even whole lexicons. Relations in the model are characterized as purely lexical, purely semantic or lexico-semantic.
- **Linguistic Metadata:** The Linguistic Metadata (LiMe) module allows for general metadata about the lexicon such as the number of entries and senses it contains.
- **Lexicographic (in development):** This module describes several aspects that are common in print lexicography, including the ordering and grouping of senses, as well as lexico-semantic restrictions, and examples.
- **Morphology (in development):** The morphology module aims to describe the inflectional and agglutinating morphology of rules both in terms of their attested form, but also as a productive phenomenon.

3.2.1 Usage in the interface

In this section we present some examples of the use of the parameters we have for retrieving an entry in the OntoLex-lemon format (as specified here: <https://www.w3.org/2016/05/ontolex/>).

We selected as the original dictionary resource the Algemeen Nederlands Woordenboek (ANW, <http://anw.inl.nl/about>). The example depicted below shows the

transformation from the ANW entry for the word “wijn” (wine) (see <http://anw.inl.nl/article/wijn>; Tiberius and Declerck, 2017) into the OntoLex-lemon format, using the Turtle syntax. We focus here on the parameters listed at the beginning of subsection 3.1:

```
:lex_wijn_182155
  rdf:type ontolex:Word ;
  lexinfo:anw_articleType "\"de\"" ; lexinfo:gender
  lexinfo:masculine ;
  lexinfo:partOfSpeech lexinfo:commonNoun, lexinfo:noun ;
  ontolex:canonicalForm :form_wijn_singular ;
  ontolex:otherForm :form_wijnen_plural ;
  ontolex:sense :sense_wijn1.0, :sense_wijn1.1, :sense_wijn1.2,
               :sense_wijn1.3, :sense_wijn1.4 .
```

Figure 3: An example of the OntoLex modelling of the ‘wijn’ entry from the AWN dictionary.

The OntoLex lexicographic module aims to close the gap between the computational use cases originally envisioned by the OntoLex Community Group and the kind of lexicographic data handled in projects such as ELEXIS. One of the principal differences that has been observed is that OntoLex has a strict and relatively restrictive definition of a lexical entry as having a single lemma and being of a single part-of-speech class. In the Lexicography module this may be handled by super-entries which give a structured and ordered grouping of an entry and its senses, e.g.,

```
:lead-1 a lexicog:SuperEntry ;
  rdf:_1 [ lexicog:describes :lead-n-1 ] ; # As in "a dog lead"
  rdf:_2 [ lexicog:describes :lead-v-1 ] . # As in "they lead"

:lead-2 a lexicog:SuperEntry ;
  rdf:_1 [ lexicog:describes :lead-n-2 ] ; # The metal
  [ lexicog:describes :lead-n-a-1 ] . # A derived adjective
```

Figure 4: The use of the OntoLex Lexicography module in the interface.

3.3 TEI Lex-0

TEI Lex-0 comprises a subset of the Text Encoding Initiative schema³ (TEI) developed with the express aim of providing a baseline encoding and target format to better facilitate the interoperability of heterogeneously encoded lexical resources. As such TEI Lex-0 situates itself both within the context of the creation lexical infrastructures such as Ermolaev and Tasovac (2012), as well as in the development of generic TEI-aware tools, including dictionary editing software. Note that although TEI Lex-0 is a subset of TEI it should be not thought of as a replacement of the Dictionary Chapter in the

³ <https://tei-c.org/>

TEI Guidelines⁴ and neither is it intended as a format that must be used for editing or managing individual resources – particularly not resources belonging to projects and/or by institutions that already have established workflows based on their own flavours of TEI. Instead it is intended to serve as a format that existing TEI dictionaries can be univocally transformed to in order to be queried, visualized, or mined in a uniform way. At the same time TEI Lex-0 has also been developed with a number of other core use cases in mind, for instance as a best-practice example for didactic purposes, and as a set of best-practice guidelines for new TEI-based projects⁵.

Preliminary work for the establishment of TEI Lex-0 started in the Working Group “Retrodigitized Dictionaries” as part of the COST Action European Network of e-Lexicography (ENeL). Upon the completion of the COST Action in 2017, the work on TEI Lex-0 was taken up by the DARIAH Working Group “Lexical Resources”. Currently, the work on TEI Lex-0 is conducted by the DARIAH WG “Lexical Resources” and falls within the ELEXIS project. According to the Github repository in which the (currently provisional) TEI Lex-0 guidelines are hosted⁶, the current status of the schema is, at the time of writing, as a work in progress. However, even though TEI Lex-0 is not currently production-ready, the core elements of the model are said to be in place. It is therefore possible to describe some of the most important features of TEI Lex-0, those that distinguish it from the TEI dictionary chapter. These include the following (a fuller description can be found at the Github repository for TEI LEX-0⁷):

- **The <entry> element:** TEI Lex-0 simplifies and unifies the encoding of dictionary entries by dispensing with the TEI elements <entryFree>, <superEntry>, and <re>. In TEI, the first of these elements is used to encode a single unstructured entry, the second a sequence of entries which are grouped together, and to embed a related lexical entry within another one. Instead in TEI Lex-0 the TEI element <entry> is used (with appropriate adjustments to its content model) in all of these cases as well as for single structured entries (this latter being its usage in the current TEI guidelines), with a recommendation to make use of the type attribute of <entry> to specify the type of entry being encoded.
- **Sense information:** TEI Lex-0 takes a much stricter approach to grouping sense-related information together than the current TEI guidelines. This affects the kinds of elements that can be children of the <entry> element, and in particular <def> which can appear under <sense> and <cit> which can only

⁴ <https://www.tei-c.org/release/doc/tei-p5-doc/en/html/DI.html>

⁵ To this end TEI Lex-0 aims to stay as aligned as possible with the subset of TEI which comprises the TEI serialization of the updated version of LMF (Lexical Markup Framework) standard (cf. Romary, 2015)

⁶ <https://github.com/DARIAH-ERIC/lexicalresources/tree/master/Schemas/TEILex0>, accessed 6-6-2019

⁷ <https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html>

appear under <sense> or <dictScrap>.

- The element <hom> is deprecated in TEI Lex-0.

3.3.1 Use of TEI Lex-0 in the interface

Within the context of the ELEXIS project TEI Lex-0 is used both as a target format, to which already existing TEI-encoded dictionaries can be converted, as well as a baseline format into which retrodigitized paper dictionaries and digital native dictionaries in other non-TEI formats will be encoded. This will ensure a sufficient level of homogeneity (both semantic and structural) amongst the resources which have been ingested within the ELEXIS platform (something which it would have been hard to guarantee with TEI), while maintaining compatibility with one of the leading standards for text encoding within the digital humanities, and one which is also becoming increasingly popular for encoding lexical resources.

Below we present some examples of the use of the parameters we have for retrieving lexical information from a resource encoded in TEI Lex-0. The following example is taken from a bilingual dictionary and illustrates the entry for the French verb *horrifier* ('horrify') in TEI Lex-0.

```
<entry xml:lang="fr" xml:id="horrifier">
  <form type="lemma">
    <orth>horrifier</orth>
  </form>
  <gramGrp>
    <pos ud:norm="VERB">v</pos>
  </gramGrp>
  <sense>
    <cit
      type="translationEquivalence" xml:lang="en">
        <quote>horrify</quote>
      </cit>
    <cit type="example">
      <quote>elle était horrifiée par la dépense</quote>
      <cit type="translation" xml:lang="en">
        <quote>she was horrified at the
          expense</quote> </cit>
      </cit>
    </sense>
  </entry>
```

Figure 5: The entry for the French word 'horrifier' represented in TEI-Lex0

The entry for ‘horrier’ is enclosed in an <entry> tag, which in the context of TEI-Lex-0 is used to encode the basic element of the dictionary microstructure; grouping all the information related to a particular linguistic entity, including further entries related to it (e.g. homographs or compound phrases). The <form> tag on the next line groups all the information on the written and spoken forms of one headword. The above entry is of the lemma type. The <gramGrp> (grammatical information group) tag groups morpho-syntactic information about a lexical item. In the context of ELEXIS, a @norm attribute is required to specify a normalized (UD) part of speech value for the entry (see introduction). Within the <sense> tag, all information relating to one word sense in a dictionary entry is grouped together, for example definitions, examples, and translation equivalents. The example entry for ‘horrier’ contains a translation in English (<cit type="translationEquivalent" xml:lang="en">) and an example (<cit type="example">) which also has a translation in English. Note that the translations have a language attribute, identifying the language of the translation.

4. Interoperability in the project architecture

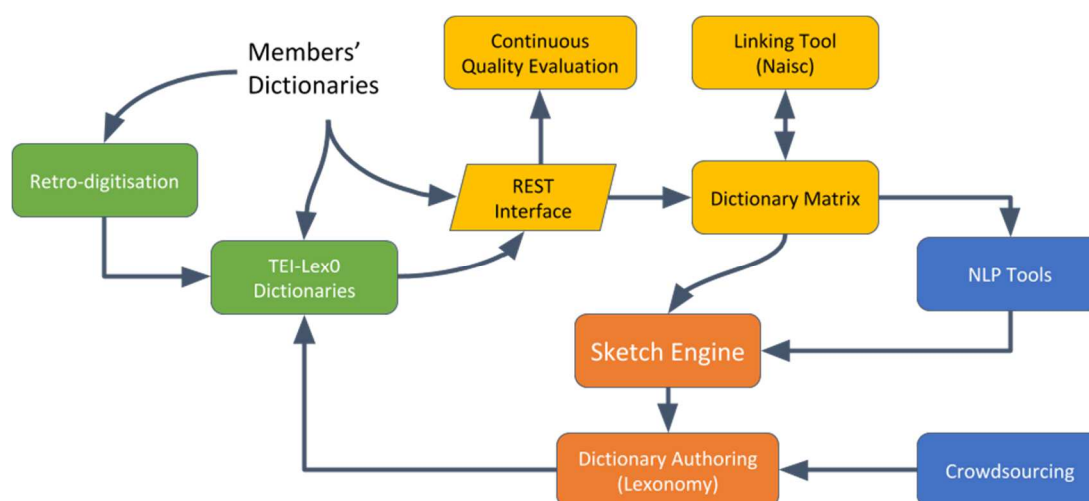


Figure 6: The tools of the ELEXIS infrastructure as an instantiation of the virtuous cycle of eLexicography

The ELEXIS architecture is shown in Figure 6, showing how the REST interface defined above plays an important role in the cycle as the primary interface point. In Figure 7, we show the various ways in which data can enter the infrastructure:

1. From a PDF source or similar OCR is applied and then a semi-automatic tool will be used to identify the structure of the dictionary and output as TEI-Lex0,
2. An existing (non-TEI) XML will be mapped to TEI-Lex0 by identifying the elements that conform to the data model of ELEXIS,

3. TEI-Lex0 documents can be taken directly,
4. Similarly, OntoLex-Lemon can be processed without any modification,
5. Other third-parties may also maintain complete control of their data by implementing the interface above on their own.

Once the data has been provided to the linking infrastructure (yellow in Figure 6), then it will be further processed for NLP applications (blue in Figure 6) and provided to the lexicographic editing interface (orange in Figure 6), which consists of the corpus management tool, Sketch Engine, and the Lexonomy tool for managing and editing lexicographic data, leading to new dictionaries (green in Figure 6).

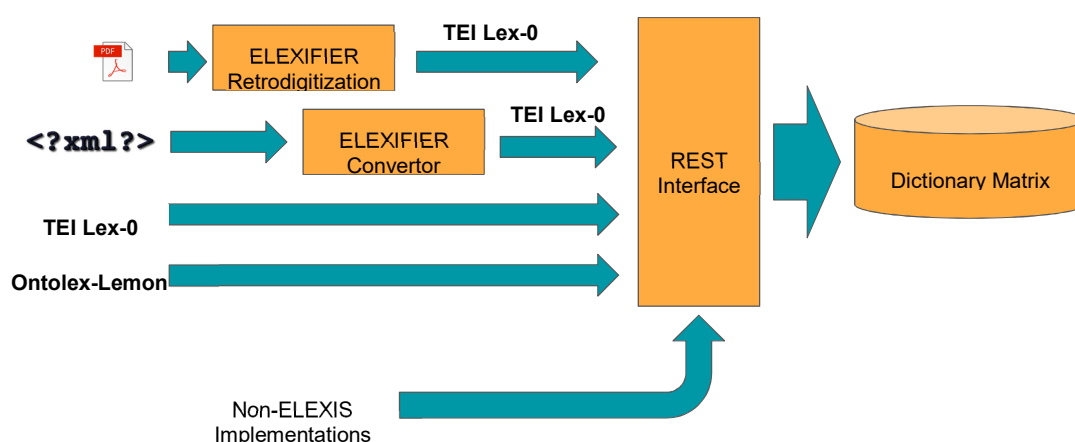


Figure7: Access routes to the ELEXIS architecture depicting the ways data may come into the Dictionary Matrix

4.1 Linking in the ELEXIS infrastructure

There is a plethora of monolingual and multi-lingual resources with a broad range of usage, such as historical dictionaries and terminological resources, available for most European languages. In order to enhance interoperability across resources and languages, ELEXIS provides services for linking resources semi-automatically across languages at various matching levels such as headword, sense and lexeme. Aligned lexical resources, such as Yago (Suchanek et al., 2007), BabelNet (Navigli & Ponzetto, 2012a) and ConceptNet (Speer et al., 2017), have shown to improve word, knowledge and domain coverage and increase multilingualism. In addition, they can improve the performance of NLP tasks such as word sense disambiguation (Navigli & Ponzetto, 2012b), semantic role tagging (Xue & Palmer, 2004) and semantic relations extraction (Swier & Stevenson, 2005).

Lexical data alignment is a challenging task, as lexical information is presented in different structures and dissimilar levels of granularity (Ahmadi et al., 2019). To this end, we are aiming to align lexicographic resources by leveraging ontological properties

and semantic similarity methods. With the current advances in neural networks and resources of significant size available in ELEXIS, we are also interested in applying statistical methods for this task.

4.2 Access to ELEXIS Interface through REST Interface

The retrodigitization tools to be developed in the ELEXIS project will be used for dictionaries that are not already in a digital format. This will apply OCR to the text and then process this text by adding XML markup in the form of TEI-Lex0. For dictionaries that are already available in a digital form, but not one that is supported directly by the project, the conversion tool developed in the ELEXIS project will be used to convert these resources to TEI-Lex0. If the dictionary is already in TEI-Lex0 or has been converted to TEI-Lex0 by one of the two methods described above, then it can be consumed directly by the interoperable interface which will be developed in the next year and reported in D2.2. If the dictionary is in another format supported by the project, in particular OntoLex-Lemon, then this can also be supported directly in the REST interface. Finally, it will be possible for other institutes to participate in the interface by implementing the interface described in this document. The implementation in this case is up-to the institute but it must conform to the specification of this document.

4.3 Using legacy and retrodigitized formats (ELEXIFIER)

The ELEXIFIER tool can take dictionaries in two distinct formats as input: (1) XML file with a custom structure/schema and (2) PDF or similar formats originating from word processors (e.g. MS Word). In the custom XML scenario XPath formalisms are used for conversion of the original dictionary to the TEI Lex0-compliant format. In the PDF scenario a more complex process is needed, similar to the one described in Romary and Lopez (2015). In the first step, text and other formatting features (font style, size, colour, etc.) are extracted from the dictionary in PDF form. In the next step, users are asked to manually annotate part of the dictionary in the Lexonomy online dictionary editing tool, according to the ELEXIS data model compatible with TEI-Lex0 standard. In the last step, the annotated text is used as the training material for machine learning algorithms that produce the entire dictionary converted to TEI-Lex0 format. The converted dictionaries can be edited further in the Lexonomy editor.

4.4 Reference implementation for TEI and OntoLex



Figure 8: A screenshot of the reference implementation of the REST interface.

A reference implementation is available for the interface at <https://github.com/elexis-eu/dictionary-service>, which allows a server to be set up based on either a JSON, OntoLex or TEI document. This interface is implemented in the Rust programming language and as such is available for a wide range of platforms and provides high performance. For JSON files these are directly loaded, however for the TEI and OntoLex it may be necessary to provide some configuration, in particular the mapping of the values used for part-of-speech in the dictionary with the Universal Dependencies categories. It is recommended that those who contribute to the process refer to the existing documentation available from the Universal Dependencies about how to map their categories.

5. Conclusion

eDictionaries are typically in very different stages of digitization, from those where the only digitization is that they have been scanned up to those that have been carefully marked-up with standards such as TEI-Lex0 or ‘linked-data native’ (Gracia et al., 2017) in OntoLex-Lemon formats. As such there needs to be a highly flexible interface for integrating lexical resources into an ambitious project such as ELEXIS. We have shown a REST interface that will integrate with the retrodigitization and conversion tools in this project to provide multiple ways of entrance into the infrastructure, which ensures that this infrastructure will be open to a wide range of lexicographers.

6. Acknowledgements

All authors are supported by the EU H2020 programme under grant agreements 731015

(ELEXIS - European Lexical Infrastructure). John McCrae is also supported by a research grant from Science Foundation Ireland, co-funded by the European Regional Development Fund, for the Insight Centre under Grant Number SFI/12/RC/2289.

7. References

- Ahmadi, S., Arcan, M. & McCrae, J. (2019). Lexical sense alignment using weighted bipartite b-matching. *2nd Conference on Language, Data and Knowledge (LDK 2019)*, p. 5.
- Bowers, J., Herold, A. & Romary, L. (2018). TEI-Lex0 Etym-towards terse recommendations for the encoding of etymological information. *JADH 2018*, p. 243.
- Cimiano, P., McCrae, J. P. & Buitelaar, P. (2014). Lexicon Model for Ontologies: Community Report.
- Ermolaev, N. & Tasovac, T. (2012). Building a lexicographic infrastructure for serbian digital libraries. *Libraries in the Digital Age (LIDA) Proceedings*, 12.
- Francopoulo, G., George, M., Calzolari, N., Monachini, M., Bel, N., Pet, M. & Soria, C. (2006). Lexical markup framework (LMF). In *International Conference on Language Resources and Evaluation-LREC 2006*. p. 5.
- Gracia, J., Kernerman, I. & Bosque-Gil, J. (2017). Toward linked data-native dictionaries. In I. Kosem, J. Kallas, C. Tiberius, S. Krek, M. Jakubíček & V. Baisa (eds.) *Electronic Lexicography in the 21st Century: Lexicography from Scratch. Proceedings of the eLex 2017 conference, 19-21 September 2017, Leiden, Netherlands*. Brno: Lexical Computing Ltd.
- Kilgariff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, 1(1), pp. 7–36.
- McCrae, J., de Cea, G. A., Buitelaar, P., Cimiano, P., Declerck, T., Gómez-Pérez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D. & Wunner, T. (2012). Interchanging lexical resources on the Semantic Web. *Language Resources and Evaluation*, 46(6), pp. 701–709.
- McCrae, J. P. & Buitelaar, P. (2018). Linking Datasets Using Semantic Textual Similarity. *Cybernetics and Information Technologies*, 18(1), pp. 109–123.
- Měchura, M. (2017). Introducing Lexonomy: an open-source dictionary writing and publishing system. In I. Kosem, J. Kallas, C. Tiberius, S. Krek, M. Jakubíček & V. Baisa (eds.) *Electronic Lexicography in the 21st Century: Lexicography from Scratch. Proceedings of the eLex 2017 conference, 19-21 September 2017, Leiden, Netherlands*. Brno: Lexical Computing Ltd.
- Navigli, R. & Ponzetto, S. P. (2012a). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193, pp. 217–250.
- Navigli, R. & Ponzetto, S. P. (2012b). Joining forces pays off: Multilingual joint word sense disambiguation. In *Proceedings of the 2012 joint conference on Empirical*

- Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, pp. 1399–1410.
- Nivre, J., de Marneffe, M. C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R. & Zeman, D. (2016). Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Paris, France: European Language Resources Association (ELRA).
- Petrov, S., Das, D. & McDonald, R. (2012). A Universal Part-of-Speech Tagset. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA).
- Romary, L. & Lopez, P. (2015). GROBID-Information Extraction from Scientific Publications. *ERCIM News*, 100.
- Speer, R., Chin, J. & Havasi, C. (2017). ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *AAAI*. pp. 4444–4451.
- Sporny, M., Longley, D., Kellogg, G., Lanthaler, M. & Lindström, N. (2014). JSON-LD 1.0.
- Suchanek, F. M., Kasneci, G. & Weikum, G. (2007). Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*. ACM, pp. 697–706.
- Swier, R. S. & Stevenson, S. (2005). Exploiting a verb lexicon in automatic semantic role labelling. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 883–890.
- Tiberius, C. & Declerck, T. (2017). A lemon Model for the ANW Dictionary. In I. Kosem, J. Kallas, C. Tiberius, S. Krek, M. Jakubíček & V. Baisa (eds.) *Proceedings of the eLex 2017 conference*. INT, Trojína and Lexical Computing, Lexical Computing CZ s.r.o., pp. 237–251.
- Xue, N. & Palmer, M. (2004). Calibrating features for semantic role labeling. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



Improving Dictionaries by Measuring Atypical Relative Word-form Frequencies

Kristian Blensenius, Monica von Martens

University of Gothenburg, Dept. of Swedish, Lundgrensgatan 1B, Gothenburg (Sweden)

E-mail: kristian.blensenius@gu.se, monica.von.martens@gu.se

Abstract

In this article, we discuss and give examples of how word-form frequency information derived from existing corpora statistics can be used to improve dictionary content. The frequency information is used in combination with rule-based morphological data based on derivational and inflectional information from the Swedish Morphological Database compiled at the University of Gothenburg, and the lexical database owned by the Swedish Academy. The method currently used in the ongoing project for updating the monolingual Contemporary Dictionary of the Swedish Academy is described, and some examples of dictionary entries identified as candidates for update based on frequency measures are given. Different aspects of morphological dictionary content are discussed and highlighted by comparison between the above-mentioned definition dictionary and a learner's dictionary. The role of headword or lemma as well as cross-referencing methods in a digital dictionary as compared to a printed dictionary is also discussed. Finally, a few examples of suggested modifications and enhancements are given.

Keywords: morphology; frequency; word forms

1. Introduction

In this article, we discuss how dictionary content can be improved by (re)using morphological information and enhancing it with corpus frequency information. Two contemporary Swedish monolingual dictionaries are used to illustrate how this method can be used to enhance dictionary content.

Morphological frequency matters have been much discussed from various perspectives, for example from the point of view of learning and producing word forms (e.g. Bybee, 1995; Hay, 2001; Dąbrowska, 2008). Hay (2001) specifically targets the question of absolute vs. relative word-form frequency, and particularly the relation between the so-called base form and a derived word form, showing that relative frequency seems to be even more important than absolute frequency when it comes to morphological decomposition.

In lexicography, morphology is instead traditionally often discussed in terms of the amount, compactness, and type of inflectional and derivational information to be presented in dictionaries (e.g. Heuberger, 2018; Svensén, 2009: 124ff.), reflecting the fact that presenting morphological information often constitutes a central component in dictionary entries. Inflectional information serves receptive functions, e.g. for finding

the lemma form and, importantly, productive functions, e.g. for finding inflected forms based on a lemma form. Finally, the morphological form of the headword is a topic discussed in the lexicographic literature, because it is often important to distinguish between the “base” form and other forms of a lemma. For example, some plural nouns like *arms* ‘weapons’ have a distinct meaning and may need to be presented as distinct headwords from their singular counterparts (see Atkins & Rundell, 2008: 325).

Whereas frequency-based lemma selection has indeed been discussed (e.g. Trap-Jensen et al., 2012), the distribution of individual *word forms* (inflectional and derivational forms) in terms of their relative frequencies has not been much considered in the lexicographic field. For that reason, our aim is, as mentioned above, to fill this gap by presenting a study of word forms in a morphological database, which we will evaluate with respect to two dictionaries.

2. The dictionaries

We limit our study to two major dictionaries of contemporary Swedish, aimed at two different user categories.

The first one is the monolingual ‘Contemporary Dictionary of the Swedish Academy’ (*Svensk ordbok utgiven av Svenska Akademien*, abbreviated SO), which is a definition dictionary primarily aimed at native speakers and advanced learners of Swedish. It is mainly a reception dictionary, but it is also production-oriented (Sköldberg, 2017: 123). SO is an edited extract of a much larger database compiled at the University of Gothenburg. The printed version of SO was published in 2009, the corresponding mobile app in 2015 and, finally, a freely available on-line web version was released in 2017. The tools used in this study have been developed as part of the revision process aiming at publishing a new up-to-date online version of SO.

The second dictionary used in our study is the present on-line version of Lexin, which is primarily a learner’s dictionary (see the Lexin introduction). Lexin consists of a monolingual Swedish core, compiled at the University of Gothenburg on behalf of the Language Council of Sweden, which is translated into a number of immigrant languages. Older versions of Lexin have been published as both monolingual and bilingual printed dictionaries.

2.1 Morphology and headword policy in the dictionaries

Both SO and Lexin provide morphological information next to the headword. For example, the verb *köpa* ‘buy’ is paired with the following inflectional information:

SO: **köpa** *köpte köpt, pres. köper*

Lexin: **köper** att köpa, köpte, har köpt, är köpt, köp!

In SO, the inflectional paradigm of *köpa* is represented by the infinitival headword, the preterite and supine forms, and finally the present form of the verb. In addition, SO provides derivational information further down in the dictionary entry in the shape of two nominalizations, *köpande* ‘buying’ and *köp* ‘purchase’. The learner’s dictionary Lexin presents the headword, and then comes the infinitive (preceded by the infinitive marker *att*), the preterite, the supine (preceded by the perfect auxiliary in the present tense, *har*), the perfect participle (preceded by the passive auxiliary in the present tense, *är*), and finally the imperative form followed by an exclamation mark.

Svensén (2009) presents a list of “the grammatical forms most used as lemma forms”, remarking that nouns are presented in the nominative singular form, verbs in the active infinitive, etc., provided that the structure of the language allows it. In the light of this, it could be noted that Lexin stands out in presenting the present form of verbs as the headword, in this case *köper*. This is not motivated by frequency¹ but by the assumption that the present form is the better basis for deriving the other forms of the verb (Gellerstam, 1999: 7f.). While on the subject, it could be pointed out that dictionaries for many other languages (e.g. Arabic and, as noted by Svensén, Latin) do not necessarily use the infinitive form as headword. Also, although the infinitive has been the conventional lemma form for verb entries in Swedish dictionaries for approximately two hundred years, the present tense was commonly used in older dictionaries (e.g. Spegel, 1712, and Schenberg, 1739; see Hannesdóttir 1998: 148, 202). In older dictionaries in general, the choice of headword sometimes looks quite arbitrary to modern eyes for other word classes, too. Adjectives, for example, which take the suffix *-t* in the neuter singular form in Swedish, are frequently presented in this headword form in older dictionaries (e.g. Schenberg, 1739).

Dictionary entries in SO and Lexin commonly include special cross-referential headword forms, such as irregular verb forms like *gick* ‘went’ pointing to the base form *gå* ‘go’ (or the present *går* ‘goes’ in Lexin). Being electronic, both dictionaries should handle headword identification (Lew, 2012) automatically in these cases, either as redirections or links. This is the case for Lexin and the app version of SO, but regrettably these referential lemmas have at the time of writing gone missing in the web version of SO. This can be taken as a reminder of the fact that digitalization has not only upsides but also downsides; even a thorough proof-reading and testing phase on one or a few platforms cannot guarantee full functionality on all existing and upcoming environments, and it is seldom in the hands of the editors to decide about, and stipulate conditions for, the availability of the dictionary on new devices.

Another instance of morphological consideration affecting the choice of headword form for the dictionary entry is cases where the canonical lemma form is hardly ever, or

¹ Our frequency investigations also show that the infinitive and the present-tense forms are almost equally frequent for most verbs in our corpora, so frequency considerations can hardly be called on to favour one form over another as headword.

never, used. An example in English is the plural lemma form of nouns such as *scissors* (cf. Svensén, 2009: 105f.). A counterpart in SO is the active preterite auxiliary verb *torde* ‘is probably, should’ used as headword (this form is not included in Lexin).

Luckily, in digital dictionaries the choice of headword form for a lexical entry is typically not an either/or choice. As mentioned, by means of clickable links or redirection the user can often reach the desired entry regardless of which word form is entered in the search box. Still, one has to take caution not to give the user the impression he or she made some kind of mistake causing redirection. It has been reported by second language teachers that Lexin users sometimes believe a redirection was caused by misspelling, when the redirection was in fact caused by a void in the dictionary. Also, behind the scenes, in the database, it is strongly advised to attach inflection information in a standardized manner to a standard base form even if that form is not the one used to head the entry as shown to the public.

3. Problems

In Section 2 above we reviewed some cases of well-known morphologically induced problems a lexicographer needs to address, such as words for which the expected base form of a lemma is out of use, or almost out of use, and the case of verbs with irregular inflection which creates a need for several “entry points”. In this section we address a couple of more intricate problems, for example how to deal with cases where a “base” headword form is indeed used but another word form is much more frequent and may have a slightly different meaning.

3.1 Word forms with a slightly different meaning than the base form

Sometimes the frequency distribution differs between word senses. Looking at, say, the plural form *blommor* ‘flowers’, we find that this form is much more frequent in several corpora than the singular form *blomma* ‘flower’, which is the headword form in SO and Lexin. Now, this does not necessarily mean that the plural *blommor* should be considered as a headword, not even for cross-referencing. Instead, the fact that the plural form is much more frequent than the singular form should make the lexicographer attend to the structure and content of the dictionary article. In this case, it is quite clear from corpus inspection that the plural form in most cases refers to flowers in the sense ‘flower plants’ (i.e. including stems and leaves) while the core sense, which is far less common in everyday language, has a more regular distribution of word-form frequencies. This structure is reflected in the article structure in Lexin, where the ‘plant’ meaning is given as the first sense. However, in SO, the first sense given for *blomma* only refers to the often brightly-coloured reproductive part of a plant. This is probably motivated by etymology (the ‘brightly-coloured flower’ is older than the ‘flower plant’), as well as by a tradition of trying to identify and present the core meaning of a word. What makes a plant a flower (in the second sense) is having flowers (in the first sense).

Another example of sense shifting with word form is the Swedish word *pengar* ‘money’, which morphologically is a regularly formed plural of the word *peng*, ‘coin’. The learner’s dictionary Lexin has two separate entries, one for *pengar* and one for *peng*. The latter entry lists *pengar* as the plural, i.e. ‘coins’, without any reference to the alternative meaning of this word form. SO only gives the entry *peng*, together with the core definition ‘coin or note’, and the usage information ‘mostly plural’. The meaning ‘money’ is given as a sub-sense of the core meaning.

3.2 Frequency of inflected forms varying with orthography of the headword

A much debated issue in Swedish from a language-planning perspective is the use of English spelling. In particular, the English plural suffix *-s* is counteracted by the normative ‘Swedish Academy Glossary’ (*Svenska Akademiens ordlista*, abbreviated SAOL). This approach can also be found in the current edition of SO. For example, the headword *skanner* ‘scanner’, spelled with *k*, is provided together with the recommended indefinite plural form *skannrar* ‘scanners’. SO also gives the *c*-spelling variant, *scanner*, as an alternative, and the recommended plural form *scannrar*. The *-s* plurals *skanners/scanners* are given as optional plural forms. Focusing on the relation between the variant spelling and the two plural forms *-rar* and *-s*, the frequency tool shows that the distribution of plural suffixes is far from even between the variant spellings. It seems that people using the more “Swedish-looking” *k*-spelling *skanner* also use the Swedish plural suffix *-rar*, whereas the (more frequent) spelling *scanner* tends to be combined with the *-s* plural suffix. This is not reflected in the article in SO. (Note that in Lexin, the plural *-s* suffix is not included as a plural variant.)

3.3 Very frequent derivations

Lexicographers’ decisions about which items to be included as lexical items with individual main entries and which ones to be registered as derivatives are often unclear (Battenburg, 1992: 69). Using a frequency test can provide some interesting results.

Creating nominalizations is a conventional way of deriving Swedish verbs. Adding *-ande* to an arbitrary Swedish verb theoretically yields both nominalizations and present participles (the latter are usually adjectival or verbal). For example, from the verb *springa* ‘run’, one can derive *springande*, which means both ‘running’ and ‘the act of running’. In SO, *-ande* forms are often included as words forms in the verb entries, to indicate nouns carrying the semantics of the verbal headword. However, the adjectival (participial) *-ande* forms are often missing in the dictionary even though the adjectival use of the word might be much more common than the nominal use. These “missing adjectives” can be found using frequency information, by examining words which have a high frequency of the *-ande* form compared to the frequency of the infinite headword

form. One example is the “nominalization” *ambulerande* ‘moving from place to place’ in relation to the verb headword **ambulera** ‘move from place to place’. Looking at this word more in detail using the corpus tool Korp (Borin et al., 2012), it can be noted that the *-ande* form is primarily an adjectival form used as a modifier in noun phrases like *ambulerande tjänsteman* ‘travelling administrator’, *ambulerande tivoli* ‘travelling amusement park’, etc.; see Figure 1 below.

The screenshot shows the Korp corpus search interface. At the top, the Korp logo is visible next to a search bar containing '125 av 237 korpusar valda — 2,13G av 13,31G token'. Below the search bar, there are tabs for 'Enkel', 'Utökad', 'Avancerad', and 'Jämförelse'. The 'Utökad' tab is selected. A search box contains the word 'ambulerande' and a 'Sök' button. Below the search box, there are checkboxes for 'i följd och även som', 'förd', 'efterled och', and 'skiftlägesberoende'. Further down, there are controls for 'KWIC' (träffar per sida: 25), 'sortera inom korpus på: förekomst', and 'Statistik: sammanställ på: ord'. There are also checkboxes for 'Visa statistik' and 'Visa ordbild'. Below these controls, there are tabs for 'KWIC', 'Statistik', and 'Ordbild'. The 'KWIC' tab is selected. The search results show a list of concordance hits for 'ambulerande'. The first hit is from 'ÅBO UNDERRÄTTELSE 2013' and shows the word 'ambulerande' in a blue box. The second hit is from 'ASTRA 1960–1979' and shows the word 'ambulerande' in a blue box. The third hit is from 'ÅLANDSTIDNINGEN 2012' and shows the word 'ambulerande' in a blue box. The fourth hit is from 'BLOGGMIX 2003' and shows the word 'ambulerande' in a blue box. The fifth hit is from 'BLOGGMIX 2006' and shows the word 'ambulerande' in a blue box. The sixth hit is from 'BLOGGMIX 2007' and shows the word 'ambulerande' in a blue box. The seventh hit is from 'BLOGGMIX 2008' and shows the word 'ambulerande' in a blue box. The eighth hit is from 'BLOGGMIX 2008' and shows the word 'ambulerande' in a blue box. The ninth hit is from 'BLOGGMIX 2008' and shows the word 'ambulerande' in a blue box. The tenth hit is from 'BLOGGMIX 2008' and shows the word 'ambulerande' in a blue box. The search results are displayed in a table with columns for the source, the concordance line, and the word 'ambulerande'.

Figure 1: Korp corpus concordance search for *ambulerande*

Comparing with the learner’s dictionary Lexin, it could be noted that only the verb form *ambulera* is included, although this form is infrequent in use.

3.4 Word forms in phrases which special syntactic functions

In SO, adjectives are typically illustrated as modifiers in noun phrases and as subject complements. For example, an adjective like *gul* ‘yellow’ is illustrated with examples such as *torrt gult gräs* ‘dry yellow grass’ and the subject-complement clause *bladen var gula redan i slutet av september* ‘the leaves were yellow already at the end of September’. Adjectives that are mostly used serving other syntactic functions are usually marked e.g. ‘typically used adverbially’, while adjectives frequently used both as attributes and

a definition and are introduced by the formula “I frasen ...” (‘in the phrase ...’). An example is the word *aftonkvisten* (lit. *the branch of evening*), which is principally only used in the definite singular in the prepositional phrase *på aftonkvisten* ‘between afternoon and evening’ (lit. *on the branch of evening*). An examination of word-form frequencies confirms this.

Another example is the lemma noun *sort* ‘kind’, where the genitive singular form, *sorts*, is almost 600% more frequent than the nominative headword form (the singular nominative is normally much more frequent than the singular genitive). A concordance study reveals that a great number of the genitive forms make up the very frequent classifying construction *en sorts* + NOUN, as in *en sorts frukt* ‘a kind of fruit’ (lit. *a kind.GEN fruit*). The fact that the genitive form is included in a collocation and the fact that only the genitive form is allowed in this collocation is not really clear, neither in SO, nor in Lexin.

4. The morphological corpus frequency tool

The morphological frequency tool stores information on word forms and frequencies for a number of corpora in a format which is easily combined with dictionary information on headwords, inflection groups, and inflected forms. The frequency information is retrieved from the Korp corpus tool and “de-lemmatized”, i.e. stripped of lemma information before being stored in a relational database which can be accessed using standard tools like MySQL Workbench. The frequency information is used in combination with rule-based morphological data based on derivational and inflectional information from Svensk Morfologisk Databas (‘The Swedish Morphological Database’, Berg & Cederholm, 2001) compiled at the University of Gothenburg and the lexical database owned by the Swedish Academy.

Part of the information is integrated in the editorial interface, for the convenience of the editors, while such tasks as retrieving lists of candidates for closer examination are carried out with the help of stored procedures.

A stored procedure in a relational database management system serves as a means to store a group of SQL statements with an assigned name, which can be called using parameters. We use stored procedures to create and examine word-form distribution tables based on joining the inflectional information from the dictionary database with frequency data from the corpus frequency database. For each word in the dictionary database the editor can enter a code indicating the inflectional paradigm (see subSection 4.1). Entering or changing the code generates a “blow-up” of all word forms with associated tags – up to over 20 forms for some verbs, including derived participles – which are stored in a table. This table, at the moment holding information on approximately 1,800,000 word forms, is immediately available for joining with corpus-frequency information for presentation in the editor interface, and it provides up-to-date information for the stored procedures. The inflectional paradigm code system and

associated tags and rules for word-form generation are a development of the system used for Svensk Morfologisk Databas.

4.1 Editorial interface

Each word in the dictionary is classified as belonging to an inflectional group, and in the process of this classification the editor is presented with frequency information for the actual word forms, which can be compared to some basic metrics indicating normal frequency distribution.

Nuvarande värden:
subst. **blomma** lnr121197

Välj böjningsgrupp:

OBS! Lista med vanligaste böjningsklasserna visas om man tömmer fältet och klickar i det!

Ordbildning (för **avstavning och segmentering**) tagg:

Ordbildning	teknisk stam	böjning	grupp	tagg
Nuvarande:	blomm	-an -or	11a	

Allmän regel för **genitiv-s**: Om ordet är mer än 1 bokstav långt och slutar på något av s, x, z, S, X, Z. +s-regel påförs inget s i SAOL/SMDB visas + endast om det finns / i regeln, här visas + framför alla ändelser

Allmän regel för **n+n**: +na efter ord som slutar på 'n' dubblar ej 'n'

börjar med kombo tvåställa **11** och treställa **11a**

nr	tagg	regel	resultat	rensat	FLASHBSAMH-frek	ff
01(11a)	NCUSNI	=	blomma	blomma	663	(2*)
02(11a)	NCUSGI	=+/s	blomma+s	blommas	3	(1+1*)
03(11a)	NCUSND	=+/n	blomma+n	blomman	101	(1)
04(11a)	NCUSGD	=+/n/s	blomma+n+s	blommans	4	(1)
05(11a)	NCUPNI	%sv+/or	blomm+or	blommor	2483	(1)
06(11a)	NCUPGI	%sv+/or/s	blomm+or+s	blommors	0	(1)
07(11a)	NCUPND	%sv+/or/na	blomm+or+na	blommorna	403	(1)
08(11a)	NCUPGD	%sv+/or/na/s	blomm+or+na+s	blommornas	3	(1)

Normalfördelning för substantiv: sg/pl: c. 75/25, bestämd form betydligt ovanligare än obestämd.

Adjektiv: sg genitiv(AQPUSGI/AQPNSGI) har normalt sett frekvensen 0, avvikelser kan tyda på substantivering

Transitiva verb, pres.ind./perf.part.utr:(VOIPA/AFOUSNI) c 90/10, avvikelser kan tyda på lexikalisering

ff= antal superlemman i fullformstabellen som uppvisar denna form treställigt + tvåställigt, *= intern homografi för något av dessa

Figure 3: Frequency information shown while editing inflectional information for a dictionary entry

The example in Figure 3 shows how the word forms generated by the inflectional code *11a*, when applied to the noun *blomma* ('flower'), are presented to the editor together with frequency information from the corpus *FLASHBSAMH* (a popular internet discussion group). These figures can be compared to the hint below the form-frequency table regarding 'normal' distribution for singular and plural forms (roughly 75% and 25%, respectively) and definite and indefinite forms (the definite forms being much less frequent).

The existence of homographic word forms can obscure this kind of comparison, so the rightmost column shows the number of homographic word forms for each form. In this case there is indeed a homograph to the singular indefinite form, the verb *blomma* ('to bloom'), which means the singular noun form *blomma* ('flower') is in fact even less frequent than shown in the table.

While editing a dictionary entry, the editor has an integrated view of 1) the updated entry, 2) the published version of SO, 3) the word forms given in the latest version of the Swedish Academy Glossary, and 4) an overview of word form frequencies for a number of corpora of contemporary Swedish. See Figure 4:

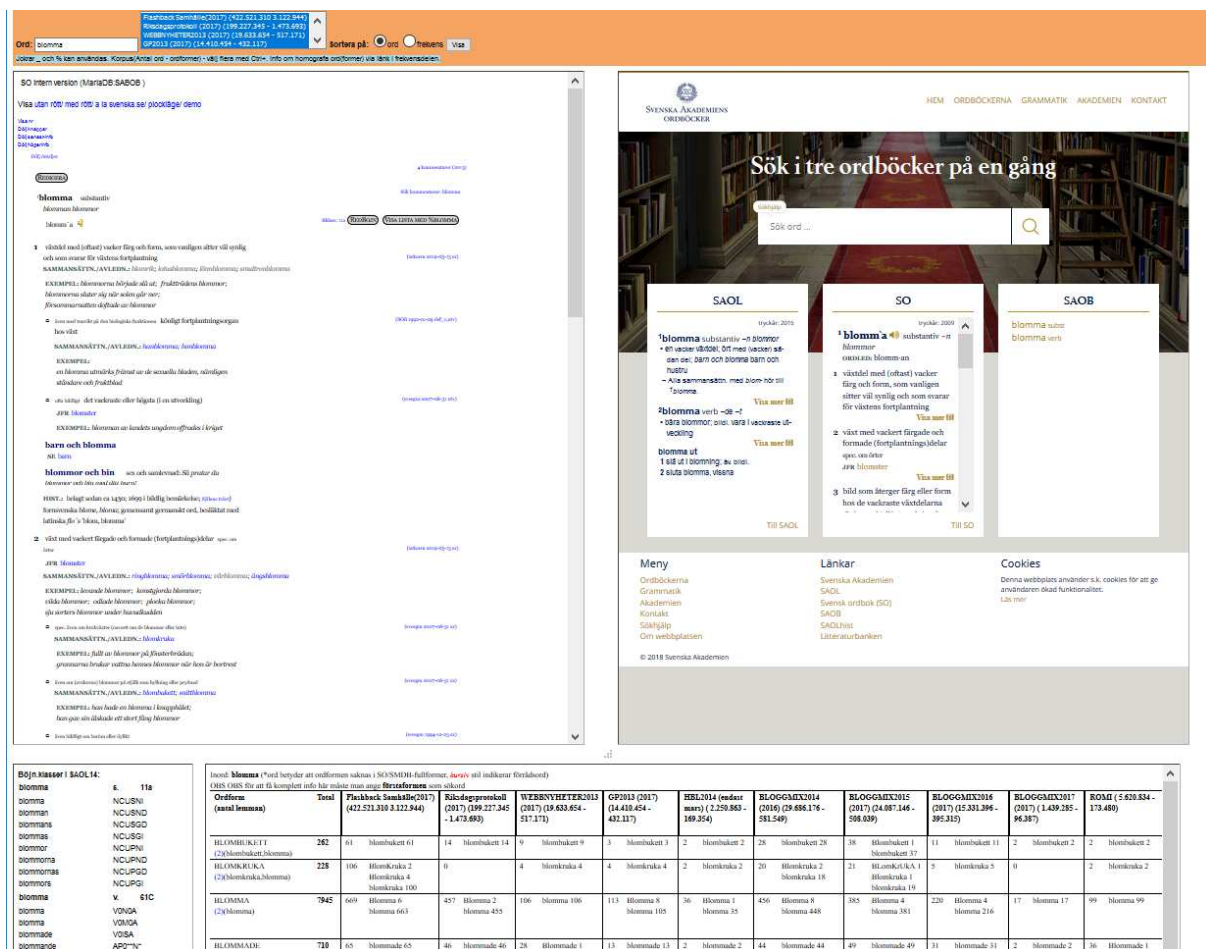


Figure 4: Editors' view of entry being edited, published dictionary entry, and word form frequencies in different corpora

The frequency-distribution view can also be used to check the relative frequency of different spellings. Cf. for example the word *kafé/café* (‘café’) in SO, which is more often spelled *café* in our corpus of contemporary Swedish (Figure 5). The former spelling variant *café* has now been upgraded to headword status, while the former headword, *kafé*, is considered a spelling variant.

Visa många frekvenser

https://k5.spraakdata.gu.se/li/red/SABOB/flerval_manykorprefrek_m_sub_utf8.php?ord=cafe&utf8&typ= 67%

Inord: cafe (*ord betyder att ordformen saknas i SO/SMDB-fullformer, *kurziv* stil indikerar förrådsord)
OBS OBS för att få komplett info här måste man ange förstaformen som sökord

Ordform (antal lemmar)	Total	Flashback Samhälle(2017) (422.521.310 3.122.944)	Riksdagsprotokoll (2017) (199.227.345 - 1.473.693)	WEBBNYHETER2013 (2017) (19.633.654 - 517.171)	GP2013 (2017) (14.410.454 - 432.117)	HL2014 (endast mars) (2.250.863 - 169.354)	BLOGGMIX2014 (2016) (29.686.176 - 581.549)	BLOGGMIX2015 (2017) (24.087.146 - 508.039)	BLOGGMIX2016 (2017) (15.331.396 - 395.315)	BLOGGMIX2017 (2017) (1.439.285 -96.387)	ROMII (5.620.834 - 173.480)	poeten.se frekvenser (68.282.974 - 907.209)
CAFE (1)(café)	7823	4079 CAFE 14 CAFE 1 CAFE 2 CAfe 1 CAfe 413 CAfe 10 CAfe 888 cafe 862 cafe 25 cafe 1862 cafe 1	21 CAFE 1 CAfe 18 cafe 2	157 CAFE 2 CAfe 10 CAfe 89 cafe 2 cafe 54	284 CAFE 12 CAfe 195 CAfe 1 cafe 1 cafe 75	50 CAFE 9 CAfe 36 cafe 5	853 CAFE 2 CAFE 2 CAfe 59 CAfe 4 CAfe 237 cafe 111 cafe 434	770 CAFE 1 CAFE 1 CAfe 42 CAfe 7 CAfe 248 cafe 56 cafe 9 cafe 406	514 CAFE 3 CAfe 23 CAfe 9 CAfe 200 cafe 23 cafe 6 cafe 250	40 CAFE 16 cafe 1 cafe 23	35 CAFE 1 CAfe 17 cafe 17	977 CAFE 1 CAFE 1 CAFE 4 CAfe 36 CAfe 3 CAfe 159 cafe 101 cafe 11 cafe 660 cafe 1
CAFEET (1)(café)	6361	5703 CAFEET 4 CAFEET 1 CAFEET 2 CAfeet 77 CAfeet 24 CAfeet 254 cafeet 1674 cafeet 3 cafeet 116 cafeet 1 cafeet 15 cafeet 3528 cafeet 4	10 cafeet 10	15 Cafeet 5 cafeet 10	14 cafeet 14	1 Cafeet 1	108 Cafeet 9 cafeet 8 cafeet 1 cafeet 90	95 Cafeet 3 CAfeet 2 CAfeet 9 cafeet 11 cafeet 1 cafeet 69	79 Cafeet 1 CAfeet 14 cafeet 1 cafeet 63	6 Cafeet 1 cafeet 5	3 Cafeet 1 cafeet 2	304 CAFEET 1 CAfeet 1 CAfeet 32 cafeet 13 cafeet 1 cafeet 7 cafeet 4 cafeet 245
CAPET (1)(café)	2720	2437 CAPET 6 CAPET 2 Cafet 179 Cafet 4 Cafet 195 cafe 553 cafe 85 cafe 1413	0	0	4 Cafet 1 cafe 3	2 cafe 2	76 Cafet 3 Cafet 1 Cafet 13 cafe 5 cafe 54	35 Cafet 6 Cafet 10 cafe 8 cafe 1 cafe 30	14 Cafet 1 Cafet 3 cafe 1 cafe 7	4 cafe 1 cafe 3	1 cafe 1	127 Cafet 11 Cafet 1 Cafet 10 cafe 21 cafe 2 cafe 82
KAFEET (1)(café)	2527	2025 KAFEET 2 Kafeet 12 Kafeet 1 Kafeet 54 kafeet 239 kafeet 18 kafeet 15 kafeet 1684	50 Kafeet 4 kafeet 46	49 Kafeet 3 kafeet 46	122 KAFEET 1 Kafeet 1 Kafeet 14 kafeet 106	27 Kafeet 4 kafeet 23	20 Kafeet 1 Kafeet 1 kafeet 18	19 Kafeet 1 kafeet 18	9 kafeet 9	1 kafeet 1	26 Kafeet 1 kafeet 25	146 Kafeet 2 Kafeet 9 kafeet 6 kafeet 1 kafeet 3 kafeet 1 kafeet 124
KAFE (1)(café)	1677	542 KAFE 2 Kafe 13 Kafe 40 kafe 98 kafe 3 kafe 386	87 Kafe 1 kafe 86	141 Kafe 1 Kafe 11 kafe 1 kafe 128	361 Kafe 34 kafe 327	57 Kafe 3 kafe 54	76 Kafe 6 Kafe 28 kafe 4 kafe 38	74 Kafe 4 Kafe 2 Kafe 13 kafe 2 kafe 2 kafe 51	62 KAFE 1 KAFE 2 Kafe 4 Kafe 22 kafe 1 kafe 32	5 kafe 5	43 KAFE 1 Kafe 2 kafe 1 kafe 39	202 KAFE 2 Kafe 34 kafe 12 kafe 9 kafe 145
CAFEER (1)(café)	1121	633 CAFEER 5 CAfeer 28 cafeer 116 cafeer 2 cafeer 19 cafeer 462 cafeer 1	3 cafeer 3	20 cafeer 1 cafeer 18	27 cafeer 27	0	97 CAFEER 1 CAfeer 1 cafeer 8 cafeer 1 cafeer 86	131 CAFEER 2 cafeer 4 cafeer 6 cafeer 1 cafeer 118	78 CAFEER 4 cafeer 1 cafeer 73	6 cafeer 6	1 cafeer 1	111 CAFEER 2 cafeer 12 cafeer 4 cafeer 6 cafeer 2 cafeer 85
KAFEER (1)(café)	777	172 Kafeer 1 Kafeer 9 kafeer 29 kafeer 4 kafeer 129	125 Kafeer 2 kafeer 123	137 Kafeer 5 kafeer 1 kafeer 131	157 Kafeer 5 kafeer 1 kafeer 151	16 kafeer 16	28 KAFEER 2 Kafeer 1 Kafeer 5 kafeer 20	38 Kafeer 2 kafeer 1 kafeer 35	33 KAFEER 1 Kafeer 1 kafeer 31	0	18 kafeer 18	49 Kafeer 2 Kafeer 3 kafeer 1 kafeer 1 kafeer 3 kafeer 3

Figure 5: Frequency distribution of *kafé* and *café* (and inflections of the two spelling variants)

4.2 Back-office SQL tool

The ‘back-office’ SQL query tool provides access to stored procedures which are used for comparing word form distribution in selected corpora. These procedures are used for identifying words the presentation of which might need to be reviewed and updated based on the actual use. For example, the syntactic examples given in the dictionary should reflect the actual use.

A call to a stored procedure can look like this:

```
CALL jmf_frekw ('BLOGGMIX2015_frekw', 'NCUSNI','NCUPNI',500,500)
```

Here, *BLOGGMIX2015_frekw* is the corpus used to extract word-form frequencies, *NCUSNI* and *NCUPNI* are the word-form tags to be compared (in this case *indefinite*

singular vs. *indefinite plural* for neuter nouns) and the last two figures set the threshold for words to be considered, the minimum frequency for each of the two word forms.

This call returns a table of words ordered by the relative percentage of the frequencies of the two word forms in the corpus (Figure 6). In this case, the top row holds the pair *minut/minuter* ('minute/minutes') with 1,017 occurrences of the singular indefinite vs. 7,648 occurrences of the plural indefinite, giving a relative percentage of 752%. The last row holds the pair *man/män* ('man/men') with a relative frequency of 1.5%. A quick check shows that the singular form *man* is homographic with other very frequent words (e.g. the generic pronoun *man* 'one'), which means our frequency information is not useful as a source of information regarding this word. The second last row displays *mamma/mammor* ('mum/mums'), having a relative frequency of 6.25% for the plural.

1 • CALL `SABOB`.`jmf_frek`('BLOGGMIX2015_frek', 'NCUSNI', 'NCUPNI', 500, 500);

Result Grid | Filter Rows: | Export: | Wrap Cell Content: |

s_nr	l_nr	grundform	ordform_bin	tagg	frek1	ordform_bin	tagg	frek2	relproc
174500	837197	minut	minut	NCUSNI	1017	minuter	NCUPNI	7648	752.0157
94882	245388	minut	minut	NCUSNI	1017	minuter	NCUPNI	7648	752.0157
57327	166291	förälder	förälder	NCUSNI	728	föräldrar	NCUPNI	3146	432.1429
2617	251962	människa	människa	NCUSNI	2258	människor	NCUPNI	8989	398.0957
151225	363107	titt	titt	NCUSNI	559	tittar	NCUPNI	1976	353.4884
123769	305585	sak	sak	NCUSNI	4681	saker	NCUPNI	15547	332.1299
114698	286657	produkt	produkt	NCUSNI	1156	produkter	NCUPNI	3805	329.1522
3197	362550	timme	timme	NCUSNI	2660	timmar	NCUPNI	8042	302.3308
65026	182549	hit	hit	NCUSNI	3427	hittar	NCUPNI	7493	218.6460
171753	406015	övning	övning	NCUSNI	668	övningar	NCUPNI	1416	211.9760
102338	261132	nv	nvhet	NCUSNI	534	nvheter	NCUPNI	1121	209.9251
102510	261480	nvhet	nvhet	NCUSNI	534	nvheter	NCUPNI	1121	209.9251
34799	118790	bild	bild	NCUSNI	7126	bilder	NCUPNI	14433	202.5400
166876	395825	vän	vän	NCUSNI	3908	vänner	NCUPNI	7763	198.6438
39521	128762	båt	båt	NCUSNI	829	båtar	NCUPNI	1579	190.4704
158291	377869	uopåift	uopåift	NCUSNI	598	uopåifter	NCUPNI	939	157.0234
2150	154213	fot	fot	NCUSNI	589	fötter	NCUPNI	920	156.1969
.....									
2625	254725	natt	natt	NCUSNI	3331	nätter	NCUPNI	771	23.1462
89743	234553	låt	låt	NCUSNI	3056	låtar	NCUPNI	699	22.8730
119599	296822	resa	resa	NCUSNI	4248	resor	NCUPNI	949	22.3399
133863	326704	slut	slut	NCUSNI	7676	slutar	NCUPNI	1383	18.0172
51185	153454	form	form	NCUSNI	4219	former	NCUPNI	748	17.7293
41753	133502	del	del	NCUSNI	18515	delar	NCUPNI	3244	17.5209
111889	280893	plats	plats	NCUSNI	8098	platser	NCUPNI	1180	14.5715
166339	394714	våg	våg	NCUSNI	5225	vågar	NCUPNI	652	12.4785
143412	346575	stund	stund	NCUSNI	6411	stunder	NCUPNI	772	12.0418
47878	146529	famili	famili	NCUSNI	4740	familier	NCUPNI	544	11.4768
155990	373079	två	två	NCUSNI	8556	tvåer	NCUPNI	615	7.1879
83830	222209	kväll	kväll	NCUSNI	9307	kvällar	NCUPNI	654	7.0270
149812	360055	tid	tid	NCUSNI	17364	tider	NCUPNI	1126	6.4847
91540	238381	mamma	mamma	NCUSNI	9242	mammor	NCUPNI	578	6.2541
2564	238441	man	man	NCUSNI	120074	män	NCUPNI	1821	1.5166

Figure 6: Stored procedure for comparing word forms.

The words in the top and bottom of the table stand out, and this could be important information for the dictionary user. For the word *minut*, the comment 'mostly in the plural form' could be added in the entry, and cases where the singular form is used might need to be analysed. Do examples like 'Give me a minute!' and 'it took him 15

minutes’ fit under the same definition? For the word *mamma*, the overwhelming use of the singular indefinite bare form in this corpus is probably an indication of this word form often functioning as a name rather than an ordinary noun. *Mamma var här i går* (‘mother was here yesterday’) does not mean that an indefinite or a generic mother was here.

SQL queries are used for establishing the “normal” distribution of word-form frequencies for each word class, eventually resulting in informational hints to the editors in the editorial interface (Figure 3 above). Finding the normal distribution is done by excluding homographic word forms from accumulated queries and the result is validated through comparison with the frequency information for some typical words (Figure 2 above).

5. Suggestions and discussion

Here, we provide examples of “candidates for change” found using our morphological frequency tool and typical considerations that arise when studying actual words and how they are presented in the dictionaries.

5.1 Word forms with a slightly different meaning than the base form:

Revise article structure or content?

When the word *blomma* ‘flower’ was identified as having a non-standard distribution of word form frequencies we reviewed the examples given in the dictionary entries in SO and Lexin. In both dictionaries the very common phrase *plocka blommor* (‘pick flowers’) is given as an example for the first sense of the word, which was surprising as the dictionaries have ordered the senses differently. The outcome was a decision to move not only this, but several syntactic examples from the first ‘colourful reproductive part of a plant’ to the second, ‘plant with flowers’ sense in the coming edition of SO.

5.2 Frequency of inflected forms varying with orthography of the headword: Change lemma form?

As for the word *skanner/scanner* with optional plurals *-rar/-s* discussed in Section 3.2, the more frequent form *scanner* will be the headword in the coming, updated version of SO. Our investigation suggests the plural *scanners* should be given as the preferred plural form for *scanner*, while *skannrar* would be the preferred plural for *skanner*, but the formal decision still has to be made.


5.3 Very frequent derivations

As for the word *ambulerande*, the present participle and *nomen actionis* of the verb *ambulera*, discussed in Section 3.3, there is already a note “(ofta pres. part.)”, ‘often (used as) present participle’ in the existing dictionary entry (see Figure 7). We also suggest adding a syntactic example illustrating this usage, as we cannot expect all

dictionary users to be familiar with the implications of the grammatical note.

ambulera verb

ambulerade ambulerat

UTTAL: ambule´ra 

- (ofta pres. part.) ständigt växla plats för sin verksamhet särskilt vid verksamhet som normalt är stationär

KONSTRUKTION: *ambulera (mellan NÅGRA)*
ambulera (NÅGONSTANS)

EXEMPEL:
som hemspråkslärare ambulerade hon mellan tio olika skolor

HIST.: belagt sedan 1768; av latinska *ambula´re* 'gå omkring'; jfr ursprung till **¹somnambul**

DET ATT ***ambulera*: ett ambulerande**

Figure 7: The verb *ambulera* lacking an example for the pres. participle *ambulerande*

5.4 Word forms in phrases which special syntactic functions

Our investigation suggests that an explicit grammatical note regarding the (almost exclusively) adverbial usage of the word *undantagslös* should be added to the dictionary, in conformance with how other similarly behaving words are presented.

5.5 Word form signalling multi-word expression

Certain words, regularly restricted to particular inflections, are almost exclusively associated with special constructions, for example the plural-only noun *döddagar* (lit. *dying-days*) in the prepositional phrase *till döddagar* ‘to my dying day/to the end of time’ or the fossilized indefinite singular *korvspad* ‘sausage stock’ in the adjective phrase *klart som korvspad* ‘plain as a pikestaff’ (see Sköldberg, 2007; Sköldberg, 2008). For these words it is, of course, essential that the special constructions they are associated with appear in the dictionary entry. Other words might occur frequently in an inflected form in common collocations that could be identified by corpora searches initiated based on word form frequency distribution anomalies. As yet, we have not had the time to do such systematic corpora searches.

5.6 Final discussion

The tools described in this paper have only been available to us for a limited time and we are still in a process of learning how to best take advantage of the new possibilities at hand. Moreover, limited personnel resources have not allowed for a thorough investigation of all words with an exceptional frequency distribution of word forms, but already looking at a few of these words has proved to us that the word-form relative frequency information gives a valuable additional aspect of knowledge to the lexicographer, providing a means to add quality to dictionary entries.

Providing an optimal toolbox for lexicographers, and giving the right amount of useful information at the right time, is a challenging task. Overloading the editorial interface with too much information can be perceived as a hindrance to the creative work of writing, but no-one is happy with getting important information too late in the process, when already having moved on mentally to the next task. The process of reviewing and enhancing dictionary content, and the tools provided for supporting this process, must therefore ideally be developed in close co-operation between system developers and editors.

6. References

- Atkins, S. B. T. & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.
- Battenburg, J. (1992). *English Monolingual Learners' Dictionaries: A User-Oriented Study*. Tübingen: Niemeyer Press.
- Berg, S. & Cederholm, Y. (2001). Att hålla på formerna. Om framväxten av Svensk morfologisk databas. In S. Allén (ed.) *Gäller stam, suffix och ord. Festskrift till Martin Gellerstam den 15 oktober 2001*. Göteborg: Meijerbergs institut för svensk etymologisk forskning, pp. 58–69.
- Borin, L, Forsberg, M. & Roxendal, J. (2012). Korp – the corpus infrastructure of Språkbanken. *Proceedings of LREC 2012*. Istanbul: ELRA, pp. 474–478.
- Bybee, J. (1995). Regular morphology and the lexicon. *Language and Cognitive Processes*, 10, pp. 425–455.
- Dąbrowska, E. (2008). The effects of frequency and neighbourhood density on adult speakers' productivity with Polish case inflections: An empirical test of usage-based approaches to morphology. *Journal of Memory and Language*, 58, pp. 931–951.
- Hannesdóttir, A. H. (1998). *Lexikografihistorisk spegel. Den enspråkiga svenska lexikografins utveckling ur den tvåspråkiga*. Göteborg: Meijerbergs institut för svensk etymologisk forskning.
- Hay, J. (2001). Lexical frequency in morphology: is everything relative? *Linguistics*, 39(6), pp. 1041–1070.
- Heuberger, R. (2018). Dictionaries to assist teaching and learning. In P. A. Fuertes-

- Olivera (ed.) *The Routledge Handbook of Lexicography*. Milton Park, Abingdon, Oxon: Routledge, pp. 300–316.
- Lew, R. (2012). How can we make electronic dictionaries more effective? In S. Granger & M. Paquot (eds.) *Electronic Lexicography*. Oxford: Oxford University Press, pp. 343–361.
- Lexin introduction: Om lexinordböckerna. Accessed at: <http://lexin.nada.kth.se/lexin/doc/Lexin.pdf>. (3 June 2019)
- Lexin: *Lexins svenska lexikon*. Accessed at: <http://lexin.nada.kth.se/lexin/>. (3 June 2019)
- Schenberg, P. (1739). *Lexicon latino-svecanum*. Norcopiæ: C.F. Broocmanno.
- Sköldberg, E. (2007). Birds of a feather and sheep's clothing: on unique constituents in Swedish. In M. Nenonen & S. Niemi (eds.) *Collocations and Idioms 1. Papers from the First Nordic Conference on Syntactic Freezes Joensuu, May 19-20, 2006*. Joensuu: Joensuun yliopisto, pp. 297–305.
- Sköldberg, E. (2008). Från vrångstrupen till fotabjället. Om presentationen av unika konstituentier i Svensk ordbok utgiven av Svenska Akademien. In Á. Svavarsdóttir, G. Kvaran, G. Ingólfsson & J. H. Jónsson (eds.) *Nordiske Studier i Leksikografi 9. Rapport fra konference om leksikografi i Norden. Akureyri 22.–26. maj 2007*. Reykjavík: Nordisk forening for leksikografi, pp. 421–432.
- Sköldberg, E. (2017). Innehållet i Svensk ordbok utgiven av Svenska Akademien – eller kampen mellan norm och bruk. In S. Bendegard, U. Melander Marttala & M. Westman (eds.) *Språk och norm. Rapport från ASLA:s symposium Uppsala universitet 21–22 april 2016*. Uppsala, ASLA: Svenska föreningen för tillämpad språkvetenskap, pp. 123–129.
- Spegel, H. (1712). *Glossarium – Sveo-Gothicum Eller Svensk-Ordabook*. Lund: A. Habereger.
- SO: *Svensk ordbok utgiven av Svenska Akademien*. (2009). Accessed at: <https://svenska.se/so/>. (3 June 2019)
- Svensén, B. (2009). *A handbook of lexicography. The theory and practice of dictionary-making*. Cambridge: Cambridge University Press.
- Trap-Jensen, L., Lorentzen, H. & Sørensen, N. H. (2014). An odd couple – corpus frequency and look-up frequency: What relationship? *Slovenščina 2.0*, 2(2), pp. 94–113.
- Trueswell, J. C. (1996). The role of lexical frequency in syntactic ambiguity resolution. *Journal of Memory and Language*, 35, pp. 566–585.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



Planning a Domain-specific Electronic Dictionary for the Mathematical Field of Graph Theory: Definitional Patterns and Term Variation

Theresa Kruse¹, Laura Giacomini^{1,2}

¹ Institute for Information Science and Natural Language Processing (IwiSt),
Universität Hildesheim, Universitätsplatz 1, D-31141 Hildesheim

² Institute for Translation and Interpreting (IÜD), University of Heidelberg,
Plöck 57a, D-69117 Heidelberg

E-mail: theresa.kruse@uni-hildesheim.de, laura.giacomini@uni-hildesheim.de

Abstract

We plan to create an electronic dictionary for the mathematical field of graph theory. The dictionary should help students to improve their usage of the mathematical terminology. Besides the alphabetical access, the dictionary will also provide thematic, onomasiological access; it will contain lemmas in German and English, related terms and equivalence statements. Presently, such a dictionary does not exist. The dictionary basis is formed by two corpora composed of textbooks, scientific papers and lecture notes, containing all the texts the students use in their graph theory course in German and English. In the current pre-lexicographic stage, our focus is on relations between terms and on patterns used in the corpus to express them. We collect the definition patterns in the corpus and plan to use them for term extraction. Thereby, we can extract the semantic relations at the same time. In this paper we explore in particular the synonymy relations from an orthographical, morphological and syntactic perspective and draw conclusions for data acquisition. It might be possible to apply our extraction methods later for creating dictionaries in other mathematical domains.

Keywords: terminology, mathematical; patterns; relations; term variation

1. An electronic dictionary for graph theory: brief overview

We plan to create an electronic dictionary for the mathematical field of graph theory. The dictionary shall be bilingual, German and English. The purpose of the dictionary is to help mathematics students to improve their academic writing regarding terminology. We extract terms from the texts using definition patterns and aim to associate with each pattern a particular semantic relation which we will then use to automatically create components of an ontology, as a backbone of the electronic dictionary.

In this paper, we first give an overview of the historical and linguistic aspects of graph theory and mathematics, respectively. The first step is to show that the language of

graph theory is a language for special purposes (Section 2). Section 3 deals with the planned dictionary itself. There will be a closer look at the target group, the composition of the corpus and at the planned structure concerning distribution, micro- and macrostructure as well as user guidance. Section 4 presents definition patterns, their creation and the semantic relations. Additionally, we introduce the topic of domain specific variants and provide a first analysis of their usage.

2. Historical and linguistic aspects of graph theory

In the following, an overview of the lexicographic aspects of mathematics is given. A complete theory of the multimodal structure of mathematical texts is still missing. Mathematical language is regarded as a symbolic language, and all conclusions are inherent to the language (Atayan et al., 2015). Nevertheless, mathematical texts have a macrostructure¹¹ in the sense of Roelcke (2010). The macrostructure consists of text types like definitions, theorems and proofs which came with the formalization of the mathematical language at the beginning of the 20th century (Atayan et al., 2015). According to Atayan et al. (2015), the language of mathematics, science and technology constitutes a linguistic variety.

The reasons for a particular term to be well-established are often historical and depend on influential publications. According to Hischer (2010), mathematical terminology uses words from the general language. That is the case for graph theory as well; for example *tree*, *complete* and *edge* also have mathematical meanings.

Graph theory is very young compared to other mathematical fields. The first problem of graph theory was the problem of the seven bridges in Königsberg, where the aim was to find a path through the city whereby every bridge is crossed only once (cf. Figure 1).

Leonard Euler proved in 1735/36 that this is not possible (Euler, 2009 (1736)). He called the mathematical field of this problem *Geometria situs* (geometry of position). More than a hundred years later Sylvester (1878) proposed the term *graph* for these structures. That was the first time the term *graph* appeared in this context. A further overview of the introduction of important terms in graph theory is given by Mulder (1992).

¹ It should be noted that the macrostructure of a language for special purposes differs from the macrostructure of a dictionary.

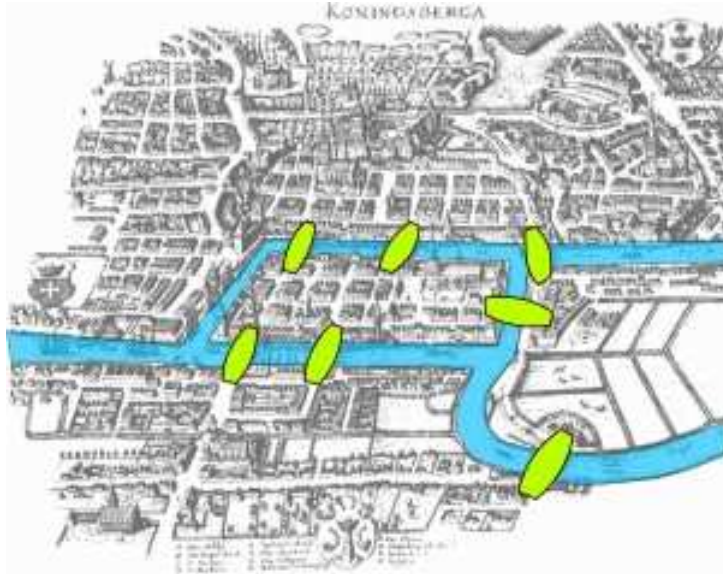


Figure 1: Map of Königsberg with the seven bridges to cross (Graphic: Bogdan Giușcă).

In this paper, the language of graph theory is regarded as a language for special purposes (LSP) because typical characteristics of LSP can be identified (Roelcke, 2010). Some of our examples only apply to the German language, as Roelcke's work is mainly targeted at German, and as, for example, German and English compounding patterns differ notably on the surface. One of Roelcke's (2010) LSP criteria is richness in compounds. In our German corpus we find examples of compounds like *Kantenzug*, *wohlquasigeordnet* or *Kantenfärbung*. A derivative is *Wohlquasigeordnetheit*. Abbreviations are also very common in mathematical texts in general: *f* stands for *function*, or *G* stands for *graph*.

Another criterion for a LSP according to Roelcke (2010) is the preference for the third person. To check this for the texts on graph theory we did an investigation on the part of the corpus which is already machine processable.² In German we searched for *ich*, *du*, *man*, *er*, *sie*, *es*, *wir*, *ihr*, *Sie*, *Leser*, *Leserin*. In English for *I*, *you*, *one*, *he*, *she*, *it*, *we*, *they*, *reader*. The results are given in Table 1.

We excluded the cases from the table in which *ihr* is used as a possessive pronoun as well as those in which *er* or *sie* refer as a pronoun to things, e.g. to a graph. As a result, the relevant subject pronouns in German are *man*, *es* and *wir* which together represent about 95 percent of all pronoun occurrences. Unlike in Roelcke's hypothesis, it is the first person plural, not the third person which dominates.³ This is a special feature of the LSP in mathematics. In English, the difference is even stronger, as one third is the use of *it* and two thirds concern *we*. The results are independent of the text

² 20,938 types and 482,604 tokens in German; 10,245 types and 378,629 tokens in English.

³ This investigation will be repeated as soon as the complete corpus is available.

type in the corpus. Obviously, this small investigation can only give a first overview of the usage of the person. Further investigation is necessary, but not part of this dictionary project.

	hits			hits	
ich	40	0.32%	I	6	0.09 %
du	0	0%	you	49	0.76%
man	2,177	17.19%	one	29	0.45%
er	2	0.02%	he	27	0.42%
sie	0	0%	she	0	0%
es	3,363	26.55%	it	2,078	32.17%
wir	6,550	51.71%	we	4,249	65.78%
ihr	0	0%	they	0	0
Sie	464	3.66%			
Leser in	70	0.55%	reader	21	0.33%
Sum	12,666	100%	Sum	6,459	100%

Table 1: Usage of personal pronouns

Nevertheless, we suppose the language of graph theory to be an LSP because we find examples for the other criteria, including recurrence and isotopy. We will make use of the latter in the creation of the pattern list.

3. Planning the dictionary

According to the model by Roelcke (2010), we want to consider the *intrafachlichen* and parts of the *interfachlichen Fachsprachwortschatz* (intra/inter domain specialized vocabulary) for the dictionary which means all the terms from the domain of graph theory and some terms from related mathematical domains.

3.1 Target group

The function theory of lexicography distinguishes between dictionaries for communicative, cognitive and interpretative situations (Fuertes-Olivera & Tarp, 2014; Tarp, 2008). The following description is based on the terminology in the taxonomy

presented by Bothma et al. (2017). They divide communicative situations into text reception and production, where the usage can be either automated or interactive.

The planned e-dictionary is primarily aimed at providing information interactively to the user in communicative as well as cognitive situations. On the one hand, users have to prepare presentations and texts in German on the basis of English texts, which is regarded as a communicative (text production) situation. The equivalents should help with that. On the other hand, the users do not simply have to translate the texts but also have to completely understand their content, which constitutes a cognitive situation. The communicative needs will be addressed by the provision of LSP equivalents. Here, the dictionary goes far beyond what can be found in general bilingual dictionaries: The latter would give both *komplett* and *vollständig* as equivalents of *complete*, while in graph theory the only acceptable and collocational equivalent is *vollständig*. The cognitive needs will be addressed by the inclusion of an ontology, such that the dictionary will support both semasiological and onomasiological access.

According to Roelcke (2010), there are some decisions to make. The target group are students, so they are semi-experts with a basic but no deeper knowledge of the subject. Furthermore, the dictionary will have a descriptive as well as a prescriptive function. The first step in dictionary creation is only descriptive, but some of our lemma selection criteria will include prescriptive elements. This is particularly true for the decisions related to variants, as we have to choose one main term for each variant. The main term should later be the main lemma. This is further investigated in Section 4.3.

3.2 The corpus

The dictionary is based on two corpora, one in English and one in German, composed of textbooks and scientific papers from the field of graph theory. Text sources are chosen in two steps due to different aspects. First, we chose all texts used in the bibliography for the lectures on graph theory at University of Hildesheim, because students attending these courses are the (first) target group of the dictionary. These texts are the lecture notes and (parts of) seven German books. The English subcorpus from this first step contains five books and 21 scientific papers.⁴

Secondly, we did a survey with 40 students asking them which sources they had been using for the preparation of their talks and asked to rate them according to their importance for the preparation. The importance could be rated on a scale from 1 (=very important) to 5 (=not important at all). The scores were the following: Internet 1.7, papers 1.74, other students 2.12, consultation-hour 2.39, books 2.93, lecture notes 3.04. The survey also had the aim to find out if further online resources needed to be included

⁴ Due to the amount of texts we will not give exact source references for the examples.

in the corpus. The Internet was used by 92% of the students and ranked highest with regard to importance compared to other resources. Wikipedia was the most common online resource, with 55% for the English and 47.5% for the German version. Other sources like forums were not relevant for the corpus due to quantitative and qualitative factors.

After a qualitative analysis, we included the relevant texts. Books with a general introduction to mathematics or algebra with no focus on graph theory were excluded. So we added two German and four English books as well as four scientific papers. Relevant scientific papers in German do not exist in this field. In total, the German corpus comprises the script of the lecture, five books on graph theory and four books of which only the parts about graph theory are chosen. At the moment not all components are fully digitized and accessible.

Using the typology of Gläser (1990), we deal with monographs and scientific articles (including abstracts) for the domain internal communication. For the domain external communication we have textbooks for academic purposes. The lecture notes shall be regarded as somewhere in between. In the English corpus, there are nine books and 26 papers. Both corpora contain approximately 500,000 tokens each, which is a relatively small but still acceptable size for an LSP-corpus.

3.3 The structure of the dictionary

For creating the dictionary, we have to consider aspects of micro- and macrostructure in the planning process. Furthermore, we will have a look at the planned access structure.

3.3.1 Microstructure

The dictionary will have a hierarchical microstructure (Wiegand, 1989). As already mentioned, many of the mathematical terms are also part of general language, so that information on pronunciation or part of speech is not needed by the target group.

The focus will be on semantic aspects. Therefore, the articles will contain definitions, abbreviations, equivalents, collocations as well as information on semantically related terms like, for example, synonyms, antonyms or hyponyms – basically all the relations which will be examined in Section 4.2 below. Additionally, there can be usage examples extracted from the corpus. An etymological indication might be interesting but depends on whether there are valid data available for the majority of the terms.

The decision about which grammatical information shall be included depends on a further analysis of the material. For example, the users have German as their L1, and therefore there is no need to include the gender of the nouns as many of the nouns are

also used in the general language. Only in the case of irregularities might it be worth including gender indications. Similarly, there is no need to include further information on morphological inflection forms.

3.3.2 Macrostructure and access structure

We use the term macrostructure in the way presented by Wiegand and Gouws (2013) and Bergenholtz et al. (2008). We strive to achieve a fully developed macrostructure which means that all elements of the macrostructure will be linked (Nielsen, 1994).

The main part of creating the macrostructure is the lemma selection. The dictionary should contain nouns, adjectives, verbs and the corresponding multi-word terms; additionally pronouns or adverbs if they appear in patterns with the mentioned items. The terms will be from the field of graph theory, in both German and English with their equivalents.

Nouns indicate, for example, parts of graphs, special kinds of graphs or graph groups having specific names, but also problems, algorithms and theorems with a proper name and terms you can associate with graphs. Adjectives mainly indicate qualities of a graph or of its parts. Verbs denote things a graph or its parts can do or things one can do with a graph.

In addition, common phrases shall be included. It will be discussed where to draw a line with regard to other parts of the mathematical language, because graph theory also includes aspects of linear algebra. This decision will be made on the basis of corpus evidence.

According to the terminology discussed in Giacomini (2015), the dictionary shall have a search interface, an alphabetical index with a list of the alphabet characters as well as a list of alphabetically ordered terminological lemma signs and a systematic index. The latter might be based on the ontology, as the user can browse it with this index. For example you can choose ‘qualities of a graph’ and find the subcategories *vertex*, *edge* and *other*. Potentially, there will be included a tool in which one can insert a graph and the corresponding qualities and articles are shown.

The articles can be addressed either by semasiological or onomasiological access (Engelberg et al., 2016). For the first case, there will be a query form where after two or three letters a drop-down menu appears offering terms fitting the query. Thereby, the user might save some time during the search process. Speech recognition can be an option if appropriate software is available, but will not be a main focus. Furthermore, there will be the possibility of searching terms with an alphabetical index. Additionally, graphic elements can be included to show graphs and their corresponding lemmas.

4. Preparing the extraction of patterns, relations and variants

4.1 Finding definition patterns

We build on the methods used by Meyer (2001) and Barnbrook (2002). We identified typical patterns for definitions. They were found by looking closely at some of the texts, finding the patterns in the definitions in the first chapter, and using them as a random sample. In the next step, the detected patterns were applied to the corpus in order to verify if a pattern generalizes.

A further step was made by looking for all possible complements the patterns could have, and so resulting in the final patterns. The list is not fixed yet, but shall be extended during the project.

4.2 Semantic relations

Given the list of patterns, we tried to associate each pattern with a particular semantic relation. In some cases, the relations were ambiguous which resulted in an adjustment of the patterns. For example, we had the pattern *X is called Y* which was used for hypernyms, attributes and synonyms. A more detailed analysis allowed us to distinguish more refined patterns of *is called* as shown in Table 2. As Table 2 shows, it might be also possible to extract several relations from the same pattern as per (6), (7) and (8).

	Pattern	Relation
(1)	If-clause N1 is called N2	N1 hyp N2
(2)	N1 is called N2 If-clause	N1 hyp N2
(3)	N is called ADJ	ADJ attr N
(4)	N1 is called N2	N1 syn N2
(5)	N1 of N2 is called N3 If-clause	(N1 of N2) hyp N3
(6)	ADJ N1 is called N2	ADJ attr N1
(7)	ADJ N1 is called N2	ADJ N1 syn N2
(8)	ADJ N1 is called N2	N1 hyp N2

Table 2: Pattern *is called*. *hyp* stands for hyperonymy, *attr* for an attributive relation and *syn* for synonymy.

The chosen relations are based on GermaNet (Hamp & Feldweg, 1997; Heinrich & Hinrichs, 2010). Some adjustments were made as not all GermaNet relations are relevant for the domain of mathematics. At the same time some relations were added.

In GermaNet there are the following relations: synonymy, antonymy, hyperonymy / hyponymy, meronymy / holonymy, causation, association, pertonymy, participle and compound relations.

We use synonymy, antonymy, hyperonymy / hyponymy, meronymy / holonymy and pertonymy in the same way as GermaNet. Causation might be interesting, but most of the examples we found had a structure like *färben* – *gefärbt* which is a pertonymy relation.

For an association GermaNet gives the example *Schließvorrichtung* – *schließen*. We use the term association in a sense more typical for mathematics, in which it describes a kind of mapping, e.g. *weight* – *edge*. Compound relations might be investigated at a later point in time.

Furthermore, we use some new relations: an attributive relation between adjectives and nouns as not every noun term can be described by any attribute. For example a *Graph* can be *zusammenhängend* (engl. *connected*) but a *Kante* (engl. *edge*) cannot.

Additionally, with each algorithm or each mathematical process, we can associate its purpose: you use the *Hierholzer-Algorithmus* to find an *Eulertour*. We call the semantic relation between *Hierholzer-Algorithmus* and *Eulertour* ‘purpose’. Eponyms shall also be indicated in the dictionary, cf. *Euler* – *Eulertour*.

Another domain-specific relation is given by alternatives, for example two different algorithms for the same purpose. Additionally, there are analogies, such as *Eckenfärbung* and *Kantenfärbung*. An open topic to investigate in this context are differences between German and English in the cases where the German language tends to use compounds which do not exist in a similar form in English. Therefore we not only have relations between single word terms, but between multi-word terms as well.

The last type of relation, e.g. combinations between verbs and nouns appearing together, cannot be found within patterns.

4.3 A closer look at variants

4.3.1 The notion of synonymous variation

In this contribution, we would also like to address the topic of synonymous variation as a phenomenon in mathematical terminology. Just like other LSP, the language of mathematics is not free from synonymy. As already seen in the previous section, synonymy is one of the semantic relations that can be identified in definitional patterns.

This study deals with homogeneous text genres. This means that synonymous variants of a term can be found in texts with comparable content and structural characteristics.

Hence, synonymous variation is not embedded in different systemic levels (like in the case of chronological or geographical variation), but rather in the same textual system. In order to adequately cover this kind of non-diasystemic synonymy, we apply the definition and the classification model proposed by Giacomini (2019) and originally developed for technical language. In this model, variation is defined as the presence, within a domain discourse, of one or more synonymous and morphologically similar terms. Synonymy is understood as a semantic function shared by words in the same or in similar contexts. The notion of functional synonymy also allows for the inclusion of near synonyms.

Despite our focus on non-diasystemic variation, we cannot exclude the presence of some register variants *a priori*. In our future work, we will be able to provide more details on this.

Lexicographic resources supporting text production should include variation inside a specific microstructural position, providing dictionary users with necessary information about variant types available for a certain term and their distribution in the reference corpus (e.g. source type, source name, author, etc.).

4.3.2 Variant location and distribution

In our comparable corpora, synonymous variants can be found in

- definitions (definitional patterns) and
- other textual components (e.g. titles, text body).

The former type of variant description is particularly relevant for its substantial contribution to the explicit and normative building of mathematical terminology. Among variants are both single-word terms and multi-word terms. We will now give some examples of definitional patterns in which the available variant pairs or chains are highlighted:

- (a) A **closed path** is called a **cycle**
- (b) A **connected forest** is called a **tree**
- (c) A **maximal independent set** is called a **basis**
- (d) Die **Elemente von V** nennen wir **Ecken** (oder **Knoten**; engl. **vertices**) **von G**, die **Elemente {u, v} in E** heißen **Kanten** (engl. **edges**) **von G**
- (e) Die **Elemente von V** nennen wir **Ecken von D**, die **Elemente (u, v) in A** heißen **Bögen** (oder **gerichtete Kanten**) **von D**

- (f) Im folgenden bezeichnen wir mit $K = K(G)$ immer die **Anzahl der Komponenten eines Graphen G**

Besides variation at the level of contents related to graph theory, definitional patterns also reveal ‘functional’ variants, i.e. variants of terms which are employed to build the definition itself, e.g. *X bezeichnen wir mit Y* and *X nennen wir Y* in German, as well as *X is called Y* and *X heißt Y* in the English-German language comparison. In these patterns, Y indicates the definiendum, X the definiens.

We consider the definiens to be per se a variant of the definiendum, independently of its form, which can be

1. the combination of a genus proximum and differentia specifica like in *closed path* (cf. example (a)), with the hypernym path specified by *closed*, or *maximal independent set* (cf. example (c)), with the hypernym *set* subsequently specified by *independent* and by *maximal*;
2. a proper synonym or paraphrase like in *Elemente von V* (cf. example (d)) or *Anzahl der Komponenten eines Graphen* (cf. example (f)).

Definitional patterns may include more than one variant. Among variants, we also count English equivalents provided by some German sources (cf. example (d)). Variation within definitions is sometimes expressed in more complex ways, for instance through the inclusion of conditional restrictions for synonymy (cf. example (g)), or cross-referencing to other passages (cf. example (h)):

- (g) Eine Menge $M \subseteq E$ von Kanten in einem Graphen $G = (V, E)$ heißt **Matching** (oder **Paarung**), wenn keine zwei Kanten aus M einen gemeinsamen Knoten besitzen
- (h) Der in der Graphentheorie übliche Name für eine **Tabelle, die einen Graphen in der oben angegebenen Weise beschreibt**, ist **Adjazenzmatrix**

We also observe the presence of concatenated definitions in successive sentences, with a term first used as a definiendum and then as the definiens of a new term, for instance in:

- (i) Das lässt sich leicht durch einen weiteren Begriff beschreiben: Ein **Graph, der als ebener Graph gezeichnet werden kann**, d.h. zu einem ebenen Graphen isomorph ist, heißt **plättbar** (oder **planar**). Ein **Würfel** ist also ein **plättbarer Graph** und wie wir oben gesehen haben ebenso alle anderen **Polyeder**

In example (i), the following complex variation structure can be identified in discourse:

- *Polyeder* is a hypernym of *Würfel*

- variants of *Polyeder* and *Würfel* are *plättbarer Graph*, *planarer Graph*, *Graph, der als ebener Graph gezeichnet werden kann* and *Graph, der zu einem ebenen Graph isomorph ist*.

This example also hints at a common feature of definitional texts: variants may be introduced for definitional purposes only (cf. *planar* as a synonymous variant for *plättbar*) without being further employed in the text. Tables 3, 4 and 5 display the corpus distribution of the synonymous variants collected so far, together with their absolute frequency.

Only a corpus-based diachronic study could provide relevant information for what concerns the origin of variation in the language of graph theory. Some cases, however, suggest the influence of the English language on German terminology, for instance for EN *adjacent* (which has a Latin origin) and DE *adjazent*, which coexists with the Germanic form *benachbart*, or EN *Chinese Postman Problem* and the loan translation DE *chinesisches Briefträgerproblem*, which coexists with some German adaptations such as *Problem des chinesischen Postboten*.

Motivation for the presence of one variant or another is also a complex aspect to handle, which would require a detailed analysis of textual structures and contents (cf. Freixa (2006) for a study on variation motivation).

4.3.3 Variant classification

In this study, we apply the classification devised by Giacomini (2017) and Giacomini (2019) for the technical language, with the following three variation types:

- orthographical variation (OV, mainly concerning changes in hyphenation and capitalisation),
- morphological variation (MV, concerning changes in lexical morphemes), and
- syntactic variation (SV, concerning changes in the order of compound elements, words, and syntagmatic structures).

According to this variation model, each pair *main term*, *variant* is analysed in terms of the combination of all three variation types, which can take the following values: OV / no OV; full MV / partial MV / no MV; SV / no SV. Among the criteria for determining which is the main term of a variant cluster, we choose frequency as the most suitable at the moment (for a discussion on the topic of main terms cf. Giacomini (2019)). We decided not to automatically choose a term introduced in a definition as the main term. This is due to the fact that distributions of variants in texts show that these terms are often not systematically employed in the argumentation following a definition.

Some of the previously listed variants will be classified in Table 3 in relation to the corresponding main term. The starting point are ten possible variation patterns resulting from the combination of the three variation types (cf. Table 3).

Information concerning the available variant patterns and the source in which they typically occur should be made available in the specialized dictionary to support users during text production.

Variation pattern			Language	Main term	Variant(s)
noOV	fullMV	SV	DE	TSP	Traveling Salesman Problem
			DE	Bogen	gerichtete Kante
OV	partMV	SV	DE	Dijkstra-Algorithmus	Dijkstras Kürzeste-Wege-Algorithmus
noOV	partMV	SV	EN	x and y are adjacent	y is a neighbour of x
OV	noMV	SV	DE	Eulerchar (S)	Euler-Charakteristik von S
noOV	noMV	SV	DE	Hamiltonkreis	Hamiltonscher Kreis
			DE	Eulertour	eulersche Tour
noOV	fullMV	noSV	DE	chordal	trianguliert
OV	partMV	noSV	EN	four colour theorem	four-color conjecture
noOV	partMV	noSV	EN	eulerian tour	Euler tour
			EN	plane graph	planar embedding
OV	noMV	noSV	DE	Eulerscher Kantenzug	eulerscher Kantenzug
			DE	Petersen-Graph	Petersen Graph
noOV	noMV	noSV	EN	Petersen graph	Petersen's graph

Table 3: Variant classification (OV: orthographical variation, MV: morphological variation, SV: syntactic variation).

4.3.4 Variant identification and extraction

Variants are either explicitly introduced in texts by means of formulations that usually put them in relation to a main term (this is mostly the case of definitions), or employed as alternatives to the main term.

As previously mentioned, variants can be also found in textual components other than definitions, for example in

- (j) Zur geschickten Konstruktion von Eulertouren in Graphen, die diese Eigenschaften besitzen, gibt es zwei verschiedene Algorithmen, den **Zwiebelschalen-Algorithmus** (**Hierholzer-Algorithmus**) und Fleurys Algorithmus

At the present stage of the project, we cannot predict the level of heterogeneity of variation description in text bodies concerned with graph theory. Our assumption, however, is that heterogeneity poses particular problems for the automatic extraction of variants from a corpus.

So far, we have identified variants by manually analysing definitional patterns and by relying on our own specialized expertise. As soon as corpus pre-processing and annotation will be completed and textual data and structures analysed more closely, rule-based and statistical approaches will be applied to detect further synonymous variants in texts (cf. Giacomini, 2019) for the model of variant extraction from technical texts).

5. Conclusion and further work

We have proven that the language of graph theory is an LSP according to Roelcke, although there are some exceptions to his criteria definitions. Therefore we have the possibility of creating an electronic LSP dictionary. This process can be automated to a considerable degree, as there are pattern structures in the mathematical language which are used to express certain semantic relations. Another aspect we have to consider in the creation process of the dictionary are orthographical, morphological and syntactic variations. They can be extracted as well.

For our future work we have to come up with an approach that allows us to decide which variant should be regarded as the main term. For this decision, we will use linguistic and technical factors. In addition, it is still necessary to investigate how to guarantee that all patterns and all variants for a term are found.

6. References

- Atayan, V., Metten, T. & Schmidt, V.A. (2015). Sprache in Mathematik, Naturwissenschaften und Technik. In *Handbuch Sprache und Wissen*. Berlin/Boston: De Gruyter, pp. 411–434.
- Barnbrook, G. (2002). *Defining Language: A local grammar of definition sentences*. Amsterdam: John Benjamins.
- Bergenholtz, H., Tarp, S. & Wiegand, H. E. (2008). Datendistributionsstrukturen, Makro- und Mikrostrukturen in neueren Fachwörterbüchern. In *Fachsprachen: Ein internationales Handbuch zur Fachsprachenforschung und Terminologiewissenschaft*. Berlin/Boston/New York: De Gruyter, pp. 1762–1832.
- Bothma, T. J. D., Prinsloo, D. J. & Heid, U. (2017). A taxonomy of user guidance devices for e-lexicography. *Lexicographica*, 33, pp. 391–422.
- Engelberg, S., Müller-Spitzer, C. & Schmidt, T. (2016). Vernetzungs- und Zugriffsstrukturen. In *Internetlexikografie. Ein Kompendium*. Berlin/Boston: De Gruyter, pp. 153–195.
- Euler, L. (2009 (1736)). Lösung eines Problems, das zum Bereich der Geometrie der Lage gehört (Solutio problematis ad geometriam situs pertinentis). In W. Velminksi (ed.) *Die Geburt der Graphentheorie: Ausgewählte Schriften von der Topologie zum Sudoku*. Berlin: Kulturverlag Kadmos, pp. 11–27.
- Freixa, J. (2006). Causes of denominative variation in terminology. A typology proposal. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 12(1), pp. 51–77.
- Fuertes-Olivera, P. A. & Tarp, S. (2014). *Theory and Practice of Specialised Online Dictionaries - Lexicography versus Terminography*. Berlin/Boston: De Gruyter.
- Giacomini, L. (2015). Macrostructural properties and access structures of LSP edictionaries for translation: the technical domain. *Lexicographica*, 31, pp. 90–117.
- Giacomini, L. (2017). An Ontology-terminology Model for Designing Technical edictionaries: Formalisation and Presentation of Variational Data. In *Proceedings of eLex*. Leiden, Netherlands, pp. 110–123. URL <https://elex.link/elex2017/wp-content/uploads/2017/09/paper06.pdf>.
- Giacomini, L. (2019). Ontology - Frame - Terminology. A method for extracting and modelling variants of technical terms. Habilitationsschrift, forthcoming.
- Gläser, R. (1990). *Fachtextsorten im Englischen*. Tübingen: Gunter Narr.
- Hamp, B. & Feldweg, H. (1997). GermaNet - a Lexical-Semantic Net for German. In *Proceedings of the ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*. Madrid, Spain, pp. 9–15. URL <https://www.aclweb.org/anthology/W97-0802>.
- Heinrich, V. & Hinrichs, E. (2010). GernEdiT - The GermaNet Editing Tool. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*. Valletta, Malta, pp. 2228–2235. URL http://www.lrec-conf.org/proceedings/lrec2010/pdf/264_Paper.pdf.

- Hischer, H. (2010). *Was sind und was sollen Medien, Netze und Vernetzungen? - Vernetzung als Medium zur Weltaneignung*. Hildesheim/Berlin: Franzbecker.
- Meyer, I. (2001). Extracting knowledge-rich contexts for terminography: A conceptual and methodological framework. In *Recent Advances in Computational Terminology*. Amsterdam/Philadelphia: John Benjamins, pp. 279–302.
- Mulder, H. M. (1992). Die Entstehung der Graphentheorie. In K. Wagner & R. Bodendiek (eds.) *Graphentheorie: Zahlen, Gruppen, Einbettungen von Graphen und Geschichte der Graphentheorie*. Mannheim/Leipzig/Wien/Zürich: Wissenschaftsverlag, pp. 296–313.
- Nielsen, S. (1994). *The Bilingual LSP Dictionary - Principles and Practice for Legal Language*. Tübingen: Gunter Narr.
- Roelcke, T. (2010). *Fachsprachen*. Berlin: Erich Schmidt.
- Sylvester, J. J. (1878). Chemistry and Algebra. *Nature*, 17, p. 284.
- Tarp, S. (2008). *Lexicography in the Borderland between Knowledge and Non-Knowledge. General Lexicographical Theory with Particular Focus on Learner's Lexicography*. Tübingen: Max Niemeyer.
- Wiegand, H. E. (1989). Arten von Mikrostrukturen im allgemeinen einsprachigen Wörterbuch. In *Wörterbücher. Dictionaries. Dictionnaires. Ein internationales Handbuch zur Lexicographie*. Berlin/New York: De Gruyter, pp. 462–501.
- Wiegand, H.E. & Gouws, R.H. (2013). Macrostructures in printed dictionaries. In *Dictionaries: An International Encyclopedia of Lexicography*. Berlin/Boston/New York: De Gruyter, pp. 73–110.

Synonymous variants in German	number of texts	number of hits
adjazent	3	212
Benachbart	7	207
Bogen	5	226
Gerichtete Kante	4	39
Chinesisches Briefträgerproblem	2	10
Briefträgerproblem	1	2
Chinesisches-Postboten-Problem	1	1
Problem des chinesischen Postboten	1	1
Chinese Postman Problem	1	1
chordal	2	18
trianguliert	4	11
Dijkstra-Algorithmus	3	6
Dijkstras-Algorithmus	1	4
Algorithmus von Dijkstra	2	3
Dijkstras Krzeste-Wege-Algorithmus	1	1
Euler-Charakteristik von S	1	1
Eulerchar (S)	1	2

Eulerscher Kantenzug	2	4
Eulerweg	1	3
offener Euler-Zug	1	1
eulerscher Kantenzug	1	1
Eulertour	3	48
eulersche Tour	1	33
Eulersche Tour	1	8
Euler-Kreis	1	8
Eulerkreis	1	5
geschlossener Euler-Zug	1	1
Hamiltonkreis	3	127
hamiltonscher Kreis	1	17
Hamiltonscher Kreis	2	7
Traveling Salesman-Tour	1	1
Königsberger Brückenproblem	4	29
Brückenproblem	2	3
Matching	7	538
Paarung	3	90
Petersen-Graph	5	37
Petersen Graph	1	2
plättbar	2	51
planar	6	73
TSP	1	18
Serien-Parallel-Graph	1	5
sp-Graph	1	4
Traveling Salesman Problem	2	14
Traveling Salesman-Problem	1	7
Rundreiseproblem	1	5
Problem des Handlungsreisenden	1	1
Vierfarbenproblem	5	15
Vier-Farben-Problem	4	12
Vier-Farben-Satz	2	7
4-Farbenproblem	1	2
Zwiebelschalen-Algorithmus	1	6
Algorithmus von Hierholzer	1	2
Zwiebelschalenalgorithmus	1	1
Hierholzer-Algorithmus	1	1
Bestimmung einer Eulertour nach	1	1
Algorithmus nach Hierholzer	1	1

Table 4: Examples for corpus distribution of the synonymous variants in German.

Synonymous variants in English	number of texts	number of hits
arc	6	301
directed edge	5	9
Chinese remainder theorem	1	17
Chinese Remainder Theorem	2	5
Euler totient function	2	5
Euler's totient function	1	1
Euler's Phi function	1	1
eulerian tour	1	53
Euler tour	1	15
Euler circuit	1	1
four colour theorem	2	20
four colour problem	2	4
four-color conjecture	1	1
Hamilton cycle	2	71
Hamiltonian cycle	3	13
if and only if	18	510
iff	1	1
Petersen graph	3	152
Petersen's graph	1	3
plane graph	3	115
planar embedding	3	12
embedding in the plane	1	1
x and y are adjacent	14	440
y is a neighbour of x	3	87
neighbor	5	10
$X \sim y$	1	1

Table 5: Examples for corpus distribution of the synonymous variants in English.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



Text Visualization for the Support of Lexicography-Based Scholarly Work

Shane Sheehan, Saturnino Luz

Usher Institute of Population Health Sciences & Informatics,
The University of Edinburgh, UK
E-mail: Shane.Sheehan@ed.ac.uk, S.luz@ed.ac.uk

Abstract

We discuss three visualisation techniques for corpus analysis, Concordance Mosaic, Metafacet and ComFre, and explore the design rationale based on a characterization of the corpus linguistic domain. The Concordance Mosaic visualization is designed for the investigation of collocation patterns. It encodes word positions in a concordance list in a manner that emphasizes quantitative analysis of frequency or collocation statistics. Metafacet provides an interface for investigating concordance lists through the lens of meta-data. When combined with the Mosaic it provides a powerful technique for investigating collocations in the context of meta-data. ComFre can be used to compare word frequencies between two corpora of different size, it has potential use as a technique for identifying terms which are representative of the corpora under investigation. The domain characterization shows how the visualizations were designed with corpus linguistic methodologies at the core. It consists of a task analysis based on the methodology outlined in Sinclairs' *Reading Concordances: An Introduction*, and the analysis of methodology case studies from language scholars.

Keywords: visualization; concordance; frequency; meta-data; collocation

1. Introduction

Concordance analysis is a core activity of scholars in a number of humanities disciplines, including corpus linguistics, classical studies, and translation studies, to name a few. Through the advent of technology and the ever increasing availability of textual data this type of structured analysis of text has grown in importance (Sinclair, 1991; Bonelli, 2010).

In concordance analysis, every corpus occurrence of a keyword of interest is displayed along with its context. The context is an ordered list of words which precede and follow the keyword. The analyst then seeks to discover the linguistic properties of the keyword and the contextual patterns which predict them by observing the frequencies of occurrence, in the keyword's context, of words (collocations), word combinations, parts of speech (colligations) or the various other lexical classifications (Sinclair, 2003; Scott, 2010).

The most widely used tool in this kind of analysis is a form of tabular visualization known as keyword-in-context (KWIC). The creation of concordances through the

keyword in context indexing technique was first proposed by Hans Peter Luhn in the 1950's (Luhn & Division, 1959). KWIC displays, enhanced in interactive systems by features such as search, context sorting and statistical analysis, are widely used not only by academics and scholars, but also by professional translators and post-editors (Karamanis et al., 2011; Doherty et al., 2012).

While these KWIC interfaces provide support for exploring the linear structure of the concordance, word frequency and other statistics rarely form any part of the visualization. This statistical information is essential to the work of the text analyst. However, in the presence of large corpora, it is difficult to explore statistical regularities armed solely with the KWIC display. External statistical tools are often used to complement the concordance. We argue that integration of this analysis step into the concordance visualization fits in well with the task structure of corpus linguists, and will be of great benefit to the text analyst.

There have been calls for the creation of more advanced concordance analysis tools (Rockwell, 2003), and advancements such as Sketch Engine have provide new analytic paths (Kilgarrieff et al., 2014). However, the adoption of visual analysis tools for concordance analysis is very limited. That does not mean that visual representations of the concordance do not exist, it is simply that they have not been adopted by analysts or integrated into analysis tools.

It has been suggested that the publication of more domain characterization papers for visualization would be beneficial for tool adoption (Munzner, 2009). It is at this level of design that relevant problems are identified, and creating visual solutions to problems that are not relevant to domain experts is wasted effort. Publication of domain characterization should also encourage wider conversation and help identify and characterize overlooked areas of investigation.

In this paper we outline the functionality of three corpus analysis tools, Concordance Mosaic, Metafacet and ComFre. Concordance Mosaic displays positional collocation statistics for any corpus word or regular expression. Interactive restructuring of a concordance browser is enabled through the interface. This restructuring combined with colour highlighting of the concordance lines creates a powerful technique for investigating significant collocation patterns.

The MetaFacet visualization enables exploration of corpora through the lens of meta-data. Keyword frequency can be investigated across any combination of meta-data attributes associated with corpus source files. The concordance browser and Mosaic can be interactively filtered by these attribute combinations, allowing investigation and comparison of lexical information across combinations such as date, author and topic.

ComFre is a tool for corpus frequency comparison, which provides a method of comparing corpora of different size in a visual and statistically valid manner.

These visualizations were designed in close collaboration with language scholars with an emphasis on translation studies. The design rationale is rooted in a domain characterization which encompasses a literature-based task analysis and ethnographic studies of methodology. Relevant portions of this domain characterization are presented following the visualization descriptions.

2. Modnlp plugins

The visualization tools are developed as plugins for the open source concordance browser included in the Modnlp toolkit. Significant contributions were also made to the core Modnlp project to better integrate the plugins and enable interactions with the concordance list. Modnlp provides a modular architecture and tools for natural language processing, it comes with an indexer, feature rich concordance browser and server implementation (Luz, 2011, 2000). Previous versions of the Modnlp software have been used by the European Parliamentary Comparable and Parallel Corpora project¹ (ECPC) and by the Translational English Corpus² (TEC). The toolkit is currently being developed as part of the Genealogies of Knowledge project³ (GoK) and the plugins are fully integrated into the GoK corpus browser.

The goal when developing these plugins is to improve the efficiency and capability of corpus linguistic methodologies and tools. Here we present the visualization plugins from a purely functional perspective to provide an overview of the capabilities and context for later discussion of the relevance to lexicography and corpus linguistics.

The English GoK corpus is used to exemplify the usage of the visualizations. This corpus is quite varied, it includes translations and re-translations of texts from antiquity as well as modern internet blogs and magazine articles. The corpus is designed to enable researchers to trace the trajectory of key concepts as they enter different cultural and temporal spaces, predominantly but not exclusively through the mediation of various forms of translation. The corpus is specialized and the examples used may not exhibit general lexical properties due to the issues of representativeness in relation to frequency (Summers, 1996).

In the discussion of the visualization functionality we do not try to analyse or interpret the linguistic properties of the words or corpus. Any analysis choice or comments on linguistic properties are to help clarify the examples and should not be viewed as an attempt to perform corpus analysis.

¹ <http://www.ecpc.uji.es/>

² <http://genealogiesofknowledge.net/translational-english-corpus-tec/>

³ <http://genealogiesofknowledge.net/>

2.1 Concordance Mosaic

The first visualization designed was the Concordance Mosaic. This visualization has the concept of keyword in context at its core. The visualization is designed to display word statistics per position extracted from a concordance list. The underlying graph based abstraction of the concordance list and an early prototype were presented in an earlier work (Luz & Sheehan, 2014).

Using the visual metaphor of the KWIC, Mosaic represents positions relative to the keyword as ordered columns of tiles. The mosaic is created using a space-filling approach introduced by Luz and Masoodian (2007), where each tile represents a word at a position relative to the keyword, and the height of each tile is proportional to the word statistic at that position. In its simplest form each tile represents the frequency of a word at a position relative to the keyword. In Figure 1 the Mosaic of the keyword “hazard” is presented along with the concordance list for the 335 occurrences in the corpus. The Mosaic is set to display column frequencies. Due to the strong visual metaphor of KWIC it should be clear the word “to” is the most frequent word immediately to the left of the keyword (K-1) and also at positions K-2 and K-3. Hovering over any tile will display a tool-tip with the word count and frequency at the position, this relieves the need for manually counting or performing additional searches to retrieve position based word frequencies.

Words with high corpus frequency tend to dominate the positional frequency distributions for most keywords. The second view Mosaic affords is a stop-word filtered view of column frequency. The columns are filtered using a threshold based on corpus frequency. In Figure 2, the stop-words are removed and column heights are no longer uniform. The reduction in a column’s height represents the density of stop-word frequency at that position. At K-1 we notice stop-words were the most frequent for any position. At K-1 the next most frequent word after stop-words is “moral”. Tile heights and thus frequency are comparable across positions, from the Mosaic we can see that “moral” at position K-1 and “run” at position K-2 have similar positional frequencies.

The mosaic and concordance browser have been presented together but we have not yet commented on the interaction. The data is linked to both interfaces, and interactions with the mosaic can be reflected on the concordance list. In Figure 2 the tile for the word “run” at K-2 has been left clicked with the mouse. This interaction colours white any position word tiles on the Mosaic that are found in concordance lines, including “run” at K-2.

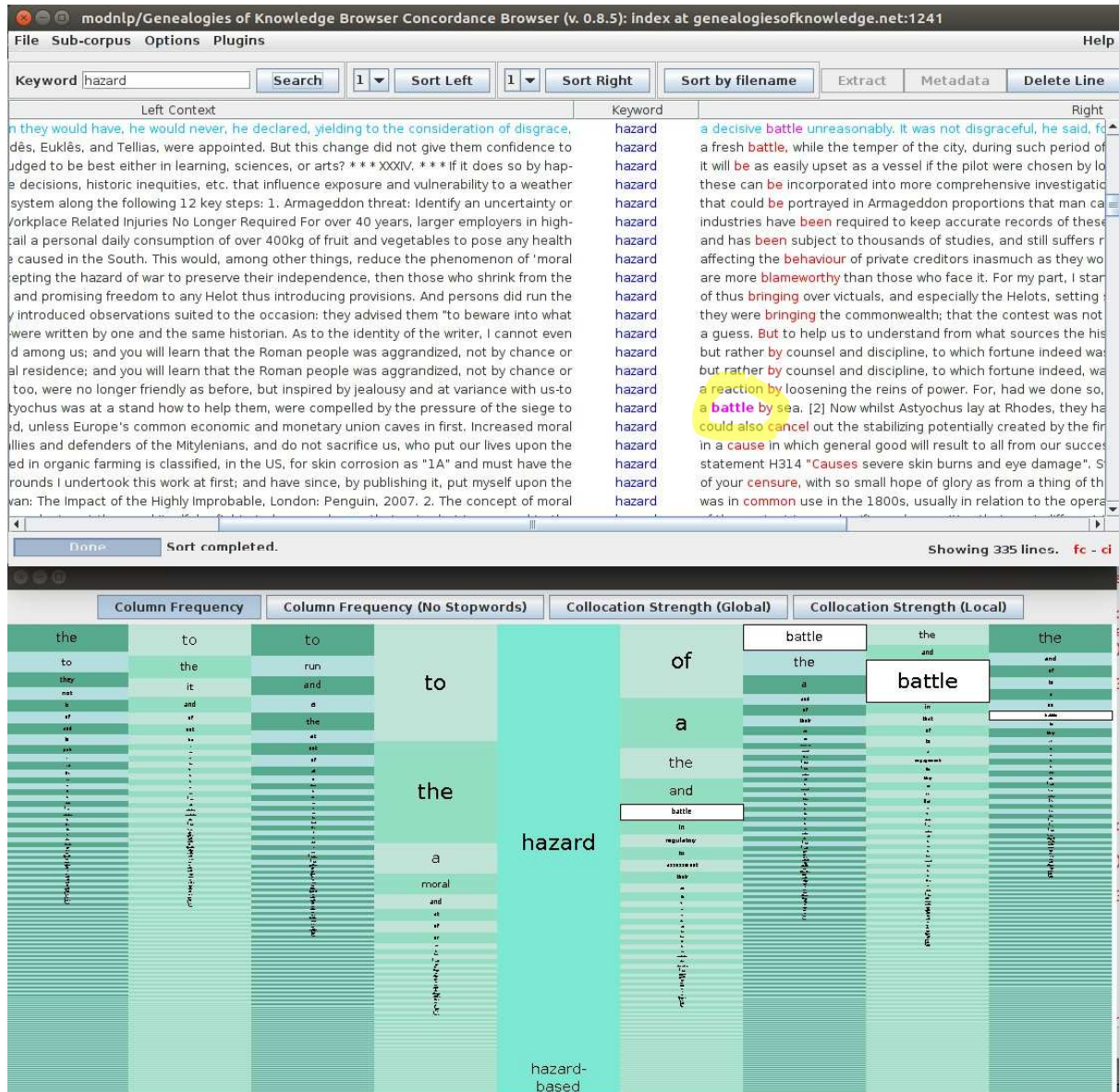


Figure 1: Concordance Mosaic for keyword “hazard”. Right click selection of “battle” at position K+3.



Figure 2: Concordance Mosaic for keyword “hazard”, stop words have been removed. Left click selection of “run” at position K-2.

Looking at the Mosaic we see that at least one concordance line with “run” at position K-2 also contains “battle” at K+2. The concordance list has been sorted at the selected position and scrolled automatically to the selected word. For emphasis the sorted position words are coloured red and the selected word coloured pink. The horizontal concordance lines for the selected word are coloured blue for easy identification. In addition, any occurrences of the selected word at other positions are also highlighted in pink, and as you investigate the entire list it is possible to get a sense of global patterns which may not be restricted to the selected position. In Figure 3 the selection of the word “to” at K-1 and a sample of its many occurrences at other positions are visible.

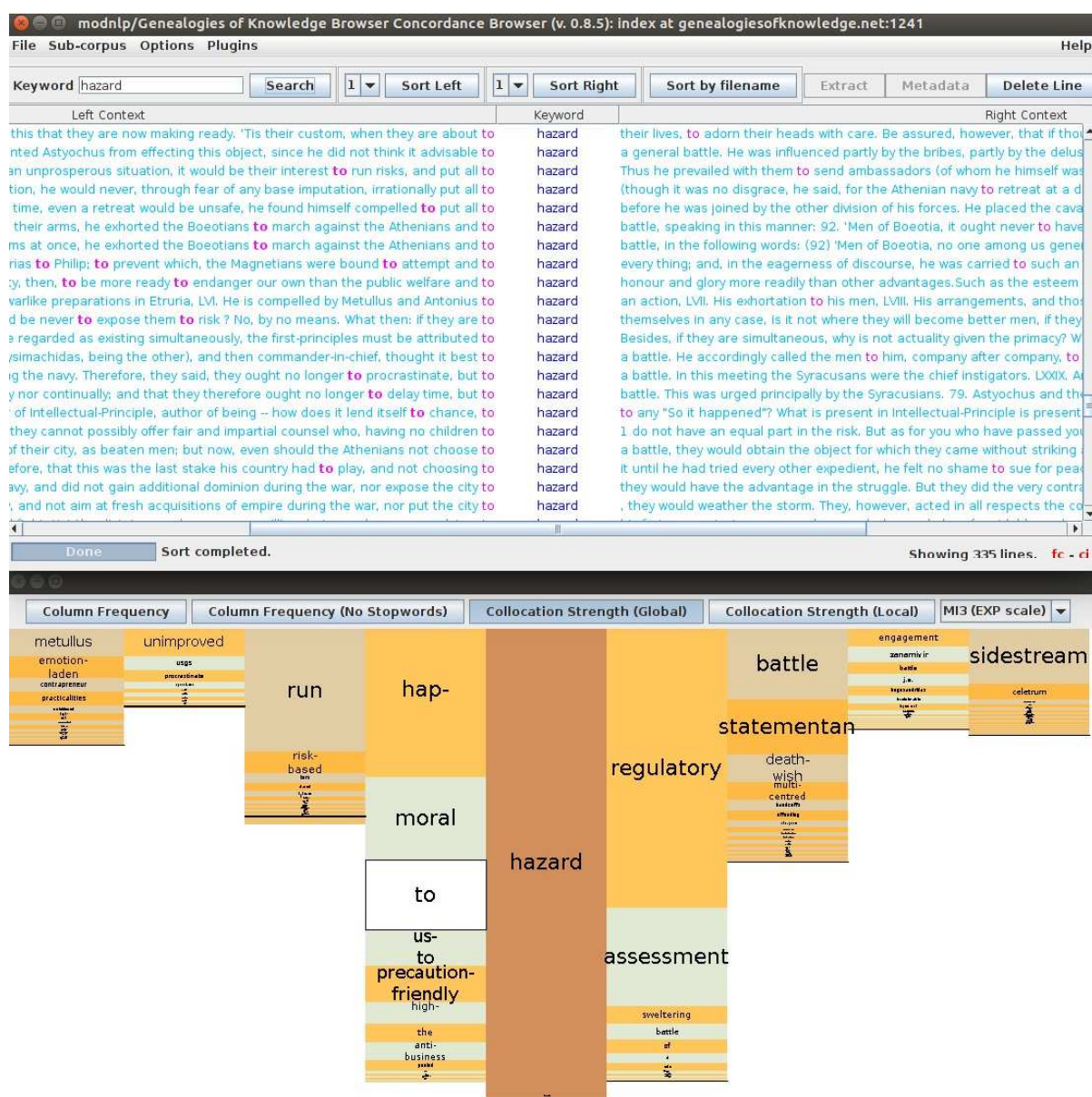


Figure 3: Concordance Mosaic for keyword “hazard”. Global view of MI3 is selected. Right click selection of “to” at position K-1.

The second click is activated by right clicking on a Mosaic tile. This interaction has the same effect on the concordance list as the left click interaction but differs in its change to the Mosaic. Right clicking on a Mosaic tile highlights other occurrences of the word at all positions in the mosaic. This is useful for getting a better sense of the frequency distribution of a word across all positions in a concordance list. In Figure 1, “battle” at K+3 is selected. Tiles representing “battle” at positions K+1 K+2 and K+4 are coloured white for easy identification. In the concordance list we can see one of these additional occurrences of “battle” at K+2 highlighted in pink.

Positional word frequency is a fundamental property of the concordance list, but other quantitative measures are used extensively to reason about collocations. Statistics such as Mutual Information (MI), Cubic Mutual Information (MI3) and Z-Score are often used to investigate collocation statistics in a window surrounding a keyword (Manning

& Schütze, 1999). This windowed approach most often groups word positions together and presents the results as a list. However we wish to preserve the positional aspect of these statistics and present them as a Mosaic. Figure 3 shows the *global collocation strength* view of Mosaic. Global in this setting means the tiles can be compared across positions and have not been scaled to fill the space. This contrasts with Figure 4 where the *local* view of collocation strength makes each column full height, and this allows easier investigation of each position but removes the ability to compare tiles across positions.

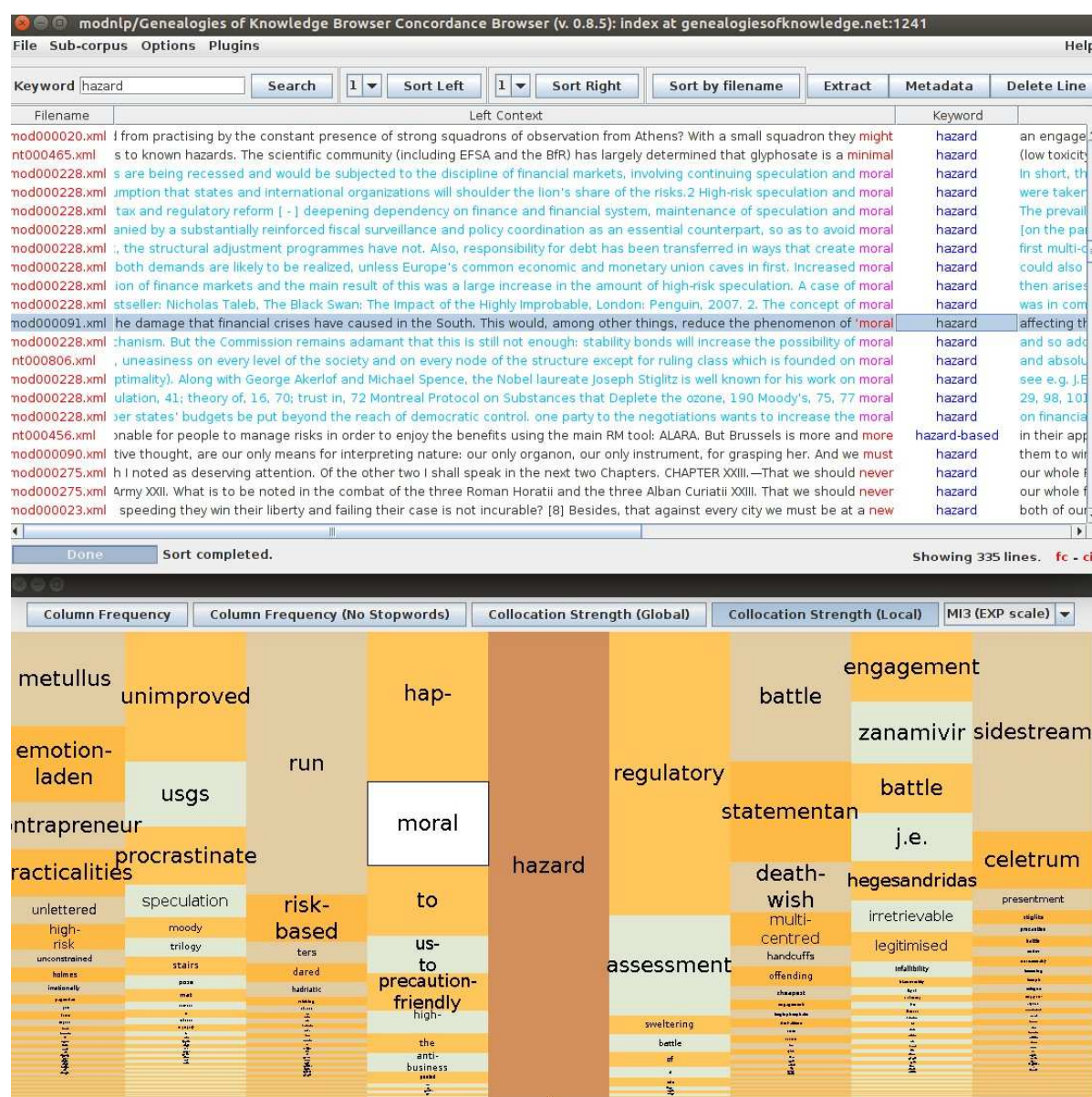


Figure 4: Concordance Mosaic for keyword “hazard”. Local view of MI3 is selected. Right click selection of “moral” at position K-1. Concordance list scrolled horizontally to reveal filenames.

In the *Global* view shown in Figure 3 the column heights give an indication of the word positions relative to the keyword where the statistical association is highest. Each individual tile’s height is proportional to the value of the statistic calculated for that

word at that position. In this example MI3 is selected as the statistic under investigation. The strongest association based on MI3 is the word “regulatory” at K+1. It may be worth noting that the stop-word “to” is shown to have a strong association at K-1.

If we investigate the concordance lines of the tile “moral” at K-1 (since it has both high frequency and MI3 score) we find that all but two of its 14 occurrences originate from the same file, see Figure 4.

2.2 Metafacet

The Modnlp concordance browser presents the file-names along with concordance lines. An interaction is available in the browser to view meta-data about each file and section on a line by line basis. However, this is a time consuming and challenging process for the corpus analyst if the meta-data of a large number of lines need to be investigated. The Metafacet plugin is a proposed solution to this issue and provides interactive filtering of the concordance list and the Mosaic using all available meta-data facets.

The Metafacet interface is quite simple, and uses a horizontal bar chart to display concordance line frequency per meta-data attribute. An attribute is a possible value that a meta-data facet can take. As an example “Plato” is an Attribute of the Facet “author”. A drop-down list is used to choose which facet is displayed and the bars are sortable by frequency or lexicographical order and the window can be filtered using a sliding scale to view a smaller portion of the attributes. This conforms to the common visualization design practice of first presenting an overview, and then more detail on demand (Shneiderman, 1996).

In Figures 5 and 6 the Metafacet interface for the concordance of “hazard” is shown for the facet “author” sorted by frequency. Figure 6 shows a window of this data focusing on the nine most frequent attributes of this facet in the concordance list. The hover interaction is shown for “Thucydides”, who is the most frequent author of the keyword “hazard” in the GoK corpus, with a total of 94 concordance lines out of a list of 335.

Metafacet when used alone provides an interface to quickly explore keyword distribution across meta-data attributes. By interactively combining it with the concordance list and Mosaic we can navigate the corpus in a new way, viewing the concordance as attributed sets of collocations that can be interactively explored. In Figure 7 the stop-word Mosaic shown in Figure 2 is filtered to remove any concordance lines with the attribute “book” from the “format” facet. Books account for the majority of the concordance lines, and removing them from the concordance significantly changes the collocation structure of the Mosaic. During interactive filtering the current selection can be kept by pressing the “Update Bars” button, and this will refresh the Metafacet window with filtered concordance.

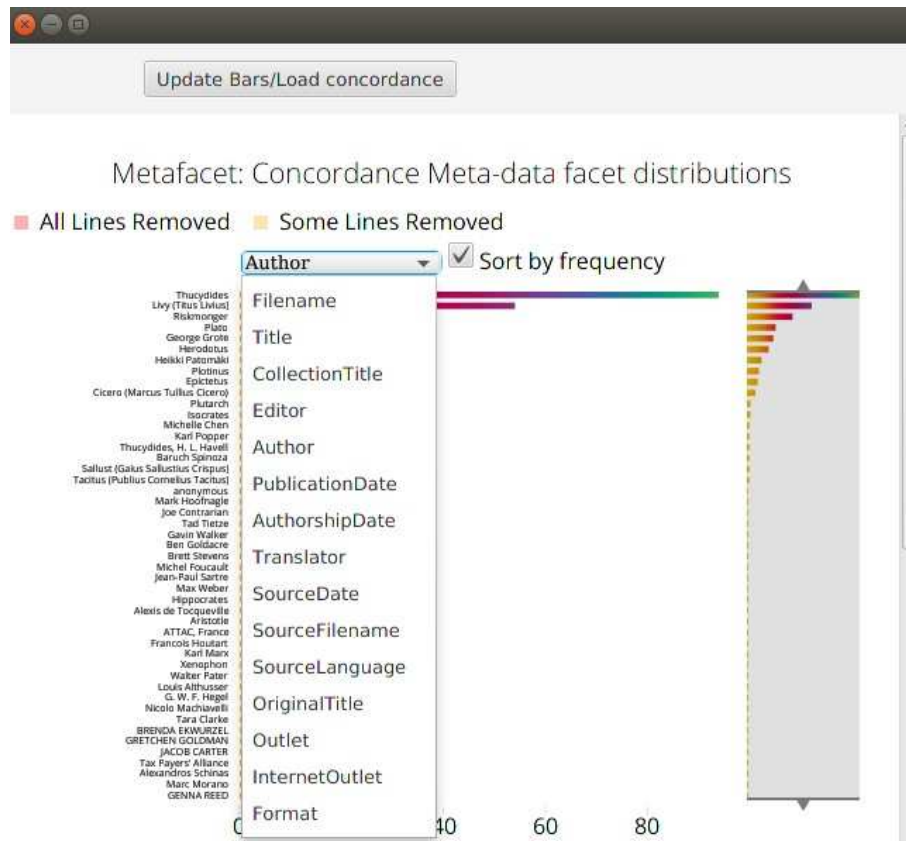


Figure 5: Metafacet interface showing all available meta-data facets. Fully zoomed out but obscured view of all authors in the concordance of “hazard”.

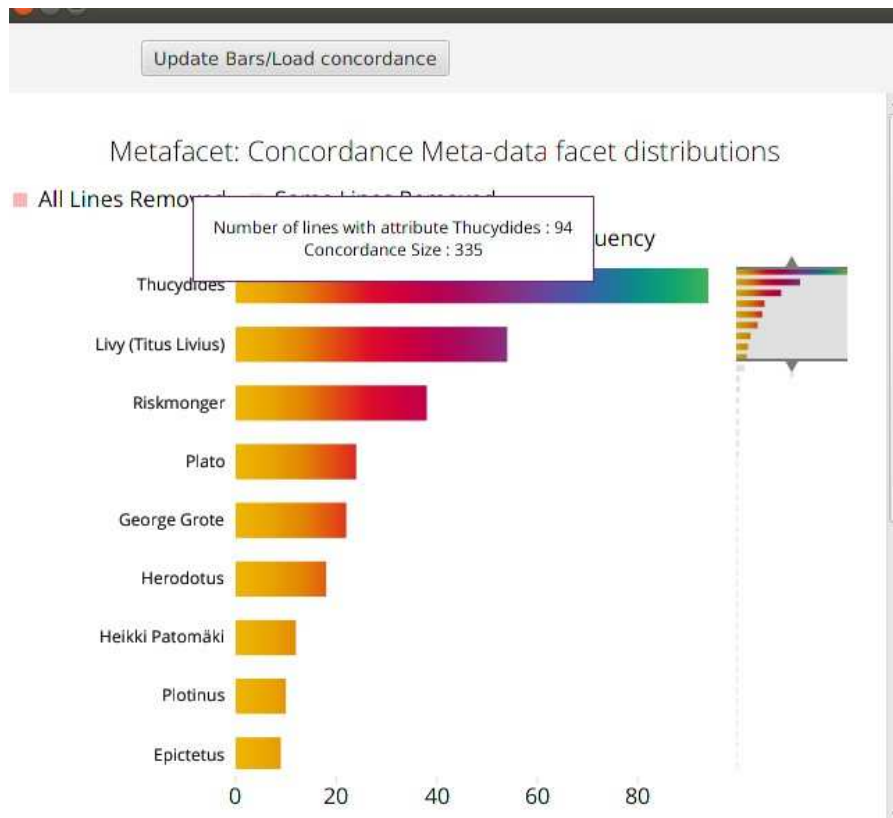


Figure 6: Metafacet zoomed to most frequent authors. Hover interaction displaying attribute name, associated concordance lines and total concordance lines for “hazard”.



Figure 7: Left click interaction filtering out any lines from the concordance associated with the attribute Format= “book”. Both Concordance Mosaic and List respect the click interaction.

Left clicking a bar removes an attribute from the concordance list, right clicking removes everything but the clicked attribute. Once an attribute or multiple attributes have been selected it is possible to switch to another facet to explore further. In Figure 8 the facet “author” is displayed after books have been removed. We can see from the red bars that the most frequent author was only found in books. The second and fourth most frequent authors are coloured yellow, this indicates that some of the lines associated with these authors have been removed but others have not. To view how much these yellow bars have been reduced the “Update Bars” button must be pressed to generate a new Metafacet for the filtered concordance. It is possible on this author facet window to add attributes back into the list by clicking on the red or yellow bars, and this would generate a filtered list where all books except those of the selected authors have been removed.



Figure 8: Viewing frequent authors after filtering out attribute Format = “book”. Partially removed attributes coloured yellow, fully removed attributes coloured red.

The combination of facets and attributes which can generate a single filtered list is limited only by the attribute crossover of the concordance lines. Finally the only author not colouring a block red or yellow in the nine most frequent is “Riskmonger”, who does not have any concordance lines associated with the attribute “book”. We stop the analysis here, but further exploration could be done to investigate the concordance lists and Mosaics for facets such as authorship/source dates and outlets. We would find that “Riskmonger” is a modern internet author who is responsible for the collocation patterns of “hazard” + “regulatory” and “assessment” at position K+1.

2.3 ComFre

The ComFre visualization is a corpus comparison tool where frequency lists can be compared visually in a statistically valid manner. The functionality of the tool has been detailed elsewhere (Sheehan et al., 2018), it has since been modified to operate as a plugin for Modnlp and is briefly presented here.

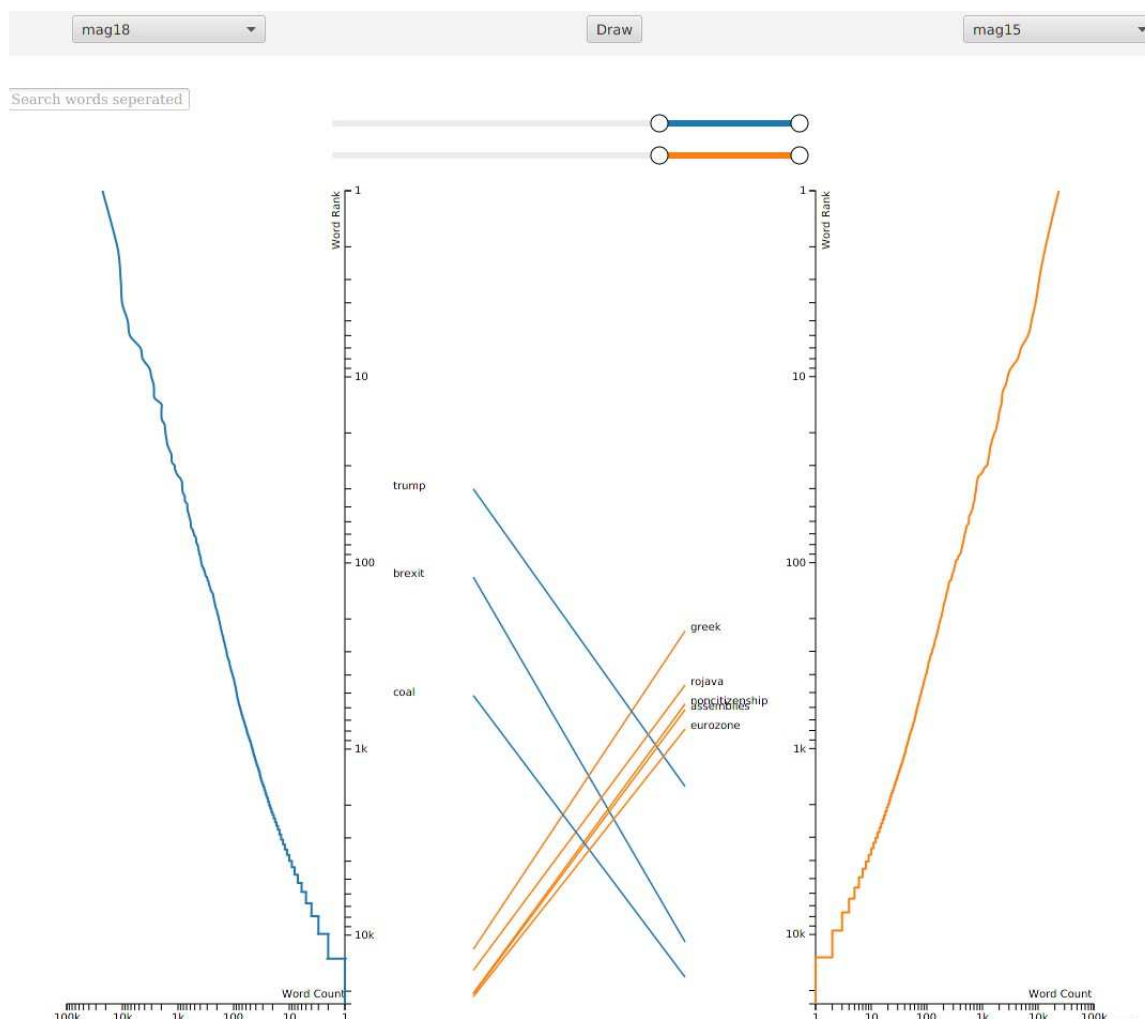


Figure 9: ComFre visualization comparing the words with the largest change in distribution rank between magazine articles from the GoK corpus authored in 2018 and 2015.

The Modnlp software has a sub-corpus selection interface which can be used to save the named sub-corpora for later reuse. ComFre makes these named sub-corpora available for comparison in dropdown lists. In Figure 9 “mag18” and “mag15” are selected for comparison, these sub-corpora are magazine articles from the GoK corpus which were authored in 2018 and 2015, respectively.

In ComFre both axis are log scaled, which should yield a linear frequency diagram if the word frequencies follow a Zipfian distribution. Scaling both ranked lists to the same height and comparing a word’s position in the distributions lets us compare sub-corpora

of vastly different size.

In Figure 9 the majority of the words have been filtered out to reveal the words with the greatest frequency changes between the two corpora. We can see that “Trump”, “Brexit” and “coal” were used much more often in the 2018 corpus, while words such as “Greek” and “eurozone” had much higher usage in 2015.

3. Domain characterization summary

This section explores the domain of corpus linguistics to identify problems and methods which will benefit from visualization. Visualizations which try to address the needs of corpus linguists are much more likely to be effective if those needs are well understood. The inclusion of domain experts in this visualization design stage is very beneficial, however just talking to users is typically not sufficient to achieve a full and accurate domain characterization. Expert users are extremely important when defining the high level goals and tasks of the domain and with ranking the importance of tasks. The characterization can be made more detailed by using methods such as examination of domain literature, contextual studies (Sedlmair et al., 2012) and needs assessments (Marai, 2018).

By performing a domain characterization, as outlined in the nested model (Munzner, 2009), the methodologies used to achieve the identified goals can be systematically investigated. The aim is to extract the low level tasks which are performed in the process of working towards the higher level goals. This analysis can be arranged as a hierarchy of goals, tasks and low level actions. The hierarchy can then be used to gain insight into the challenges faced by corpus linguists and how they have been previously addressed.

At its core domain characterization for visualization design is about identifying real problems which are relevant to the domain under investigation. This process is fluid and iterative, a level of domain understanding must be reached before work can begin on a visualization, but the design process should be reviewed as opportunities to refine the problems and domain characterization emerge.

The analysis presented here is not a full detailing of our characterization efforts. Rather, it is a presentation of some of the clearer insights and how they relate to the design choices which can be observed in the created visualization tools.

3.1 Literature-based domain analysis

Consultation and collaboration with the language scholars of the GoK project who interrogate corpora as an essential part of their analytical work lead to the natural discussion of visual tools to support analysis.

These collaborations revealed how integral the KWIC-based concordance display is to

the work of the text analyst. These visual representations provide an essential view of the context in which the keyword occurs. However, examining the relative frequencies of the words which surround the keyword is also a commonly performed task using these tools, for which it would appear these tools are not well suited. In practice, the analyst usually complements the textual information provided by the KWIC display with lists of words sorted by frequency of occurrence in the sub-corpus under examination, as well as other statistics. Different processes and sub-tasks mediate the analysis as a whole.

To study this type of concordance analysis in a practical context we turned to a reference work entitled *Reading Concordances: An Introduction* (Sinclair, 2003). This book is intended as a tutorial on how to look for certain linguistic properties of a keyword (such as word sense, phrasal usage, part of speech and many others) using a KWIC concordance list. The reader is invited to perform eighteen tasks which introduce the key practical actions and usage of linguistic knowledge required to make decisions about the properties of a word or collocation. For each of these tasks we performed a hierarchical task analysis (Annett, 2003) by combining or splitting the steps into a series of actions and sub-actions.

Each of the eighteen tasks was analysed and tagged to assist with the classifying and counting of the actions and sub-actions. Before explaining the exact meaning of the tags, an example of the tagging procedure for task 4 is given. This tagging procedure can allow a visualisation researcher with limited knowledge in the problem domain to extract meaningful actions.

Task 4 is concerned with identifying literal and metaphorical usage phrases. The preamble to the task provides some linguistic insight explaining that “some idiomatic phrases in English are recognizable because they contain a word which is not found anywhere else, like *at loggerheads*”. They may also be recognizable because the literal meaning is absurd. But others are more subtle and don’t have the aforementioned identifying marks. As an example the phrase *he got cold feet* seems to be a literal way of saying that his feet are cold. How do we as readers know when it means he is cowardly? The task studies the example of the phrase “free hand”. A concordance of 30 lines is provided and a set of twelve directions in how to analyse the concordance are given to the reader. An answer key is also provided which expands on the analysis and the insights that can be gained.

The first direction tells the reader to look at the position directly to the left of the phrases which have been sorted alphabetically “and list them in order of frequency. Can you associate any of the SINGLETONS with any of those that recur?” (Sinclair, 2003: 21) We tag this action with the *frequency* tag, *word position* tag, *group* tag and *expert decision* tag. The key gives a breakdown of the words at the position and notes that “her, your” are in the same word class as “his” and that “completely, fairly, totally” are in the same word class as “relatively”.

Step two asks the reader to

“Look again at the five lines where N—1 is an adverb of degree. What is the word at N—2? Then consider the two lines where N—1 is one. What is the word at N—2? Can you associate these seven lines with the two big groups of a and his . . . ?”

The positional notation N—2 means the set of words two positions to the left of the keyword. The same tags are applied to this action as word position, exact frequency counts and linguist knowledge are used. The answer key states

“Where N - 1 is an adverb of degree, N—2 is a; so these five lines join the group of the indefinite article. Where N—1 is the word one, in no. 25 N - 2 is her and so this line joins those with possessive adjectives. The other one, no. 24, has only at N - 2 , which is unlike all the other lines in this sample, so we will fit it in later on.”

Step three starts by explicating that in the previous step 28 of the 30 lines were extracted and divided into two groups based on “choice of determiner in front of the noun hand” the reader is then told “here the difference is not just the type of determiner; consider the meaning of free hand in the two types of line and comment on the distinction in meaning.” This task is tagged with *Similar Meaning*, *expert decision* and *read context*. For this examples the meanings of the keyword must be analysed by reading the contexts and using linguist knowledge to compare the meanings The answer key explains that when a possessive adjective is the determiner the word “free” means “available” and the word “hand” is a part of the human body. When the determiner is a the phrase “a free hand” it means “an unrestricted opportunity”.

Skipping forward to step seven the reader is narrowing in on the linguistic patterns which are used to determine literal or metaphorical usage of the phrase “free hand”. The reader is asked to group concordance lines according to whether the verb is active or passive and to examine if this accounts for the use of the word “given” exclusively before “a free hand”. Tags *group*, *read context* and *expert decision* all apply. Step 8 then combines all of the previous analysis to describe an algorithm for determining metaphorical or figurative usage of the phrase “free hand”. Many of the lines which have been discarded as not matching any patterns are not included in the construction of the algorithm.

Condition 1 of the algorithm is that there is a form of the word “give” or a word with similar meaning to the left of the phrase. If not is there an occurrence of the verb “have” or “get”, or one with a similar meaning and use?

Condition 2 is that the indefinite article precedes the core phrase, either directly or with only an adverb of degree in between.

If both conditions hold the phrase “free hand” means “to be set a task without restrictions on resources or methods to accomplish it”.

Steps nine to twelve examine all that had not previously examined in the concordance. The word frequencies and patterns to the right of the keyword are analysed and used to help account for the lines which could not be explained by the left context analysis.

This example should help clarify how the tags were assigned to the individual steps of the tasks. There was a significant amount of variation across the tasks, but the core actions could be described with a relatively small set of tags.

The actions and sub-actions generalize the descriptive analysis steps into operations which are common to many of the tasks. Taking an overview of our classifications of these actions we created the hierarchy shown in Figure 10.

At the first level of the hierarchy, the primary actions (second level) are split into quantitative and qualitative groups. Qualitative actions are classified on the criteria that a decision, in which it would be possible for experts to disagree, needs to be made to complete the action. These experts could be human users or algorithmic classification processes. Quantitative actions may form a part of a qualitative action, for example, frequent patterns must be identified before they can be classified as phrasal or non-phrasal usage (Sinclair, 1991).

The quantitative actions are those in which the steps involved in the action can be clearly stated, and, given the classifications have already been made, the results will be the same when performed by a reliable analyst. For example, for a concordance word frequencies at a specific word position can be accurately and repeatably determined. The quantitative actions often make use of the results of a qualitative action, such as estimating the frequency of words to the left of a meaning group where the group has to first be identified by expert decision.

The second level of the hierarchy contains the primary actions. These are the actions which most often describe the spirit of the instructions given in the eighteen tasks. Deeper into the hierarchy the sub-actions required to perform these primary actions are presented.

At the third level of the hierarchy the *area of analysis* is displayed, this is the level at which we perform the primary action. Looking first at the quantitative actions, we found that in three of the primary actions (filter, frequency and estimate frequency) a word’s position relative to the keyword is the area at which the actions are applied. A fourth quantitative action, frequent patterns, has an area of analysis, estimate frequency, which is one of the other primary actions. This means the action is performed on a collection of results from estimate frequency actions i.e. the analysis is performed on frequency estimations across word positions. It is worth noting that in four of the five quantitative tasks identified the word position or multiple word positions is the area at which the action is performed. The final action identified, *significant collocates*, uses

the results of statistical analysis of the keyword and its context from the corpus under investigation. This analysis is usually undertaken as a separate piece of analysis, which has its results reported as a list of frequent collocations with a keyword.

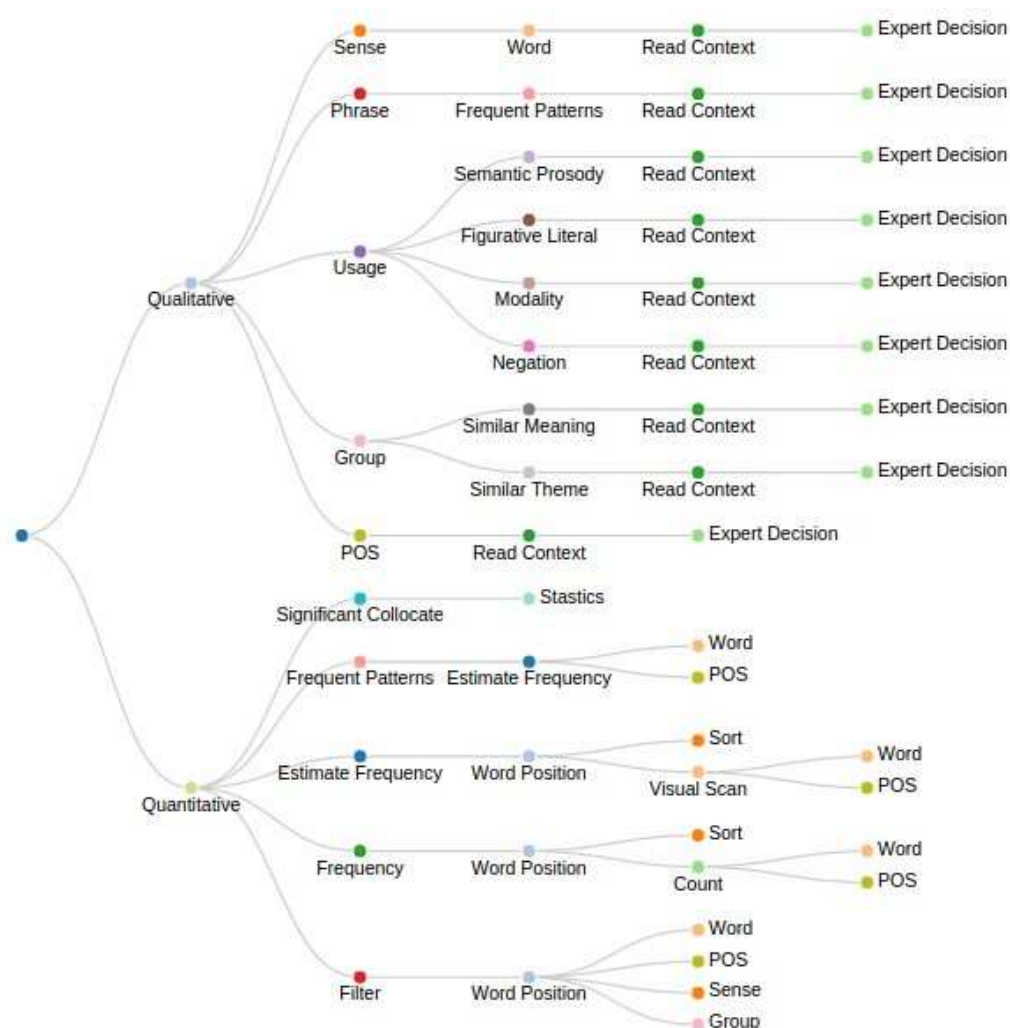


Figure 10: Hierarchical visualization of concordance-based corpus analysis actions.

Turning to the qualitative actions and, again, looking at the area of analysis at level three, we see that the analysis always occurs at the sentence level, which is implied by the read context action. This is in contrast with quantitative actions where positions are the most common area of analysis, and for qualitative actions it appears the horizontal structure of the KWIC list is emphasized while the qualitative actions make better use of the vertical alignment. Each of the actions requires an expert (or algorithm) who evaluates the context of individual occurrences of the keyword and makes a classification decision based on the semantic and syntactic content of the concordance line. This *Expert Decision* can often be the result of a combination of reading the individual contexts (the linear structure of the text) and performing some of the quantitative actions (positional statistics of the text). In essence, the *Expert Decision* action encapsulates the process of using the information extracted by other primary actions to answer questions about the keyword using linguistic knowledge.

Tag	No. of tasks in which an action appears	Total action appearances
expert decision	18	60
estimate frequency	16	34
read context	16	31
frequent patterns	15	21
frequency	14	18
word position	13	24
POS: Part of speech	11	23
filter	11	18
sense	10	19
group	7	9
significant collocate	5	7
usage	5	6
phrase	5	6

Table 1: Action counts from task analysis. Total numbers of actions found in the 18 tasks and numbers of the 18 tasks which feature the action.

While most of the tags represent actions, a few additional tags were chosen to help clarify and add information about the tasks and sub tasks. The tags *word*, *semantic prosody*, *Similar Meaning* and others are not themselves actions, but are useful in clarifying the objective or operation of the sub-actions. The part of speech (*POS*) tag is both a primary action tag and a clarifying tag. The POS primary action is to determine the part of speech of a word occurrence. The POS clarifying tag represents the use of part of speech information in another action. The purely clarifying tags are omitted from the analysis of tag frequency.

We recorded the distribution of the tags according to the number of tasks in which it appeared and the total number of actions which received the tag, as shown in Table 1. At a high level, this table tells us that both qualitative actions enabled by reading concordance lines and quantitative actions which require positional statistics are necessary for the style of concordance analysis outlined by Sinclair (2003).

3.1.1 Influence on visualization design

The structure that the task analysis and tag weightings add to the descriptive methodological steps was very useful for the early visualization design. The initial prototype of the Concordance Mosaic followed directly from this analysis. By focusing on frequent yet difficult aspects of the methodology we were able to create an interface which was likely to be of interest to corpus linguists. This gave us the opportunity to

engage with domain experts in the iterative development of tools and methodology starting with a useful prototype.

3.2 Methodological descriptions for GoK case studies

During the development of the visualization tools many interactions with GoK researchers occurred in situations such as progress meetings, design reviews and informal meetings. One set of interactions which made significant contributions to identifying relevant domain problems is presented here.

This takes the form of an initial presentation and follow up observation session with one GoK researcher. In the initial meeting a simplified methodology for a case study was described and a visualization which could be useful was suggested. The follow up observation session took place a number of months later after the Mosaic interface had been improved and made available to the researchers.

3.2.1 Methodology presentation

In the methodology discussion meeting a brief presentation outlining an example methodology and its challenges was given by a member of the GoK project to help with the initial definition of visualization goals for the project. The methodology was explained in the form of a case study. The case study made use of the portion of GoK English corpus which was available at the time. The task was defined as comparing the patterning around the keyword “*citizen**”. The * represents a regular expression search for continuations of the word citizen such as citizens and citizenship. The patterns identified were compared across two large sub-corpora.

- **Sub-corpus 1** A sub-corpus of modern English translations from Classical Greek (1850 onwards);
- **Sub-corpus 2** A sub-corpus of translated and non-translated texts written by contemporary authors, published between 1992 and the present day.

The method itself consisted of two techniques. The goal of the first technique is the identification of explicit definitions of “citizenship” contained within each sub-corpus. To find these definitions the researcher wants to compile a list of frequently used verbs and prepositions at position “keyword+1”. To achieve this the GoK corpus browser is used. Sub-corpus 1 was selected using the sub-corpus selection tool, the regular expression “citizen*” was searched and the concordance was sorted at position “keyword+1”. The researcher then spends time scrolling through the concordance and compiling a list of relevant frequent words at the position of interest, Figure 11 shows the concordance window sorted and scrolled to the preposition *as*. With this list in hand more accurate searches can be run such as:

- citizenship+“(is/as/was/defined/conceived/are/equals /considered/appears/means)”
- citizenship+“(has/should/must/will/may)”
- citizen+“(is/as)”
- citizens+“(are/as)”



Figure 11: Visualization proposed by GoK researcher

By reading the concordance lines generated by these new searches definitions can be extracted. Some examples of the definitions found are:

- Citizenship is a status bestowed on those who are full members of a community.
- As well as enjoying rights, citizens are required to undertake responsibilities such as paying taxes, and jury or military service.
- Citizenship should be based purely on residency
- US citizenship has represented a safe haven from oppressive regimes around the world

The second technique is the observation of patterns in the kinds of adjectives used to modify “citizenship”, as well as constructions such as “citizens+of+*”. The researcher explained that this technique is more difficult and time consuming using a concordance browser. To quote the researcher.

“Specifically, it is difficult to get a quick overview of such patterns using the concordancer given that the number of lines returned for my searches is quite large:

e.g. 4420 hits for “citizen*” in my sub-corpus of translations from Classical Greek.”

The researcher had some experience with linguistic visualization having used early versions of Mosaic and in the past had used word clouds, such as Wordle (Viegas et al., 2009), to present research results. There are some challenges to overcome to use word clouds for the methodology. The first which the researcher noticed is that stop-words dominate the frequency distributions of the word positions, so some technique has to be applied to get meaningful results. The suggested technique was to use a stop-word list to filter the visualization. The concordance would need to be processed to extract the words at particular positions for visualization, since the concordance is structured as a list of aligned text extracts. The result of the researchers reasoning was an interface for displaying positional word clouds with the option to exclude stop-words. The presentation included a mock-up of what a visualization to solve this problem would look like, as shown in Figure 12. The mock-up displays a word cloud for either a full concordance or a chosen word position, and has the option to remove stop-words. Looking at the mock-up in Figure 12 the words modifying citizen are presented in a manner that emphasizes frequency and provides an overview on a single screen of a position relative to the keyword.

At the end of the presentation the idea and its feasibility were discussed and some questions were asked to clarify the methodology. The notes taken were later discussed with the researcher and the following questions and answers were prepared.

- What is the domain in which the case study is situated?

“Translation and Reception studies. How have we received classic Greek texts? How has translation shaped this reception? The role of translation is often overlooked.”

- Is this methodology (excluding the proposed visualization) typical of the field?

“Translation Studies as a discipline tends to encourage close qualitative analysis of a small selection of examples chosen from specific texts to illustrate a particular argument.

Corpus analysis enables the translation scholar to identify and investigate with significantly greater ease differences between and patterns within translations, taking into account the full length of each work as a complete text.

Corpus analysis has been extensively used in translation studies before (e.g. within the TEC project and many others) but the field has tended to focus mainly on more micro-level linguistic concerns, rather than the socio-political implications of translators’ word-choices etc.”



Figure 12: Visualization proposed by GoK researcher.

- How did the idea for this example arise?

“GoK seeks to understand the constellation of concepts related to the body politic across time and space. Citizenship is a lexical item in that constellation. Comparing meaning, frequency and usage of related terms is an exploratory process used to discover obvious patterns.”

3.2.2 Methodology presentation: Design influence

The presentation helped confirm that the tasks and actions identified in the task analysis were relevant to at least one linguistics researcher. The early design of Mosaic did not take into account the need for removal of stop-words to make the Mosaic more usable. Th researcher identified this flaw but did not notice the equivalence between a mosaic column and a word cloud. By removing the stop stop-words from the Mosaic you present the same information as a positional word cloud with a greater visual emphasis on word position and frequency. This was a very beneficial meeting, and led to the addition of this “No Stop-word” view of Concordance Mosaic.

3.2.3 Methodology observation: Case study of “the people”

After a significant amount of time follow-up observation sessions were organized to gain further insight into the methodologies of the researcher who gave the presentation. This took place after the development and release of the mature Concordance Mosaic, but prior to the development of Metafacet.

Prior to the observation session a spreadsheet was created with the headings filenames, date, translator, people, citizens, commons, Athenians, public. The meta-data information related to filename, date and translator were added to the table. The remaining headings are keywords which will be investigated as part of this study. The spreadsheet used in the study can be seen in Figure 13. Partitioning the frequencies by date, file or translator is equivalent for this sub-corpus, as each file has a unique author and date.

	A	B	C	D	E	F	G	H
1	Filename	Date	Translator	people	Citizen	commc	Atheni	public
2	mod000023.xml	1629	Hobbes	167				
3	mod000098.xml	1848	Dale	158				
4	mod000148.xml	1873	Wilkins	27				
5	mod000020.xml	1874	Crawley	145				
6	mod000019.xml	1881	Jowett	185				
7	mod000214.xml	1910	Havell	29				
8	mod000016.xml	1919	Smith	182				
9	mod000048.xml	1998	Lattimore	211				
10			Total	1112	551	151	8310	405

Figure 13: The spreadsheet which was used in the study of “the people” in translations of “Thucydides” from the GoK corpus.

The first steps of the study focused on the keyword frequencies in the entire sub-corpus.

- The sub-corpus of “Thucydides” was selected.
- The keyword “people” was searched and the total frequency in the corpus was recorded
- Regulator expressions for the other “citizens?”, “commons?”, “Athenians” and “public” were searched and the total frequency in the sub-corpus was recorded.

The researcher commented, after the keyword frequencies had been recorded, that the keyword “Athenians” is much more frequent than other keywords. This is unexpected and will need to be investigated.

The next step was to gather the keyword frequencies for individual files.

- Make a sub-corpus selection for each individual file. Record in the spreadsheet the number of lines returned for the keyword “people”.

The analysis now turns from keyword frequency to the identification of collocation patterns. Mosaic was used extensively to identify collocation patterns and frequency of occurrence. The steps observed were:

- Make a sub-corpus selection for the first file.
- Perform a search for the first keyword “people” in the concordance browser.
- Open the Mosaic visualization and remove stop-words.
- Examine word frequencies.
- Open a document for taking notes and record in it the most frequent collocations directly to the left of the keyword. The words “common and “Athenian” were recorded.
- Return to the sorted concordance list and check if any continuations (such as “Athenians”) are present.
- Record the counts for the frequent collocated words. (common 8, Athenian 6).
- Open the frequency mosaic with stop-words included.
- Record in notes “lots of hits for the+people (i.e. unmodified)”
- Similar analysis for second file.
- Frequent collocates directly to left of “people” (common 34, Athenian 5).
- Record “A few more different adjectives modifying this noun:entire, experienced, free, dynamic, adventurous.”
- Similarly for the third file the noted collocates were (Athenian 13, whole 13, common 5).

The recording was ended and the researcher explained how the analysis would progress. The collocation pattern method is repeated and would continue in the same manner for each file and keyword. The next stage of the analysis would be to analyse the frequency patterns using the table. Possibly making bar charts in a spreadsheet application. Temporal patterns are expected. Identified patterns will be investigated using qualitative analysis, which involves reading the concordance lines related to the identified patterns. Understanding the meaning of the concept of “the people” at

different times is the goal.

This analysis is performed in the context of the knowledge the researcher has about the corpus and texts. She states that it is interesting that there are

“No translations 1919-1998, during period of huge cultural change in Britain. Possible reasons for this include Suffrage, war or technological revolution. The researcher explained that information about the authors and texts will influence the analysis. Some examples of information which is relevant are “the political leanings of the translators which is established relevant knowledge” and “certain texts are partial translations, abridged versions etc.”

Any differences identified, temporal or otherwise, must take into account translator style, politics and more.

Some questions were asked the researcher to elicit more information about the methodology

- How did you come up with this methodology?

“Playing around with the corpus tools, generating concordances for interesting keywords, trying to find patterns in the data.”

- How did you choose the keywords?

“Obvious keywords associated with the concept of “the people”. The idea for the study emerged through reading the literature on citizenship.”

- Would this methodology be useful for other researchers in the field?

“Other scholars using the GoK software to investigate the role of translation in the evolution of political and scientific discourse use similar methods. Other projects developing other corpora may also adopt some aspects of the methodology.”

- What are barriers to the adoption of your methodology?

“Not sure. Perhaps better documentation of the corpus software, detailing what it can and can’t do, with lots of example analysis. The publication of case-studies by members of the team will also help demonstrate the potential of the tools.”

- Mosaic was used in this analysis, is this typical when you investigate collocation patterns?

“Yes. Mosaic will be very useful for this case-study and any investigation of collocations, because it tells you in very quick and transparent way which

are the most common collocates in each word position for a given keyword.”

- You did not make use of collocation strength in your analysis, do you intend to?

“No. The collocation strength Mosaic is not immediately clear, and so (to be brutally honest) would tend to slow down analysis rather than speed it up.”

- Have you used this methodology for other studies?

“The collocation pattern aspect of this study is unique in my work. I have in previous studies studied keyword frequency in larger sub-corpora where there are multiple files for each author and date. I can show you an example for the concept of “Statesman”.”

3.2.4 Methodology observation: Case Study of “Statesmanship”

An unpublished paper on a case study of the concept of “Statesmanship” was supplied by the researcher and the major conclusions and analysis were described.

In the GoK corpus the term “statesman” was found to exist “almost exclusively (90%) in translations from Classical Greek”. This pattern was not observed for other similar keywords such as “governor”, “leader”, “ruler” and “citizen”, which are more evenly distributed across all language pairs. The analysis which arrived at this conclusion was a simple keyword frequency comparison across the translation facets of the corpus. This involved selecting each sub-corpus individually and recording the number of concordance lines for the keywords in each sub-corpus.

The frequency of the keyword “statesman” in the sub-corpus of Classical Greek translations was analysed. A spreadsheet with an entry for each of the 261 files in the sub-corpus was created and meta-data (the author, the title, the translator and the date) was entered for each file. This was done manually and was time consuming. The researcher explained that in this form “the information could easily be (re)sorted according to each of these meta-data facets and patterns more easily identified”. The number of concordance lines for each file was found by selecting a sub-corpus of a single file and searching for “statesman”. Performing this action for each of the 261 files was also time consuming. A sample of the completed spreadsheet can be seen in Figure 14.

By examining the spreadsheet and generating bar charts, such as Figure 15, the faceted distributions of “Statesman” can be understood. “statesman” seemed to be “bursty”, to use the author’s term, and to exhibit a temporal pattern.

The frequency of “statesman” in these corpora suggest most recent translations (1950-2012) of ancient Greek texts use “statesman” much less frequently. This is surprising because the corpus contains several recent re-translations

(published within the last seventy years) of classical texts such as Aristotle's Politics or Plato's Dialogues which in earlier English-language interpretations included the keyword "statesman" very prominently.

Filename	Author	Title	Translator	Date	Hits for statesm*
mod000023	Thucydides	History of the Peloponnesian War	Thomas Hobbes	1843	1
mod000149	Herodotus	Histories	Henry Cary	1847	0
mod000098	Thucydides	The history of the Peloponnesian war by Thucyd	Henry Dale	1848	0
mod000179	Plato	Apology	Henry Cary	1848	0
mod000180	Plato	Crito	Henry Cary	1848	0
mod000181	Plato	Gorgias	Henry Cary	1848	0
mod000182	Plato	Phaedo	Henry Cary	1848	0
mod000026	Hippocrates	Oath	Francis Adams	1849	0
mod000027	Hippocrates	Airs, Waters, Places	Francis Adams	1849	0
mod000035	Hippocrates	Law	Francis Adams	1849	0
mod000186	Plato	Republic	Henry Davis	1849	0
mod000178	Plato	Statesman	Georges Burges	1850	79
mod000212	George Grote	History of Greece Vol. 7		1851	2
mod000213	George Grote	History of Greece Vol. 8		1851	17
mod000211	George Grote	History of Greece Vol. 6		1851	19
mod000152	Plato	Republic	John Llewelyn Davies	1852	6
mod000177	Plato	Laws	Georges Burges	1852	17
mod000150	Thucydides	THE HISTORY OF THE PLAGUE OF ATHENS; Translat	Charles Collier	1857	1
mod000147	Herodotus	Histories	George Rawlinson	1858	0
mod000188	Plato	Gorgias	E. M. Cope	1864	32
mod000163	Plato	Apology	Benjamin Jowett	1871	0
mod000164	Plato	Crito	Benjamin Jowett	1871	0
mod000165	Plato	Phaedo	Benjamin Jowett	1871	0
mod000172	Plato	Theaetetus	Benjamin Jowett	1871	1
mod000169	Plato	Meno	Benjamin Jowett	1871	12
mod000170	Plato	Sophist	Benjamin Jowett	1871	19
mod000153	Plato	Republic	Benjamin Jowett	1871	33
mod000168	Plato	Laws	Benjamin Jowett	1871	39
mod000167	Plato	Gorgias	Benjamin Jowett	1871	41
mod000171	Plato	Statesman	Benjamin Jowett	1871	100
mod000148	Thucydides	Speeches from Thucydides	Henry Musgrave Wilkins	1873	8
mod000020	Thucydides	The History of the Peloponnesian War	Richard Crawley	1874	2
mod000252	G. W. F. Hegel	Hegel's Logic (Part One of Hegel's Encyclopaedia	William Wallace	1874	2

Figure 14: A sample from the spreadsheet used in the study of "statesman" in translations of Classical Greek from the GoK corpus. The full spreadsheet contains 261 lines of analysis.

Some clarifying questions were asked and answered:

- You mentioned the process of completing the spreadsheet was time consuming, how long did it take?

"Probably around 5-6 hours because of the amount of manual processing required. It would take a lot longer if I were to investigate more than one keyword."

- Where did the idea for this study and methodology come from?

"This was exploratory. I was not trying to establish anything in particular, only to understand whether the term "statesman" was used, how frequently (in comparison with other semantically related terms), and if any obvious patterns could be found from these initial quantitative analyses.

The terms “statesman” and “citizenship”, which I have investigated previously, are very closely related concepts, especially in classical Greek thought.”

- Were the visualization tools used in this case study?

“My focus on the use of a single keyword (“statesman”) and alternative word choices did not require and collocation pattern analysis. This is more typical of translation studies research. The corpus tools lend themselves particularly well to the analysis of collocations (this is one of their clear advantages), and this is why I want to push my research in this direction with my next case study.”

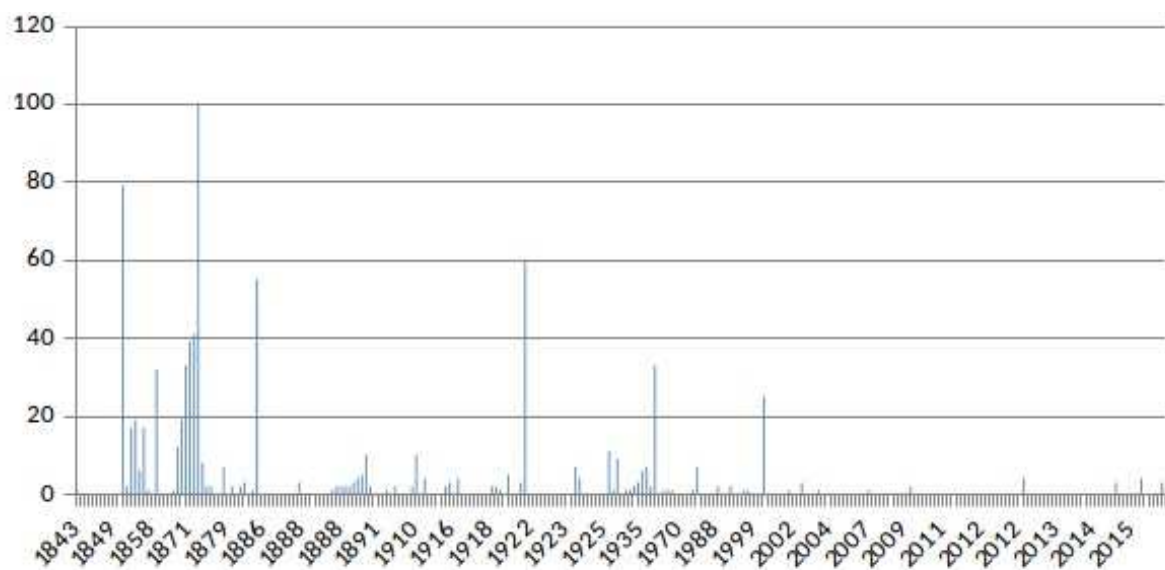


Figure 15: Bar chart examining temporal spread in translations of ancient Greek.

- Are there any areas of your methodology where you current or new visualization tools could be beneficial?

“Constructing the spreadsheets is time consuming. A tool which can help identify patterns in the dispersion of a concept according to different meta-data facets would be extremely helpful, at least for the kinds of research I intend to carry out as part of this project.”

3.2.5 Case study observation: Influence on visualization design

The most significant outcome of the two case studies was the emergence of the obvious need for a method to support the analysis of concordance lists through the lens of metadata. This observation session led to further discussion and needs assessment for a meta-data analysis tool which eventually became Metafacet.

Another problem identified was that in the version of Mosaic available to the researchers at that time only a single collocation statistic was available, and it was based on Mutual Information. The researcher did not know exactly what the scaling scheme for the collocation strength of Mosaic View was, and so could not accurately interpret or use it for publication. This led to the creation of optional scaling schemes based on well-known collocation metrics. More collocation measures are still being added to the tool.

4. Discussion and conclusions

We have presented three visualization techniques for corpus analysis. We hope that they can be adopted where appropriate by lexicographers and the wider corpus linguistic community. In addition, discussion of the tools and techniques by the community is welcomed.

We would be glad to hear any ideas, comments or criticisms of our ideas, understanding and designs. We believe the problems the tools address are general enough to have wide applicability in corpus linguistics, but we do not doubt that specific domains, such as lexicography, will have nuanced requirements that may need specialized interactions or entire redesigns to make them useful enough to be widely adopted.

The domain characterization detailed here can be another point of discussion, perhaps leading to more specialized future work on specific domain problems. We believe it is extremely important to provide a rationale for design decisions and to engage with domain experts when designing or modifying a tool or technique. Future work in this area will take the form of modifications which are identified during further domain exploration, and new visualization techniques where entire new problem areas are uncovered.

5. Acknowledgements

This paper is supported by European Union's Horizon 2020 research and innovation programme under grant agreement No 825153, project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media). The results of this paper reflect only the author's view and the Commission is not responsible for any use that may be made of the information it contains. Pierre Albert has been funded through the INCA project. We thank the INCA project members in Ireland for granting us access to the trainee data.

6. References

- Annett, J. (2003). Hierarchical task analysis. *Handbook of cognitive task design*, 2, pp. 17–35.
- Bonelli, E. T. (2010). Theoretical overview of the evolution of corpus linguistics. *The*

- Routledge handbook of corpus linguistics*, p. 14.
- Doherty, G., Karamanis, N. & Luz, S. (2012). Collaboration in Translation: The Impact of Increased Reach on Cross-organisational Work. *Computer Supported Cooperative Work (CSCW)*, 21(6), pp. 525–554.
- Karamanis, N., Luz, S. & Doherty, G. (2011). Translation practice in the workplace: contextual analysis and implications for machine translation. *Machine Translation*, 25(1), pp. 35–52. URL <http://dx.doi.org/10.1007/s10590-011-9093-x>.
- Kilgariff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, 1(1), pp. 7–36. URL <http://dx.doi.org/10.1007/s40607-014-0009-9>.
- Luhn, H. & Division, I.B.M.C.A.S.D. (1959). *Keyword-in-context Index for Technical Literature (KWIC Index)*. ASDD Report. International Business Machines Corporation, Advanced Systems Division. URL <http://books.google.ie/books?id=Dk7pAAAAMAAJ>.
- Luz, S. (2000). A Software Toolkit for Sharing and Accessing Corpora Over the Internet. In M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperidis & G. Stainhauer (eds.) *Proceedings of the Second International Conference on Language Resources and Evaluation: LREC-2000*. pp. 1749–1754.
- Luz, S. (2011). Web-based corpus software. In A. Kruger, K. Wallmach & J. Munday (eds.) *Corpus-based Translation Studies – Research and Applications*, chapter 5. Continuum, pp. 124–149.
- Luz, S. & Masoodian, M. (2007). Visualisation of Parallel Data Streams with Temporal Mosaics. In E. Banissi et al. (eds.) *Procs. of the 11th International Conference on Information Visualisation*. Zurich: IEEE Computer Society, pp. 197–202.
- Luz, S. & Sheehan, S. (2014). A Graph Based Abstraction of Textual Concordances and Two Renderings for their Interactive Visualisation. In *Proceedings of the International Working Conference on Advanced Visual Interfaces, AVI '14*. New York, NY, USA: ACM, pp. 293–296.
- Manning, C. D. & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts: The MIT Press.
- Marai, G. E. (2018). Activity-Centered Domain Characterization for Problem-Driven Scientific Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 24(1), pp. 913–922.
- Munzner, T. (2009). A Nested Model for Visualization Design and Validation. *IEEE Transactions on Visualization and Computer Graphics*, 15(6), pp. 921–928.
- Rockwell, G. (2003). What is Text Analysis, Really? *Literary and Linguistic Computing*, 18(2), pp. 209–219.
- Scott, M. (2010). What can corpus software do. *The Routledge handbook of corpus linguistics*, pp. 136–151.
- Sedlmair, M., Meyer, M. & Munzner, T. (2012). Design Study Methodology: Reflections from the Trenches and the Stacks. *IEEE Transactions on Visualization and Computer Graphics*, 18(12), pp. 2431–2440.

- Sheehan, S., Masoodian, M. & Luz, S. (2018). COMFRE: A Visualization for Comparing Word Frequencies in Linguistic Tasks. In *Proceedings of the 2018 International Conference on Advanced Visual Interfaces, AVI '18*. New York, NY, USA: ACM, pp. 42:1–42:5. URL <http://doi.acm.org/10.1145/3206505.3206547>.
- Shneiderman, B. (1996). The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings the IEEE Symposium on Visual Languages*. pp. 336–343.
- Sinclair, J. (1991). *Corpus, concordance, collocation*. Describing English language. Oxford University Press.
- Sinclair, J. (2003). *Reading Concordances: An Introduction*. Longman Publishing Group. URL <http://books.google.ie/books?id=Ms9nQgAACAAJ>.
- Summers, D. (1996). Corpus lexicography—the importance of representativeness in relation to frequency. *Longman Language Review*, 3, pp. 6–9.
- Viegas, F. B., Wattenberg, M. & Feinberg, J. (2009). Participatory Visualization with Wordle. *IEEE Transactions on Visualization and Computer Graphics*, 15(6), pp. 1137–1144.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



Validating the OntoLex-*lemon* Lexicography Module with K Dictionaries' Multilingual Data

Julia Bosque-Gil^{1,3}, Dorielle Lonke², Jorge Gracia³,

Ilan Kernerman²

¹ Ontology Engineering Group, Universidad Politécnica de Madrid, Spain

² K Dictionaries, Tel Aviv, Israel

³ Aragon Institute of Engineering Research (I3A), University of Zaragoza, Spain

E-mail: jbosque@funizar.es, dorielle@kdictionaries.com, jogracia@unizar.es,

ilan@kdictionaries.com

Abstract

The OntoLex-*lemon* model has gradually acquired the status of *de-facto* standard for the representation of lexical information according to the principles of Linked Data (LD). Exposing the content of lexicographic resources as LD brings both benefits for their easier sharing, discovery, reusability and enrichment at a Web scale, as well as for their internal linking and better reuse of their components. However, with *lemon* being originally devised for the lexicalization of ontologies, a 1:1 mapping between its elements and those of a lexicographic resource is not always attainable. In this paper we report our experience of validating the new *lexicog* module of OntoLex-*lemon*, which aims at paving the way to bridge those gaps. To that end, we have applied the module to represent lexicographic data coming from the Global multilingual series of K Dictionaries (KD) as a real use case scenario of this module. Attention is drawn to the structures and annotations that lead to modelling challenges, the ways the *lexicog* module tackles them, and where this modelling phase stands as regards the conversion process and design decisions for KD's Global series.

Keywords: Linguistic Linked Data; RDF; multilingual; OntoLex-*lemon*; K Dictionaries

1. Introduction

Linked data (LD) technologies are increasingly adopted in lexicography, whether in academic research and development, the industry, or combining both (see for instance Klimek and Brümmer (2015), Declerck et al. (2015), Abromeit et al. (2016), Parvizi et al. (2016), Bosque-Gil et al. (2016a) and Kaltenböck & Kernerman (2017)). LD refers to a set of best practices for exposing, sharing and connecting data on the Web (Bizer et al., 2009). The adoption of LD in lexicography enhances the tendency to standardize the ways of representation and query of lexical content at a Web scale. Connections can also be established to other LD resources, so that lexicographic data can be enriched with different types of complementary information, such as additional translations, definitions, examples of usage, etc.

The *de-facto* standard for representing ontology lexica, the *lemon* model (McCrae et al., 2012) and its more recent version, *OntoLex-lemon*¹ (McCrae et al., 2017), have been the preferred choice by developers to convert lexicographic resources into LD. Early experiences in using *lemon* show that the model is highly effective as regards the accounting for the core lexical information in lexicographic resources (Klimek & Brümmer, 2015; Declerck et al., 2015; Abromeit et al., 2016; McCrae et al., 2019). However, there are various situations in which no perfect match is available between the elements of the model and those found in lexicographic entries, or in which the model falls short of capturing certain peculiarities of lexicographic works, e.g. the order of senses in an entry, details on the morphological features of word-forms when used for a specific sense, etc. In this context, the W3C OntoLex community group² has analysed the main issues regarding the representation of lexicographic information as LD and is releasing this year an updated module to represent lexicographic data that extends the *lemon* core model – the *lexicog* module.³

In this paper we analyse the application of the *lexicog* module for LD-based representation of the Global series of K Dictionaries (KD).⁴ The main contribution of this work is twofold:

1. This pioneering experience serves to validate this new module with an actual use case as well as to introduce some recommendations for future applications.
2. By focusing on KD's data, we examine how the limitations of the *OntoLex-lemon* model already reported in the literature (Klimek & Brümmer, 2015; Bosque-Gil et al., 2016b) are successfully addressed by the module.

The rest of this paper is structured as follows: Section 2 provides an overview of KD's Global series and elaborates on the motivation for its conversion to LD, as well as a summary of previous conversions of these data to LD and the challenges encountered in this process. In Section 3 the *lexicog* module is introduced. Section 4 briefly presents the different stages of LD generation, and where the modelling with *lexicog* stands with respect to the whole conversion of KD's data to the Resource Description Framework (RDF), along with the technologies and the design decisions we adopted. Section 5 addresses some of the limitations previously detected in the literature on the conversion of KD's data and provides a modelling solution in terms of *lexicog*. Concluding remarks and future lines of work are presented in Section 6.

¹ <https://www.w3.org/2016/05/ontolex/>

² <https://www.w3.org/community/ontolex/>

³ The *lexicog* module and report are available at <http://www.w3.org/ns/lemon/lexicog#> and <http://www.w3.org/2019/09/lexicog/> respectively.

⁴ <http://www.lexicala.com/>.

2. K Dictionaries' Global series

In this section we briefly describe the dictionary data that we used to validate the *lexicog* module, which stems from the Global series of KD. This series is based on the monolingual lexicographic cores of 25 different languages and their bilingual and multilingual versions, and includes nearly 100 language pairs and numerous multilingual variations. We discuss the motivation of converting it into LD and describe preliminary conversions that were performed in the past.

2.1 Converting KD's data to the RDF: motivation and overview

The Global series of KD (Kernerman, 2009, 2011, 2015) has been conceived as a cross-lingual, multi-layer mosaic of lexicographic resources that evolve within a single systemic framework, sharing a common technical macrostructure and a common entry microstructure that is able to accommodate and adapt to particular characteristics of different languages. All the language sets share the same XML schema (DTD), wherein certain languages can feature additional orthographic scripts (e.g., have Kanji, Hiragana, Katakana and Romaji for Japanese, or encompass diverse inflected verb forms, for example, perfective/imperfective for Polish and Russian). Each language resource is created on its own, based on deep corpus analysis from which stem its editorial style guide, headword list, lexical deciphering and mapping, and diverse semantic and syntactic attributes. The result is a detailed monolingual core that might contain overlapping elements, such as definitions alongside sense disambiguation elements, synonyms or antonyms, etc., which can then be used selectively to customize that data to the needs of particular target audiences and usages. This core is ready to be complemented by translation equivalents (of the senses, examples of usage, and multiword units) for developing bilingual versions, which are juxtaposed and form a multilingual network revolving around the initial monolingual set. Eventually, the translations (and other components) of each language network can also be interlinked to each other and exponentially multiply the cross-lingual connections.

Since its inception in 2005, 25 language cores were created, and altogether nearly 100 language pairs are available so far, besides numerous multilingual combinations. Rather than aim to compile any specific dictionary product, the idea was to develop multifunctional data sets that can be applied in different forms and media, either independently or in conjunction with other data, whether intended to publish a print dictionary, develop an online or a mobile dictionary, offer lexical services, or be incorporated in NLP applications. The advent in recent years of Linguistic LD and Semantic Web technologies has opened new horizons to enhance this strategic approach of creating well-structured, detailed and extensive lexicographic data rather than single dictionary products, by reinforcing and further expanding existing data, and improving interoperability between content from the Global series and other multilingual data on the Web, attaining reciprocal enrichment of the Global series by external resources (on

the one hand), and enhanced incorporation of data from the Global resources into external ones (on the other hand). To put this notion into practice it became necessary to first transform KD's Global data from its original XML (hierarchical) format to an RDF structure (knowledge graph), for smoother linking to external resources. Thus, KD decided to apply the best-known LD standard model for representing lexicographic content, first in the form of *lemon*, then conforming to *OntoLex-lemon*, and most recently in line with its up-to-date *lexicog* module.

The motivation of KD to focus and invest in this venture can thus be explained by the invaluable upgrade this should carry for its resources, through facilitating their interoperability and enhancing depth, precision, and cross-linguality. Such improved features are needed to deal with the emerging multilingual single digital market, primarily in Europe and eventually all over the world, which calls for multiple adaptations of content and technology, international standards, multi-disciplinarity, etc. LD methods are at the forefront of the current generation of powerful language technology solutions, at the heart of human-machine interaction. Providing quality cross-lingual lexical data, with the LD-driven option of linking to other sources, greatly increases the appeal and uniqueness of the KD resources and places KD in a leveraged position to other competing dictionary APIs.

The new API of KD, renamed Lexicala API, provides access to the Global (and other KD) data in JSON, with the first touches of JSON-LD. It constitutes a vital step in an innovative trend of turning passive dictionary products into active lexical data services that interoperate with real-world computational linguistics applications. Two ongoing H2020 projects employ Lexicala API as part of their solutions: Lynx⁵ will integrate KD (as well as terminological and other) resources with data from the legal domain in the heart of its Legal Knowledge Graph platform for multilingual compliance services; and Elexis⁶ will make use of the API to receive KD content for its future European lexicographic infrastructure. Making KD resources available in state-of-the-art RDF conforming to world-class standards will both help to enhance the operation of Lynx and Elexis platforms, and those of a multitude of future applications, and to reinforce and expand KD content through interaction with more LD resources.

2.2 Previous representations of KD's data as RDF

The current conversion of KD's multilingual Global series is not the first effort to convert this data to RDF. In 2014 KD became involved in the first attempt to convert Global data from XML format to RDF, adhering to the *lemon* model and focusing on the German monolingual dataset (Klimek & Brümmer, 2015). The next massive step was taken in the two-year project carried out in 2015-2017 as part of a EUREKA

⁵ <http://lynx-project.eu/>

⁶ <https://elex.is/>

bilateral framework between KD and Semantic Web Company (SWC), called Linked Data Lexicography for High-End Language Technology Application (LDL4HELTA).⁷ As part of the LDL4HELTA project, the Global data for three languages (English, German and Spanish) was converted to RDF in line with the OntoLex-*lemon* model (Bosque-Gil et al., 2016b).

In the first work (Klimek & Brümmer, 2015), the authors identified some gaps in the *lemon* model with regard to representing KD's data, for instance, the way to link a compound phrase defined inside of a sense group to that same sense. The lack of an ontology to provide ontological references for `lemon:LexicalSenses` was also highlighted. This point is strongly connected to the original aim of the *lemon* model to serve to lexicalize ontologies, not to represent lexicographic resources in the Web of Data. In addition, the authors identified some gaps in the LexInfo⁸ catalogue of grammar categories (typically used in conjunction with *lemon*) and created their own custom vocabulary to capture the values of KD's DTD attributes. In the later conversion of the series to OntoLex-*lemon* (Bosque-Gil et al., 2016b), some problems that were identified in the previous conversion were no longer relevant, as both the model and its modules had evolved to cover more cases (e.g. now the *vartrans* module allows to represent lexical relations).

It is worth noting that, whereas in the first two attempts the conversion was carried out under the strict principle of round-tripping, i.e. aiming to obtain full and complete 1:1 data transformation from XML to RDF and from RDF back to XML – so the RDF structure had to convey each and every detail of the complex features of the original XML structure – the current work was released from this obligation. The reasons for applying such a demand in the first place were, on the one hand, to serve as validation of perfect transfer from XML to RDF while, on the other hand, to be able to benefit from the potential enrichment of the data in RDF when linking to other data resources and importing such new data back to the existing resource in XML. Removing this restriction has helped to liberate and enhance the data flow from one format to another, and emphasized the autonomous status of each model and the fact that every format should behave freely and reflect its autonomous characteristics that are different from the other.

However, OntoLex-*lemon* proved to be not exhaustive enough to cover the representation requirements of the original resource. Four kinds of challenging situations were detected in the modelling of KD's multilingual data:

1. Cases in which solely applying the OntoLex-*lemon* model would lead to a loss of structural information reflecting lexical distinctions. For example, entries *not* originally conceived as dictionary entries in KD data are treated equally as

⁷ <https://www.eurekanetwork.org/project/id/9898>

⁸ <https://lexinfo.net/ontology/2.0/lexinfo.owl>

original entries in the RDF representation (i.e. as `ontolex:LexicalEntry`). This highlighted a lack of elements for representing the components of a lexicographic entry in cases in which there is no 1:1 mapping with *OntoLex-lemon* classes and properties. In KD data, we encounter several examples of this type of situation: compounds, synonyms, antonyms, and translations. Compounds are defined inside the dictionary entry as one of its components and do not occur as lemmas (i.e. in their own dictionary entry). Synonyms and antonyms, even though they are usually independent lemmas in that same resource, are embedded in dictionary entries and they do not necessarily have their corresponding dictionary entry in that resource (but could occur as dictionary entries in another KD dictionary). In addition, a translation of a headword into another language is treated as an `ontolex:LexicalEntry` (Bosque-Gil et al., 2016b), too, but just as a synonym, and the source data in its current state does not guarantee for the word to be a lemma in the dictionary of the target language. This fact called for a distinction between an original dictionary entry and the `ontolex:LexicalEntry` newly created in the process, thus recording the outcome of the headword selection step in the compilation of the dictionary. In lexicographic resources other than KD, the same gap would surface in those cases in which a dictionary entry needs to be split into more than one `ontolex:LexicalEntry`, each with a different part of speech, in order to be *OntoLex-compliant*.

2. Cases in which *OntoLex-lemon* or LexInfo falls short of covering the representation needs that KD's dictionary entries give rise to. This concerned the representation of examples and translations of examples, which are fairly common elements in other dictionaries as well (Bosque-Gil et al., 2017).
3. Cases in which *OntoLex-lemon* does contain elements to cover a particular type of information, but there are no specific guidelines on how to use them in the process of conversion of lexicographic data to RDF (without involving ontology lexicalization). For example, the representation of lexicographic definitions with the *OntoLex* core (e.g. with `skos:Concept` or `ontolex:LexicalConcept`), the encoding of geographical usage restrictions on senses, or the modelling of selectional restrictions for predicate arguments.
4. Mismatches between LexInfo elements and KD's DTD tags and values.

Since situations of types (1) and (2) were also generalizable to other lexicographic resources, *lexicog* was proposed as an extension of *OntoLex-lemon* (Bosque-Gil et al., 2017). For cases of type (3), the *OntoLex* Community, in its bi-weekly telcos on lexicography, discussed a series of practices for the use of *OntoLex-lemon* elements in the conversion of lexicographic data to RDF.⁹ These practices emerged as solutions to

⁹ <https://www.w3.org/community/ontolex/wiki/Lexicography>

a list of issues detected in the literature. A series of guidelines, with more examples and recommendations, are also planned as future steps in the OntoLex community. Cases of type (4) were addressed in 2016 by creating a custom ontology for KD, which is currently under revision and update.

3. The *lemon* lexicography module: *lexicog*

The *lemon* model has been extensively used for representing lexicographic data. However, some limitations were detected in several preliminary experiences, as reported in Section 1.

Such issues were collected and analysed by the W3C OntoLex community group with the aim of reaching some agreement that allows for a better and more interoperable migration of existing dictionaries into LD. As a result of this community effort, the OntoLex-*lemon* lexicography module (*lexicog*) was defined as an extension of the OntoLex-*lemon* model.¹⁰ The module is targeted at the representation of dictionaries and any other linguistic resource containing lexicographic data, and addresses structures and annotations commonly found in lexicography.

The *lexicog* module overcomes some limitations of OntoLex-*lemon* when modelling lexicographic information as LD. It aims at capturing the underlying original structure and annotations of the lexicographic entry in a way that keeps the purely lexical content separate from the lexicographic one, minimizing information loss and allowing queries restricted to the lexical layer. By being able to keep record of the original dictionary arrangement as RDF, the module does not impose a certain view on the lexicon and thus becomes agnostic to the standpoint of the lexicographer. For that purpose, new ontology elements have been added that reflect the dictionary structure (e.g., sense ordering, entry hierarchies, etc.) and complement the OntoLex-*lemon* lexicon.

Figure 1 depicts the main classes and relations defined in the *lexicog* module. We refer to the specification document for more details, but we give here an overview of its main modelling ingredients:

- LexicographicResource, which represents a collection of lexicographic entries in accord with the lexicographic criteria followed in the development of that resource.
- Entry, a structural element that represents a lexicographic article or record as it is arranged in a source lexicographic resource.
- LexicographicComponent, which is a structural element aimed at representing the (sub)structures of lexicographic articles providing information about entries,

¹⁰ A record of the discussed issues and intermediate design decisions can be found at <https://www.w3.org/community/ontolex/wiki/Lexicography>.

senses or subentries. Lexicographic components can be arranged in a specific order and/or hierarchy.

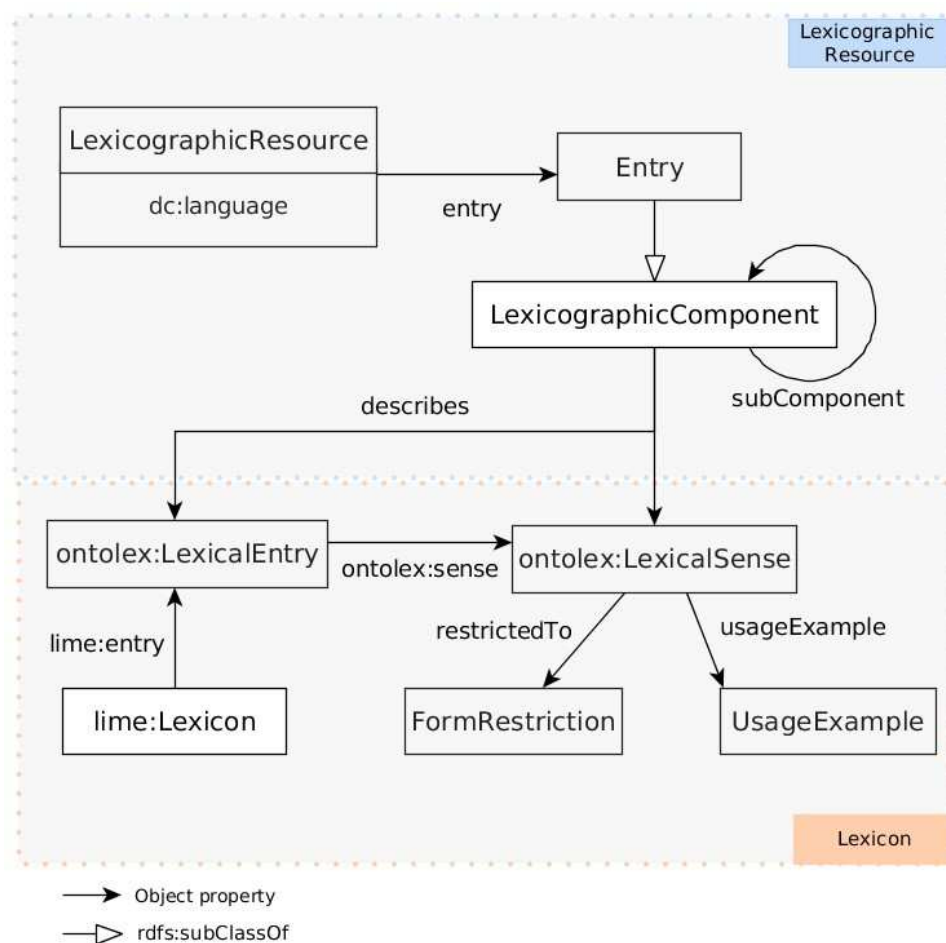


Figure 1: Scheme of the lexicography module (taken from the “OntoLex-*lemon* Lexicography Module” W3C community group final report).

The three above elements account for the basic structure of the LexicographicResource. To that end, a property entry relates a LexicographicResource to each of its entries. An Entry in turn can group several LexicographicComponents. We can indicate that the components belong to an entry by simply using the RDF native mechanisms for containers.¹¹ In particular, the `rdfs:member` property can be used if the order of the components is not relevant, and `rdfs:ContainerMembershipProperty` (`rdf:_1`, `rdf:_2`, ...) when the order of the components needs to be represented. Notice that an Entry is a particular subclass of LexicographicComponent used to represent the main “entry point” in the dictionary, i.e., the headword or the root of the lexicographic record.

The lexicographic components only reflect the structure of the dictionary and do not encode any lexical content themselves. To associate them to their corresponding lexical information (e.g. lexical entries or lexical senses), the property “describes” is used. Such

¹¹ https://www.w3.org/TR/rdf-schema/#ch_containervocab

elements belonging to a lexicon are taken from OntoLex, in particular:

- `ontolex:LexicalEntry`, which consists of a set of forms that are grammatically related and a set of base meanings that are associated with all of these forms.
- `ontolex:LexicalSense`, which represents the lexical meaning of a lexical entry when interpreted as referring to the corresponding ontology element.
- `lime:Lexicon`, or a collection of lexical entries for a particular language or domain.

These classes can be further connected with many other elements that describe the lexicon and that can be found in the OntoLex-*lemon* specification. Particularly, the `ontolex:Form` class, to account for the grammatical realization of a lexical entry (typically by means of its written representation) and the `ontolex:LexicalConcept` class, that can be used to store definitions through the property `skos:definition`.

Finally, we mention the `UsageExample` class of the *lexicog* module, which is intended to represent a textual example of the usage of a sense in a given lexicographic record.

4. Methodology

4.1 Incremental approach and steps taken

The process of converting KD data into LD was carried out with an incremental approach, starting with the very basics of a single entry (headword, senses, part of speech, definitions) and proceeding with more complicated elements (synonyms, translations, examples of use, compounds, etc.), validating the results of the conversion after each iteration. This approach allowed for constant validation and error elimination, and facilitated the technical conversion process. Prior to converting actual data, some groundwork was necessary. For this purpose, the DTD of KD's XML data was examined, and each XML path in KD data was manually defined as a corresponding OntoLex, *lexicog* or LexInfo element. Next, a URI naming strategy was established, following the previous conversion of the Global series (Bosque-Gil et al., 2016b). In addition, the DTD was revised and edited where possible, adhering to the standards set by LexInfo and OntoLex and prioritizing smooth conversion and adaptable results.

After setting the foundations for conversion, the following steps were taken for each iteration:

- Identifying a few entities in *lexicog* to test, and manually creating an example RDF entry with real KD data. Only a handful of components comprising a complete dictionary entry were selected for each iteration, to simplify each step and govern the results more easily. In order to maintain that the conversion was carried out exhaustively and accurately, logs were kept, and the URI naming

strategy was under constant revision and scrutiny.

- Writing and running a conversion script. The manually constructed example had a vital part in determining the conversion script. The RDF conversion pipeline relies on already existing conversion of XML data into JSON, adding LD elements and restructuring the JSON document to comply to the triple relations encompassed in the JSON-LD structure. In each iteration, the conversion was applied to all of the resources of the Global series, resulting in a collection of JSON-LD documents, with each dictionary entry represented by its own JSON document and reflecting an RDF graph introducing only the components that were the focus of the current iteration, on top of previously covered components.
- Validating output RDF. The final step for each iteration was validating the results.
- The method of validation selected to this end is twofold, consisting of the JSON Schema as an initial means of validation, and a SPARQL endpoint and query service for querying the RDF output.
- Repeat for the next components.

These steps allowed for constant appraisal and control. Further iterations were conducted with taking into consideration any conclusions drawn on their predecessors, and the workflow enabled simultaneous work on the theoretical conversion alongside writing the conversion script by all team members. In particular the JSON schema was very important, as this provided exhaustive validation as part of the pipeline prior to the querying phase.

4.2 Performing the validation

The validation process consists of two parts: the first, initial validation is conducted by defining a JSON schema and validating the JSON-LD documents against it as part of the conversion pipeline; the second, final validation is uploading the RDF output onto a SPARQL query service, e.g. any triple store supporting JSON-LD, and querying the data to certify that all of the input data was properly converted.

The selection of JSON schema for initial validation of the JSON-LD documents was a natural one; designed to validate JSON documents, the schema can be tailored to specific needs and ensure that the JSON document is well-structured and includes only desired elements. The same principles can be applied to JSON-LD, harnessing the advantages of JSON schema to control the triple structure and ensure that URIs are well-defined. The main points of validation offered by the schema are the following:

1. The JSON schema checks that the predicates are in place, that is, that there will not be a JSON object nested inside another JSON object where no relation

exists between them. Together with the context, which can be validated both manually and automatically, the schema basically checks that the correct triple relations occur, and that there are no relations that should not occur.

2. It checks that all necessary information is present, and that nothing was left out during conversion.
3. It checks that the JSON does not contain anything that should not be there, insofar that if something is not specified in the schema but appears in the document, it constitutes an error.
4. It checks that the URIs are well-defined by defining regular expressions according to the URI naming pattern.

By checking these four points, the schema corroborates both the triple relations and the URIs, essentially providing complete structural validation. A JSON-LD document that validates against the JSON schema is trusted to represent a correct RDF graph. Including JSON schema as part of the conversion pipeline ensures that the RDF output is valid, adding another layer of security prior to the querying phase, and establishing that the data stored on the triple store is well-structured and complete.

The chosen serialization, JSON-LD, was selected due to it being a standard and widely used format for structured data among the target sector of API users. Its native compatibility with JavaScript allows for flexibility and customization when converting proprietary data. Its inherently nested structure prevents redundancy and verbosity, and being the main format for API responses it can be easily parsed and manipulated. Furthermore, by defining clear and intuitive aliases for RDF classes, properties and predicates, it has the advantage of being human, as well as machine readable.

The JSON schema, while applicable only to the JSON-LD serialization, encompasses all of the relevant principles of RDF validation, which can be derived directly and applied to any other means of validation used for validating other serializations.

5. Applying *lexicog* to KD's multilingual data

The *lexicog* module draws a distinction between the lexical layer, captured mainly by OntoLex, and the structural elements that describe the lexicon and can be arranged as desired in a particular lexicographic work. We will adhere to this distinction in this section as well and first present problematic cases of KD of type (1) (see Section 2.2), concerning the distinction between a dictionary entry and an `ontolex:LexicalEntry` and the grouping of dictionary entries, and will follow with the representation of examples and their translations as LD.

5.1 lexicog:Entry and ontolex:LexicalEntry

One of the shortcomings of OntoLex-*lemon* concerned the lack of a way to capture what was originally a dictionary entry in the resource and differentiate it from an ontolex:LexicalEntry created on the fly during the conversion process, which may or may not have their corresponding dictionary entry in the resource (or in a work of the same series, i.e. the Global series from KD). In addition, a lime:Lexicon gathers a collection of ontolex:LexicalEntry elements, which, in turn, can share the language and come from different lexicographic resources from the same series (see Gracia et al., 2018). The lime:Lexicon class is thus not intended to uniquely represent the lexicographic resource as it was conceived originally, but as a collection of lexical entries belonging to the lexicon of a language.

In KD's data, compounds, synonyms, antonyms and translations are defined or described inside a dictionary entry of another lemma (in the case of compounds, inside the dictionary entry of one of their components). In order to represent their definition, form, inflection or pronunciation according to the OntoLex core, they need to be treated as ontolex:LexicalEntry elements, which causes the distinction between original dictionary entries and embedded lexical entries to be lost.

Example 1.1 shows an extract of the dictionary entry *arte* 'art' in Spanish, with its translation into Dutch and the definition of the compound *artes plásticas* 'visual, plastic arts'. This example, in addition to a description of the headword (shortened due to space constraints) provides a synonym in its first sense (*inspiración* 'inspiration'). Below the section devoted to translations, the compound *artes plásticas* is defined.

By applying *lexicog* to example 1.1, we instantiate different elements to represent lexical entries and dictionary entries respectively. Example 1.2 shows an extract of the RDF Turtle representation of example 1.1. The elements in blue refer to the lines in the RDF that mark this distinction. While the Spanish and Dutch lexica gather any unit of the lexicon that is described in the original dictionary (as a dictionary entry or as an embedded entry), represented as ontolex:LexicalEntry, a lexicog:LexicographicResource is intended to group only dictionary entries through lexicog:Entry. This way, the RDF reflects that *artes plásticas* is a unit of the lexicon but it is not a lemma in this dictionary.

lexicog:Entry serves a structural function to only capture the structure of the resource as a result of the lexicographic selection process, and it does not bear lexical information. To close this gap, the property lexicog:describes links dictionary entries (as structures) to the lexical units in the lexicon. If the RDF representation were also to reflect that *artes plásticas* or *inspiración* are lexical entries "defined" inside the dictionary entry of *arte*, the *lexicog* module would provide elements to establish this structural connection. In this case, however, reflecting the whole microstructure of the entry was not a requisite for the expected output; we limit ourselves to capture the semantic relations

between these different lexical entries (translation, synonymy) through the elements of the OntoLex-*lemon* model, following previous approaches (Bosque-Gil et al., 2016b) based on the *vartrans* module.

```
<DictionaryEntry identifier="DE00005536" version="1">
  <HeadwordCtn>
    <Headword>arte</Headword> [...]
  </HeadwordCtn>
  <SenseBlock>
    <SenseGrp [...]>
      <Synonym>inspiración</Synonym> [...]
      <TranslationCluster [...]>
        <Locale lang="nl">
          <TranslationBlock>
            <TranslationCtn>
              <Translation>kunst</Translation> [...]
            </TranslationCtn>
          </TranslationBlock>
        </Locale> [...]
      </TranslationBlock>
    </TranslationCluster>
    <CompositionalPhraseCtn version="1"> [...]
      <CompositionalPhrase>artes
        plásticas</CompositionalPhrase> [...]
    </CompositionalPhraseCtn> [...]
  </SenseGrp> [...]
</SenseBlock>
</DictionaryEntry>
```

Example 1.1: An extract of the dictionary entry *arte* ‘art’ in Spanish from KD’s Global series with its translation into Dutch and the compound *artes plásticas* ‘visual plastic arts’.

```
@prefix base: <http://lexicala.com/id/global/> .
@prefix lime: <http://www.w3.org/ns/lemon/lime#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix lexicog: <http://www.w3.org/ns/lemon/lexicog#> .
@prefix ontolex: <http://www.w3.org/ns/lemon/ontolex#> .
```

```
:mlds-ES3 a lexicog:LexicographicResource;
  dc:language "es" ;
  lexicog:entry :ES_DE00005536 .
```

```

:ES_DE00005536 a lexicog:Entry ;
    lexicog:describes :lexiconES/arte-n .

:lexiconES/arte-n a ontolex:LexicalEntry .
:lexiconES/artes-plásticas-n a ontolex:LexicalEntry .
:lexiconES/inspiración-n a ontolex:LexicalEntry .
:lexiconNL/kunst-n a ontolex:LexicalEntry .

:lexiconES a lime:Lexicon; lime:language "es" ; lime:entry :lexiconES/arte-
n, :lexiconES/artes-plásticas-n, :lexiconES/inspiración-n .

:lexiconNL a lime:Lexicon;
    lime:language "nl";
    lime:entry :lexiconNL/k
unst-n.

```

Example 1.2: RDF Turtle representation of example 1.1

5.2 Nested entries

There are other types of information in KD's Global series that require the RDF version of the dictionary to reflect structural aspects. In KD's DTD, the element `NestEntry` works as a container grouping together several dictionary entries. Example 1.3 in XML shows the entry of the verb *besuchen* 'to visit' in German. The element `NestEntry` groups together three different dictionary entries: *besuchen* (v. 'visit'), *Besuch* (n. 'visit') and *Besucher* (n. 'guest, visitor') that are related, although the nature of relation is not *explicitly* stated. These containers group together derivations or, in some cases, verbs that share a lemma but not the subcategorization value and are not homonyms.

Example 1.4 shows the RDF rendering of example 1.3 in Turtle serialization. In *lexicog*, grouping is reflected by creating a `lexicog:LexicographicComponent` and indicating that other components, namely, the three dictionary entries *besuchen*, *Besuch* and *Besucher* (as `lexicog:Entry` elements) are contained in that component. This is captured by the property `rdfs:member`.

```

<Entry HomNum="" hw="besuchen" identifier="EN00002666" pos="verb">
  <NestEntry>
    <DictionaryEntry identifier="DE00003297" version="1">
      <HeadwordCtn>
        <Headword>besuchen</Headword> [...]
      </HeadwordCtn> [...]
    </DictionaryEntry>
    <DictionaryEntry identifier="DE00003298" version="1">
      <HeadwordCtn>

```

```

        <Headword>Besuch</Headword> [...]
    </HeadwordCtn>
    [...]
</DictionaryEntry>
<DictionaryEntry identifier="DE00003299" version="1">
    <HeadwordBlock>
        <HeadwordCtn>
            <Headword>Besucher</Headword> [...]
        </HeadwordCtn>
        [...]
    </HeadwordBlock>[...]
</DictionaryEntry>
</NestEntry>
</Entry>

```

Example 1.3: An extract of the German entry *besuchen* ‘visit’ with a NestEntry container that groups the dictionary entries *Besuch* ‘n. visit’ and *Besucher* ‘guest, visitor’.

(Continuation)

```

:lexiconDE/besuchen-v a ontolex:LexicalEntry .
:lexiconDE/Besuch-n a ontolex:LexicalEntry .
:lexiconDE/Besucher-n a ontolex:LexicalEntry .

:lexiconDE a lime:Lexicon; lime:entry :lexiconDE/besuchen-
v, :lexiconDE/Besuch-n, :lexiconDE/Besucher-n.

:mlds-ES3
lexicog:entry :DE_DE00003297, :DE_DE00003298, :DE_DE00003299 .

:DE_DE00003297 a lexicog:Entry;
lexicog:describes :lexiconDE/besuchen
-v .

:DE_DE00003298 a lexicog:Entry ;
lexicog:describes :lexiconDE/Besuch-n.

:DE_DE00003299 a lexicog:Entry ;
lexicog:describes :lexiconDE/Besucher-
n .

:DE_EN00002666 a lexicog:LexicographicComponent ;
rdfs:member :DE_DE00003297, :DE_DE00003298, :DE_DE00003299 .

```

Example 1.4: RDF rendering of the NestEntry structure presented in example 1.3 in Turtle serialization

5.3 Usage Examples

The data from KD's Global series provides, for each sense of a headword, a usage example in the source language and the translations of the headword in the target language. The examples, in turn, are also translated to the target language and serve as example of usage for the translation. Example 1.5 shows another excerpt of the entry *arte* in Spanish. Inside the SenseGrp encapsulating the information of the first sense, there is an element TranslationCluster with Locale groups that include the headword translations for *arte* in its first sense: *kunst* (Dutch) and *kunst* (Norwegian). Below the translations follows an ExampleCtn with the example of usage of *arte* in that sense, *La música, la danza y la pintura son formas de arte* 'Music, dance and painting are art forms'. This example is in turn translated to Dutch and Norwegian.

```
<SenseGrp identifier="SE00007455" version="1">
  [...]
  <TranslationCluster identifier="TC00017354" text="manifestación humana con intención
    estética" type="def">
    <Locale lang="nl">
      <TranslationBlock>
        <TranslationCtn>
          <Translation>kunst</Translation> [...]
        </TranslationCtn>
      </TranslationBlock>
    </Locale>
    <Locale lang="no">
      <TranslationBlock>
        <TranslationCtn>
          <Translation>kunst</Translation> [...]
        </TranslationCtn>
      </TranslationBlock>
    </Locale> [...]
  </TranslationCluster>
  <ExampleCtn type="sid" version="1">
    <Example>La música, la danza y la pintura son formas de
    arte.</Example>
    <TranslationCluster identifier="TC00017355" [...]>
      <Locale lang="nl">
        <TranslationBlock>
          <TranslationCtn>
            <Translation>Muziek, dans en schilderen zijn vormen van kunst.</Translation>
          </TranslationCtn>
        </TranslationBlock>
```

```

</Locale>
<Locale lang="no">
  <TranslationBlock>
    <TranslationCtn>
      <Translation>Musikk, dans og maling er kunst
      typer.</Translation> </TranslationCtn>
    </TranslationBlock>
  </Locale> [...]
</TranslationCluster>
</ExampleCtn>
</SenseGrp>

```

Example 1.5: An extract of the Spanish entry *arte* with translations into Dutch and Norwegian examples and translations of the examples.

While the *lemon* model provided a class `lemon:UsageExample` and a property `lemon:example`, used previously in the literature to capture this information (Klimek & Brümmer, 2015), these are no longer included in the *OntoLex-lemon* model. Previous conversions of KD's data (Bosque-Gil et al., 2016b) proposed a custom class in order not to instantiate both *lemon* and *OntoLex-lemon* in the same resource. If an example is to be linked to a sense, the property `skos:example` would suffice to include the example as a string at the sense level. For cases in which the example has additional information, or has elements linkable to it, the *lexicog* module offers the class `lexicog:UsageExample` to link an `ontolex:LexicalSense` to an element representing the example. A `lexicog:UsageExample` can be further linked to other elements and described with data-type properties.

Example 1.6 shows the RDF Turtle representation of example 1.5. As showed in example 1.2, the headword and the translations belong to different lexica, one per language.

(Continuation)

```

:lexiconES/arte-n a ontolex:LexicalEntry ;
  ontolex:sense :lexiconES/arte-n-SE00007455-
  sense .
:lexiconNL/kunst-n a ontolex:LexicalEntry;
  ontolex:sense :lexiconNL/kunst-n-arte-n-
  SE00007455-sense .
:lexiconNO/kunst-n a ontolex:LexicalSense;
  ontolex:sense :lexiconNO/kunst-n-arte-n-
  SE00007455-sense .

:lexiconNO a
  lime:Lexicon;
  lime:language "no";

```

```

lime:entry :lexiconNO/
kunst-n .

:lexiconES/arte-n-SE00007455-sense a ontolex:LexicalSense ;
lexicog:usageExample :lexiconES/arte-n-SE00007455-sense-
TC00017355-ex .

:lexiconNL/kunst-n-arte-n-SE00007455-sense a ontolex:LexicalSense ;
lexicog:usageExample :lexiconES/arte-n-SE00007455-sense-
TC00017355-ex .

:lexiconNO/kunst-n-arte-n-SE00007455-sense a ontolex:LexicalSense ;
lexicog:usageExample :lexiconES/arte-n-SE00007455-sense-
TC00017355-ex .

:tranSetES-NL/arte-n-SE00007455-sense-kunst-n-arte-n-SE00007455-sense-TC00017354-trans
a vartrans:Translation ;
vartrans:source :lexiconES/arte-n-SE00007455-sense;
vartrans:target :lexiconNL/kunst-n-arte-n-SE00007455-
sense; dc:source :mlds-ES3 .

:tranSetES-NO/arte-n-SE00007455-sense-kunst-n-arte-n-SE00007455-sense-TC00017354-trans
a vartrans:Translation ;
vartrans:source :lexiconES/arte-n-SE00007455-sense;
vartrans:target :lexiconNO/kunst-n-arte-n-SE00007455-
sense dc:source :mlds-ES3 .

:lexiconES/arte-n-SE00007455-sense-TC00017355-ex a
lexicog:UsageExample ; rdf:value "La música, la danza y la pintura son
formas de arte."@es ; rdf:value "Muziek, dans en schilderen zijn
vormen van kunst."@nl ; rdf:value "Musikk, dans og maling er
kunsttyper."@no .

```

Example 1.6: RDF Turtle representation of an extract of the Spanish entry *arte* with translations into Dutch and Norwegian, examples, and translations of the examples.

Each `ontolex:LexicalEntry` has an `ontolex:LexicalSense`, which is the bridge between the linguistic description and the semantic layer, following the notion of *semantics by reference* embraced in *lemon*.¹² The example is recorded through an instance of `lexicog:UsageExample` linked to the senses via `lexicog:usageExample`. Note that this instance has different values, each for the realization of that example in a different language.

¹² Due to the lack of ontology entities to act as references for `ontolex:LexicalSenses`, the semantics provided by definitions will be captured through `ontolex:LexicalConcepts` and the property `skos:definition`. However, the instantiation of the OntoLex core, beyond *lexicog*, is out of the scope of this paper, and we refer the reader to the examples provided at the *lexicog* documentation page.

6. Conclusions and future work

In this paper we have presented work on applying the new *lexicog* module of OntoLex-*lemon* to KD’s multilingual data as a real use case scenario for the extension. We have shown that *lexicog* addresses the gaps previously identified in the literature (Klimek & Brümmer, 2015; Bosque-Gil et al., 2016b, 2017) as regards the loss of structural and implicit lexical information in the original resource, and provides elements to capture data frequently found in lexicographic records, such as usage examples, translations, or annotations on morphosyntactic features. In addition, and to serve as a basis for future transformations of lexicographic data, we framed the modelling with *lexicog* in the whole conversion process of KD to LD. We have detailed the incremental approach followed in the conversion process and outlined the different steps performed as part of the validation process for the resulting RDF.

The next step will be to process the data in a triple store, serving both to further validate the flawless conversion from XML to RDF and to prepare the data for linking to other external LD resources. Then, the actual linking to such external data resources can take place. Future work includes linking between different KD monolingual cores, creating one interconnected, fully cross-lingual graph, as well as linking to external sources, thus enhancing the data and providing even more elaborate and enriched data to Lexicala API users and for various research and development purposes.

The task of linking dictionaries, by associating a translation of a headword in the source language dictionary core to its corresponding entry in the target language dictionary core, is an ambitious and elaborate one. The main hindrance is automating the process, managing to link a translation equivalent to the correct senses across languages, which is ultimately related to word sense disambiguation, and has been previously attempted with KD data as part of the LDL4HELTA project and the Translation Inference Across Dictionaries shared tasks and workshops (TIAD).¹³ The conversion of KD monolingual cores to LD has laid the groundwork for this type of graph, and provided further ideas for carrying out this goal in the future.

In the meantime, linking KD data to other sources should be significantly facilitated by the current conversion. Linking KD data with external, annotated or enriched resources, will greatly enhance both its commercial appeal and potential for further research, and can serve as a detailed and efficient resource for language processing and parsing tasks in the realm of computational linguistics, thus expanding the outreach of LD-compliant lexicographic data yet further.

¹³ <http://2019.ldk-conf.org/tiad-2019/>

7. Acknowledgements

This work has been supported by the Spanish Ministry of Education, Culture and Sports through the FPU program, and by the European Union's Horizon 2020 research and innovation programme through the projects Lynx (grant agreement No 780602), Elexis (grant agreement No 731015) and Prêt-à-LLOD (grant agreement No 825182). It has been also partially supported by the Spanish National projects TIN2016-78011-C4-3-R (AEI/ FEDER, UE) and DGA/FEDER.

8. References

- Abromeit, F., Chiarcos, C., Fäth, C. & Ionov, M. (2016). Linking the Tower of Babel: Modelling a Massive Set of Etymological Dictionaries as RDF. In *Proceedings of the 5th Workshop on Linked Data in Linguistics: Managing, Building and Using Linked Language Resources (LDL-2016)*. pp. 11–19.
- Bizer, C., Heath, T. & Berners-Lee, T. (2009). Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 5(3), pp. 1–22.
- Bosque-Gil, J., Gracia, J. & Gómez-Pérez, A. (2016a). Linked data in lexicography. *Kernerman Dictionary News*, (26), pp. 19–24. https://www.kdictionaries.com/kdn/kdn24_2016.pdf.
- Bosque-Gil, J., Gracia, J. & Montiel-Ponsoda, E. (2017). Towards a Module for Lexicography in OntoLex. In *Proc. of the LDK workshops: OntoLex, TIAD and Challenges for Wordnets at 1st Language Data and Knowledge conference (LDK 2017), Galway, Ireland*, volume 1899. Galway (Ireland): CEUR-WS, pp. 74–84. http://ceur-ws.org/Vol-1899/OntoLex{__}2017{__}paper{__}5.pdf.
- Bosque-Gil, J., Gracia, J., Montiel-Ponsoda, E. & Aguado-de Cea, G. (2016b). Modelling Multilingual Lexicographic Resources for the Web of Data: the K Dictionaries case. In *Proc. of GLOBALEX'16 workshop at LREC'16, Portoroz, Slovenia*.
- Declerck, T., Wand-Vogt, E. & Mörth, K. (2015). Towards a Pan European Lexicography by Means of Linked (Open) Data. In I. Kosem, M. Jakubíček, J. Kallas & S. Krek (eds.) *Proceedings of eLex 2015. Biennial Conference on Electronic Lexicography (eLex2015), electronic lexicography in the 21st century: Linking lexical data in the digital age*. Ljubljana/Brighton: Trojina, Institute for Applied Slovene Studies, Ljubljana.
- Gracia, J., Villegas, M., Gómez-Pérez, A. & Bel, N. (2018). The apertium bilingual dictionaries on the web of data. *Semantic Web*, 9(2), pp. 231–240.
- Kaltenböck, M. & Kernerman, I. (2017). Introducing LDL4HELTA: Linked data lexicography for high-end language technology application. *Kernerman Dictionary News*, (25), pp. 2–3. https://www.kdictionaries.com/kdn/kdn25_2017.pdf.
- Kernerman, I. (2009). KD's BLDS: a brief introduction. *Kernerman Dictionary News*, (17), pp. 1–2. http://www.kdictionaries.com/kdn/kdn17_2009.pdf.

- Kernerman, I. (2011). From dictionary to database: Creating a global multi-language series. In I. Kosem & K. Kosem (eds.) *Electronic Lexicography in the 21st Century. New Applications for New Users. Proceedings of eLex*, pp. 113–121. <http://elex2011.trojina.si/Vsebine/proceedings/eLex2011-14.pdf>.
- Kernerman, I. (2015). A multilingual trilogy: Developing three multi-language lexicographic datasets. *Dictionaries: Journal of the Dictionary Society of North America*, 36(1), pp. 136–149. http://elex.link/elex2015/proceedings/eLex_2015_24_Kernerman.pdf.
- Klimek, B. & Brümmer, M. (2015). Enhancing lexicography with semantic language databases. *Kernerman DICTIONARY News*, 23, pp. 5–10. https://www.kdictionaries.com/kdn/kdn23_2015.pdf.
- Krek, S., Kosem, I., McCrae, J. P., Navigli, R., Pedersen, B. S., Tiberius, C. & Wissik, T. (2018). European Lexicographic Infrastructure (ELEXIS). In *The XVIII EURALEX International Congress*. p. 159.
- McCrae, J., Aguado-de Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., Gómez-Pérez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D. & Wunner, T. (2012). Interchanging lexical resources on the Semantic Web. *Language Resources and Evaluation*, 46, pp. 701–719.
- McCrae, J., Bosque-Gil, J., Gracia, J., Buitelaar, P. & Cimiano, P. (2017). The OntoLexLemon Model: Development and Applications. In I. Kosem et al. (eds.) *Electronic lexicography in the 21st century. Proc. of eLex 2017 conference, in Leiden, Netherlands*. Lexical Computing CZ s.r.o., pp. 587–597. <https://elex.link/elex2017/wp-content/uploads/2017/09/paper36.pdf>.
- McCrae, J., Tiberius, C., Khan, F. A., Kernerman, I., Declerck, T., Krek, S. Monachini, M. & Ahmadi, S. (2019). The ELEXIS Interface for Interoperable Lexical Resources. Deliverable, ELEXIS-European Lexicographic Infrastructure.
- Parvizi, A., Kohl, M., González, M. & Saur', R. (2016). Towards a Linguistic Ontology with an Emphasis on Reasoning and Knowledge Reuse. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



Towards the Automatic Construction of a Multilingual Dictionary of Collocations using Distributional Semantics

Marcos Garcia, Marcos García-Salido, Margarita Alonso-Ramos

Universidade da Coruña, CITIC, Grupo LyS, Dpto de Letras,

Fac. de Filoloxía. 15071, A Coruña

E-mail: {marcos.garcia.gonzalez,marcos.garcias,margarita.alonso}@udc.gal

Abstract

This paper presents the method used to create a multilingual online dictionary of collocations of English, Portuguese, and Spanish. This resource is built automatically and contains three types of collocations: verb–object (e.g., “[to] issue [an] invoice”), adjective–noun (“deep shame”), and nominal compounds (“cigarette packet”). We take advantage of dependency parsing and statistical association measures to compile collocations of each language, and then we align them with their equivalents in the other languages by means of compositional methods which use cross-lingual models of distributional semantics. Collocations are extracted from large and assorted corpora, and the cross-lingual models are mapped using unsupervised approaches. For each collocation in a given language, the system shows different equivalents in the other languages, ranked by a confidence value. Besides the multilingual perspective, the resulting dictionary can also serve as a monolingual resource to retrieve the collocates of a given base, thus being a useful application to both native speakers and language learners. The dictionary will be published as an online tool, and all the resources generated in this research will be freely available.

Keywords: collocations; distributional semantics; dictionary; multilinguality

1. Introduction

One of the main characteristics of collocations is that the selection of one of its elements is unpredictable. In this regard, when learning English, one should know that a horse *gallops* but a dog *scampers*, even if both verbs convey basically the same meaning. In a multilingual scenario, this unpredictability is even more important, because a collocation equivalent in a target language is often non-congruent, i.e., it is not the direct translation of both lexical units of the source combination (Nesselhauf, 2003). For instance, while in English an *invoice* is *issued*, in Portuguese a *factura* (‘invoice’) is *emitida* (literally ‘emitted’). Thus, mastering the use of collocations and other formulaic sequences presents advantages for processing and improves the production performance of language learners (Millar, 2010).

Dictionaries with collocational information are becoming more frequent, allowing both native speakers and language learners to produce idiomatic combinations in different

domains (Benson et al., 1986; Crowther et al., 2009; Bosque, 2006; Alonso-Ramos et al., 2010). However, multilingual resources of collocations and other multiword expressions, such as idioms, are scarce, but they are very useful to command such structures in other languages (Alonso-Ramos, 2015). In this respect, building multilingual dictionaries of collocations is a hard task which requires a huge effort from expert lexicographers in different languages (Orenha-Ottaiano, 2017).

Taking the above into account, this paper presents the steps to automatically create a multilingual dictionary of collocations of English, Portuguese, and Spanish. The dictionary includes three types of collocational patterns: (i) verb–object (*obj*) such as the “[to] issue [an] invoice”; (ii) adjective–noun (*amod*), e.g., “deep shame”, and (iii) nominal compounds (*nmod*) such as “cigarette packet” (or “packet of cigarettes”).

Broadly speaking, the method consists of the following steps: first, we compile large corpora in each of the three languages, and analyse them using natural language processing (NLP) tools to obtain morphosyntactic and syntactic information (Gamallo et al., 2018; Straka & Straková, 2017). Then, we apply different statistical association measures (AMs) to automatically select collocation candidates from the corpora (Evert et al., 2017; Garcia et al., 2019). After that, we use cross-lingual models of distributional semantics to apply compositional strategies able to identify equivalents of a given collocation in other languages (Garcia et al., 2017; Gamallo & Garcia, 2019). Instead of using parallel corpora, the cross-lingual models can be generated with monolingual resources, thus avoiding the need of obtaining large parallel texts for each language pair (Artetxe et al., 2018). The resulting dictionary provides, for each collocation in a source language, a set of equivalents in the target ones, ranked by a confidence value which represents the translation probabilities. The dictionary will be published as an online tool, and all the resources generated in this research will be freely available.

The rest of this paper is organized as follows. Section 2 briefly presents some previous work concerning different methods to extract collocations from corpora. Then, the approaches to both identify monolingual and multilingual collocations are introduced in Section 3, which also discusses some shortcomings and further lines of research. Finally, Section 4 summarizes the main properties of the online dictionary, while Section 5 contains the conclusions of our study.

2. Methods to extract collocations with a lexicographic aim

In order to create the lexicographic resources to release a multilingual dictionary, our work takes advantage of different NLP methods aimed at identifying monolingual collocations from corpora as well as at finding their equivalents in other languages.

The first approaches to extract collocations from corpora consisted of applying AMs to short sequences of ngrams (Smadja, 1993). Using similar approaches, other studies defined patterns of part-of-speech tags to identify specific constructions (Krenn & Evert,

2001), while the use of syntactic dependencies was evaluated in articles such as Lin (1999) or Seretan and Wehrli (2006). Besides classical association measures such as pointwise mutual information or t-score, several authors have proposed directional AMs to capture the asymmetry of collocations (Gries, 2013; Carlini et al., 2014).

With a view to comparing the performance of different AMs, studies such as Pecina (2010), Evert et al. (2017), or Garcia et al. (2019) performed different evaluations of various measures to extract collocations in several languages. The results, however, differ with respect to the collocation type as well as to the interpretation of collocations, which involves divergent annotations on each gold-standard data.

With regard to the multilingual identification, the first studies exploited parallel corpora to find bilingual translations of collocations and other multiword expressions (Smadja, 1992; Kupiec, 1993; Haruno et al., 1996). More recently, the use of syntactic analysis was also proposed to restrict the search to predefined patterns (Wu & Chang, 2003; Seretan & Wehrli, 2007).

Other studies tackled this problem using comparable and unrelated corpora in two languages, by performing word-to-word translations of each component of the collocations — and other similar constructions — (Grefenstette, 1999; Baldwin & Tanaka, 2004; Delpech et al., 2012). Similar approaches, which improve the word-to-word translation by taking advantage of distributional models were presented in Morin and Daille (2012) and Garcia (2018). Finally, recent articles investigate the use of contextualized compositional models as well as weighted additive vectors to improve the identification of equivalents of multiword expressions in different languages (Gamallo & Garcia, 2019; Garcia et al., 2019).

Concerning dictionaries with collocational information, the majority of the publications are monolingual resources mostly focused on language learners. In this respect, English is the most represented language among the three targets (Benson et al., 1986; Crowther et al., 2009; Rundell, 2011), but there also exist dictionaries for Spanish, oriented to both native speakers and language learners (Alonso-Ramos, 2004; Bosque, 2004, 2006). For Portuguese, the COMBINA-PT project has generated a database of different multiword expressions, including collocations (Mendes et al., 2006), while *Syntax Deep Explorer* provides an online tool to retrieve co-occurrence information from large corpora (Correia et al., 2016). Moreover, the work presented in Larens (2016) describes the creation of a collocational database of Brazilian Portuguese.

From a multilingual perspective, some dictionaries with collocational information have been published for various language pairs, such as English–Russian (Benson & Benson, 1993), German–French (Ilgenfritz et al., 1989), or Italian–German (Konecny & Autelli, 2014). Several articles and projects have also carried out research aimed at creating multilingual dictionaries of collocations (Grefenstette et al., 1996; Nerima et al., 2003; Konecny & Autelli, 2014; Alonso-Ramos, 2015; Garcia et al., 2017; Orenha-Ottaiano, 2017). Concerning the three languages of our study, Alegro et al. (2010) presents a

bilingual dictionary of adjectival collocations in English and Portuguese. However, to the best of our knowledge there is no freely available multilingual resource of collocations for English, Portuguese, and Spanish, so our research aims at contributing to this area with an online tool and free resources in the three target languages. It is worth mentioning, however, that online dictionaries such as *Linguee*¹ contain not only monolexical entries, but also some multiword expressions (including several collocations). In this regard, the main difference with respect to our work is that we extract the equivalents from comparable and unrelated corpora instead of parallel data.

3. Automatic extraction of collocations

This section presents the different steps to automatically generate equivalents of collocations in various languages. First, we explain the processes used to obtain collocation candidates in one language, and then we introduce the approach to obtain their equivalents in other languages. Finally, we briefly discuss some features and shortcomings of the proposed strategies.

3.1 Monolingual extraction

We understand collocations as phraseological combinations of two lexical units (LUs) which are directly linked by a syntactic relation (Hausmann, 1989; Mel'čuk, 1995). The internal structure of these expressions is not symmetrical, since while one of the LUs is freely selected due to its meaning (the base), the selection of the other component (the collocate) is restricted by the former (Mel'čuk, 1996). Thus, a base such as *shame* may select the collocates *deep* or *intense* (but not *strong* or *heavy*) in order to convey the meaning 'intense'.

Aimed at identifying the syntactic relation between two lexical units we employ dependency parsing, which establishes binary relations between the different words of a sentence (Tesnière, 1959; Kübler et al., 2009). To capture the collocability of two syntactically related words we use various association measures which assign numerical values that allow us to rank the *attraction* or *repulsion* of the word pairs (Evert, 2008).

With this in mind, our method to extract monolingual collocation candidates is as follows: First, we obtain large amounts of corpora for each language (in our case, English, Portuguese, and Spanish). So far we have been working with texts from different sources, such as the Wikipedia, the Europarl (Koehn, 2005), OpenSubtitles (Lison & Tiedemann, 2016), as well as text from other genres such as essays, literature, and web pages. These corpora are first processed using LinguaKit to identify sentence boundaries and to provide tokenization, lemmatization and PoS-tagging (Gamallo et al., 2018). Then, the corpora are enriched with syntactic information using UDPipe

¹ <https://www.linguee.com/>

models (Straka & Straková, 2017), which are based on *Universal Dependencies* annotation (Nivre, 2015).² It is important to note that the use of dependency parsing also allows us to identify long distance dependencies which are not captured in a short span of text.

Once we have the processed data for each language, we select as candidate collocations those pairs of lemmas that belong to the following dependency relations, structured as base-collocate tuples: *obj* (*invoice,issue*), *amod* (*shame,deep*), *nmod/compound* (*cigarette,packet*). We use lemmas instead of tokens (i.e., we represent the different inflected forms of a word by a single entry) to reduce the data sparseness.³

Over these candidates, we apply different association measures (e.g., *t-score*, *log-likelihood*, *Dice*) to rank each list of pairs. From the results of previous studies, we use different AMs for each dependency relation (Garcia et al., 2019). Moreover, and since most frequent candidates tend to be phraseological, these ranks are combined with frequency data to select the top-*n* combinations (Krenn & Evert, 2001). At the end of this process we have, for each language, large sets of collocation candidates for the three mentioned patterns.

3.2 Bilingual equivalents

In order to obtain equivalents in various languages of a given collocation in a source we use compositional semantics strategies by means of cross-lingual distributional models.

3.2.1 Cross-lingual distributional models

Monolingual models of distributional semantics (also known as *word embeddings*) use contextual information to represent words as *n*-dimensional vectors, so words occurring in similar contexts tend to have similar vectors (Landauer & Dumais, 1997). Likewise, cross-lingual models represent the words of different languages in the same vector space, thus allowing for the computation of distributional similarities between those different languages (Rapp, 1999; Ruder et al., 2019).

To build our collocational database we have used two different approaches to obtain cross-lingual models of distributional semantics. On the one hand, we have used MultiVec (Bérard et al., 2016) to train bilingual models using parallel data from the Europarl and OpenSubtitles corpora (Koehn, 2005; Lison & Tiedemann, 2016). On the other hand, and taking into account that large amounts of parallel data from different domains are scarce, we have also trained monolingual models using *word2vec* (Mikolov

² <https://universaldependencies.org/>

³ Note that, for instance, a single verb in several Romance languages (including Portuguese and Spanish) may have more than 50 different inflected forms.

et al., 2013), and then mapped into a shared vector space with *vecmap* (Artetxe et al., 2018). The latter approach obtains high-quality cross-lingual models by means of unrelated corpora, so it allows us to use a large variety of texts from different genres which in turn generate better word embeddings.

We train the distributional models converting the original tokens of each corpus into *lemma_PoSTag* entries. This strategy alleviates both the sparseness produced by morphological variation as well as the potential ambiguity of words with different morphosyntactic categories which have the same lemma (e.g., *plane_NOUN*, *plane_VERB*, *plane_ADJ*). Besides, using these linguistically-enriched models allows us to select only those base and collocate candidates which belong to a specific part-of-speech.

In sum, cross-lingual models of distributional semantics allow us to obtain distributionally similar words in a target language for a given input in a source language. For instance, if we search for the most similar nouns (in English) to *adversário* (in Portuguese), we may get words such *adversary*, *foe*, or *opponent*.

3.2.2 Compositional semantics methods

A collocation encodes semantic information from both the base and the collocate, so that they are semantically compositional expressions. Nevertheless, it is worth noting that a collocate may convey a particular meaning in each specific combination (Mel'čuk, 1995). For instance, different adjectives such as *heavy* and *strong* convey basically the same meaning in collocations such as “heavy rain” and “strong coffee”, while the verb *[to] pay* has a different meaning in “pay attention” and “pay the bills”. The bases, however, have a stable meaning across the different combinations in which they appear. With that in mind, the semantic properties of collocations should be taken into account when searching for equivalents in other languages.

The approach that we use to find multilingual equivalents has been evaluated in various languages and relations with high precision results (Garcia et al., 2017; Garcia, 2018). On the one hand, we rely on the previously extracted monolingual collocations to select candidates which have some degree of collocability (or are at least frequent combinations) in each language. On the other hand, we select as candidate translations those collocations with a high degree of similarity between the input and target bases. The procedure is as follows: given an input collocation in a source language (e.g., *lío tremendo*, ‘huge mess’ in Spanish), we select its base (*lío*) and retrieve the *n* most similar words with the same part-of-speech in the target language: e.g., “trouble”, “mess”, etc. in English (where *n* = 5 and the similarity is computed by their cosine distance). Then, we select those collocations in the target language with the candidate bases (e.g., “little trouble”, “deep trouble”, “huge mess”, “fine mess”, etc.). After that, we compute the similarity between the source collocate and the target ones (e.g., “tremendo” *versus* “little”, “deep”, “huge”, and “fine”). If the cosine distances between

both the source and target bases and collocates are higher than a given threshold, they are selected as potential equivalents, and the average similarity between both components is set as the translation confidence value (e.g., “lío tremendo–huge mess”: 0.72).

This strategy follows the base–collocate structure of collocations by selecting in the target language only candidates with very similar bases. Also, it allows us to identify not only word-to-word translations between the collocates, since we use distributional similarity to compute the distance between the different candidates (Morin & Daille, 2012). Finally, using previously extracted collocations (instead of artificially generating new instances) avoids the creation of unconventional combinations in the target languages.

3.3 Discussion

Even though the proposed approaches effectively obtain equivalents in various languages with high precision (about 90%, depending on the scenario), it is worth noting that the results and error analyses carried out in different studies have pointed to some issues that could be improved in further research (Garcia et al., 2017, 2018).

On the one hand, using statistical data (frequency and various association measures) to rank the monolingual collocation candidates may result in non-phraseological expressions such as free combinations (e.g., “buy [a] beer”) or quasi-idioms (“big deal”) (Mel’čuk, 1995). In this regard, we do not consider this circumstance a serious problem as long as the equivalents in the other languages (if any) are valid. However, and with a view to refine the monolingual identification, several strategies can be implemented to improve the ranking of the candidates and to automatically identify non-compositional expressions (Pecina, 2010; Cordeiro et al., 2019).

On the other hand, and even if distributional semantics models are able to identify some non-congruent bilingual equivalents, several collocates convey a very different meaning (with respect to their most frequent one) in some specific collocations. In these cases, finding appropriate candidates without using parallel corpora may be a difficult task: for instance, both the Portuguese and Spanish translations of the English verb “[to] pay” will probably belong to the economic field (*pagar* ‘[to] pay’, *cobrar* ‘[to] earn’, etc.), so our approach may not identify *prestar atenção/atención* (literally ‘[to] lend attention’) as equivalents of “[to] pay attention”. There is, however, recent research which could improve the extraction of these cases: as mentioned in Section 2, Garcia et al. (2019) propose a compositional strategy to find bilingual collocation equivalents using weighted additive vectors. Besides, in Gamallo and Garcia (2019) the authors use contextualized word embeddings based on syntactic dependencies to represent the meaning of composite expressions. In this regard, combining both approaches could be an interesting line of research for further work.

Finally, the performance of our current approaches is also influenced by one of the main shortcomings of standard distributional methods, which represent in the same vector different senses of the same word. To overcome this issue (apart from the mentioned strategy of Gamallo and Garcia (2019)), studies such as Iacobacci et al. (2015) have implemented sense-based distributional models, while recent research in NLP obtains pre-trained contextual representations, where the vector of a given word is based on the other words which occur in the same sentence (Weir et al., 2016; Devlin et al., 2019).

4. Towards a multilingual dictionary of collocations

This section illustrates how we leverage the multilingual resources generated by the methods presented above to create an online tool with monolingual and multilingual collocational information. This tool is not a finished multilingual dictionary of collocations, but rather an instrument to help language users by exploiting our database. In this regard, it is worth remembering that this database is automatically constructed and freely available, and that it can be updated both with new information obtained from corpora as well as with manual annotation from lexicographers.

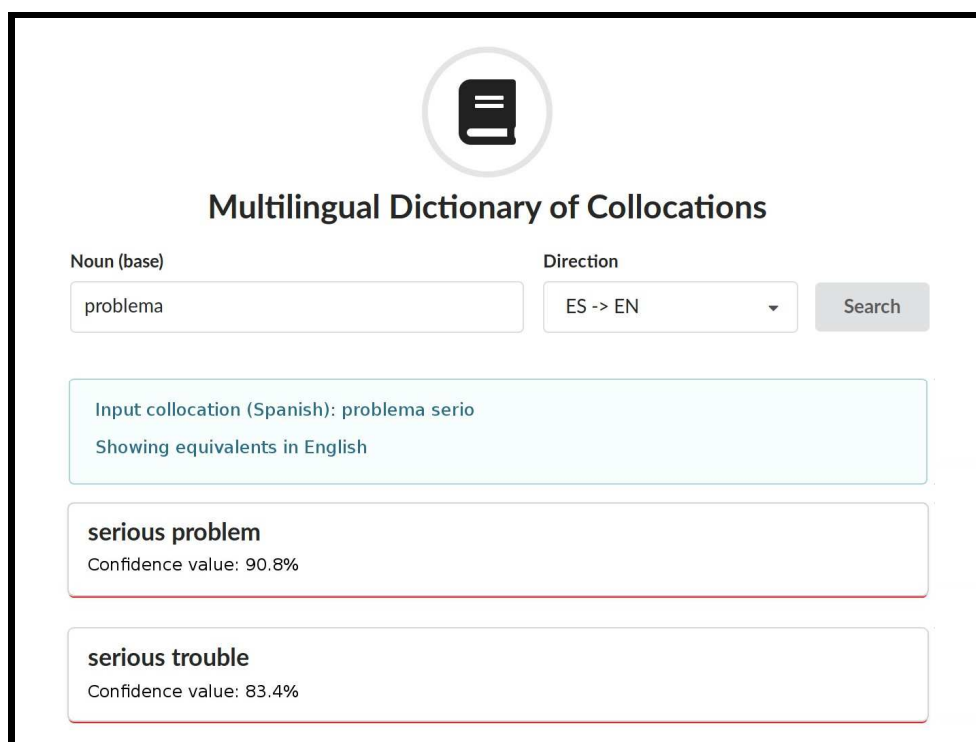
The query interface is based on a *source–target* structure, so that the user should first select the desired translation direction (e.g., English→Portuguese, Portuguese→Spanish, etc.). As in other resources, the basic units of the dictionary are nouns (Lea & Runcie, 2002). In our case, however, the selection of nouns as the main unit derives from the fact that they are the bases of the three considered patterns. Nevertheless, the same strategy can be applied to other collocational patterns such as verb–adverb (e.g., “really want”, where the verb is the base), or adjective–adverb (e.g., “extremely powerful”, in which the adjective is the base), among others.


Thus, after selecting a source and target language, the user introduces a noun (by its lemma) in the search box (e.g., “wine” in English→Portuguese). As the input query is performed, the dictionary will show, in three columns, the highest ranked combinations in the source language with the given base. In the previous example, the verb–object column may include “drink wine”, “produce wine”, or “export wine”; adjective–noun collocates such as “red wine”, “white wine”, or “varietal wine”, and “bottle [of] wine”, “glass [of] wine”, or “wine grape” as nominal compounds. The user can expand one specific column to search for other collocations in the desired pattern. Besides, the tool allows for clicking in a particular collocation to see a few usage examples extracted from corpora. At this point, the dictionary can be also seen as a database of collocations in a specific language.

Continuing with the multilingual tool, the user can select a collocation in the source language to search for equivalents in the target one (e.g., adjective–noun: “red wine”). Then, the dictionary will show the collocation equivalents in the target language, sorted by the confidence value obtained using the compositional strategies presented in Section

3.2. Following the previous example, the Portuguese equivalents (and their confidence values) of the English collocation “red wine” may be *vinho tinto*: 0.95 (‘red wine’), *vinho rosé*: 0.86 (‘rosé wine’), or *vermute tinto*: 0.83 (‘red vermouth’), among others. Again, the tool allows the user to expand the number of bilingual equivalents as well as to see real examples in corpora, this time in the target language.

It is worth mentioning that the database is automatically enlarged with new entries built by transitivity. Therefore, in those cases where it has a specific collocation translated in two language directions it infers the third one if it has not been extracted. As an example, let us say that we obtained the English→Portuguese and English→Spanish equivalents of “red vermouth” (*vermute tinto* and *vermú rojo*, respectively), but not the Portuguese→Spanish translation. Thus, the tool will infer that *vermú rojo* may be a suitable translation of *vermute tinto*. In these cases, the inferred equivalents are presented using a slightly different colour to inform the user of this fact.





Multilingual Dictionary of Collocations

Noun (base)

Direction

ES -> EN
▼

Input collocation (Spanish): **problema serio**

Showing equivalents in English

serious problem

Confidence value: 90.8%

serious trouble

Confidence value: 83.4%

Figure 1: Example of the online interface of the dictionary.

Figure 1 shows an example of the online interface. The inserted noun (top row) is *problema* (‘problem’) and the translation direction Spanish→English. The figure includes the second visualization of the tool, after selecting the input collocation *problema serio* (adjective– noun). It displays the top two translations together with their confidence values, and allows the user to click on any of them to see real examples.

The current version of the online tool (together with the multilingual database) presents two issues that can be addressed in future research. First, as our approach relies on

lemmatized instances of syntactic dependencies, we do not pay particular attention to the surface structures allowed by each collocation. Thus, the dictionary provides the users with base-collocate data, but it does not explicitly inform, for instance, whether a noun requires a determiner or not (e.g., **“take the advantage” versus “have a look”*). The second peculiarity concerns the order in which both LUs are shown to the users. In each pattern, collocations are presented in their canonical structure in the three languages (e.g., adjective–noun pairs are shown as noun–adjective in Portuguese and Spanish), but while some of them are mostly used only in a particular pattern (e.g., *“football manager” versus **“manager of football”**), others may appear in both ways (e.g., *“energy consumption – consumption of energy”*).⁴ Both problems are partially addressed with the usage examples of each collocation, but further work could also focus on these issues in order to improve the representation of each combination.

5. Conclusions and further work

This paper has presented a set of methods to automatically create a database of collocation equivalents in English, Portuguese, and Spanish. This database is used to supply an online dictionary which aids language users in the selection of both monolingual and multilingual combinations of a given noun.

To extract candidate collocations we employ dependency parsing and statistical association measures applied to large monolingual corpora. We have focused on the following three collocational patterns: verb–object, adjective–noun, and nominal compounds. To identify bilingual equivalents of a given collocation in a source language, we use compositional distributional methods which rely on pre-extracted collocations in the target languages. The cross-lingual distributional models can be directly learned using parallel corpora, or mapped after monolingual training with unrelated resources.

Apart from presenting the different strategies to build the collocation database, this study also discusses some shortcomings of the proposed approaches, aimed at improving both the monolingual extraction and the multilingual alignment in further work.

Finally, the paper presents the main structure and functionalities of the online tool, which can be useful for language users in a monolingual scenario (searching for collocates in a particular language) and in a multilingual one (to find equivalents in other languages). It is worth noting that all the resources created in this research will be freely available.

⁴ A different issue occurs in some constructions which may have a different meaning with respect to their structure, such as *“coffee cup”* and *“cup of coffee”*.

6. Acknowledgements

This research was supported by a 2017 Leonardo Grant for Researchers and Cultural Creators (BBVA Foundation), by Ministerio de Economía, Industria y Competitividad (project with reference FFI2016-78299-P), and by the Galician Government (Xunta de Galicia grant ED431B-2017/01). Marcos Garcia has been funded by a *Juan de la Cierva incorporación* grant (IJCI-2016-29598), and Marcos García-Salido by a post-doctoral grant from Xunta de Galicia (ED481D-2017-009). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

7. References

- Alegro, A., Mobaid, R. & Brezolin, A. (2010). *Happy Couples. Dicionário de Colocações Lexicais Adjetivas*. Disal.
- Alonso-Ramos, M. (2004). DiCE: Dicionario de Colocaciones del Español. Universidade da Coruña. <http://dicesp.com>.
- Alonso-Ramos, M. (2015). Discovering hidden collocations in a bilingual Spanish–English dictionary. In I. Kosem, M. Jakubiček, J. Kallas & S. Krek (eds.) *Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference*. Institute for Applied Slovene Studies/Lexical Computing Ltd, pp. 170–185.
- Alonso-Ramos, M., Nishikawa, A. & Vincze, O. (2010). DiCE in the web: An online Spanish collocation dictionary. In *ELexicography in the 21st Century: New Challenges, New Applications: Proceedings of ELex 2009*, volume 7. Presses univ. de Louvain, pp. 369–374.
- Artetxe, M., Labaka, G. & Agirre, E. (2018). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pp. 789–798.
- Baldwin, T. & Tanaka, T. (2004). Translation by Machine of Complex Nominals: Getting it Right. In *Second ACL Workshop on Multiword Expressions: Integrating Processing*. Association for Computational Linguistics, pp. 24–31.
- Benson, M. & Benson, E. (1993). *Russian-English dictionary of verbal collocations*. John Benjamins Publishing.
- Benson, M., Benson, E. & Ilson, R. (1986). *The BBI combinatorial dictionary of English: A guide to word combinations*. John Benjamins Publishing.
- Bosque, I. (2004). *Redes. Diccionario combinatorio del español contemporáneo*. SM.
- Bosque, I. (2006). *Diccionario combinatorio práctico del español contemporáneo. Las palabras en su contexto*. SM.
- Bérard, A., Servan, C., Pietquin, O. & Besacier, L. (2016). MultiVec: a Multilingual and Multilevel Representation Learning Toolkit for NLP. In N.C.C. Chair), K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo,

- A. Moreno, J. Odijk & S. Piperidis (eds.) *The 10th edition of the Language Resources and Evaluation Conference (LREC 2016)*. Paris, France: European Language Resources Association, pp. 4188–4192.
- Carlini, R., Codina-Filba, J. & Wanner, L. (2014). Improving collocation correction by ranking suggestions using linguistic knowledge. In *Proceedings of the third workshop on NLP for computer-assisted language learning*. Uppsala: LiU Electronic Press, pp. 1–12.
- Cordeiro, S., Villavicencio, A., Idiart, M. & Ramisch, C. (2019). Unsupervised Compositionality Prediction of Nominal Compounds. *American Journal of Computational Linguistics*, 45(1), pp. 1–57.
- Correia, J., Baptista, J. & Mamede, N. (2016). Syntax Deep Explorer. In J. Silva, R. Ribeiro, P. Quaresma, A. Adami & A. Branco (eds.) *Computational Processing of the Portuguese Language*, volume 9727 of *Lecture Notes in Computer Science*. Springer, pp. 189–201.
- Crowther, J., Dignen, S. & Lea, D. (eds.) (2009). *Oxford Collocations Dictionary for student's of English*. Oxford University Press.
- Delpech, E., Daille, B., Morin, E. & Lemaire, C. (2012). Extraction of Domain-Specific Bilingual Lexicon from Comparable Corpora: Compositional Translation and Ranking. In *Proceedings of COLING 2012*. The COLING 2012 Organizing Committee, pp. 745–762.
- Devlin, J., Chang, M.W., Lee, K. & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186.
- Evert, S. (2008). Corpora and collocations. In A. Lüdeling & M. Kytö (eds.) *Corpus Linguistics. An international handbook*, volume 2. Berlin: Mouton de Gruyter, pp. 1212–1248.
- Evert, S., Uhrig, P., Bartsch, S. & Proisl, T. (2017). E-VIEW-affiliation—A large-scale evaluation study of association measures for collocation identification. In *Proceedings of eLex 2017—Electronic lexicography in the 21st century: Lexicography from Scratch*. pp. 531–549.
- Gamallo, P. & Garcia, M. (2019). Unsupervised Compositional Translation of Multiword Expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, at the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019). Association for Computational Linguistics, pp. 40–48.
- Gamallo, P., Garcia, M., Pineiro, C., Martinez-Castaño, R. & Pichel, J. C. (2018). LinguaKit: a Big Data-based multilingual tool for linguistic analysis and information extraction. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*. IEEE, pp. 239–244.
- Garcia, M. (2018). Comparing bilingual word embeddings to translation dictionaries

- for extracting multilingual collocation equivalents. In S. Markantonatou, C. Ramisch, A. Savary & V. Vincze (eds.) *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, volume 3 of *Phraseology and Multiword Expressions*, chapter 12. Language Science Press, pp. 319–342.
- Garcia, M., García-Salido, M. & Alonso-Ramos, M. (2017). Using bilingual word-embeddings for multilingual collocation extraction. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*. Association for Computational Linguistics, pp. 21–30.
- Garcia, M., García-Salido, M. & Alonso-Ramos, M. (2018). Discovering bilingual collocations in parallel corpora: A first attempt at using distributional semantics. In I. Doval & M.T. Sánchez Nieto (eds.) *Parallel corpora for contrastive and translation studies: New resources and applications*, volume 90 of *Studies in Corpus Linguistics*. John Benjamins Publishing Company, pp. 267–279.
- Garcia, M., García-Salido, M. & Alonso-Ramos, M. (2019). A comparison of statistical association measures for identifying dependency-based collocations in various languages. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, at the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019). Association for Computational Linguistics, pp. 49–59.
- Grefenstette, G. (1999). The World Wide Web as a resource for example-based machine translation tasks. In *Proceedings of the ASLIB Conference on Translating and the Computer*, volume 21.
- Grefenstette, G., Heid, U., Schulze, B., Fontenelle, T. & Gerardy, C. (1996). The DECIDE project: Multilingual collocation extraction. In *Euralex 96 Proceedings*. Göteborg, pp. 93–108.
- Gries, S.T. (2013). 50-something years of work on collocations. *International Journal of Corpus Linguistics*, 18(1), pp. 137–165.
- Haruno, M., Ikehara, S. & Yamazaki, T. (1996). Learning bilingual collocations by word level sorting. In *Proceedings of the 16th Conference on Computational Linguistics*, volume 1 of *COLING 1996*. Association for Computational Linguistics, pp. 525–530.
- Hausmann, F. J. (1989). Le dictionnaire de collocations. *Wörterbücher, Dictionaries, Dictionnaires*, 1, pp. 1010–1019.
- Iacobacci, I., Pilehvar, M. T. & Navigli, R. (2015). SensEmbed: Learning Sense Embeddings for Word and Relational Similarity. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, pp. 95–105.
- Ilgenfritz, P., Schneider, G. & Stephan-Gabinel, N. (1989). *Langenscheidts Kontextwörterbuch Französisch-Deutsch*. Langenscheidt.
- Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*. AAMT, AAMT, pp. 79–86.

- Konecny, C. & Autelli, E. (2014). *Kollokationen Italienisch - Deutsch*. Helmut Buske.
- Krenn, B. & Evert, S. (2001). Can we do better than frequency? A case study on extracting PP-verb collocations. In *Proceedings of the ACL Workshop on Collocations*. Association for Computational Linguistics, pp. 39–46.
- Kupiec, J. (1993). An algorithm for finding noun phrase correspondences in bilingual corpora. In *Proceedings of the 31st Annual Meeting on Association for Computational Linguistics*, ACL 1993. Association for Computational Linguistics, pp. 17–22.
- Kübler, S., McDonald, R. & Nivre, J. (2009). *Dependency Parsing*. Morgan and Claypool Publishers.
- Landauer, T. & Dumais, S. (1997). A solution to Plato’s problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2), pp. 211–240.
- Larens, J. (2016). *Colocações do Português Brasileiro: Tipologia, Categorização, e Construção de uma Base de Dados*. Ph.D. thesis, Universidade Federal do Ceará.
- Lea, D. & Runcie, M. (2002). Blunt Instruments and Fine Distinctions: a Collocations dictionary for students of English. In *Proceedings of the Tenth EURALEX International Congress. Copenhagen: Center for Sprogteknologi*, volume 2. pp. 819–829.
- Lin, D. (1999). Automatic identification of non-compositional phrases. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL 1999. Association for Computational Linguistics, pp. 317–324.
- Lison, P. & Tiedemann, J. (2016). OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In N.C.C. Chair), K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk & S. Piperidis (eds.) *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Paris, France: European Language Resources Association, pp. 923–929.
- Mel’čuk, I. (1995). Phrasemes in language and phraseology in linguistics. In M. Everaert, E. J. van der Linden, A. Schenk & R. Schreu (eds.) *Idioms: Structural and psychological perspectives*. Hillsdale: Lawrence Erlbaum Associates, pp. 167–232.
- Mel’čuk, I. (1996). Lexical Functions: a Tool for the Description of Lexical Relations in a Lexicon. In L. Wanner (ed.) *Lexical Functions in Lexicography and Natural Language Processing*, volume 31 of *Studies in Corpus Linguistics*. John Benjamins Publishing, pp. 37–102.
- Mendes, A., Antunes, S., Nascimento, M.F.B.d., Casteleiro, J. M., Pereira, L. & Sá, T. (2006). COMBINA-PT: A Large Corpus-extracted and Hand-checked Lexical Database of Portuguese Multiword Expressions. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*. European Language Resources Association, pp. 1900–1905.
- Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). Efficient estimation of word

- representations in vector space. In *Workshop Proceedings of the International Conference on Learning Representations (ICLR) 2013*. Scottsdale, Arizona. arXiv preprint arXiv:1301.3781.
- Millar, N. (2010). The processing of malformed formulaic language. *Applied Linguistics*, 32(2), pp. 129–148.
- Morin, E. & Daille, B. (2012). Revising the Compositional Method for Terminology Acquisition from Comparable Corpora. In *Proceedings of COLING 2012*. The COLING 2012 Organizing Committee, pp. 1797–1810.
- Nerima, L., Seretan, V. & Wehrli, E. (2003). Creating a multilingual collocations dictionary from large text corpora. In *10th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 131–134.
- Nesselhauf, N. (2003). The Use of Collocations by Advanced Learners of English and Some Implications for Teaching. *Applied linguistics*, 24(2), pp. 223–242.
- Nivre, J. (2015). Towards a Universal Grammar for Natural Language Processing. In *International Conference on Intelligent Text Processing and Computational Linguistics*, volume 9041 of *Lecture Notes in Computer Science*. Springer, pp. 3–16.
- Orenha-Ottaiano, A. (2017). The compilation of an online Corpus-Based bilingual Collocations Dictionary: motivations, obstacles and achievements. In *Proceedings of eLex 2017–Electronic lexicography in the 21st century: Lexicography from Scratch*. Lexical Computing, pp. 458–473.
- Pecina, P. (2010). Lexical association measures and collocation extraction. *Language Resources and Evaluation*, 44(1-2), pp. 137–158.
- Rapp, R. (1999). Automatic Identification of Word Translations from Unrelated English and German Corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*. College Park, Maryland, USA: Association for Computational Linguistics, pp. 519–526.
- Ruder, S., Vulić, I. & Søgaard, A. (2019). A Survey of Cross-Lingual Word Embedding Models. *Journal of Artificial Intelligence Research*. arXiv preprint arXiv:1706.04902.
- Rundell, M. (ed.) (2011). *Macmillan Collocations Dictionary*. Macmillan.
- Seretan, V. & Wehrli, E. (2006). Accurate collocation extraction using a multilingual parser. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*. pp. 953–960.
- Seretan, V. & Wehrli, E. (2007). Collocation translation based on sentence alignment and parsing. In *Actes de la 14e conference sur le traitement automatique des langues naturelles*, TALN 2007. IRIT Press, pp. 401–410.
- Smadja, F. (1992). How to compile a bilingual collocational lexicon automatically. In *Proceedings of the AAAI Workshop on Statistically-Based NLP Techniques*. AAAI Press, pp. 57–63.
- Smadja, F. (1993). Retrieving collocations from text: Xtract. *Computational*

- Linguistics*, 19(1), pp. 143–177.
- Straka, M. & Straková, J. (2017). Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Association for Computational Linguistics, pp. 88–99.
- Tesnière, L. (1959). *Eléments de syntaxe structurale*. Librairie C. Klincksieck.
- Weir, D., Weeds, J., Reffin, J. & Kober, T. (2016). Aligning Packed Dependency Trees: A Theory of Composition for Distributional Semantics. *Computational Linguistics*, 42(4), pp. 727–761.
- Wu, C. C. & Chang, J. S. (2003). Bilingual collocation extraction based on syntactic and statistical analyses. In *Proceedings of the 15th Conference on Computational Linguistics and Speech Processing*. Association for Computational Linguistics and Chinese Language Processing, pp. 1–20.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



SkELL Corpora as a Part of the Language Portal

Sõnaveeb: Problems and Perspectives

Kristina Koppel¹, Jelena Kallas¹, Maria Khokhlova²,

Vít Suchomel^{3,4}, Vít Baisa^{3,4}, Jan Michelfeit³

¹ Institute of the Estonian Language, Estonia

² St. Petersburg State University, Russia

³ Lexical Computing Ltd., Czech Republic

⁴ Masaryk University, Czech Republic

E-mail: kristina.koppel@eki.ee, jelena.kallas@eki.ee, m.khokhlova@spbu.ru,
vit.suchomel@sketchengine.co.uk, vit.baisa@sketchengine.co.uk,
jan.michelfeit@sketchengine.co.uk

Abstract

The paper provides an analysis of the quality and presentation of authentic corpus sentences from Sketch Engine for Language Learning (SkELL) corpora (Baisa & Suchomel 2014), based on the example of Sõnaveeb (Wordweb), a new language portal being developed in the Institute of the Estonian Language. Currently Sõnaveeb contains a total of 150,000 Estonian headwords; about 70,000 of them have Russian equivalents. Authentic corpus sentences are displayed for both languages. In some cases (e.g. terms, derived forms, compounds and multi-word expressions), corpus sentences are the only source of usage examples that are available on the portal.

We describe the parameters of Good Dictionary Examples (GDEX) (Kilgarriff et al., 2008) configurations for Estonian and for Russian used for the compilation of etSkELL 2018 and ruSkELL 1.6 corpora, give an overview of an evaluation of the GDEX configuration for Estonian, and outline the requirements for the user-friendly presentation of SkELL corpora as a part of the language portal.

Keywords: GDEX; SkELL; learner corpus; Estonian; Russian

1. Introduction

Despite the fact that most modern dictionaries are corpus-based, displaying authentic corpus data in dictionary portals is still quite a new trend in e-lexicography. There are some dictionaries (e.g. the 5th edition of LDOCE¹, Wordnik²) that offer automatically-retrieved corpus sentences alongside manually-selected examples (Cook, 2014) but as it became evident in a survey about lexicographic practices in Europe (Kallas et al., 2019), most dictionary websites do not offer automatically-retrieved corpus sentences nor a

¹ <http://ldoce.longmandictionariesonline.com/main/Home.html> (3 June 2019).

² www.wordnik.com (3 June 2019).

link to corpus data. The survey revealed that if links are offered, they are generally automatic URLs pointing to the Corpus Query System (CQS) for the headword. The user cannot specify which elements they want to retrieve from the corpus (e.g. example sentences with metadata/without metadata). Only after the user has entered the CQS, can they change the query.

One way to optimize this process is to display corpus data not from general corpora but from corpora that consist of pre-filtered examples instead. As an example of such corpora, Sketch Engine for Language Learning (SkELL) corpora can be used. SkELL corpora were initially intended for language learning purposes (e.g. for teachers or students to efficiently find out how a word is used in a language), but they can also be seen as a source of clean and processed examples (which is especially the case when we speak of web data).

The principle of SkELL corpora is to prepare roughly 1 billion tokens of clean sentences from various resources. This is achieved either by compiling several trusted resources (in the case of English) or extensive filtering of web-based corpora (in the case of Estonian and Russian). De-duplication (i.e. the removal of the same or even similar text fragments) is a part of the process. Cleaned data (sentences) are evaluated with the example extraction tool GDEX (Good Dictionary Examples, Kilgarriff et al., 2008) and then sorted by GDEX scores. The scores correspond to sentence values and vary from 0 (the worst) to 1 (the best). Its computation is based on a formula that deals with a variety of formal classifiers, paying attention to various features (see Section 3). The formula itself is described in the GDEX configuration files³. Unlike other corpora, SkELL corpora do not contain whole documents (assuming language learners do not need them) but only sentences with the highest scores (i.e. most suitable as examples in a dictionary according to the heuristic) are taken into the resulting corpus. By treating sentences separately, the intersentential context is lost, but this approach makes it possible to sort the corpus by GDEX score and to have all searches GDEX-sorted by default.

The family of SkELL corpora is led by English SkELL, which was released in 2014. Later, Russian (2016), Czech (2017), Italian (2018), German (2018) and Estonian (2018) were added. The English SkELL is used most extensively (150,000 page views per month)⁴. October, November, March and April are the most active months every year (which synchronizes well with academic year cycles).

SkELL corpora can be searched through a simple user interface⁵, which is a simplified version of the CQS Sketch Engine (Kilgarriff et al., 2004). In SkELL's interface, users

³ <https://www.sketchengine.co.uk/user-guide/user-manual/concordance-introduction/gdex/> (3 June 2019).

⁴ Statistics based on Google Analytics (3 June 2019).

⁵ <https://skell.sketchengine.co.uk> (3 June 2019).

can use Sketch Engine's most important features: concordances, word sketches and similar words (i.e. the thesaurus). Compared to more advanced CQSs, the output in SkELL's interface is limited: up to 40 sentences and similar words are shown; in word sketches only simplified grammar relation names are presented. The data accessed via SkELL's interface give a quick overview of examples, word distribution, collocations and the thesaurus.

2. Corpus sentences in the language portal Sõnaveeb

In Sõnaveeb⁶ (Wordweb) – a new language portal of the Institute of the Estonian Language – SkELL corpus data are displayed directly via API from two different Corpus Query Systems. Estonian sentences are queried from the etSkELL 2018 corpus via the CQS KORP API. Russian sentences are queried from the ruSkELL 1.6 corpus via the CQS Sketch Engine JSON API⁷.



Figure 1. Headword *Patarei vangla* 'Patarei prison' in Sõnaveeb.

This is the first time in Estonian lexicography that users get direct access to automatically-retrieved authentic sentences. The main motivation was to provide usage examples for headwords that do not have in their entries any example sentences

⁶ <https://sonaveeb.ee/> (3 June 2019).

⁷ <https://www.sketchengine.eu/documentation/json-api-query/?highlight=API> (3 June 2019).

compiled by lexicographers, as is the case with many terms, derived forms, compounds and multi-word expressions (MWEs). Figure 1 shows the MWE *Patarei vangla* ('Patarei prison') in Sõnavaab, with its definition, and the *hea teada* 'good to know' comment. SkELL sentences are the only usage examples of the word displayed in the bottom right corner of the page.

Figure 2 shows the Russian headword *планета* 'planet' in Sõnavaab with its domain label *ASTRONOOMIA* 'Astronomy' and definition. SkELL sentences are the only usage examples in Sõnavaab for the Russian headwords.

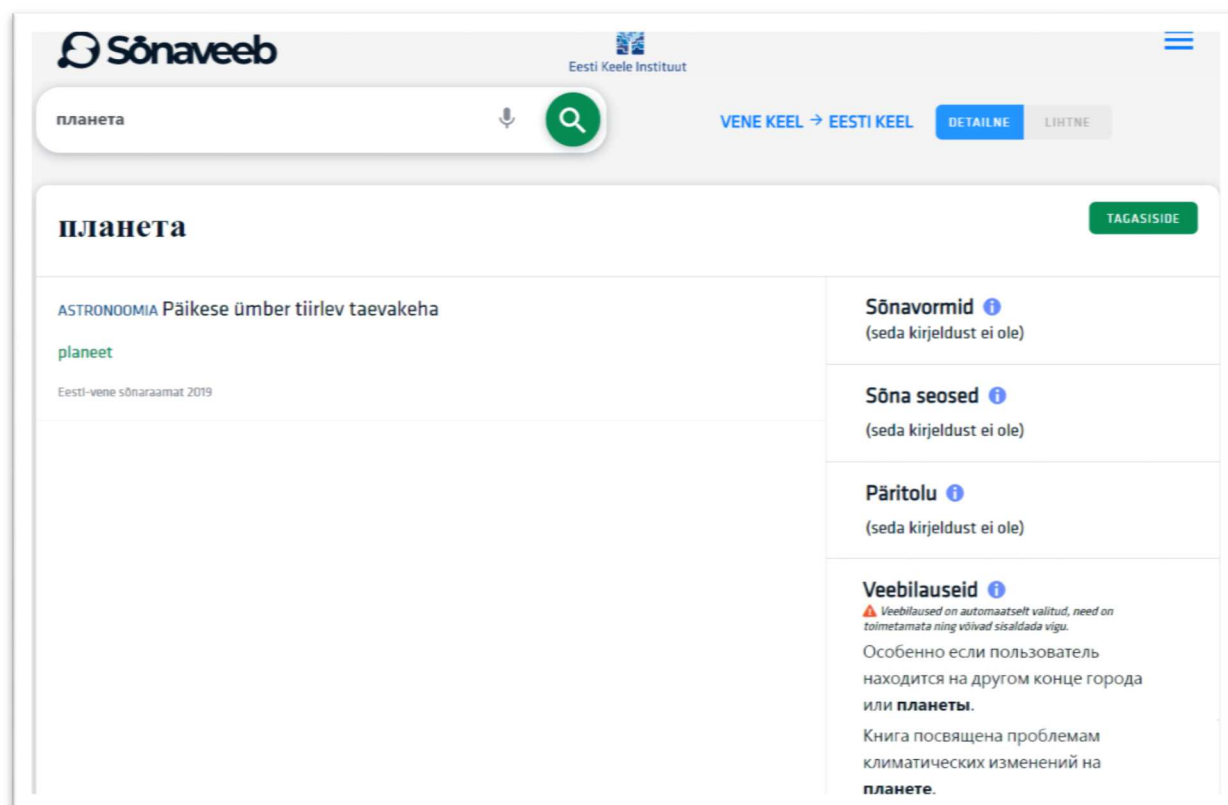


Figure 2. Headword *планета* 'planet' in Sõnavaab.

Up to 26 sentences are displayed for both languages. Only the first two sentences are displayed by default. The rest of the sentences can be opened by clicking the *Näita rohkem* 'Show more' option. For Russian, sentences are displayed according to GDEX scores: the highest scored sentences are shown first. This was also the case for Estonian sentences at first, but it soon became evident that in some cases the first two SkELL sentences can include errors, e.g. in lemmatization and POS-tagging (see Section 5 for more information). The problem of displaying the same (two) inaccurate sentence(s) for the same headword constantly was solved by displaying random sentences instead. This approach ensures that the presentation of corpus sentences is as dynamic as searches on the web, where the user will get a different set of web sentences each time consulting the same word. Obviously, this approach does not guarantee that the sentences will not contain errors. It still depends a lot on the quality of lemmatization

and morphological analysis.

In the next section, we describe the GDEX configurations for Estonian and Russian used for the compilation of etSkELL 2018 and ruSkELL 1.6 corpora. In Section 4 we present and analyse the results of the evaluation of the Estonian GDEX configuration. In Section 5 we address the main problems with displaying authentic corpus sentences and offer solutions.

3. Good Dictionary Examples (GDEX) and SkELL corpora

Good Dictionary Examples (GDEX) (Kilgarriff et al., 2008) is a function in the Sketch Engine (Kilgarriff et al., 2004) that ranks corpus sentences according to predefined criteria, assigning a numerical score (GDEX score) to each sentence, which separates good candidates from bad ones. This mechanism can be seen as a kind of filter, as it helps lexicographers to work with more relevant citations even though they have not been manually annotated. GDEX scores measure the lexical and syntactical features of the sentence and sort concordances according to how perfectly they meet all the relevant criteria. As a result, GDEX offers a list of example sentences with the best candidates presented first (at the top of the list).

In order to get a list of good example candidates, one needs to define a GDEX configuration that takes into account various criteria, e.g. sentence length and word frequencies (Kosem et al., 2019). The configuration can be seen as a formula that uses a number of parameters⁸:

```
formula: >

(50 * is_whole_sentence() * blacklist(words, illegal_chars))

+ 50 * optimal_interval(length, 10, 14)

* greylist(words, rare_chars, 0.1)

) / 100

variables:

illegal_chars: ([<|\]\[\>/\\^@]{\*\#\#>«"~_}())

rare_chars: ([A-Za-z0-9A-Я'.!,?)(;:-])
```

The classifier *is_whole_sentence* gives the highest value of 1 to ‘true’ sentences, e.g. ones that begin with a capital letter and end with a punctuation mark (a full stop,

⁸ The given example does not list all of the parameters; the whole description can be found in the manual <https://www.sketchengine.eu/syntax-of-gdex-configuration-files/>.

question mark or exclamation mark). The formula assigns values from 0 to 1. The most appropriate length of a sentence can be described as an argument in the *optimal_interval* classifier. In the above example, it varies from 10 to 14 and assigns the value 1 to such sentences. Along with a formula (which is a mandatory part of the configuration), a user can define the variables. *Illegal_chars* represents a list of characters that a sentence should not have, otherwise it will receive a low score. In the example we put restrictions on meaningless combinations of punctuation marks. *Rare_chars* represent a set of symbols that when they appear in a sentence will receive a penalty. For example, a Russian text in Cyrillic with many Latin characters is not seen as a good example.

GDEX configurations have been developed for several languages (see e.g. Kosem et al., 2019) and can be optimised using a special interface called the GDEX editor⁹ (Figure 3). The GDEX editor is meant for the evaluation of candidate sentences selected according to a GDEX configuration. This system evaluates sentences using two versions of the configuration and assigns two scores and ranks; based on this information, the configuration developers can mark apt sentences and thus assess which set of parameters is more suitable for the task. Writing these formal rules can be seen as an iterative process.

Old rank	Rank	Sentence	Old score	Score	Flag
1	1	На каком этапе брать деньги и сколько – на ваше усмотрение.	0.98	0.98	?
2	2	Помимо платы за рейс по обязательной таксе извозчик имел право брать чаевые до рубля включительно.	0.97	0.97	?
3	4	Поехали самостоятельно, у гида ничего брать не стали.	0.97	0.97	?
4	3	Парамонов . Ну так что, будем брать пример с наших уважаемых отечественных йогов?	0.97	0.97	?
5	6	Можно вместо всего этого одним молоком брать 1 раз в неделю ходить.	0.96	0.96	?
6	8	Нужно ли журналисту брать согласие субъекта на распространение его персональных данных?	0.96	0.96	?
7	9	Вашему партнеру страшно брать на себя ответственность за вас.	0.96	0.96	?
8	7	Для небольших производств, можно не брать, а использовать специальный инструмент для выполнения этой операции.	0.96	0.96	?
9	5	Застройщиков обяжут брать деньги на строительство не у частных инвесторов, а в банках и продавать готовое жилье.	0.95	0.97	?
10	10	Постараюсь брать с Вас пример и поехать знакомиться.	0.94	0.94	?

Figure 3. GDEX editor interface for evaluating candidate sentences for the Russian headword **брать** ‘to take’.

⁹ <https://gdexed.sketchengine.eu/> (3 June 2019).

Once the GDEX configuration is developed for a particular language, it can be used for the development of a SkELL corpus.

3.1 Parameters of good dictionary examples for Estonian and etSkELL

2018

The first version of the GDEX configuration for Estonian was developed in 2014. It was used for extracting example sentences into the Estonian Collocations Dictionary (ECD) database (Kallas et al., 2015). The ECD is aimed at learners of Estonian as a foreign or a second language at the upper intermediate and advanced levels (CEFR levels B2-C1).

The latest version of the GDEX for Estonian (GDEX 1.4) (Koppel, 2017) was used to compile the Estonian Corpus for Learners 2018 (etSkELL)¹⁰, which is used in both the etSkELL interface¹¹ and in the language portal Sõnaveeb. The process of corpus compilation was two-part: all sentences of the Estonian National Corpus 2017¹² (1.1 billion words) were first filtered using hard classifiers of GDEX 1.4, which resulted in filtering out about 83% of the sentences. The remaining 17% of the sentences were then scored using soft classifiers and compiled into the etSkELL 2018 corpus. The corpus consists of sentences from various media texts, fiction and scientific texts, Estonian Wikipedia and Estonian textbooks.

Table 1 gives an example of the volume of the etSkELL 2018 corpus. All occurrences of a token are accounted for in the structure size, while the lexicon size consists of a count of unique items.

etSkELL structure sizes		etSkELL lexicon sizes	
sentences	24,811,421	lower-case words	1,853,989
words	248,203,200	lower-case lemmas	813,498

Table 1. etSkELL 2018 corpus structure and lexicon sizes.

The parameters of GDEX 1.4 (Koppel, 2017) were fine-tuned based on the analysis of two datasets: one containing selected sentences from the examples of ECD offered by the original GDEX configuration, and one containing rejected or non-selected sentences

¹⁰ DOI: 10.15155/3-00-0000-0000-0000-07335L

¹¹ <https://etskell.sketchengine.co.uk/> (3 June 2019).

¹² DOI: 10.15155/3-00-0000-0000-0000-071E7L

of ECD. The most important parameters are described below.

- **Sentence length.** The average length of an example sentence in ECD is 9 to 10 tokens. Whereas three-word sentences are frequently used in Estonian and are very common in Estonian learners' dictionaries (based on the analysis of example sentences in the Basic Estonian Dictionary (BED) (2014)), the allowed sentence length is set at 4–20 tokens. The optimal interval is set at 6–12 tokens.
- **Word length.** The average word length of the sentences in ECD is six characters. As Estonian has a rich word formation system and some compounds can be quite long, e.g. *kiiruisutamismeistrivõistlused* 'speed skating championships' (30 characters), the maximum word length is set at 20 characters.
- **Low frequency words.** Two different classifiers in the Estonian configuration refer to low frequency words. Firstly, no word forms with a frequency of less than five are allowed in the examples. Following the example of Slovene configuration (Kosem et al., 2013), a classifier penalizing lemmas with a frequency of less than 1,000 was added. This classifier also helps to reduce the probability of complex compounds and rarer proper names occurring in the top ranked examples, which often happened with the previous configuration.
- **Number of elements in the sentence.** The average number of occurrences of certain elements (commas, numerals, proper names, adverbs, verbs, pronouns and conjunctions) in the sentences was determined. Each of the listed elements are grouped together in a classifier in GDEX 1.4, and they share the same weight due to shared characteristics. As a result, if a sentence includes more than one adverb, one pronoun, one proper name, one numeral, one conjunction, one comma, or two verbs, it gets penalized.
- **Sentence initial tags.** 54% of the selected sentences in ECD start with a substantive, 12% with an adjective, 11% with a pronoun and 8% with a verb. None of the selected sentences start with an interjection, abbreviation, genitive attribute or punctuation mark; hence sentences starting with the previously listed tags get heavily penalized.
- **Sentence initial words and word sequences.** Certain words and two-word sequences are not allowed to occur at the beginning of sentences. These are mostly anaphoric words and word sequences that refer to previous sentence(s) and are therefore context dependent, e.g. *pigem* 'rather', *teisisõnu* 'in other words', *seda enam* 'even more' and *teisest küljest* 'on the other hand'.
- **Non-finite constructions.** In order to avoid syntactically complex sentences, certain non-finite constructions are penalized. These constructions occur often, for example, in bureaucratic jargon and formal style, which can be difficult to understand for language learners.

- **Weights.** In GDEX 1.4, weights are assigned to soft classifiers. Optimal interval and word frequency have turned out to be the most distinguishing features of good examples, followed by penalizing anaphors (including certain pro-adverbs and demonstrative pronouns), so they are assigned the highest weight.

The results of the evaluation of the GDEX 1.4. are presented in Subsection 4.1.

3.2 Parameters of good dictionary examples for Russian and ruSkELL 1.6

The first version of Russian GDEX configuration GDEX 1.1. (Apresjan et al., 2016) was used for the compilation of ruSkELL 1.5¹³. However, the preliminary evaluation revealed that ruSkELL 1.5 still contained quite long sentences (up to 150 words). Some sentences did not begin with capital letters; there were also one-word sentences, and sentences containing obscene lexis. It was decided to develop the next version of GDEX configuration 1.2., which would partially solve these problems. GDEX 1.2 for Russian was used for the compilation of the ruSkELL 1.6 corpus, which has been implemented for querying sentences in the language portal Sõnaveeb. It was made on the basis of ruSkELL 1.5, and just re-sorted with the new GDEX 1.2 configuration, favouring average-length sentences with mid-frequency words which are more suitable for learners. Only the top 68 million sentences were used, providing the corpus with 975 million words, or 1,224 million tokens (see Table 2 for details).

ruSkELL 1.6 structure sizes		ruSkELL 1.6 lexicon sizes	
sentences	68,224,440	lower-case words	7,810,025
words	975,584,449	lower-case lemmas	7,403,227

Table 2. ruSkELL 1.6 corpus structure and lexicon sizes.

When writing GDEX configuration rules for Russian, we used several restrictions in order to get more precise results (i.e. more readable sentences).

The most important parameters are described below.

- **Sentence length.** When it comes to the selection of good dictionary examples,

¹³ <https://www.sketchengine.eu/ruskell-examples-and-collocations-for-learners-of-russian/> (3 June 2019).

readability should also be taken into account. According to the Russian Frequency Dictionary (Sharoff¹⁴), the average sentence length is 10.38 words. An analysis of dictionary entries in MAS (Jevgen'jeva, 1981-1984) showed that citations are longer and consist of 13 words. We came to the conclusion that the optimal sentence length would vary from 7 to 16 tokens and thus the allowed length was set at 6–20 tokens.

- **Blacklists.** We defined a number of variables that impose restrictions on sentence content. We filtered out emoticons and other combinations with punctuation marks (e.g. slashes, parentheses and quotes); they should not be used in “good” sentences. But not only characters can lower GDEX scores. An obscene lexicon should not be used in corpora for learners, and thus one of the blacklists includes such words.
- **Greylists.** Unlike the previous lists, greylists describe the elements whose presence in a sentence leads to lower scores. For Russian texts, we had to limit the usage of Latin characters and of words written in capitals. Such sequences may include trademarks, company or other proper names that would not make much sense to language learners. Also, the presence of digits can be seen as a drawback in a sentence.
- **Sentence initial words and word sequences.** Following the Estonian configuration (Koppel 2017), we also prepared a list of words and word sequences that are not allowed to appear at the beginning of sentences. On the one hand, as was stated above, such elements have mostly an anaphoric nature and refer to previous sentences. On the other hand, they can be a trace of a formal language that we prefer to avoid in the corpus, e.g. *vo-pervykh* ‘firstly’, *dalee* ‘then’, *todga* ‘hence’, *sleduet otmetit* ‘it should be noted’ and *kak sledstvie* ‘as a result’.

The evaluation of the GDEX 1.2 configuration for Russian has not been carried out yet, but will occur soon.

4. Users’ attitudes towards authentic corpus sentences

4.1 Evaluation of the GDEX 1.4. configuration for Estonian

In 2019 an evaluation (Koppel, 2019b) of the GDEX 1.4 output was completed by students of Tallinn University and the University of Tartu who speak Estonian at the B2–C1 proficiency levels, and by lexicographers working at the Institute of the Estonian Language. The purpose of the evaluation was to determine whether, according to the

¹⁴ <http://www.artint.ru/projects/frqlist.php> (3 June 2019).

two types of evaluators, authentic and unedited corpus sentences would be suitable example sentences in the language portal.

The GDEX 1.4 output evaluation consisted of two tasks. The first assessment task involved using the open source platform Pybossa¹⁵, which is used to carry out simple crowdsourcing projects and analyse the data collected. The aim of the first assessment task was to rate the suitability of sentences in general. The follow-up assessment task was performed in the Google Forms environment, and its purpose was to identify the reasons why the evaluators considered certain sentences not suitable for the dictionary.

For evaluation, we selected 40 random headwords from the ECD, the dictionary aimed at B2-C1 level learners: ten for each part of speech (substantive, verb, adjective and adverb). Then we took a random selection of sample sentences for each headword which included:

- one corpus sentence that meets the criteria of GDEX 1.4;
- one corpus sentence that does not meet the criteria of GDEX 1.4;
- one unfiltered corpus sentence;
- one example sentence compiled by a lexicographer.

All corpus sentences were taken from the Estonian National Corpus 2017; the dictionary example was taken from the Dictionary of Estonian 2019 (DicEst).

In the first assignment, there were 160 sentences in total, which were divided into four smaller tasks, in which each assignment contained all four types of sentences. The assessment task in Pybossa was set up so that each sentence had to be rated by five different lexicographers and five different language learners. The task was sent to seven lexicographers and 31 students, of whom five lexicographers and nine language learners responded (when one sentence had been rated by five different lexicographers and five different language learners, it was no longer displayed for the next evaluator).

Language learners were asked to assess the sentences based on their Estonian language skills; lexicographers were asked to assess if the sentences were suitable for a dictionary aimed at learners at the B2-C1 language proficiency levels.

One sentence was displayed to the evaluators at a time, preceded by the question “Is this sentence suitable as an example of the word X?” The response options were “yes”, “no” and “I don't know”. Neither the definition nor the source of the sentence was displayed to the evaluator (Figure 4).

¹⁵ <https://pybossa.com/> (3 June 2019).

Kas see lause sobib sõna **inimtühi** näitelauseks?

Inimtühjal tänaval võib keegi sulle sama nähtamatult, nagu on helkurvestita politseinik, joosta sebrale.

Jah Ei Ei oska hinnata

Lahendad praegu ülesannet number 1. Oled lahendanud 0 ülesannet 160 -st.
 Sa peaksid lahendama 40 ülesannet.
 Kui sul tekib mingeid kommentaare, siis täida tagasiside küsimustik.

Figure 4. Sentence assessment task in Pybossa

While the purpose of the first assessment task was to establish quantitatively whether different types of sentences were in the lexicographers' and language learners' opinions suitable example sentences in a learner's dictionary, the purpose of the follow-up survey was to identify why evaluators considered some good corpus sentences not suitable and some bad corpus sentences suitable. For this reason the evaluators were asked to re-assess three types of sentences in the follow-up survey:

- Corpus sentences that met all the criteria of GDEX 1.4 but most evaluators did not think were suitable (or they did not know).
- Corpus sentences that did not meet the criteria of GDEX 1.4 but most evaluators thought were suitable (or they did not know).
- Dictionary examples that most evaluators did not think were suitable (or they did not know).

The request to participate in the follow-up survey was sent to the same evaluators who had participated in the first assessment task (five lexicographers and nine language learners), and we received replies from five lexicographers and five learners. In the follow-up survey, lexicographers were asked to re-evaluate 18 of the previously mentioned three types of sentences, and language learners were asked to re-evaluate 20 sentences, of which 11 sentences overlapped.

The final results of the two assessment tasks showed that, according to most lexicographers and language learners, as many as 96% of the dictionary examples and 85% of corpus sentences chosen as good examples by GDEX 1.4. were considered to be suitable example sentences. Only 6% of the sentences that were discarded by GDEX 1.4 were considered suitable, meaning that 94% of the bad candidates had been filtered out successfully. As for unfiltered corpus sentences, 60% of those were considered

unsuitable. When evaluators were asked their reasons for considering a sentence unsuitable, the most common arguments were that the sentences included anaphora and hence needed more context, or that the sentences were colloquial, too long or too short.

The results of the evaluation show that even more attention should be paid to anaphora, either by raising the penalty or by adding more words to the greylist. It also makes sense to invest more effort into figuring out the ideal range of sentence length, as short sentences tend to lack context and long sentences were mostly considered unsuitable.

4.2 User feedback on the presentation of corpus sentences in Sõnaveeb

Dictionary users are accustomed to the fact that all data presented in a dictionary are controlled and edited by a lexicographer, and are hence correct. In contrast, corpus sentences in Sõnaveeb are authentic, unedited and may include errors. Since Sõnaveeb's launch in February 2019, the lexicographers working in the Institute of the Estonian Language have received feedback from users in which they have said that they find some of the corpus sentences are inappropriate or incorrect. At the beginning, no clear warning of the authenticity of the sentences was displayed by default. The user could only read the information about the source of the sentences (etSkELL 2018 corpus/ ruSkELL 1.6 corpus) by moving the cursor over the information button. After receiving user feedback, the editors of Sõnaveeb decided to use the same strategy as in Merriam-Webster's¹⁶ and Collins'¹⁷ dictionary portals, and added an explicit note saying that the sentences were chosen automatically, were unedited and might contain errors. The user feedback also indicated that users, especially language professionals, want to see the metadata of each sentence, e.g. author, title, and year. This information would indicate whether the word is archaic, colloquial, to which genre it belongs to, etc.

5. Problems and possible solutions

Several problems have arisen with displaying authentic corpus sentences, and it is difficult to eliminate them with the help of a tool operating solely on a rule-based method. Some of these problems are language independent, and some are language specific. The most typical problems are described below.

1. Polysemous words. When choosing the sentences, the polysemy of the headword is not taken into account. For example, the query for the Estonian polysemous headword *leht* ('newspaper', 'leaf', 'webpage') provides sentences in which the word occurs in different meanings.

¹⁶ <https://www.merriam-webster.com/> (3 June 2019).

¹⁷ <https://www.collinsdictionary.com/> (3 June 2019).

2. Lexical and POS-homonymy. When choosing the sentences, the homonymy of the headword is not taken into account. For example, the query for Estonian homonymous headword *tamm* ('aok'; 'dam'; 'king') provides sentences in which the word occurs in different meanings.
3. Lemmatization and POS-tagging errors. These arise particularly in the case of grammatical homonymy. For example, the query for the Estonian grammatical homonym *joon* (*joon*-n 'stripe-Substantive', *joon*-v '(I) drink-Verb', provides sentences in which this word occurs as a noun in nominative case, as well as the first person singular of the verb *jooma* 'to drink' in present indicative. The Russian grammatical homonym *дома* (*дом*-n 'house- Substantive', *дома*-d 'at home- Adverb') can be either the genitive singular or the nominative/accusative plural of the noun *дом* 'house' or the adverb *дома* 'at home'. The query gives examples for both lemmata without distinguishing between them.
4. Machine-translated sentences. Machine-translated sentences get crawled from bilingual web pages that match the predefined parameters of GDEX but may sometimes be ungrammatical.
5. Absence of information for low frequency words. It is difficult to find example sentences for low frequency words. For example, the noun *kalla* 'arum lily' does not appear in the etSkELL 2018 corpus.
6. Multiword expressions. The selection includes sentences where the headword is actually part of an MWE, e.g. in the output for the keyword *tulema* 'to come', sentences with the MWE *toime tulema* 'to manage' might appear.
7. A certain type of problem comes from the source texts (either mistakes, typos or errors of recognition), e.g. the Russian *на* instead of the preposition *на* 'on' (here we should note that in Cyrillic they have similar graphic forms: *на* vs *на*). Hence we can try to filter out such examples, defining a separate blacklist for typical errors.
8. Along with Russian, there are other Slavic languages (Ukrainian and Belorussian) that use Cyrillic. Although the corpus was cleaned up with the right encoding, it still has irrelevant examples in other languages. One of the possible solutions is to prepare a list of frequent words in Cyrillic that are not Russian, in order to filter out such sentences.

Finding suitable example sentences for different meanings of polysemous words could have been facilitated if the corpus had been semantically annotated and queries could be based on using the same semantic types as used in the dictionary. The semantic types developed by Margit Langemets (2010) that have been used in the compilation of BED and DicEst could be applied for the Estonian language.

An additional way of solving the problem of polysemy and lexical homonymy is to consider the collocations of the headword, so that the example sentences with the most frequent collocations appear in the output. For Estonian, the database of the Estonian Collocations Dictionary could be incorporated. For example, if the headword *tamm* ('oak'; 'dam'; 'king') has three homonyms and the user chooses the meaning of 'dam', the sentences with the collocations *tamm puruneb* 'the dam collapses' and *tammi ehitama* 'to build a dam' would appear first.

One other possible way of solving the problem of POS-homonymy (e.g. *noor-a* 'young-Adjective' and *noor-s* 'young person-Substantive'), is to query sentences via API using lempos instead of lemma. This is already done in Sõnaveeb. It helps in cases of POS homonyms, but becomes an obstacle in the case of errors in lemmatization and POS-tagging. It is a very frequent problem, especially in the case of grammaticalization and lexicalization, when a morphological analyser defines lemma and POS on the basis of an outdated lexicon. For example, the headword *tasuta-d* 'for free-Adjective' used to be analysed as the substantive *tasu* 'fee' in abessive case in dictionaries, and is still analysed as a substantive by POS-taggers. But all dictionaries published in Sõnaveeb consider it to be an adjective. Since we query sentences via API using lempos, the system does not find such a lempos *tasuta-a* 'for free-Adjective' and the query shows completely erroneous results.

Detection of machine-translated texts seems to be a topic in its own right (see e.g. Aharoni et al., 2014; Nguyen-Son et al., 2019 for more). In order to avoid automatically translated sentences occurring in the output, machine generated texts should be automatically detected and rejected at the stage of corpus crawling. The best way to do that is to combine multiple approaches. Firstly, there is a need to identify problematic sites and remove them completely from the corpus. If such sites are already known from previous corpus crawling, the crawler can be instructed to avoid them altogether in crawling the new corpus. Secondly, the crawling should start from trustworthy sources. It is also important to keep track of the distance of a site from these trustworthy sources¹⁸. According to our experience, sites having too long names could be avoided¹⁹. Documents coming from web sites not available one month after crawling from the corpus should be removed²⁰. It would also help to build a classifier that recognizes computer-generated texts. For languages where syntactic analysis is possible, it can be used to reveal suspicious sentences. But even in this case some problems remain unsolved: 1) a large part of human-produced text uses unorthodox

¹⁸ For example, www.eki.ee is a seed with distance 0; a site referenced by www.eki.ee, has a distance of 1; a site referenced by sites with distance n but not sites closer to seeds than n has a distance of $n + 1$.

¹⁹ Too long is ≥ 40 characters (or ≥ 50 characters to reduce false positives) according to our unpublished experience with reviewing the content of random sites with long names.

²⁰ According to our unpublished experience with checking sources of computer-generated text in the corpus, the life of spam sites is short. The reason may be they become useless once blacklisted by search engines.

syntax, so we don't know what "faulty" is, and 2) neural machine translation produces syntactically perfect sentences, and it is difficult to detect them.

Absence of information for low frequency words is a possible bias in the corpus crawling procedure. It makes sense to combine queries from different sources, e.g. when a word is not found in the (smaller) learner corpus, the query will be made on the basis of a (large) general corpus (e.g. Estonian NC).

In addition, the origin of source texts must be taken into account when creating a new learner corpus: this would make it possible to give priority to sentences from Estonian Wikipedia and periodicals rather than sentences from blogs and forum posts.

In addition, it is obvious that one corpus cannot satisfy the needs of all users. One possibility is to apply additional filters (e.g. vocabulary lists of different language proficiency levels). For Estonian, special GDEX configurations aimed at different CEFR (Common European Framework of Reference) levels of Estonian L2 proficiency (Koppel, 2019a) and CEFR vocabulary lists (Kallas & Koppel, 2018a, 2018b, 2018c) have been developed, but have not yet been used to compile SkELL corpora for different CEFR levels.

6. Conclusions and Future Work

In this paper we analysed different issues that are connected with the quality and presentation of authentic corpus sentences as a part of an (academic) language portal. Most recent dictionaries in Estonia are corpus-based but have traditionally not included authentic corpus data in their online versions. Sõnaveeb is the first of its kind where its users can read authentic corpus sentences without leaving the language portal's interface.

The important question is what kind of corpora are more suitable for this purpose. In the paper we argue that one type of corpora that might be used is Sketch Engine for Language Learning, or SkELL corpora. SkELL corpora were initially intended for learning purposes but they can be seen as a source of good dictionary examples, as they contain only sentences which are ranged as 'good' according to the GDEX system. In order to compile such corpora, Good Dictionary Examples (GDEX) (Kilgarriff et al., 2008) configurations for Estonian (GDEX 1.4.) and for Russian (GDEX 1.2) were developed and later used for the compilation of etSkELL 2018 and ruSkELL 1.6 corpora, which are used as sources of authentic corpus sentences in the new language portal Sõnaveeb. Estonian sentences are queried from the etSkELL corpus via the Corpus Query System Korp API. Russian sentences are queried from the ruSkELL 1.6 corpus via the Corpus Query System Sketch Engine JSON API.

The evaluation of the GDEX 1.4 configuration for Estonian showed that as many as 85% of corpus sentences chosen as good examples by GDEX 1.4 were also evaluated as "good" by users (lexicographers and Estonian language learners). Only 6% of the

sentences that were discarded by GDEX 1.4 were considered suitable, meaning that 94% of the bad candidates had been filtered out successfully. The main reasons for considering a corpus sentence unsuitable were anaphora, colloquialisms and sentence length (too long or too short). User feedback has also revealed that some users get confused if they see inappropriate or incorrect sentences as a part of the portal. For this reason it was decided to add a clear note that says that the sentences are chosen automatically and that they may contain errors. Some users also pointed out that they need the description of a source of sentences, e.g. author, title, and year. These parameters help to understand whether a word is archaic, colloquial, which genre it belongs to, etc.

However, there are also problems that are difficult to eliminate solely by using the GDEX system. These are lemmatization and POS-tagging errors in corpus data, homonyms, polysems, low frequency words, sentences with inappropriate content, machine-translated sentences etc. Finding suitable example sentences for different meanings of polysemous words could have been facilitated if the corpus had been semantically annotated and queries could be based on using the same semantic types as used in the dictionary. One way of solving the problem of polysemy and homonymy is to consider the headword's typical collocates, so that the example sentences with the most frequent collocates of the headword appear first in the output. The problem of POS homonymy can be solved by querying sentences via API on the basis of lemmas instead of lemmas.

Some problems with mistakes and inconsistencies in corpora can be solved during the compilation. Various reference databases can be applied, e.g. a database of common spelling mistakes and a database of frequent foreign or dialectal lexis.

For the purpose of customization, there is a need to compile special SkELL corpora for each CEFR (Common European Framework of Reference) level. Special GDEX configurations aimed at different CEFR levels of Estonian L2 proficiency (Koppel, 2019a) have already been developed. This will allow us to show different sets of sentences for users with different Estonian L2 proficiencies.

In order to facilitate the development of GDEX configurations and to decrease the number of incorrect and/or inappropriate sentences shown in Sõnaveeb, we plan to use crowdsourcing methods. Users will be given an opportunity to vote on each sentence via the portal, and after a sentence receives a certain number of downvotes, the system would not show them again, while the upvoted sentences could be displayed first. This approach will help to create two datasets – one with upvoted sentences and the other with downvoted sentences – which then could be used for the development of learning algorithms (as patterns or features). Implementing crowdsourcing would also make the language portal more interactive.

7. Acknowledgements

The creation and development of the portal was funded by the Digital Focus Program of the Ministry of Education and Research (2018–2021) and by EKI-ASTRA program (2016–2022). The creation of the dictionary and terminology database Ekilex was funded by EKI-ASTRA program (2016–2022). Software development has been provided by OÜ TripleDev.

This work has been partly supported by the Ministry of Education of CR within the LINDAT-Clarín infrastructure LM2015071. This publication was written with the support of the Specific University Research provided by the Ministry of Education, Youth and Sports of the Czech Republic.

This work was supported by the grant of the President of Russian Federation for state support of scholarly research by young scholars (Project No. MK-2513.2018.6).

The research received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

8. References

- Aharoni, R., Koppel, M. & Goldberg, Y. (2014). Automatic detection of machine translated text and translation quality estimation. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 289-295.
- Apresjan, V., Baisa, V., Buivolova, O., Kultepina, O. & Maloletnjaja, A. (2016). "RuSkELL: Online Language Learning Tool for Russian Language." *Proceedings of the XVII EURALEX International Congress*, Tbilisi, Georgia, pp. 292-299.
- Baisa, V. & Suchomel, V. (2014). SkELL: Web Interface for English Language Learning. *Eighth Workshop on Recent Advances in Slavonic Natural Language Processing*. Brno: Tribun EU, 2014, pp. 63-70.
- BED = *Eesti keele põhisõnavara sõnastik 2019* (1. trükk 2014). [The Basic Estonian Dictionary 2019, BED] Eesti Keele Instituut. Sõnaveeb 2019 [Wordweb 2019]. Available at: <https://sonaveeb.ee> (3 June 2019).
- Cook, P., Rundell, M., Lau, J. H., & Baldwin, T. (2014). Applying a word-sense induction system to the automatic extraction of diverse dictionary examples. *Proceedings of the XVI EURALEX International Congress*, pp. 319-328.
- ECD = *Eesti keele naabersõnad 2019*. [The Estonian Collocations Dictionary 2019] Eesti Keele Instituut. Sõnaveeb 2019 [Wordweb 2019]. Available at: <https://sonaveeb.ee> (3 June 2019).
- DicEst = *Eesti keele sõnaraamat 2019*. [The Dictionary of Estonian 2019] Eesti Keele Instituut. Sõnaveeb 2019 [Wordweb 2019]. Available at: <https://sonaveeb.ee> (3 June 2019).
- etSkELL 2018 = Sketch Engine for Estonian Language Learning 2018. Accessed at:

- <https://etskell.sketchengine.co.uk/> (3 June 2019)
- GDEX editor*: Accessed at <https://gdexed.sketchengine.eu/> (3 June 2019)
- Kallas, J., Kilgarriiff, A., Koppel, K., Kudritski, E., Langemets, M., Michelfeit, J., Tuulik, M. & Viks, Ü. (2015). Automatic generation of the Estonian Collocations Dictionary database. In I. Kosem et al. (eds.) *Proceedings of the eLex 2015 conference*, Herstmonceux Castle, United Kingdom, pp. 1-20.
- Kallas, J., Koeva, S., Kosem, I., Langemets, M. & Tiberius, C. (2019). *Lexicographic practices in Europe: a survey of user needs*. Available at: https://elex.is/wp-content/uploads/2019/02/ELEXIS_D1_1_Lexicographic_Practices_in_Europe_A_Survey_of_User_Needs.pdf (3 June 2019).
- Kallas, J. & Koppel, K. (2018a). *Eesti keele B1-taseme sõnavara*. [B1 Estonian Vocabulary List.] Eesti Keele Instituut.
- Kallas, J. & Koppel, K. (2018b). *Eesti keele A2-taseme sõnavara*. [A2 Estonian Vocabulary List.] Eesti Keele Instituut.
- Kallas, J. & Koppel, K. (2018c). *Eesti keele A1-taseme sõnavara*. [A1 Estonian Vocabulary List.] Eesti Keele Instituut.
- Kilgarriiff, A., Husák, M., McAdam, K., Rundell, M. & Rychlý, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus. *Proceedings of the 13th EURALEX International Congress*. Barcelona: Institut Universitari de Linguística Aplicada, Universitat Pompeu Fabra, pp. 425-432.
- Kilgarriiff, A., Rychlý, P., Smrž, P. & Tugwell, D. (2004). The Sketch Engine. *Proceedings of the 11th EURALEX international congress*. Lorient, France: Université de Bretagne Sud, pp. 105-115.
- Koppel, K. (2017). Heade näitelausete automaattuvastamine eesti keele õppesõnastike jaoks [Automatic detection of good dictionary examples in Estonian learner's dictionaries]. *Eesti Rakenduslingvistika Ühingu aastaraamat* [Papers in Applied Linguistics], 13, pp. 53-71. DOI:10.5128/ERYa13.04.
- Koppel, K. (2019a). Eesti keele kui teise keele õpikute lausete analüüs ja selle rakendamine eri keeleoskustasemete sõnastike näitelausete automaatsel valikul. [Parameters of CEFR-graded coursebook sentences and their use for automatic detection of good dictionary examples]. *Eesti Rakenduslingvistika Ühingu aastaraamat* [Papers in Applied Linguistics], 15, pp. 99-119. DOI:10.5128/ERYa15.06.
- Koppel, K. (2019b). Leksikograafide ja keeleõppijate hinnangud automaatselt tuvastatud korpuslausete sobivusele õppesõnastiku näitelauseks [Suitability of automatically selected example sentences for learners' dictionaries as tested on lexicographers and language learners]. *Lähivõrdlusi. Lähivertailuja*, 29, [forthcoming].
- KORP*: Accessed at: <https://korp.keeleressursid.ee/> (3 June 2019).
- Kosem, I., Gantar, P. & Krek, S. (2013). Automation of lexicographic work: An opportunity for both lexicographers and crowd-sourcing. In I. Kosem et al. (eds.) *Proceedings of the eLex 2013*, Tallinn, Estonia, pp. 17-19.
- Kosem, I., Koppel, K., Zingano Kuhn, T., Michelfeit, J. & Tiberius, C. (2019).

- Identification and automatic extraction of good dictionary examples: the case(s) of GDEX. *International Journal of Lexicography*, 32(2), pp. 119-
<https://doi.org/10.1093/ijl/ecy014>.
- Langemets, M. (2010). *Nimisõna süstemaatiline polüseemia eesti keeles ja selle esitus eesti keelevaras*. [Systematic polysemy of nouns in Estonian and its lexicographic treatment in Estonian language resources] Tallinn: Eesti Keele Sihtasutus.
- LDOCE = Longman Dictionary of Contemporary English. Accessed at: <http://ldoce.longmandictionariesonline.com/main/Home.html> (3 June 2019).
- Nguyen-Son, H-Q., Thao, T. P., Hidano, S. & Kiyomoto, S. (2019). Detecting Machine-Translated Paragraphs by Matching Similar Words. arXiv preprint arXiv:1904.10641.
- ruSkELL1.6*: <https://www.sketchengine.eu/ruskell-examples-and-collocations-for-learners-of-russian/> (3 June 2019).
- Sketch Engine*. Accessed at: <https://www.sketchengine.eu/documentation/api-documentation/> (3 June 2019)
- Sõnaveeb* = Sõnaveeb 2019 [Wordweb 2019]. Accessed at: <https://sonaveeb.ee> (3 June 2019)
- Wordnik*: Accessed at: <https://www.wordnik.com/> (3 June 2019).

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



A Corpus-Based Lexical Resource of Spoken German in Interaction

Meike Meliss¹, Christine Möhrs²,

Maria Ribeiro Silveira², Thomas Schmidt²

¹ Leibniz-Institut für Deutsche Sprache, Mannheim /
Universität Santiago de Compostela (Spanien)

² Leibniz-Institut für Deutsche Sprache, Mannheim

E-mail: meliss@ids-mannheim.de / meike.meliss@usc.es, moehrs@ids-mannheim.de,
silveira@ids-mannheim.de, thomas.schmidt@ids-mannheim.de

Abstract

This paper presents the prototype of a lexicographic resource for spoken German in interaction, which was conceived within the framework of the LeGeDe-project (LeGeDe=Lexik des gesprochenen Deutsch). First of all, it summarizes the theoretical and methodological approaches that were used for the initial planning of the resource. The headword candidates were selected by analyzing corpus-based data. Therefore, the data of two corpora (written and spoken German) were compared with quantitative methods. The information that was gathered on the selected headword candidates can be assigned to two different sections: *meanings* and *functions in interaction*.

Additionally, two studies on the expectations of future users towards the resource were carried out. The results of these two studies were also taken into account in the development of the prototype. Focusing on the presentation of the resource's content, the paper shows both the different lexicographical information in selected dictionary entries, and the information offered by the provided hyperlinks and external texts. As a conclusion, it summarizes the most important innovative aspects that were specifically developed for the implementation of such a resource.

Keywords: online lexicography; spoken German; corpus-based

1. Introduction

The lexicographic resource described in this article in its conception and implementation was conceived and created in the research project “Lexik des gesprochenen Deutsch” (=LeGeDe) between 2016 and 2019 at the Leibniz Institute for the German Language (IDS) in Mannheim¹. The cooperation between the Department of Pragmatics and the Department of Lexical Studies at the IDS enabled a connection

¹ The resource is created within the framework of the third-party funded research project LeGeDe financed by the Leibniz Association (Leibniz Competition 2016, Funding line: 1: Innovative projects). Project website: <http://www1.ids-mannheim.de/lexik/lexik-des-gesprochenen-deutsch.html>.

of the corresponding professional competence necessary for the creation of a corpus-based lexicographic resource of spoken German in interaction during the project period. The creation of such a corpus-based electronic resource of spoken German, based on the one hand on research on the peculiarities of spoken vs. written language use, and on the other hand on important experience in the field of electronic lexicography (cf. Klosa & Müller-Spitzer, 2016), was the project's main objective. Both from the research's point of view on spoken language and from a lexicographical perspective, a completely new form of lexicographic language description and presentation needed to be developed. Furthermore, it was necessary to generate novel lexicographic types of information with audio-features that refer to the function of lexical units in interactional contexts, for which so far hardly any lexicographical models exist. The lexicographical prototype is intended to primarily serve as a knowledge repository and vocabulary documentation (<https://www.owid.de/legede/>). The resource addresses scientists, interactional linguists, and lexicologists as its primary target group (cf. Meliss et al., 2018b, 2019). Nevertheless, we are convinced that learners of German can also benefit from the resource if the experts take the corresponding intermediate position. For this purpose, quantitative and qualitative methods were developed with which the specifics of the spoken-language lexicon of German could be identified, analysed, and prepared for lexicographical application on the basis of oral corpora created at the IDS (cf. the program area "Oral corpora").

In this paper we present the most important challenges and results of the LeGeDe-project. Therefore we introduce in section 2 the project's background (research questions, aims, and objectives), and in section 3 we show the relevant information about our corpus-based database. In section 4 we present some relevant results of two empirical studies on expectations we carried out at the beginning of the project. The information on lexicographical implementation is presented in section 5, using illustrative examples. In our concluding remarks (cf. section 6), we emphasize the innovative aspect of the LeGeDe-resource and give a brief outlook on further research and work areas.

2. Research questions and objectives

The LeGeDe-project is based on the following four main assumptions and observations:

- (i) There are differences at several linguistic levels between spoken and written German. With regard to the lexicon, the divergences can have an effect on both the lexical inventory and the relation with its form, meaning, and use (cf. Deppermann et al., 2017; Fiehler, 2016; Imo, 2007; Schwitalla, 2012).
- (ii) The way existing dictionaries codify the characteristics of the spoken German lexicon is deficient in several ways (cf. e.g. Meliss, 2016; Meliss et al., 2019; Moon, 1998; Trap-Jensen, 2004). There are currently hardly any corpus-based lexicographic projects that aim to develop a lexicon of spoken language. Only one small project on interjections (cf. Hansen & Hansen, 2012) was

carried out on Danish. The results of two LeGeDe-surveys on the expectations and requirements of a lexicographic resource for the specifics of spoken German (cf. Meliss et al., 2018b, 2019), carried out in cooperation with the project “Empirische Methoden”, confirm that the lexicographical codification of spoken language and its interactional features are not satisfactorily taken into account in the currently existing dictionaries (cf. Meliss, 2016: 195; Eichinger, 2017: 283). Despite some recent advances in corpus-based lexicography of spoken language (cf. Verdonik & Sepesy Maučec, 2017; Hansen & Hansen, 2012; Siepmann, 2015), experience with spoken language data in lexicography has so far been rather rare. Therefore, the LeGeDe-resource can hardly rely on existing models that could serve as guidance for the compilation of a suitable list of headwords and for the lexicographical modelling and implementation.

- (iii) The need for information on typical spoken vocabulary has increased in general and in various areas of application, e.g. in learning and teaching areas (especially in secondary education and in the areas of German as a foreign and/or second language) as well as in the research and publication area in connection with the production of suitable study materials (cf. Handwerker et al., 2016; Imo & Moraldo, 2015; Meliss & Möhrs, 2018; Moraldo & Missaglia, 2013; Reeg et al., 2012; Sieberg, 2013). For example, in the “Common European Framework of Reference for Languages” (=GeR), among other items on the assessment grid for oral communication and the parameter “interaction” for level C1, it was explicitly noted that a learner should be able to choose an appropriate turn from a repertoire of means of discourse in order to make his utterance appropriate (cf. Trim et al., 2001: 37).
- (iv) In addition, the results of the empirical studies, carried out in the LeGeDe-project show that more than 70% of L1 and L2 speakers of German expressed a need for a dictionary on specifics of spoken German. This observation confirms the basic assumption of an increasing demand for such a resource.

These basic assumptions are the starting points for conceptual considerations in order to develop our lexicographical resource and lead to the following essential theoretical, methodological, and application-oriented aspects, which arose when dealing with the topic in the project work:

- development of quantitative and qualitative methods to identify spoken-language lexical elements and their specific characteristics in interactional contexts in comparison to the lexicon of written language (cf. Meliss & Möhrs, 2017),
- preparation of a list of headword candidates and selection of suitable lemmas for the prototype of the LeGeDe-resource (cf. Meliss et al., 2018a),
- development of further (corpus-)linguistic methods for analysing and structuring

spoken language data, also for structuring automatically generated corpus-based data (cf. Möhrs et al., 2017),

- determination of the peculiarities of spoken language usage at different levels (form, content/function, conversational setting etc.), in our project with a focus on lexical specifics,
- development of innovative forms of lexicographical information, which refer to the function of lexical units in interactional contexts (taking into account transcripts and their associated audios).

3. Database of the LeGeDe-project

The studies on the research object of the LeGeDe-project are carried out exclusively on the basis of the “Research and Teaching Corpus of Spoken German” (=FOLK: cf. Schmidt, 2014a; Kupietz & Schmidt, 2015). FOLK is the largest corpus of conversational German, which was developed at the IDS and is integrated in the “Database for spoken German” (=DGD: cf. Schmidt, 2014b). FOLK primarily contains authentic data from interactive conversations (cf. Schmidt, 2017). Included are conversation recordings and transcripts (partly also video recordings) from German-speaking regions in various private, institutional, and public contexts. The data can be categorized by the following characteristics: oral media, authentic, spontaneous, mostly of the standard language, and up-to-date. Currently, FOLK is available in DGD version 2.12 with almost 250h/2.4 million tokens and 306 different speech events.² As a corpus analysis tool, the DGD offers a variety of possibilities for indexing oral data according to linguistic and interactional characteristics, and is constantly further developed and equipped with innovative corpus technology functionalities. Structured token searches can be realized via the user interface and searched via four annotation levels (cGAT transcript, normalization, lemmatization, PoS). In addition, metadata on speakers and on the conversation event can be retrieved for the conversations. The size of the corpus, the data it contains from authentic interaction, and the annotation of the data provide a reliable basis for lexicological and interactional analysis.

Since 2018, the use of the tool *Lexical Explorer* (cf. Batinić-Lemmenmeier, in press), an application developed during the LeGeDe-project, allows further access to FOLK as well as to the GeWiss (“Gesprochene Wissenschaftssprache”) corpus. With this tool, quantitative corpus data on spoken German can be explored with the help of frequency tables regarding the distribution across word form variation, co-occurrences, and metadata.

² The samples analyzed in the LeGeDe-project were based on DGD Version 2.11.

4. Empirical studies: expectations on a dictionary of spoken German

Two empirical studies were carried out at the beginning of the LeGeDe-project. The main goal of these studies was to shed light on people's expectations on the planned lexicographical online-resource. In the first study, selected experts were polled in the form of a guided interview. In the second, a broader online survey was conducted, which aimed to reach a wider range of potential users.³ With our two conducted surveys (interview and online survey) we intended to learn about expectations with regard to as many different lexicographical aspects as possible. In addition, sociodemographic data were also collected, and questions concerning the personal handling and use of (online) dictionaries together with the specific handling of the spoken-language lexicon were asked.

In our first study, we interviewed 17 experts from different linguistic areas. Each interview consisted of 30 questions mainly in an open question format, so the analysis of the greater part of the data was performed with qualitative methods. A smaller number of questions were presented in a closed format, so these data could be analysed with quantitative methods and be compared to answers from the online survey. Nevertheless, when viewing the results of the interview and especially when comparing them to data from the online survey, it must be considered that these are data from only 17 participants. The purpose of the online survey was also to ask for the opinions of a wider range of potential users and beneficiaries (e.g. linguists, teachers of German, domestic or abroad) of the planned resource. For this questionnaire, which contained 35 questions, we mainly used closed question formats. Altogether 333 participants completed the online survey.

In the following sections we present particular results relevant for basic considerations for the implementation of the LeGeDe-resource as well as results directly concerning it.

4.1 Target group of the planned resource

The question of the target group is fundamental for the lexicographical implementation of the collected data. Since the LeGeDe-resource initially functions as a knowledge store and vocabulary documentation, the presentation of the data is primarily geared towards a scientifically interested group of users (including conversation researchers, interaction linguists, corpus linguists, lexicologists, lexicographers).

However, the results of our empirical surveys on the question *For which target group*

³ In Meliss et al. (2018b, 2019), the results from both studies are summarized either from a general or from a L1 vs. L2 perspective.

could a dictionary of spoken German be of particular interest?⁴ have also shown that users in certain learning situations – especially in speech production situations – could benefit from the LeGeDe-resource. For this purpose it would be necessary that the experts (scientists, teachers, etc.) take on a corresponding mediating position. Based on the data provided by the LeGeDe-resource, language teaching material for the concrete treatment of specific lexical phenomena in spoken interaction could be developed from an application-oriented perspective for German as a foreign or mother tongue language (cf. Meliss et al., 2018b: 132; 2019: 116).

4.2 Headword candidates

We first look at the results for the following question of the online survey: *What kind of headwords would you expect in a dictionary of spoken German?* Different observations result from the answers given by the test subjects: (i) Most of the online survey respondents (87.8%) expect headwords which have a different meaning and functionality in spoken interaction than in written use. (ii) In a dictionary of spoken German, respondents to the online survey expect headwords to have a formulaic use (79.5%) as well as headwords with a special combination potential (e.g. patterns, specific units, etc.; 74.6%). (iii) Headwords that are exclusively spoken (77.7%) and those that occur particularly frequently in spoken interaction (71.6%) are also desired by the test subjects of the online survey. (iv) Slightly more than half of the online survey participants also expect lexical units that can be characterized by formal phonetic contraction (57.5%). A look at the respondents' responses to "Miscellaneous" shows that, among other things, headwords with a different spectrum of linguistic variation are also desired.

According to the experts' assessment, lexical units with a different combination potential in spoken vs. written language are the most desired headwords (94.1%). In their opinion, this includes constructions, lexical expressions, syntagmatic combinations, formulas, etc., as well as multiword lemmas. The experts also listed lexical units with differences in meaning or function as important headwords.

4.3 Information on the headword candidates

This section looks at the answers of our online survey to the following question: *In your opinion, what information should be offered in a dictionary of spoken German?* From the five different answer possibilities to this question: *Definitely* (1), *Useful, but not absolutely necessary* (2), *Not useful, but nevertheless desirable* (3), *Unnecessary* (4) to *I don't know* (5), options 1-4 are shown in Fig. 1.

⁴ The interview and the survey were conducted in German. Questions and answers are translated into English for better comprehension.

The results of the online survey – visualized here by the median and the arithmetic mean (=AM) – show that a broad spectrum of information, namely on pronunciation, meaning/function in context, formal peculiarities and features in combinatorics and word formation, together with the range of corpus data, metadata on the conversation situation, and comparative information (written vs. spoken) was equally evaluated by the respondents with the answers *Definitely* or *Useful, but not absolutely necessary*. It is also notable that the participants of the online survey on the topic of information provision also asked for information on frequency and style, index, and diatopic distribution. The evaluation of the answers must be considered in conjunction with those from the question about possible headwords (cf. section 4.2).

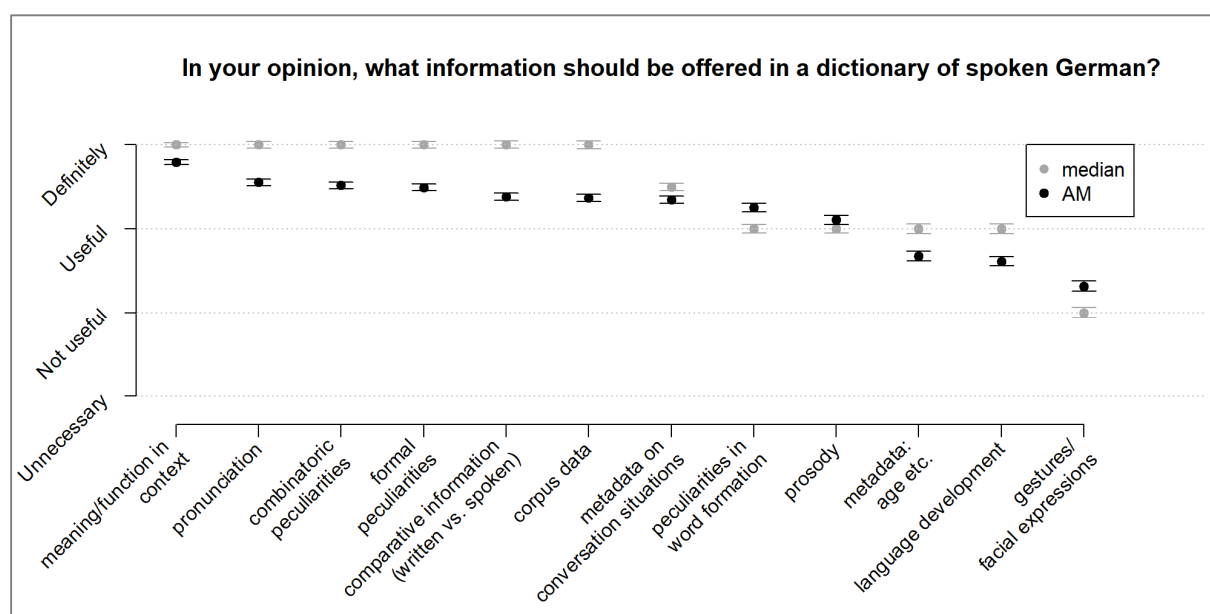


Figure 1: Distribution of expectations regarding the information provided (online survey).

When using the results of the expert interviews for comparison, it becomes clear that the information on pronunciation, meaning in context, special features in form, special features in combinatorics, supply of corpus documents, metadata on the conversation situation, and prosody were rated equally highly as *Definitely* (cf. Fig. 2). In addition, the experts – similar to the respondents to the online survey – also mentioned information on linguistic variation.

A comparison of the two surveys allows the following conclusions to be drawn: There are similarities in the following points: (i) Most of the information is rated by all respondents as necessary and useful without major differences. (ii) An exception is information on metadata, such as age, language development, and gestures/facial expressions, which have been classified as *Not useful, but nevertheless desirable*. Differences between the two surveys lie mainly in the information provided on prosody (for the experts *Definitely*, for the respondents of the online survey *Useful, but not absolutely necessary*). This divergence can be explained by the higher degree of specific conversational linguistic expertise of the interviewees from the expert interviews (cf.

Meliss et al. 2018b: 126-128, 2019: 104-106).

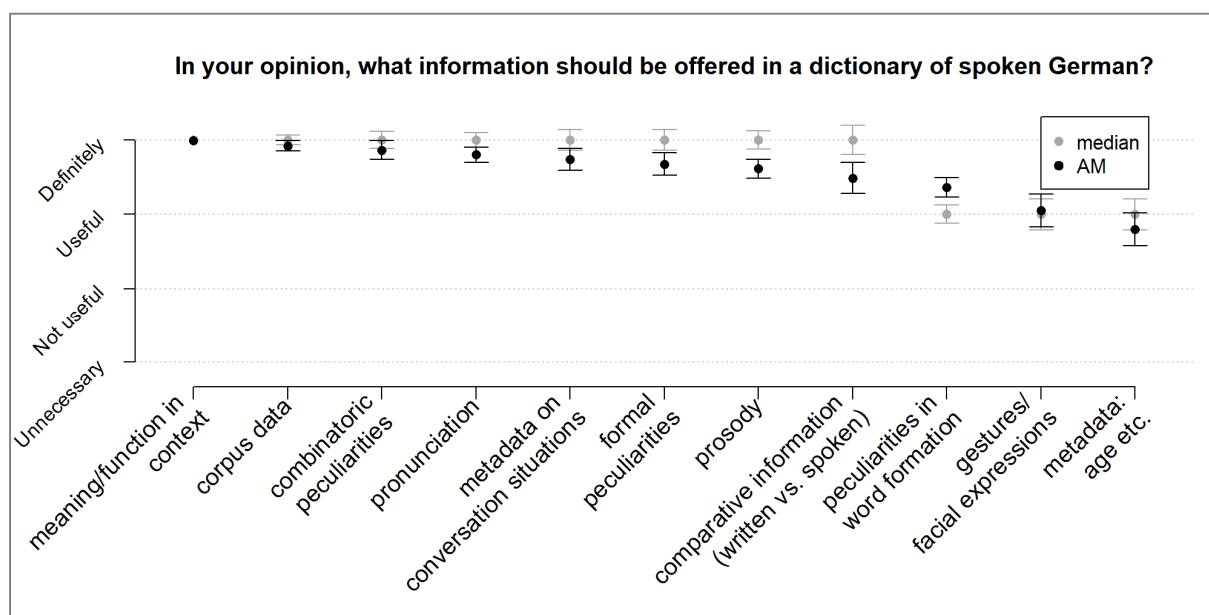


Figure 2: Distribution of expectations regarding the information provided (expert interviews).

In the following, the information included in the prototype of the resource is explained in more detail. Many of the expected aspects could be taken into account in the lexicographical implementation.

5. The LeGeDe-resource

The LeGeDe-resource offers an extensive range of information for each headword. The result is a complex lexicographical structure. In the following sections we explain the design and implementation of five aspects: (i) the identification of headword candidates, and the lemmas described in the dictionary (cf. 5.1), (ii) the range of information for each headword (cf. 5.2), (iii) the outer texts (cf. 5.3), (iv) the linking of the dictionary articles with the DGD, and (v) the possibility of further corpus analysis (cf. 5.4).

5.1 Headword candidates

One of the key research and methodological issues that the LeGeDe-project has addressed is related to the identification of typically spoken lexical peculiarities, and thus to the comparison with dictionaries based on written language. In direct relation to the distinctive features of lexical peculiarities on written and spoken language in interaction, a list of headword candidates for the LeGeDe-resource is drawn up (cf. Meliss et al., 2018a). As typical phenomena of spoken language, these candidates are used in spontaneous interaction and thus are clearly distinguishable from written language aspects.

The considerations regarding one-word lemmas, which have a specific meaning and function in interaction (e.g. interjections), have to be complemented with the integration of multiword expressions and constructions with specific functions in interaction (e.g. *was weiß ich* [engl. *I don't know*], *keine Ahnung* [engl. *no idea*], *guck mal* [engl. *look!*]) as headword candidates (cf. e.g. Bergmann, 2017; Günthner, 2017, Helmer & Deppermann, 2017; Helmer et al., 2017; Imo, 2007; Zeschel, 2017).

Hence, a corpus-based and interpretative method was developed in the LeGeDe-project (cf. Meliss et al., 2018a) to create a list of headwords, with which the most important candidates of the typical spoken lexicon could be uncovered in interaction (cf. with regard to the expectations on the headword candidates, the results are shown in section 4.2). For the comparison with the written language, the German reference corpus (=DEREKO, version 2017 I, cf. Kupietz & Keibel, 2009; Kupietz et al., 2018) was used. The method applied is briefly explained below.⁵

Since we wanted to use DEREKO as a representation of current written language, we have excluded data that contain the conceptually spoken language presented in Wikipedia discussions as well as the subcorpus “Sprachliche Umbrüche” from the years 1945 to 1968. One of the steps was to calculate the difference in lemma distribution in the two corpora by using different effect measures (odds ratio, %diff, relative risk, binary protocol of relative risk and frequency classes) and measures of statistical significance (log likelihood ratio and chi square). The lemma comparison table has been integrated into a tool we developed to quickly and easily filter and sort the data. With the help of this tool, the headword candidates can be dynamically evaluated, executed, and explored, and the parameters can be adapted to the needs of the lexicographers. After examining the results of different measurements of the frequency comparison, we opted for the difference of the “frequency classes” (“Häufigkeitsklasse” = HK; cf. Keibel, 2008, 2009), a measurement which is relatively intuitive to understand and frequently used in German lexicography (cf. e.g. Klosa, 2013). The most common word in a corpus is in frequency class 0, whereas the word(s) in class 1 is (are) about half as common as the most common word(s) in class 0, the words in class 2 are about half as common as those in class 1, etc. We calculated the difference of the frequency classes of a lemma in the two corpora as “difference of the frequency classes” ($fc_diff = fc(dereko) - fc(folk)$). After sorting the lemma list by descending fc_diff , we extracted about 320 one-word lemmas whose fc_diff was at least 2. The manual check of these candidates enabled us to see if they were suitable headword candidates in the one-word lemma range for our resource. Table 1 shows the top 25 candidates for which we can define different headword groups.

⁵ For details cf. Meliss et al. (2018a).

No.	Lemma	FOLK HK	DeReKo HK	HK Diff
1	ah	4	14	10
2	okay	4	14	10
3	ach	4	13	9
4	ja	0	8	8
5	irgendetwas	6	14	8
6	gucken	5	13	8
7	oh	5	13	8
8	halt	4	12	8
9	irgendwie	4	12	8
10	du	2	9	7
11	danke	7	14	7
12	nachher	7	14	7
13	kriegen	5	12	7
14	na	5	12	7
15	nein	2	10	7
16	also	2	8	6
17	Hey	8	14	6
18	runter	7	13	6
19	wieso	7	13	6
20	cool	7	13	6
21	Ahnung	7	13	6
22	Mama	7	13	6
23	drin	6	12	6
24	sozusagen	6	12	6
25	dein	5	11	6

Table 1: TOP 25 of one-word lemmas from a statistical point of view (FOLK, Release 2.11, cf. *Lexical Explorer*: “Study corpus vs. DeReKo”, Study corpus HK = <9, DeReKo HK = <15, HK Diff = >1, Filter = 1).

Headword candidates as one-word lemmas are defined on the basis of this method. Manual analysis is used to record information on very different grammatical, semantic, and interactional linguistic aspects. For each one-word lemma, a sample of 300 hits is drawn from FOLK. Of these, 100 valid (i.e. clear audio) hits are analysed and coded in detail. The range of information on the headwords is explained in more detail in section 5.2.

The further step to analyse the sample of each selected headword according to formal, semantic, syntactic and functional criteria shows, among other things, whether there are any occurrences of the lemma in the data that refer to one of the meanings of the

one-word lemma (e.g. ‘abwarten’ [engl. *to wait*] as one of the basic meanings of the lemma *gucken* [engl. *to look*], selected as one of our headwords (cf. no. 6 in Table 1). The results on the meaning-based analysis of one-word lemmas lead to a dictionary article “Bedeutungen” [engl. “Meanings”] (= module 1), which we describe in more detail in section 5.2.2.

In addition, the detailed analysis work on the sample shows the possibility of the occurrence of units with a special interactional function. These can be one-word or multi-word units related to the list of identified headword candidates (e.g. *halt* as a ‘modal particle’ cf. no. 8, or *guck mal* as a ‘discourse marker’ cf. no. 6 in Table 1). Section 5.2.3 describes the lemmas with interactional functions in more detail.

5.2 Range of information on the headwords

In the following, the central lexicographic information sections (overview, module 1, module 2) on selected headwords, which have been edited accordingly, will be presented (with regard to the expectations on the headword candidates, cf. the results of the studies in section 4.3).⁶

5.2.1 General overview

For each headword, general overview information is available and offers, in a descriptive form, meaning- and function-oriented information (e.g. *eben* [engl. *just*], cf. Fig. 3).

A clear modular division of the information enables the presentation of lexical-semantic information on the one hand, which is oriented to the respective meaning of the corresponding senses of a lemma (= module 1), and on the other hand of function-specific interactionally oriented information (= module 2). For both areas, specific lexicographical information was used or newly developed for description purposes, which offers completely new insights and formats in addition to traditional dictionary information.

Different cross-connections between the two modules are made explicit by an internal link. An extended external information offer is provided by a link that leads to further lexicographic resources (e.g. DWDS) on the one hand and to FOLK and the *Lexical Explorer* on the other hand. In addition, the calculated corpus-based frequency class difference between the headwords in the respective corpora (written: DEREKO, spoken: FOLK) is visualized (cf. Fig. 4).

⁶ This figure and also the following (with excerpts from the resource) are based on the beta version of the resource (last update: 5 August 2019).

⁷ *eben* as an adverb can generally be translated as *just* in English. For *eben* as modal or discourse particles, there are contexts in English in which *just* could also be used. But a clear lexical equivalent of the particles *eben* does not exist in English.

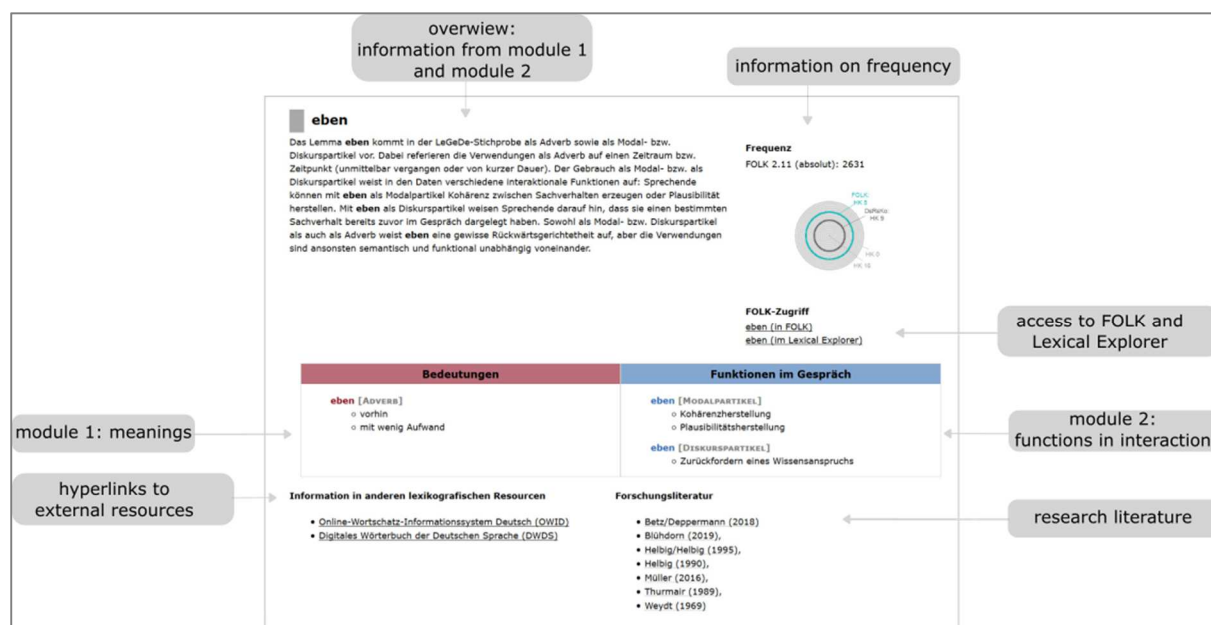


Figure 3: Overview article of the lemma *eben* (screenshot).

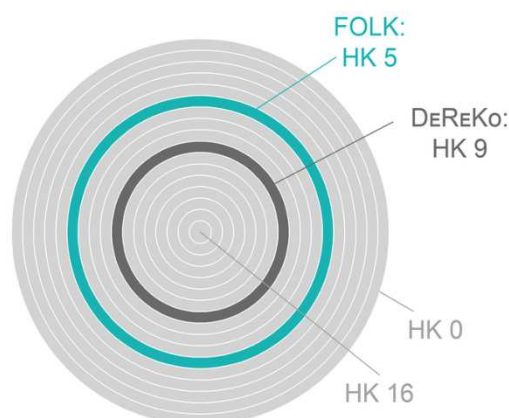


Figure 4: Visualization of the frequency classes of *eben* in FOLK and DEREKO.

An optional short reference to the related research literature enables an insight into relevant sources for each headword.

5.2.2 Module 1: Meanings

In addition to relevant general **sense-independent information (1)** (e.g. word class: adverb, verb, noun; morpheme structure in case of lexical compounds or affix constructions: *Ahnung*: Basis: *ahn-* (=verbal stem), *-ung* (=derivative suffix) [engl. *idea/knowledge*]); formal variation (e.g. *gucken*: <kucken>/[kucken]); research literature) the data of module 1 is mainly supplemented by information on meaning and combinatorics. The different information items of module 1 (see (1)-(9)) are

subsequently explained using the verb *wissen* [engl. *to know*] as an example and presented at a glance in Fig. 5.

(i) The **sense-related information** of each lemma is semantically identified by a **short label (2)** for disambiguation of meaning and by short **semantic paraphrases (3)**. In addition, a **transcript box (4)** with a transcript title, a short description of the context, an illustrative transcript excerpt (+audio), and an optional commentary is offered (cf. Fig. 5).

(ii) **Formal peculiarities**: There is also the possibility of pointing out formal peculiarities in a short comment. Different aspects can be commented on, such as the distinctive combinatorial behaviour with modal verbs, the use of certain verbal modes or the connection with certain particles or deictic expressions.

(iii) **Combinatorics (5)**: In conjunction with information on combinatorics, we distinguish between **structural patterns (6)**, **fixed phrases/collocations (7)**, and **interactional units (8)**. Transcript boxes illustrate the corresponding phenomena (cf. Fig. 5).

a. The **structural patterns (6)** (= Strukturmuster) are offered in an abstract, formulaic way (e.g. <jemand weiß, dass/ob etwas der Fall ist//was der Fall ist> [engl. <someone knows (that/if s.th. is the case//what is the case)>]). The individual arguments, from which the structure patterns are composed, are explained with regard to their semantic role, their syntactic function, and the possibilities of morphosyntactic realization. The information in the transcript boxes illustrates the use of the patterns with a short transcript excerpt.

b. **Fixed phrases/collocations (7)**: Under this broad generic term, we subsume different types of more or less fixed lexical units (e.g. collocations, routine formula, proverbs) without further specification or terminological precision. These lexical units and collocations (e.g. *Bescheid wissen* [engl. *to be in the know*], *man weiß es ja nie* [engl. *you never know*]) are described, if considered relevant, in their semantic and/or formal properties and they are also individually illustrated with a transcript box.

c. The listing of **interactional units (8)**, which could be documented in the LeGeDe-sample in direct relation to specific meanings of certain lemmas (e.g. *keine Ahnung* [connection to *Ahnung* in the sense ‘Wissen’ [engl. ‘knowledge’], *ich weiß nicht* [connection to *wissen* in the sense of ‘to be informed’]), enables a direct cross-connection to interactional functions (e.g. ‘Unsicherheitsmarker’: *ich weiß nicht* [engl. ‘epistemic hedge’: *I don’t know*]), which are described in module 2.

(iv) **Other peculiarities (9)**: Furthermore, it is possible to point out interesting data in relation to selected metadata and their frequency.

(1) formal information

(2) short label

(3) semantic paraphrase

(4) transcript box

(5) information on combinatorics

(6) structural patterns

(7) fixed phrases/collocations

(8) interactional units

(9) other peculiarities

wissen [VOLLVERB]

Form	Varianz

Forschungsliteratur

Bergmann (2017), Erman (2001), Günthner (2017), Helmer/Deppermann (2017), Helmer/Deppermann/Reineke (2017), Helmer/Reineke/Deppermann (2016), Reineke (2016), Zeschel (2017)

informiert sein ▾

Jemand besitzt Information über einen Sachverhalt infolge eigener Erfahrung/Wahrnehmung oder durch Mitteilung von anderen und äußert diesbezüglich Sicherheit. Die häufige Negation von **wissen** führt zum Ausdruck von Nicht-Wissen und/oder Unsicherheit. In Kombination mit (temporalen) Satzadverbien kann die Kernbedeutung 'informiert sein' verschoben werden zu 'sich (nicht mehr/hoch) erinnern'.

[1] wissen zum Ausdruck von 'informiert sein'

Belegkontext

Bei einem Telefongespräch erzählt SL ihrer Freundin OW von einem Skype-Gespräch, dass sie vor kurzem mit ihrer ehemaligen Gastfamilie geführt hat. SL berichtet nach einer Nachfrage von OW von dem kleinen Sohn ihrer ehemaligen Gastmutter.

MI 0761 SL ja sie is (.) w sie war jetzt heute glaub ich zum ersten mal mit thomas im (.) kindergarten weil der (.) ((Sprechersatz)) jetzt

MI 0762 (0.56)

MI 0763 SL nach den sommerferien geht er ja in_n kindergarten

MI 0764 (0.24)

MI 0765 SL für en hal/ben tag

MI 0766 OW hmhm

MI 0767 (0.88)

MI 0768 SL und da sind jetzt während der sommerferien aber schon so tage wo dann die mütter mit den kindern hinkomm können damit die sich schon

MI 0769 (0.62)

MI 0770 SL daran gewöhnen an die umgebung und dann schon **wissen** wo alles is und so

MI 0771 (2.19)

MI 0772 SL da war sie jetzt glaub ich heute zum ersten mal

(Privat, Telefongespräch - FOLK_E_00321_Y_03)

[...]

Kombinatorik

Jemand weiß, dass/ob etwas der Fall ist / was der Fall ist [STRUKTURHINTER]

- jemand:** die Person, die informiert ist
Syntaktische Funktion: Subjekt
Anmerkung: Das Subjekt kann weggelassen werden, wenn der Sprechende von sich redet.
- etwas:** der Sachverhalt, über den jemand informiert ist
Syntaktische Funktion: Akkusativkomplement
Realisierungsformen: häufige pronominale Realisierung (**das, es**); dann auch durch Stellung im Satz und Satzakkzent besonders betont; zahlreiche satzformige Realisierungen des Sachverhaltes: Diese abhängigen Sätze werden entweder eingeleitet mit **dass/ob/was** oder erscheinen uneingeleitet.

das weiß der Prüfer

MI 0821 RK du musst den zweiten gang nehmen obwohl du hier vorne gleich wieder in den ersten muss das weiß ich (.) un das weiß auch der prüfer ^h aber du fährst zu hochtourig rechts weiter

(Institution: Fahrprüfung - FOLK_E_00466_T_04)

wissen, dass

MI 0898 AW weiß nur dass es so heißt dass es halt diese (.) tests sin (.) welche schulform er geeignet is

(Institution: Meeting in einer sozialen Einrichtung - FOLK_E_00016_T_00)

[...]

Bescheid wissen [FESTE VERBUNDEN / KOLLOKATIONEN]

Mit dieser Wendung geben Sprechende zu verstehen, dass sie selber oder eine andere Person tatsächlich über einen nicht näher benannten Sachverhalt informiert sind oder stellen dies in Frage.

Bescheid wissen

MI 0676 MS so un wann kommt die andre praktikantin jetzt

MI 0677 SZ häh

MI 0678 HM drei woche später da im plan steht_s dr[in +++ ++]

MI 0679 AW ((stöhnt))

MI 0680 MS also drei woche s'päter

MI 0681 (1.46)

MI 0682 SZ aber dann könnten mer [der ja äh] äh ham_mer da ne adresse oder was oder ne nummer der der mal (.) die schon vorwarnen oder weiß die schon **bescheid**

MI 0683 AW [mir scheißegal]

MI 0684 (0.26)

MI 0685 AW [die kommt am sie [bten neunten wenn du_s genau willst (.) die frau [weingarten]

MI 0686 SZ [dass die vielleicht f]

MI 0687 SZ [dass die erschte da feschte sch[u]he oder wa[s]

(Institution: Meeting in einer sozialen Einrichtung - FOLK_E_00024_T_08)

[...]

Unsicherheitsmarker: ich weiß nicht [INTERAKTIONALE EINHEIT]

Mit **ich weiß nicht** drücken Sprechende aus, dass sie sich hinsichtlich der Wahrheit, Genauigkeit, Gültigkeit bzw. Angemessenheit einer Aussage unsicher sind und markieren die Vagheit dieser Angabe.

[...]

Über 90% der Belege der LeGeDe-Stichprobe konnten dieser Lesart zugeordnet werden. In dieser Bedeutungsvariante wird **wissen** nur selten in öffentlichen Interaktionssituationen belegt. Außerdem konnte **wissen** als Teil eines interaktionalen Musters in einem Viertel der Belege in Verbindung mit dieser Lesart dokumentiert werden. In diesen Fällen werden die Argumente des Strukturmusters nicht immer alle realisiert und es treten zusätzliche interaktionale Funktionen auf.

Figure 5: Excerpt of *wissen*⁸ ('informiert sein') [engl. *to know*, 'to be informed'] (Screenshot).

⁸ The author of the lexicographic article *wissen* (module 1: meanings) is Meike Meliss (member of the LeGeDe-Team).

5.2.3 Module 2: Functions in interaction

Module 2 describes the function of one- and multiword lemmas in spoken interaction (e.g. *eben*, *keine Ahnung*, *ich weiß nicht*). The different information items (1-7) will be explained using the example of the interactional unit *ich weiß nicht* [engl. *I don't know*] (cf. Fig. 6).

The general **cross-functional information (1)** is divided into categorical (modal particles (Thurmain, 1989), discourse particles (Willkop, 1988), etc.) and formal information regarding the elements involved in complex forms (e.g. *ich weiß nicht* [engl. *I don't know*]: verbal phrase; *keine Ahnung* [engl. *no idea*]: nominal phrase). Furthermore, formal information concerning the elements of multiword units (e.g. *ich weiß nicht*: “Phrase aus dem Personalpronomen *ich*, [...]” [engl. phrase formed from the personal pronoun [...]]) and information on possible formal variants is offered on phonetic (e.g. *eben*: [ebent]; “Epithese eines stimmlosen [t]” [engl. “epithesis of an unvoiced [t]”]), and compositional levels (e.g. *(ich) weiß nicht/weiß (ich) nicht*). A list of documented possibilities for combinatorics with an optional comment completes the general information together with a reference on the relevant research literature.

Each particular function that can be assigned to a lemma is **labelled (2)** accordingly. For example, *ich weiß nicht* as a multiword unit has a functional spectrum of different possibilities (‘Unsicherheitsmarker’, ‘Markierung potenzieller Unangemessenheit’; [engl. ‘epistemic hedge’ or ‘display of potential inappropriateness’]). A **short description (3)** of the functions should help to differentiate the various possibilities. A **transcript box (4)** with the corresponding transcript excerpt, title, context, and comment is used for illustration.

In **abstraction of function (5)** generic information is offered. This information refers to findings which go beyond the occurrences in individual transcripts and therefore point at conspicuous features that have been revealed in the sample across the transcripts. These features are explained more in depth but in a comprehensive manner every user of each target group is able to grasp.

In addition to formal, categorical, combinatorial, and functional information, module 2 is enriched by information on **syntax and sequence realization (6)** and **prosody (7)** which are both illustrated by short transcript excerpts.

(1) cross-functional information →

(2) label →

(3) short description →

(4) transcript box →

(5) abstraction of the function →

(6) syntax/sequence →

(7) prosody →

ich weiß nicht [VERBALSYNTAGMA]

Form	Varianz	Kombinatorik	Forschungsliteratur
Phrase aus dem Personalpronomen ich , der 1. Pers. Sg. des Verbs wissen (weiß) und der Negationspartikel nicht	weiß nicht / weiß ich nicht		Bergmann (2017), Helmer/Deppermann (2017), Helmer/Deppermann/Reineke (2017), Imo 2007

Unsicherheitsmarker ▶

Mit **ich weiß nicht** zeigen Sprechende an, dass sie sich bezüglich der Wahrheit/Genauigkeit einer Aussage unsicher sind und markieren die Vagheit dieser Angabe (vgl. dazu auch **wissen** in der Bedeutung **informiert sein**).

Markierung potenzieller Unangemessenheit ▼

Mit **ich weiß nicht** zeigen Sprechende an, dass sie sich hinsichtlich der Angemessenheit einer Äußerung unsicher sind.

[1] **ich weiß nicht** zur Markierung potenzieller Unangemessenheit eines Vorschlags

001 **VK** du musst irgendwas von dir verkaufen;

002 (2.01)

003 **SK** aber WAS-

004 **VK** "h na-

005 (0.36)

006 **VK** du HAST doch hier genu[chi];]

007 **NK** [was SOLL ich denn verkauf[en];]

008 **VK** [ich] WEIß nich=-

009 **VK** =verkauf doch hier diese beiden STRAßEn da.

010 **VK** die (.) turnstraße un DIEse,

011 **VK** un VERKAUF se an irgendjemanden.

012 (0.45)

013 **SK** "h (-) "h Ich verkauf sie mit HAUS an die bank.

Privat: Spielinteraktion mit Kindern - FOLK_E_00011_T_04

Belegkontext

Vater (VK) und Tochter (SK) spielen das Brettspiel Monopoly und tauschen sich dabei über die nächsten Spielzüge aus. Die Tochter (SK) muss ihren nächsten Spielzug tätigen, damit das Spiel weitergehen kann.

Analyse

Der Vater (VK) weist die Tochter (SK) darauf hin, dass sie etwas verkaufen muss und sie fragt ihren Vater daraufhin, was sie denn verkaufen soll (Z. 001-007). Auf die Frage antwortet der Vater zunächst mit "ich WEIß nich=-" (Z. 008), schließt daran jedoch unmittelbar den Vorschlag an, zwei ganz bestimmte Straßen ("die (.) turnstraße un DIEse," Z. 010) an jemanden zu verkaufen (Z. 009-011). Mit **ich weiß nicht** zeigt der Vater somit nicht an, dass er nicht weiß, welche Straßen die Tochter verkaufen soll: Er markiert mit **ich weiß nicht**, dass er sich nicht sicher ist, ob sein Vorschlag in der Folgeäußerung für die Tochter angemessen oder erwünscht ist. Diese nimmt im daraufhin den Vorschlag des Vaters an und äußert die Absicht, die Straßen mit Haus an die Bank zu verkaufen (Z. 013).

Mit **ich weiß nicht** zeigen Sprechende Unsicherheit darüber an, ob die Äußerung für das Gegenüber potenziell (un)angemessen, (un)erwünscht, zutreffend o.Ä. ist. **ich weiß nicht** bezieht sich somit nicht auf einer epistemischen Ebene auf (Un-)Kenntnis/(Un-)Sicherheit bezüglich eines Sachverhaltes, sondern auf einer pragmatischen Ebene auf (Un-)Kenntnis/(Un-)Sicherheit bezüglich der Sprechhandlung. In dieser Funktion tritt **ich weiß nicht** oft vor heiklen Aussagen wie Vorschlägen, Meinungsäußerungen, positiven/negativen Bewertungen oder Positionierungen auf. Zur Markierung einer möglicherweise unangemessenen Äußerung geht **ich weiß nicht** teils mit Verzögerungsmarkern wie Atmen und Pausen sowie mit Formulierungsarbeit durch Abbrüche und Neuansätze einher.

Syntax-Sequenz-Realisierung

ich weiß nicht wird turn- und äußerungsinitial verwendet.

ich weiß nicht turn- und äußerungsinitial

001 **NK** was SOLL ich denn verkauf[en];]

002 **VK** [ich] WEIß nich=-

003 **VK** =verkauf doch hier diese beiden STRAßEn da.

Privat: Spielinteraktion mit Kindern - FOLK_E_00011_T_04

[...]

Prosodie

ich weiß nicht wird prosodisch desintegriert verwendet und bildet eine eigene Intonationskontur.

ich weiß nicht prosodisch desintegriert, eigene Intonationskontur

001 **OH** aber ich GLAUB,

002 (0.29)

003 **OH** er hat dann AU-

004 **OH** er hat EH ni_mehr so viel motivation;

005 **OH** "h <call> ich WEIß nich >=-

006 **OH** =wird vorher: au noch mal mit ihm REDEn einfach;

007 **OH** bevor wir_s [FESTsachen so=- ne,]

008 **KL** [JA (-) klar;]

Privat: Spielinteraktion mit Kindern - FOLK_E_00011_T_04

Figure 6: Excerpt of *ich weiß nicht*⁹ ('Markierung potenzieller Unangemessenheit') [engl. *I don't know*, 'display of potential inappropriateness'] (Screenshot).

⁹ The author of the lexicographic article *ich weiß nicht* (module 2: functions in interaction) is Katja Arens (a member of the LeGeDe-Team).

5.2.4 Links between module 1 and module 2

There is a very crucial connection between the two modules. In module 1, information about patterns and constructions is offered in the “combinatorics” section. Among these constructions are those with an underlying structural pattern and a syntactically functional approach. In addition, there are patterns or constructions in our data which have a special function in conversation. These “interactional units” are listed in module 1 in the section “combinatorics” (e.g. *ich weiß nicht* [engl. *I don't know*] as part of the dictionary article to *wissen* in the sense of ‘informiert sein’ [engl. ‘to be informed’]), but they are described in more detail in module 2. There are offered separate dictionary articles for the construction *ich weiß nicht* with a description of the function ‘Unsicherheitsmarker’ [engl. ‘epistemic hedge’] (cf. 5.2.2., iii.c).

The semantic connection from module 2 to module 1 can also be illustrated by using the example of the multiword unit *ich weiß nicht* [engl. *I don't know*]. The short functional description in module 2 informs the user of the basic meaning contained in this pattern offering a reference to the sense ‘informiert sein’ [engl. ‘to be informed’] of the verb *wissen* and links to module 1 accordingly (cf. Fig. 6).

5.3 Links within the resource: Outer texts

The dictionary user is offered four different types of outer texts. A section “About the LeGeDe-project” (“Über LeGeDe”) provides a detailed reference about the project in general and to conceptual considerations about the LeGeDe-resource. In the “Usage instructions” section, a dictionary user learns in a guided tour how to navigate the resource and what types of information are offered. Very central terms used in our dictionary articles can be looked up in a “Glossary”. Especially for grammatical terms, links to “grammis” (=Grammatisches Informationssystem) are offered via the glossary entries. Technical terms from the field of interactional linguistics – e.g. “Bezugsäußerung” [engl. *reference expression*], “Diskursmarker” [engl. *discourse marker*] or “Sequenz” [engl. *sequence*] – are explained and supplemented with research literature. From the glossary as well as from the dictionary articles we indicate very fundamental “Research literature”. This can be viewed at a glance in a literature list.

5.4 Connection between the LeGeDe-resource and the DGD

A link to a lemma in the DGD database is offered in the overview article (cf. Fig. 3). Besides that, many details about a headword are supplemented in the dictionary article with authentic examples taken from the FOLK corpus (cf. Section 3). For each transcript excerpt there is the possibility to access the DGD database directly. For this purpose, it is necessary to create a personal account. After registering with the database, it is possible to view the transcript excerpts from the dictionary directly in the database, listen to the audio of the transcripts, sometimes even view video material and continue

researching the database. Via the overview article, the user is also able to search for a lemma in the *Lexical Explorer* (cf. section 3 and Batinić-Lemmenmeier in press).

6. Concluding remarks

As has been shown, the innovative aspects of the LeGeDe-resource are numerous. Since the project could not rely on any previous models, the simple fact of having created a lexicographic prototype to represent the specifics of the German spoken-language lexicon, using a corpus of spoken language in interaction as a basis, can be considered as a ground-breaking result. The conceptual considerations were based on assumptions from research, the LeGeDe-project work, and the results of the studies carried out during the LeGeDe-project. Concrete innovative aspects of the resource include the following:

- (i) Data basis: The resource is based exclusively on corpus-based data.
- (ii) Method: The corpus-based data have been quantitatively determined and qualitatively analysed and structured by a methodological approach developed by the team.
- (iii) List of headword candidates: The list of headword candidates was compiled using a specially developed corpus-based method of frequency comparison between two corpora: DEREKO as the reference corpus for written German and FOLK as the reference corpus of spoken interactional German.
- (iv) Range of information: The information offered on the lemmas is multimodular. The dictionary user finds a combination of traditional lexicographic information with an innovative offer of information which is developed specifically for the description of interactional functions. This is the first time that a proposal for new lexicographic information has been developed for the presentation of lexical phenomena of spoken language in interaction, which makes it possible to adequately structure and describe the specific phenomena for lexicographic purposes.
- (v) Authentic corpus evidence: Authentic corpus evidence is initially offered via selected transcript excerpts that provide an interface to the audio files and detailed information on the metadata. This makes the LeGeDe-resource one of the few lexicographic resources that has a direct, non-automatically generated link to the corresponding corpus data.
- (vi) Multimedia: The resource's multimedia character is characterized by the fact that, in addition to the transcripts, audio files and, in some cases, corresponding video files are available for the corpus data via access to the DGD. The link to the *Lexical Explorer* offers the possibility of extended analysis options.

- (vii) Consideration of empirical expectations: The completely new conception of a lexicographic resource for the representation of linguistic specifics enabled the concrete consideration of certain empirically raised expectations of future users of such a new resource.

Not all aspects could be considered and implemented into the developed prototype during the project duration. Thus, there are certainly still many interesting possibilities for further research and development, for example in the area of the phenomenon classes (word formation, deixis, vagueness, etc.) or the access possibilities via an extended search in order to respond to the corresponding expectations of the participants in our surveys. Although the resource and the analyses are very detailed and complex, we hope that experts can take a mediating position in order to also make the contents accessible to different kinds of L1- and L2-learners.

7. References

- Batinić-Lemmenmeier, D. (in press). Lexical Explorer: Extending access to the Database of Spoken German for user-specific purposes. In: *Corpora*, 15(1).
- Bergmann, P. (2017). Gebrauchsprofile von *weiß nich* und *keine Ahnung* im Gespräch. Ein Blick auf nicht-responsive Vorkommen. In H. Blühdorn et al. (eds.), pp. 157–182.
- Blühdorn, H., Deppermann, A., Helmer, H. & Spranz-Fogasy, T. (eds.) (2017). *Diskursmarker im Deutschen. Reflexionen und Analysen*. Göttingen: Verlag für Gesprächsforschung.
- Deppermann, A., Proske, N. & Zeschel, A. (eds.) (2017). *Verben im interaktiven Kontext. Bewegungsverben und mentale Verben im gesprochenen Deutsch*. (Studien zur deutschen Sprache 74). Tübingen: Narr.
- Eichinger, L. M. (2017). Gesprochene Alltagssprache. In Deutsche Akademie für Sprache und Dichtung & Union der deutschen Akademien der Wissenschaften (eds.) *Vielfalt und Einheit der deutschen Sprache. Zweiter Bericht zur Lage der deutschen Sprache*. Tübingen: Stauffenburg, pp. 283–331.
- Fiehler, R. (2016). Gesprochene Sprache. In A. Wöllstein (ed.) *Duden – Die Grammatik*. Berlin: Dudenverlag, pp. 1181–1260.
- Günthner, S. (2017). Diskursmarker in der Interaktion – Formen und Funktionen univverbierter *guck mal*- und *weißst du*-Konstruktionen. In H. Blühdorn et al. (eds.), pp. 103–130.
- Handwerker, B., Bäuerle, R. & Sieberg, B. (eds.) (2016). *Gesprochene Fremdsprache Deutsch*. (Perspektiven Deutsch als Fremdsprache 32). Baltmannsweiler: Schneider.
- Hansen, C. & Hansen, M. H. (2012). A Dictionary of Spoken Danish. In R. V. Fjeld & J. M. Torjusen (eds.) *Proceedings of the 15th EURALEX International Congress. 7-11 August 2012. Oslo, Norway*. Oslo: Department of Linguistics and Scandinavian Studies, University of Oslo, pp. 929–935.
- Helmer, H. & Deppermann, A. (2017). ICH WEIß NICHT zwischen Assertion und

- Diskursmarker: Verwendungsspektren eines Ausdrucks und Überlegungen zu Kriterien für Diskursmarker. In H. Blühdorn et al. (eds.) *Diskursmarker im Deutschen. Reflexionen und Analysen*. Göttingen: Verlag für Gesprächsforschung, pp. 131–156.
- Helmer, H., Deppermann, A. & Reineke, S. (2017). Antwort, epistemischer Marker oder Widerspruch? Sequenzielle, semantische und pragmatische Eigenschaften von *ich weiß nicht*. In A. Deppermann et al. (eds.) *Verben im interaktiven Kontext. Bewegungsverben und mentale Verben im gesprochenen Deutsch*. (Studien zur deutschen Sprache 74). Tübingen: Narr, pp. 377–406.
- Imo, W. (2007). Zur Anwendung der Construction Grammar auf die gesprochene Sprache – Der Fall „*ich mein(e)*“. In V. Ágel & M. Hennig (eds.) *Zugänge zur Grammatik der gesprochenen Sprache*. (Germanistische Linguistik 269). Tübingen: Niemeyer, pp. 3–34.
- Imo, W. & Moraldo, S. M. (eds.) (2015). *Interaktionale Sprache und ihre Didaktisierung im DaF-Unterricht*. (Deutschdidaktik 4). Tübingen: Stauffenburg.
- Keibel, H. (2008). *Mathematische Häufigkeitsmaße in der Korpuslinguistik: Eigenschaften und Verwendung*. Mannheim: Institut für Deutsche Sprache. Elektronische Ressource.
- Klosa, A. & Müller-Spitzer, C. (eds.) (2016). *Internetlexikografie. Ein Kompendium*. Berlin/Boston: de Gruyter.
- Klosa, A. 2013. ‘The Lexicographical Process (with Special Focus on Online Dictionaries)’. In R. H. Gouws et al. (eds.) *Dictionaries. An international encyclopedia of lexicography. Supplementary volume: Recent developments with focus on electronic and computational lexicography*. (Handbücher zur Sprach- und Kommunikationswissenschaft 5). Berlin: de Gruyter, pp. 517–524.
- Kupietz, M. & Keibel, H. (2009). The Mannheim German Reference Corpus (DEREKO) as a basis for empirical linguistic research. In M. Minegishi & Y. Kawaguchi (eds.) *Working Papers in Corpus-based Linguistics and Language Education*, no. 3. Tokyo: Tokyo University of Foreign Studies (TUFS), pp. 53–59.
- Kupietz, M. & Schmidt, T. (2015). Schriftliche und mündliche Korpora am IDS als Grundlage für die empirische Forschung. In L. M. Eichinger (ed.) *Sprachwissenschaft im Fokus. Positionsbestimmungen und Perspektiven*. (Jahrbuch des Instituts für Deutsche Sprache 2014). Berlin: de Gruyter, pp. 297–322.
- Kupietz, M., Lungen, H., Kamocki, P. & Witt, A. (2018). The German Reference Corpus DEREKO: New Developments – New Opportunities. In N. Calzolari et al. (eds.) *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki: European Language Resources Association (ELRA), 2018. pp. 4353–4360.
- Meliss, M. (2016). Gesprochene Sprache in DaF-Lernerwörterbüchern. In B. Handwerker, R. Bäuerle & B. Sieberg (eds.) *Gesprochene Fremdsprache Deutsch*. (Perspektiven Deutsch als Fremdsprache 32). Baltmannsweiler: Schneider, pp. 179–199.

- Meliss, M. & Möhrs, C. (2017). Die Entwicklung einer lexikografischen Ressource im Rahmen des Projektes LeGeDe. *Sprachreport 4/2017*, pp. 42–52.
- Meliss, M. & Möhrs, C. (2018). Lexik in der spontanen, gesprochensprachlichen Interaktion: Eine anwendungsorientierte Annäherung aus der DaF-Perspektive. *GFL*, 3/2018, pp. 79–110.
- Meliss, M., Möhrs, C., Batinić, D. & Perkuhn, R. (2018). Creating a List of Headwords for a Lexical Resource of Spoken German. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (eds.) *Proceedings of the 18th EURALEX International Congress 2018. 17-21 July 2018*. Ljubljana, Slovenia, pp. 1009–1016.
- Meliss, M., Möhrs, C. & Ribeiro Silveira M. (2018). Erwartungen an eine korpusbasierte lexikografische Ressource zur “Lexik des gesprochenen Deutsch in der Interaktion”: Ergebnisse aus zwei empirischen Studien. *Zeitschrift für Angewandte Linguistik*, 68(1), pp. 103–138.
- Meliss, M., Möhrs, C. & Ribeiro Silveira M. (2019). Anforderungen und Erwartungen an eine lexikografische Ressource des gesprochenen Deutsch aus der L2-Lernerperspektive. *Lexicographica*, 34, pp. 89–121.
- Meliss, M., Möhrs, C. & Batinić, D. (2017). LeGeDe – towards a Corpus-Based Lexical Resource of Spoken German. In I. Kosem, C. Tiberius, M. Jakubíček, J. Kallas, S. Krek & V. Baisa (eds.) *Electronic Lexicography in the 21st century. Proceedings of eLex 2017 Conference. 19-21 September 2017. Leiden, the Netherlands*. Brno: Lexical Computing CZ s.r.o., pp. 281–298.
- Moon, R. (1998). On Using Spoken Data in Corpus Lexicography. In T. Fontenelle, P. Hilgsmann, A. Michiels, A. Moulin & S. Theissen (eds.) *Proceedings of the 8th EURALEX International Congress. 4-8 August 1998. Liège, Belgium*. Liège: English and Dutch Departments, University of Liège, pp. 347–355.
- Moraldo, S. M. & Missaglia, F. (eds.) (2013). *Gesprochene Sprache im DaF-Unterricht. Grundlagen – Ansätze – Praxis*. (Sprache – Literatur und Geschichte 43). Heidelberg: Winter.
- Reeg, U., Gallo, P. & Moraldo, S. M. (eds.) (2012). *Gesprochene Sprache im DaF-Unterricht. Zur Theorie und Praxis eines Lerngegenstandes*. (Interkulturelle Perspektiven in der Sprachwissenschaft und ihrer Didaktik 3). Berlin: Waxmann.
- Schmidt, T. (2014a). The Research and Teaching Corpus of Spoken German – FOLK. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis (eds.) *Proceedings of the 9th International Conference on Language Resources and Evaluation. 26-31 May 2014. Reykjavik, Iceland*. Reykjavik: European Language Resources Association, pp. 383–387.
- Schmidt, T. (2014b). The Database for Spoken German – DGD2. In N. Calzolari et al. (eds.) *Proceedings of the 9th International Conference on Language Resources and Evaluation. 26-31 May 2014. Reykjavik, Iceland*. Reykjavik: European Language Resources Association, pp. 1451–1457.
- Schmidt, T. (2017). DGD – Die Datenbank für Gesprochenes Deutsch. Mündliche Korpora am Institut für Deutsche Sprache (IDS) in Mannheim. *Zeitschrift für Germanistische Linguistik*, 45(3), pp. 451–463.

- Schwitalla, J. (⁴2012). *Gesprochenes Deutsch. Eine Einführung*. (Grundlagen der Germanistik 33). Berlin: Schmidt.
- Sieberg, B. (2013). *Sprechen Lehren, Lernen und Verstehen. Stufenübergreifendes Studien- und Übungsbuch für den DaF-Bereich*. Tübingen: Julius Groos.
- Siepmann, D. (2015). Dictionaries and Spoken Language: A Corpus-Based Review of French Dictionaries. *International Journal of Lexicography*, 28(2), pp. 139–168.
- Thurmair, M. (1989). *Modalpartikeln und ihre Kombinationen*. Tübingen: Niemeyer.
- Trap-Jensen, L. (2004). Spoken Language in Dictionaries: Does It Really Matter? In G. Williams & S. Vessier (eds.) *Proceedings of the 11th EURALEX International Congress. 6-10 July 2004. Lorient, France*. Lorient: Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud, pp. 311–318.
- Trim, J., North, B., Coste, D. & Sheils, J. (2001). *Gemeinsamer Europäischer Referenzrahmen für Sprachen: Lernen, Lehren, Beurteilen. Niveau A1, A2, B1, B2, C1, C2*. Berlin: Langenscheidt.
- Verdonik, D. & Sepesy Maučec, M. (2017). A Speech Corpus as a Source of Lexical Information. *International Journal of Lexicography*, 30(2), pp. 143–166.
- Willkop, E.-M. (1988). *Gliederungspartikeln im Dialog*. München: Iudicium.
- Zeschel, A. (2017). *Denken und wissen im gesprochenen Deutsch*. In A. Deppermann, N. Proske & A. Zeschel (eds.) *Verben im interaktiven Kontext. Bewegungsverben und mentale Verben im gesprochenen Deutsch*. (Studien zur Deutschen Sprache 74). Tübingen: Narr, pp. 249–336.

Websites (5 August 2019)

- DEREKO*. Accessed at: <http://www1.ids-mannheim.de/kl/projekte/korpora.html>.
- DGD*. Accessed at: <https://dgd.ids-mannheim.de>.
- DWDS*. Accessed at: <https://www.dwds.de/>.
- FOLK*. Accessed at: <http://agd.ids-mannheim.de/folk.shtml>.
- grammis*. Accessed at: <https://grammis.ids-mannheim.de/>.
- LeGeDe project website*: <http://www1.ids-mannheim.de/lexik/lexik-des-gesprochenen-deutsch.html>.
- LeGeDe-Resource*: Accessed at: <https://www.owid.de/legede/> (available from September 2019).
- Lexical Explorer*. Accessed at: <https://www.owid.de/lexex/>.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



Automating Dictionary Production: a Tagalog-English-Korean Dictionary from Scratch

Vít Baisa^{1,2}, Marek Blahuš¹, Michal Cukr¹, Ondřej Herman^{1,2},
Miloš Jakubíček^{1,2}, Vojtěch Kovář^{1,2}, Marek Medved^{1,2},
Michal Měchura^{1,2}, Pavel Rychlý^{1,2}, Vít Suchomel^{1,2}

¹ Lexical Computing

² Masaryk University

Brno, Czech Republic

{firstname.lastname}@sketchengine.eu

Abstract

In this paper we present lexicographic work on a Tagalog-English-Korean dictionary. The dictionary is created entirely from scratch and all of its content (besides audio pronunciation) is initially generated fully automatically from a large web corpus that we built for these purposes, and then post-edited by human editors. The full size of the dictionary is 45,000 entries, out of which 15,000 most frequent entries are manually post-edited, while the remaining 30,000 entries are left only as automated. The project is currently ongoing and will be finished in December 2019. The dictionary will be part of the online platform run by the Naver Corporation¹ and freely available.

Keywords: Sketch Engine; Lexonomy; post-editing lexicography; dictionary; corpus; Tagalog; Filipino; English; Korean

1. Introduction

This dictionary project is the first in the series of three, the latter two are focusing on Urdu and Lao but otherwise follow the same scheme. The goal of the project is a modern, digital, corpus-based dictionary from Tagalog (Filipino) (as source language) to English and Korean (as target languages, treated equally). The key novel aspect of the dictionary building is that the contents of the dictionary will be created fully automatically using advanced natural language processing tools and a large web corpus of Tagalog, and most of the 45,000 target entries will remain automatic. Only 15,000 most frequent entries will be post-edited. In this paper we present the dictionary as a whole and specifically focus on two major methodological issues: the automatic drafting of the dictionary and the manual post-editing.

¹ Available at <https://dict.naver.com/>

2. On Tagalog

Tagalog is the most widely used language of the Philippines, where it is spoken by 24 million native speakers, along with additional 45 million second-language speakers who use its standardized form as the national language, officially called Filipino. It is an Austronesian language whose vocabulary has been influenced by a variety of foreign languages, most significantly English and Spanish. In spite of continuous efforts by the Philippine government to advance Filipino dating back to the 1935 constitution, and despite it being a compulsory part of the curriculum, the language is not used in all official domains; national law, business and government websites, for instance, are usually available only in English. Terminology in many fields has been reported to be inconsistent or missing, and code switching is a common practice. We have found this limiting with regard to the dictionary's coverage of certain topics.

The first comprehensive dictionary of Tagalog was compiled by Paul Klein, a Czech Jesuit missionary, in the beginning of the 18th century. His *Vocabulario de la lengua tagala*, inspired by earlier work by Franciscan friar Pedro de San Buenaventura, has itself become an inspiration for subsequent dictionaries of the same name, resulting in repeated reeditions until these days. Modern Tagalog is written using the Filipino Alphabet, which includes all the 26 letters of the ISO basic Latin alphabet, along with the Spanish Ñ and the Ng digraph.

While vocabulary is centred around root words and the division between parts of speech is much more blurry than in Indo-European languages, it is still possible to distinguish nouns, verbs, adjectives, and adverbs, although typically only according to the applied affixes or the position in the sentence. Verbs are the most variable part of speech – they are subject to a system of over 80 affixes, and their form determines the semantic role (“focus”) that the topic word plays in the sentence. There is no best choice for verbal lemma, because even in the infinitive there are still several possible lemmas per root word, each differing by the focus. If we listed only the root word in the dictionary (and redirect all the inflected forms to it), we would lose many important semantic distinctions, such as *bumili* (“to buy”) and *magbili* (“to sell”), which would be conflated within a single entry for the root word *bili* (the broad concept of “exchange”), without the possibility of providing an explanation of the differences in meaning (and the respective translations). On the other hand, inflection in other parts of speech is very limited; instead, the language makes use of particles.

3. Dictionary structure

The structure of the dictionary is simple but comprehensive, each entry in the dictionary consists of:

- a headword,
- a list of inflected forms,

- a recorded pronunciation,
- a division into senses, with each sense comprising:
 - a disambiguating gloss,
 - where appropriate, one picture,
 - 1–10 collocations,
 - 1–10 synonyms, antonyms and related words,
 - three post-edited examples and up to 10 more (fully automatic),
 - English translations of the headword and one example and
 - Korean translations of the headword and one example.

4. Automatic dictionary drafting and post-editing

The automation procedure entirely relies on data, tools and methods we made available in Sketch Engine (Kilgarrieff et al., 2014), a leading corpus management system. For Sketch Engine, we have crawled a 230-million-token corpus of Tagalog from the web and this has served as the basis for all the lexicographic work. While from the perspective of dictionary building the corpus is merely a needed by-product serving as empirical evidence for the automatic dictionary drafting, it represents a valuable linguistics resource as such (made available to general public through Sketch Engine), and to the best knowledge of the authors it is the biggest corpus for Tagalog as of July 2019.

The corpus was automatically part-of-speech tagged using Stanford PoS tagger (Toutanova et al., 2003)² and lemmatized using an in-house improved version of a Tagalog stemmer³. We also developed a sketch grammar so that related Sketch Engine's functions (mainly word sketches and thesaurus) become available.

For the post-editing phase of the 15,000 entries we used Lexonomy [Měchura, 2017], an open-source dictionary writing and editing tool. The editorial workflow consisted of isolated steps where editors were always post-editing only particular entry parts. In the following we explain in detail how individual entry parts were automatically generated and later post-edited. A dependency schema of the individual steps is provided in Figure 1.

² Model was obtained from <https://github.com/matthewgo/FilipinoStanfordPOSTagger>

³ Available at <https://github.com/crlwingen/TagalogStemmerPython>

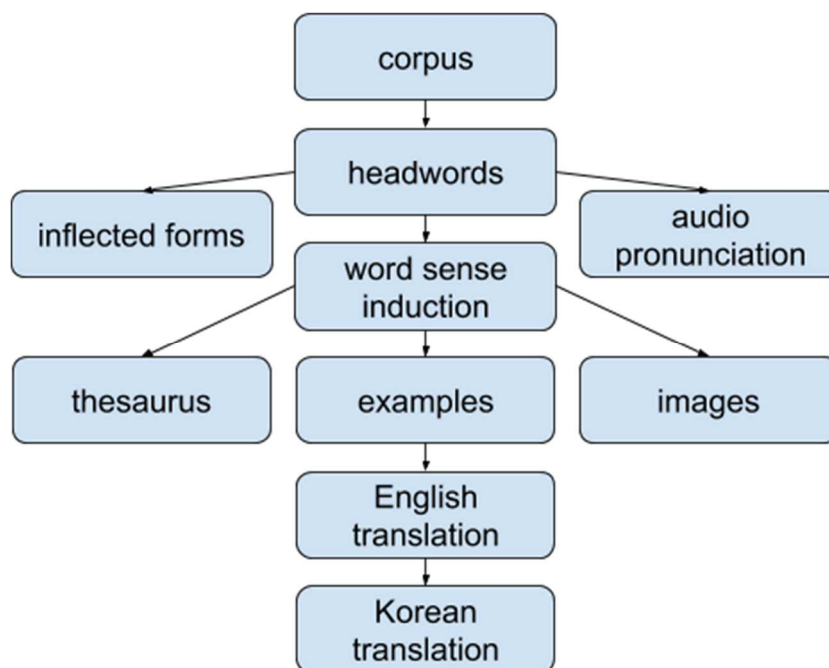


Figure 1: A dependency structure of all post-editing tasks.



Figure 2: Flagging inside of Lexonomy can be carried out with keyboard shortcuts or mouse clicks in the headword list.

4.1 Headwords

We have taken 45,000 most frequent corpus lemmas according to the document frequency. The editors have been validating them and removing non-words, foreign words, non-lemmas and proper nouns as well as correcting automatic part-of-speech tagging. The decision diagram for this task is given in Figure 7. The flagging feature was used for this task within Lexonomy (see Figure 2).

4.2 Inflected forms

Inflected forms were generated from the corpus based on the automatic lemmatization. Editors reassigned word forms to correct lemmas where necessary. Lexonomy features a built-in lay-by that behaves like an internal clipboard and is useful for moving entry parts across different entries.

4.3 Pronunciation

This is the only part of the entry that is done fully manually since there is no post-editing of automatic text-to-speech output possible. On the other hand, it turned out to be also one of the simplest tasks.

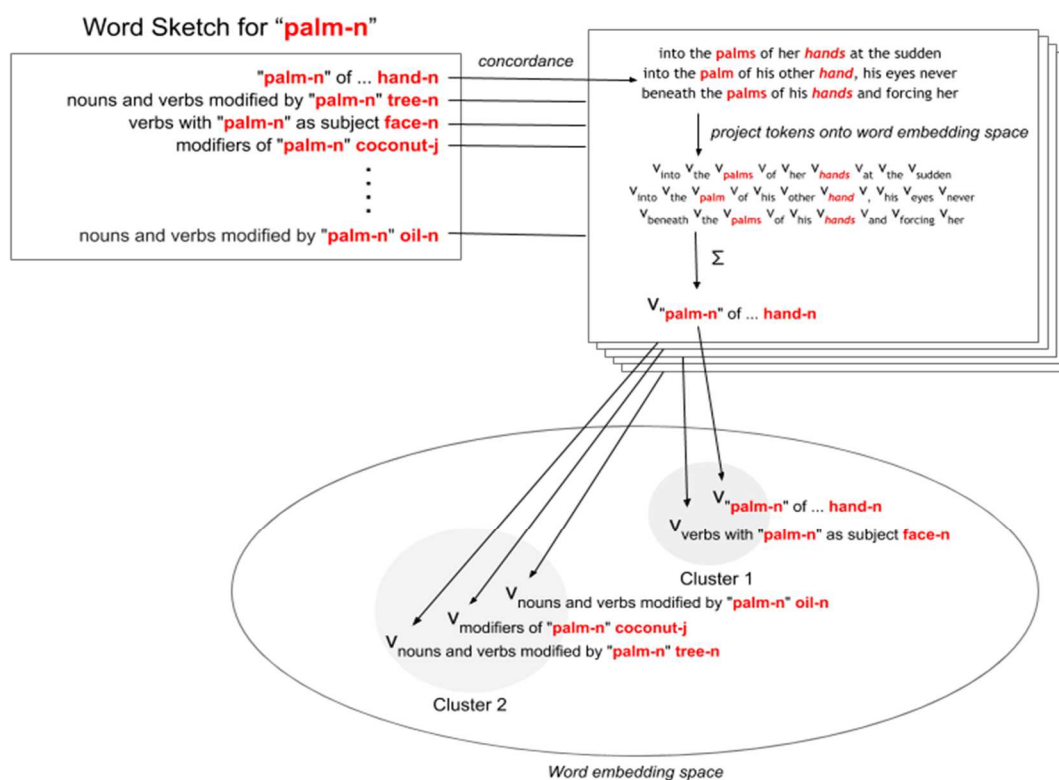


Figure 3: Workflow of a word sense induction algorithm that exploits word-sketch-based collocations and word embeddings.

We provided the editors with a recording tool that they used in a small acoustic chamber. The tool prompted them to press a key to start a three-second recording window and then read a headword, after which the recording was automatically replayed to them and they had the option to revise it or move to the next headword.

In this scenario, the editors were able to record about 900 headwords per working day (8 hours). Afterwards, the recordings were automatically trimmed for silence and normalized using the Sox tool.⁴

4.4 Word sense division

Word sense clusters have been induced using a method that combines word sketches with word embeddings. The algorithm is to be presented in a separate paper in detail, but principally works as follows:

- for an input headword, take all its collocations, filtered by frequency (at least 5) and logDice score (Rychlý, 2008) (at least 2),
- for each collocation, take vectors of all words within a short window (4 words) across all collocation occurrences in the corpus and calculate the average of these vectors.
- cluster vectors obtained in previous step using HDBSCAN clustering (McInnes et al., 2017).

The workflow is also illustrated in Figure 3. The word embeddings were calculated on the source corpus using FastText (Bojanowski et al., 2016). The result of this procedure is a set of clusters, each consisting of one or more collocations, each being represented by a set of concordance lines in the sources corpus. Having each sense represented by a set of concordance lines is a very important principle that allowed us to proceed with many subsequent actions (e.g. example selection) on a per-sense level.

Editors were subsequently lumping and splitting the automatically induced clusters. Each cluster consisted of associated collocations and was backed by a set of concordance lines allowing users to inspect the underlying corpus evidence. For this task we have developed a custom editing widget for Lexonomy that is given in Figure 4. For every cluster, the editors may move the whole cluster or individual collocations into another sense or create a new sense. Alongside the senses, the editors were also post-editing English translations of the headword in each of its senses as well as assigning disambiguating glosses for each sense.

⁴ Available from <http://sox.sourceforge.net/>.

bago_{ADJECTIVE}
Senses:
▶ sense 1 named: ✖
▶ sense 2 named: ✖

Translations:
 ✖
 ✖
 ✖

cluster 1

Mark all:

example usage	actions					collocate	relation to headword	concordance
<i>mga bagong bayani</i>	<input type="button" value="1"/>	<input type="button" value="2"/>	<input type="button" value="NEW"/>	<input type="button" value="MIXED"/>	<input type="button" value="ERROR"/>	bayani _{NOUN}	nouns modified by "bago"	
<i>ang bagong prinsesa</i>	<input type="button" value="1"/>	<input type="button" value="2"/>	<input type="button" value="NEW"/>	<input type="button" value="MIXED"/>	<input type="button" value="ERROR"/>	prinsesa _{NOUN}	nouns modified by "bago"	
<i>bagong superhero</i>	<input type="button" value="1"/>	<input type="button" value="2"/>	<input type="button" value="NEW"/>	<input type="button" value="MIXED"/>	<input type="button" value="ERROR"/>	superhero _{NOUN}	nouns modified by "bago"	

cluster 2

Mark all:

example usage	actions					collocate	relation to headword	concordance
<i>mga bagong sibol na</i>	<input type="button" value="1"/>	<input type="button" value="2"/>	<input type="button" value="NEW"/>	<input type="button" value="MIXED"/>	<input type="button" value="ERROR"/>	sibol _{NOUN}	nouns modified by "bago"	
<i>mga bagong halaman</i>	<input type="button" value="1"/>	<input type="button" value="2"/>	<input type="button" value="NEW"/>	<input type="button" value="MIXED"/>	<input type="button" value="ERROR"/>	halaman _{NOUN}	nouns modified by "bago"	
<i>bagong puno ng</i>	<input type="button" value="1"/>	<input type="button" value="2"/>	<input type="button" value="NEW"/>	<input type="button" value="MIXED"/>	<input type="button" value="ERROR"/>	puno _{NOUN}	nouns modified by "bago"	

Figure 4: A custom editing widget created for the purposes of post-editing word sense induction in Lexonomy.

4.5 Disambiguating glosses

Disambiguating glosses (in Tagalog) were initially assigned when post-editing the word sense induction. Afterwards, they were reviewed by another editor and amended if necessary.

4.6 Pictures

Pictures have been automatically searched for in three online databases that offer API to access copyright-free images, namely Wikimedia Commons (Wikidata and Wiktionary) ⁵, PixaBay ⁶ and Google Image Search ⁷ (only if no pictures were found in the previous two image sources).

The content of both Wikimedia Commons and PixaBay is copyright-free for the purposes of this project (being licenced as either CC0, CC-BY or CC-BY-SA). For Google Image Search we limited the search to pictures allowing commercial use with modifications.

⁵ See <https://commons.wikimedia.org/wiki/ Commons:API>

⁶ See <https://pixabay.com/service/about/api/>

⁷ See <https://www.googleapis.com/customsearch/v1>

Initially, each sense was accompanied with ten images. Regrettably, English turned out to be the only reliable search language for all three engines we used. Afterwards, editors were selecting the best picture (up to three) out of the candidates, obtaining new images if necessary, as seen in Figure 5.

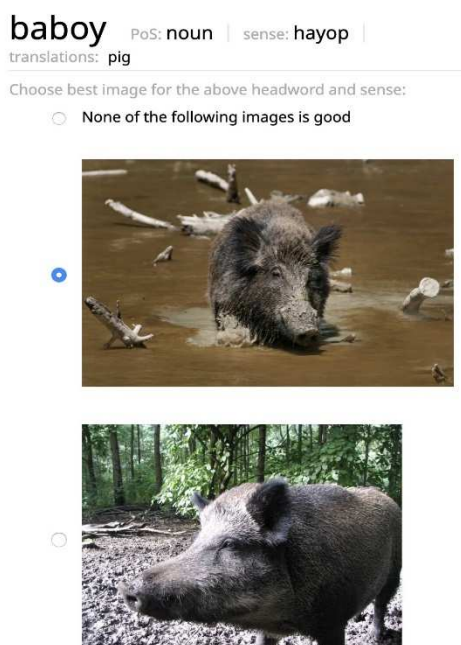


Figure 5: Post-editing interface for the selection of images matching the given word sense.

4.7 Collocations

Collocations were initially obtained using word sketches in the word sense induction phase. However, because the clustering algorithm generally identified good representatives to separate clusters, there are typically many unclustered but still salient collocations. Therefore we have been adding, after the word senses were post-edited, all high-scoring collocations if they were not clustered automatically. The goal was that for every grammatical relation in the word sketches, the top three collocations must be reviewed and added to the right sense if necessary.

It is important to emphasize the difference between using collocations as the vehicle for word sense induction (yielding clustered collocations) and making sure that all salient collocations are part of the entry and that this is not guaranteed by the word sense induction itself.

4.8 Synonyms, antonyms and related words

Semantically related words were obtained using Sketch Engine’s built-in thesaurus. We took advantage of having the word senses already post-edited and calculated the thesaurus on the sense level by adding another positional attribute indicating the sense

(based on the post-edited collocation occurrences). Early investigations have shown that the dominant sense prevails when looking up the thesaurus disregarding sense (i.e. just based on lemma and part of speech combination). On the other hand, such a sense-disregarding thesaurus tends to yield better results for the dominant sense (but not for other senses) because only a fraction of the collocations was typically clustered even for the dominant sense.

Therefore the editors were provided with the following data for each sense:

- top 10 items from a sense-disregarding (default) thesaurus
- top 10 items from a sense-based thesaurus.

Editors were then classifying all items into synonyms, antonyms, other related words (that are neither of the previous) and unrelated words in the post-editing phase.

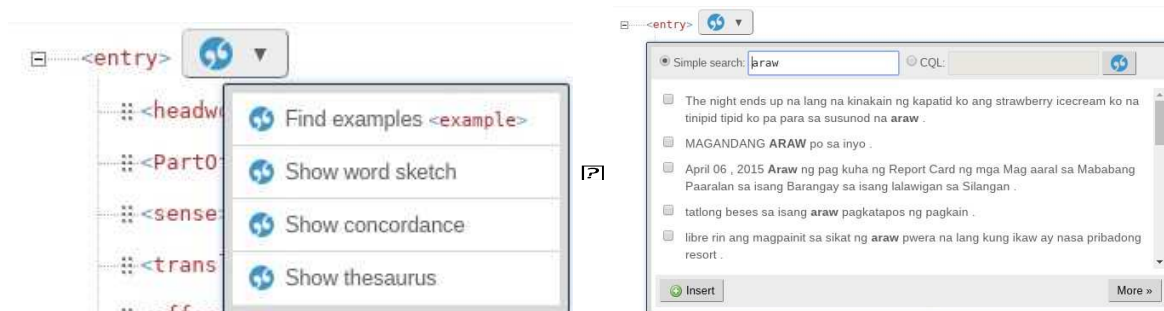


Figure 6: Retrieving additional examples from within Lexonomy by calling Sketch Engine API.

4.9 Examples

Examples were generated using the GDEX functionality of Sketch Engine [Kilgarrieff et al., 2008]. Editors selected the best of them or replaced them with new ones from the corpus using the pull model for interaction between Lexonomy and Sketch Engine (see Figure 6).

4.10 Translations

Translations to English were obtained automatically using Google Translate (which gives only one translation in the API) and Microsoft Bing (which can give multiple translations in the API), the results were merged and afterwards post-edited by translators. This happened during the post-editing of word sense for the translation of headwords/senses, and when post-editing the examples for the translation of the examples.

In the next phase, translations to English were validated by independent translators to assure their quality.

Translations from English to Korean were also carried out by post-editing machine translation output using the same commercial services. To be able to translate the isolated headwords/senses, the translators were carrying out that task together with translating the examples.

5. Editorial team and its post-editing workflow

Our team of editors consisted of seven adult native or near-native speakers of Tagalog, all with roots in the Philippines. They came from various social groups and had various occupations and educational backgrounds. All of them spoke both Tagalog and English, some also mastered another local language. In the recruitment process, we preferred the candidates *not* to be linguists, because the goal was to extract all the linguistic knowledge from the corpus and use human editors only to provide feedback on the quality of the machine-generated output and manually post-edit a selected portion of the entries. The Korean translations were commissioned to a professional Korean translator.

Work was distributed to the editors in batches in order to better account for individual needs. Before each new activity (such as headword annotation, proofreading of inflected forms, word sense division etc.), editors participated in a short training. For each activity, the content of the first batch was the same for everyone in order to check comprehension of the task, measure inter-annotator agreement and establish an average processing time per entry for each editor. The tasks were explained to the editors with as little linguistic terminology as possible, and the interface of the task-specific custom editors developed in Lexonomy was designed in order to reinforce this principle. For instance:

- In word sense division, the field to enter a disambiguating gloss was labelled in simple words: “sense name”.
- In the list of collocations, the longest–commonest match representing a collocation was titled “example usage”.
- Instead of being asked to regroup collocations among clusters (and actually feel that they are doing lumping and splitting), the editors were told to assign a sense number to each collocation in a list. This design choice has saved much clicking and the task could often be completed in a single pass.

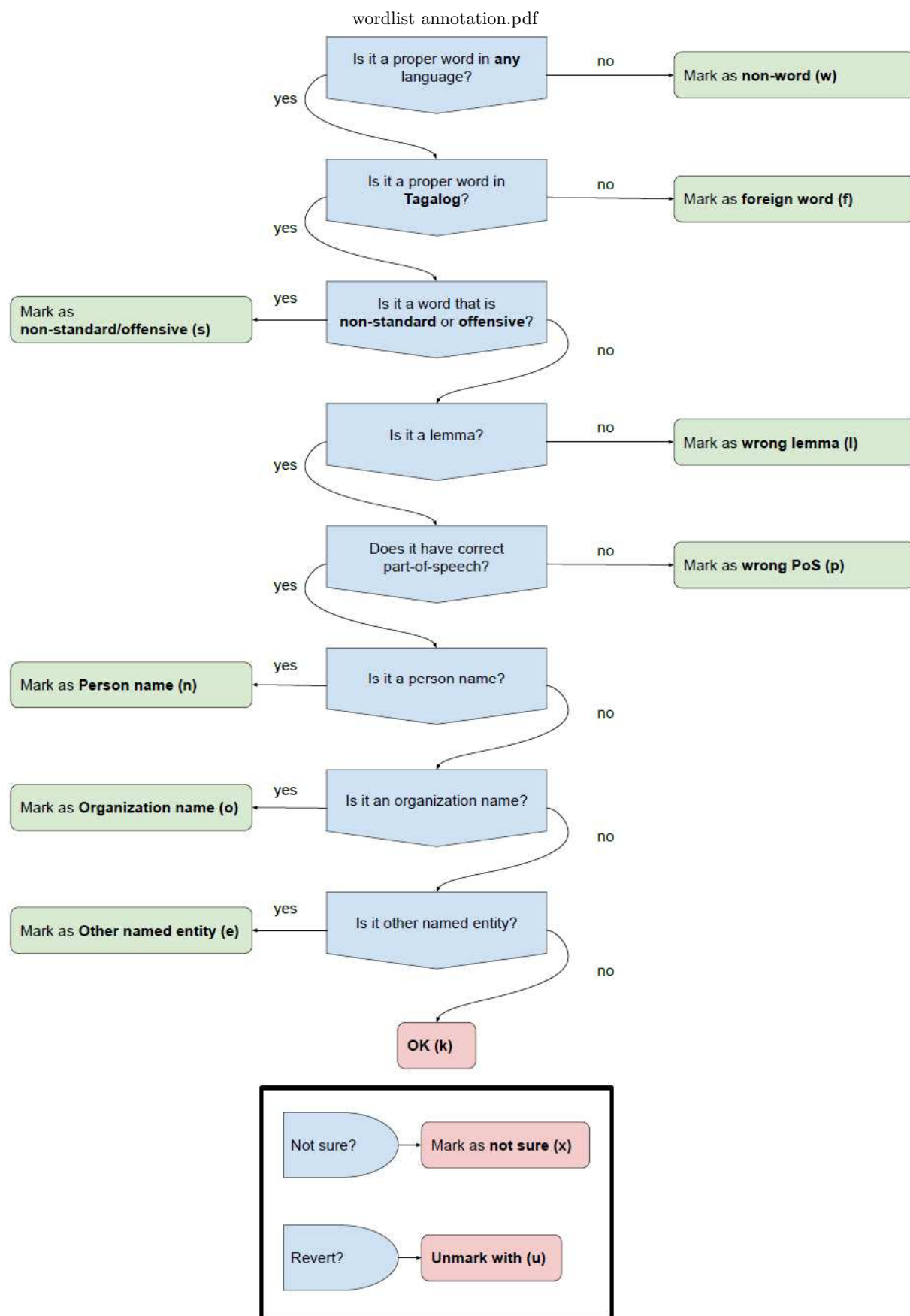


Figure 7: A diagram showing the decision process when filtering an automatically produced headword list.

- The fact that the listed collocations were grouped in clusters was not commented on at all, as the only purpose thereof was to speed up the editors' work (putting collocations that presumably belong to the same sense next to each other) and no knowledge of the underlying logic was required on their part.

Inter-annotator agreement was measured after the completion of the first batch for each task. When that was done, we would usually summon the editors again and confront them about the patterns of disagreement in their output. At that moment, we would improve the written guidelines for the task (and possibly even reinvent the annotation process if necessary) which had only been sketched or non-existent until then. Following this meeting, editors would each be given a different set of data in order to speed up the process and cut on costs, but a small percentage of data would routinely be placed in two sets (either belonging to different editors, or subsequent sets belonging to the same editor) in order to monitor agreement and consistency throughout the whole process. Contact among editors was not discouraged – after all, they would spend time together during the training and some had already known each other before the start of the project – but attention had to be paid to prevent unwanted interdependence, particularly when all editors were working on the same set of data. On the other hand, we welcomed the creation of a chat group by the editors, which they could use for seeking and giving advice among themselves, both regarding the project's technical aspects and the linguistic uncertainties they had encountered during their independent work.

Only items (headwords, word senses, example sentences) that had been accepted in one postprocessing phase could advance into the following one. In spite of that, the editors would still occasionally discover wrong items at a later stage (such as being asked to review possible inflected forms of a word that is in fact not a lemma). This has served as an extra level of quality control and for each task, editors were instructed what to do when they come across such a case. As soon as the first headwords were completing their passage through the whole post-editing process and first entries emerging, we focused our attention back on the data that had been discarded or not yet available at the earlier stages: in close cooperation with the editors, we tried to fix errors in the lemmatization process manifested by the appearance of wrong lemmas in the list of headword suggestions. We would also consider any new headwords (or inflected forms) that may have emerged if we had increased the size of the source corpus since the start of the work. Any newly discovered headwords would then be submitted into the same pipeline as their predecessors, until there was no valid input left to be processed.

6. Conclusions

In this paper we report on a newly created Tagalog-English-Korean dictionary. The dictionary is fully corpus-based and the key novel aspect of its development is that the whole dictionary was initially created in a fully automatic way and afterwards manually post-edited where necessary. The post-editing phase presents many new challenges and

is far from being a finalized approach, but clearly shows its viability, affordability and performance benefits as for the time taken to produce the dictionary, which was about 9 months.

Overall the biggest challenge in this approach is to maintain solid data and user management rather than assuring sufficient quality of the automated outputs. The post-editing requires a lot of back-and-forth and trial-and-error, each being sensitive to careful data preparation and processing as well as being very communication intensive. More automation is clearly required to make these procedures robust, less error-prone and more affordable for less technically skilled lexicographers.

As for the automated tasks, it is worth mentioning that word sense induction turned out to be less of an issue than anticipated. The algorithm used tends to perform rather well for high-frequency polysemous words (but of course a more thorough evaluation should definitely be performed which was outside scope of our very practically motivated project). Throughout the tasks the importance of the size and quality of the corpus and its annotation was heavily manifested. We struggled a lot to crawl the at least 600-million word corpus, which we do not consider to be very big (although as far we know the biggest one for Tagalog). It was very obvious that a bigger corpus and better part-of-speech tagger and lemmatizer would improve the quality of the automated outputs as well as simplify some of the post-editing tasks a lot.

To summarize the issues we faced, data and user management were the major ones, then, less seriously, the corpus and its annotation, while all the automation procedures worked more or less as expected and did not cause any major issues.

Two more dictionaries are now in the pipeline following the same approach, where the source languages are Urdu and Lao. We continuously improve the tools and workflow, and will report on the other two dictionaries in a separate paper.

7. Acknowledgements

This work is part of a joint project with Naver Corporation. This work has been partly supported by the PhD Funding Scheme of Lexical Computing at Masaryk University. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. This work has been partly supported by the Grant Agency of CR within the project 18-23891S and the Ministry of Education of CR within the OP VVV project CZ.02.1.01/0.0/0.0/16_013/0001781 and LINDAT-Clarin infrastructure LM2015071.

8. References

- Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. (2016). Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.
- Kilgariff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P.

- & Suchomel, V. (2014). The sketch engine: ten years on. *Lexicography*, 1(1), pp. 7–36.
- Kilgariff, A., Husák, M., McAdam, K., Rundell, M. & Rychlý, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus. In *Proceedings of the XIII EURALEX International Congress (Barcelona, 15-19 July 2008)*, pp. 425–432.
- McInnes, L., Healy, J. & Astels, S. (2017). hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, 2(11), p. 205.
- Měchura, M. B. (2017). Introducing lexonomy: an open-source dictionary writing and publishing system. In I. Kosem et al. (eds.) *Electronic Lexicography in the 21st Century: Lexicography from Scratch. Proceedings of the eLex 2017 conference, 19-21 September 2017*.
- Rychlý, P. (2008). A lexicographer-friendly association score. *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN*, pp. 6–9.
- Toutanova, K., Klein, D., Manning, C. D. & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pp. 173–180. Association for Computational Linguistics.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



An Open Online Dictionary for Endangered Uralic Languages

Mika Hämäläinen, Jack Rueter

Department of Digital Humanities, University of Helsinki
E-mail: mika.hamalainen@helsinki.fi, jack.rueter@helsinki.fi

Abstract

We describe a MediaWiki-based online dictionary for endangered Uralic languages. The system makes it possible to synchronize edits done in XML-based dictionaries and edits done in the MediaWiki system. This makes it possible to integrate the system with the existing open-source Giellatekno infrastructure that provides and utilizes XML formatted dictionaries for use in a variety of NLP tasks. As our system provides an online dictionary, the XML-based dictionaries become available for a wider audience and the dictionary editing process can be crowdsourced for community engagement with a full integration to the existing XML dictionaries. We present how new automatically produced data is encoded and incorporated into our system in addition to our preliminary experiences with crowdsourcing.

Keywords: online dictionary; endangered languages; Uralic languages

1. Introduction

Open-source resources have been developed in the past for a number of endangered Uralic languages in the Giellatekno infrastructure (Moshagen et al., 2014). Giellatekno is the North Sami word for *language technology*, and work in the infrastructure at what today is known as the Norwegian Arctic University originally highlighted rule-based and finite-state descriptions of Sami languages in cooperation with the language communities. In addition to the Giellatekno research portion, a complementing implementational actor *Divvun* ‘correction’ has been established by the Sami Parliament for developing orthographic and morphological spellcheckers, keyboards, syntax checkers, machine translation, etc. Naturally, other Nordic languages are included in the infrastructure as well as minority languages of the Barents Sea and even larger Circum Polar Regions. The list of language projects amounts to over one hundred, with around 50 active projects. Some finite-state language descriptions now hosted date back to work in the early 1980s, while others are only now emerging.

Finite-state description with rule-based solutions at Giellatekno caters to languages with complex morphology. The philosophy at Giellatekno-Divvun includes multiple reuse of resources, i.e. by originally developing analysers for linguists, we are able to

produce almost simultaneously basic spellcheckers ^{1,2}, and, at the same time, we can develop work with intelligent computer assisted language learning ³. In late 2012 and early 2013 a project involving the development of online morphology-savvy dictionaries and click-in-text dictionaries was also spearheaded at Giellatekno for several well described languages, for example North Sami ⁴ and South Sami⁵.

With the start of the Kone Foundation Language Programme, in Finland (2013–2017), it was decided that new minority language projects such as Livonian⁶, Olonets-Karelian⁷, Izhorian, Hill Mari⁸, Erzya-Mordvin⁹, Moksha-Mordvin, Komi-Zyrian¹⁰ and Tundra Nenets¹¹ could readily be included among the online morphology-savvy dictionaries with spell relax mechanisms (see also Rueter, 2014). What was special about the newly introduced languages was that the online dictionary XML databases simultaneously served as the source for XSL transformation and transducer generation. Thus, basic information included in the XML files consisted of lemma, derivational stem, part-of-speech and specific inflectional type information, which was complemented by translations into Finnish and possibly other languages. Subsequent work with XML dictionaries has introduced additional languages, e.g. Skolt Sami¹², Udmurt, Komi-Permyak and Meadow Mari.

These XML resources featured in many of the Uralic language projects, however, are not easily available for people who are unfamiliar with technically advanced presentations, as they are provided in source code format.

We present a MediaWiki-based multilingual online dictionary for endangered Uralic languages. The dictionary not only makes the lexicographic resources available for ordinary users, but it makes dictionary editing possible in a crowd-sourced fashion with an XMLMediaWiki synchronization (Hämäläinen & Rueter, 2018). This means that any edits made in the original XML files in the Giellatekno infrastructure will be synchronized to the online dictionary, and vice-versa.

¹ <http://divvun.no/>

² <http://divvun.org/>

³ <http://oahpa.no/davvi/>

⁴ <https://sanit.oahpa.no/>

⁵ <https://baakoeh.oahpa.no/>

⁶ <http://sonad.oahpa.no/>

⁷ <http://sanat.oahpa.no>

⁸ <http://muter.oahpa.no/>

⁹ <http://valks.oahpa.no>

¹⁰ <http://kyv.oahpa.no/>

¹¹ <http://vada.oahpa.no>

¹² <http://saan.oahpa.no>

The lexicographic entries in our online dictionary have been automatically enhanced with a multitude of Semantic MediaWiki tags. In the past, Semantic MediaWiki has been shown to be a viable way of integrating semantic web compatible information with an online dictionary (Laxström & Kanner, 2015). Our online dictionary also provides an API access to its resources. Over the API, lexicographic entries can be retrieved in JSON format and the FST transducers can be used both for morphological analysis and generation.

In this paper, we provide insight on the functionalities of our MediaWiki-based online dictionary system. Furthermore, we describe how lexicographic information newly obtained by using language technology approaches is incorporated into the online dictionary.

Currently, we support 13 endangered Uralic languages such as Skolt Sami, Komi-Zyrian, Udmurt and Erzya. We have initially experimented with crowd-sourcing for Skolt Sami and Erzya with positive results.

2. Related work

In the modern era, developing accessible and easy to use dictionaries for endangered languages has become one of the important research interests in language documentation and revitalization. Some of the work focuses more on building a new dictionary out of scratch, whereas others focus on making already existing paper dictionaries accessible for a wider audience in a much more modern fashion. In this section of the paper, we describe some of the contemporary work on this topic.

Work with endangered languages in North America has shown that the language novice must be provided for. The communities are small, and unfamiliarity with lexicographic tradition can easily be detrimental to the novice’s language learning experience. The new language learner cannot be expected to know where a dictionary entry lies nor automatically adopt the normative orthography. When the language user either lacks the keyboard or the knowledge to spell correctly, spell relax strategies can be implemented in online and mobile morphology-savvy solutions. Morphologic awareness and spell relax are used in catering to the Tsimshianic and Salish novice in dictionary use and language technology (Littell et al., 2017). On an entirely separate front, work has also been done to provide the St. Lawrence Island Yupik community with unhindered access to language materials online. This, once again, has been accomplished using a morphologically aware dictionary. In this separate rendering of the same kind of system, however, a strategy of multiple input methods catering to different writing systems (Hunt et al., 2019) has been introduced. The work here is tailored, and a strong tie is maintained between a language and its community. These endangered languages fall into the category of low-resourced languages.

‘Low-resourced language’, however, is a term used for almost any language with a lower internet presence than English. In (Nasution et al., 2018), in contrast, the Malaysian languages dealt with are relatively small in comparison to the majority languages encompassing them. The approach is to address a group of closely related languages simultaneously – an underlying multilingual or language-independent infrastructure. Pivot languages are used as means of enriching bilingual lexical resources. The authors discuss drawing upon bilingual dictionary input, and the difficulty of selecting the right bilingual dictionaries to start from.

One part of the strategy is to use cognates found through pivot-languages for locating translation candidates. Cognates are subsequently paired with multiple synonyms, and these synonym continua are established in many-to-many translation blocks. This is one of the places where native speaker editors are employed in the evaluation of automatically generated much needed lexica. Since the focus is on a larger language populations, outlines are made of actual expenses incurred in editing bilingual lexical resources, i.e. expenditures based on 10 and 30-second increments in an eight-hour day.

Low resource endangered languages do not necessarily have the native speaker-editor population to draw upon. Therefore, language-independent approaches are merited even here.

3. The MediaWiki-based dictionary

The main motivation behind the use of MediaWiki is to make the Giellateknko XML dictionaries authored for a multitude of endangered Uralic languages available for the general public. This is done in a synchronized way so that edits done in both the XMLs and the MediaWiki can be synchronized. This will ensure the availability of the latest version of the data for all users.

Uralic languages are known for their highly inflectional morphology. This makes the use of traditional dictionaries difficult, as a language learner will have to successfully inflect a word form he has encountered in a text to its lemma form in order to find it in a dictionary.

To alleviate this problem, our online dictionary includes finite-state morphological analysers (cf. Beesley & Karttunen, 2003) that will lemmatize the user input before querying the lexicographic database. In this way, the user can find the lemma and its translations even when it comes to morphologically complex word forms. These analysers are generated from the XMLs that can be edited in the MediaWiki system (cf. Rueter & Hämäläinen, 2017).

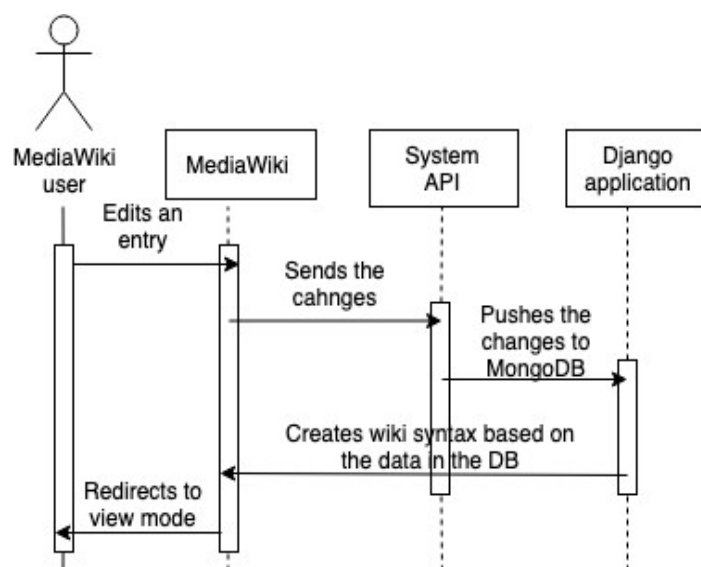


Figure 1: A diagram showing an edit on the MediaWiki side.

The synchronization of editing has been done in such a way that the up-to-date data is available for both the people working with the XMLs and on the MediaWiki. This is important as technically savvy people find XML-based editing more powerful whereas non-technical people would have problems working with the markup, where even adding a wrong character might render the whole XML syntax invalid. Figure 1 shows the process from the point of view of the person doing edits on the MediaWiki. Whenever the user is done with editing an entry in the dictionary, a Django-based synchronization system is informed. The Django system keeps an up-to-date backup in JSON format of all the entries in the dictionary. The edited entry is sent by a MediaWiki extension as JSON to the Django-based system, which updates its own database with the updated entry and re-formats the data in MediaWiki syntax to store it in the MediaWiki dictionary for visualization to the dictionary users.

Editing the XMLs is a slightly more complicated process, as shown in Figure 2. We have decided to build the XML editing on top of Git as it provides versioning and it makes it possible to compare the different versions and resolve potential conflicts in an easy to use fashion, especially due to the availability of a myriad of Git tools with a graphical user interface. The process starts by the lexicographer using a custom Git script to pull the latest version of the XML from the Django system running on the server of the MediaWiki system.

Once the lexicographer is done with the edits of the XMLs, he can push the changes to the master branch of the GitHub repository. This will initiate a pull on the MediaWiki server and the Django-based system starts a background process to first update its own internal database with the changes in the XML files, and then generate and update MediaWiki syntax for the updated entries.

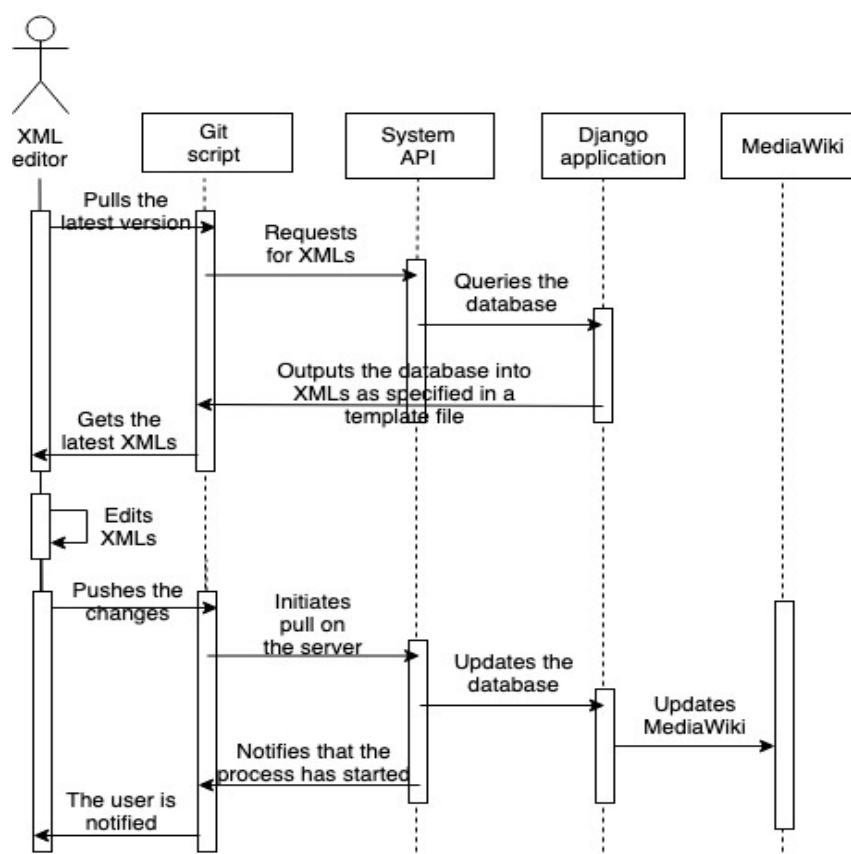


Figure 2: A diagram showing an edit on the XML side.

4. Representing the new information

This section of the paper is dedicated to describing how the data obtained by automatic language-technology methods for Uralic languages has been incorporated to our MediaWiki-based online dictionary system. Making the new data available on a system that also serves for non-academic usage is important not only for revitalization of the endangered Uralic languages, but also for community involvement.

Hämäläinen et al. (2018) presents work on combining dictionaries automatically for Skolt Sami, Erzya, Moksha and Komi-Zyrian based on the XML dictionaries also available on our MediaWiki dictionary. As all of the dictionaries are multilingual, meaning that every entry in a minority language has translations into multiple majority languages (most typically Finnish, English and Russian), it is possible to combine translation entries for all of the four minority languages. This is based on two assumptions, firstly the XML structure has meaning groups, which means that translations are grouped by senses, and secondly if a meaning group has translations into two different languages, the languages will make a semantic distinction and therefore translations that do not refer to the same meaning are not combined.

In practice, the approach takes an entry in Skolt Sami, such as *blin*, which has translations into Finnish *ohukainen* and *blini* and in English *pancake* and compares it to an entry in Komi-Zyrian, which in addition to the same translations as in the Skolt

Entry, also has the synonyms *räiskäle* in Finnish and *crepe* in English. As there is an overlap between the entries, the method extends the Skolt Sami entry with the additional synonyms from the Komi-Zyrian entry.

In order to incorporate these results into our MediaWiki dictionary, it is important to introduce a new attribute to the XML structure, namely an ID for each individual meaning group. When the meaning groups can be identified, the linking of the dictionary entries can be done on the system level. Currently, a hand-curated set of the automatic results presented in Hämäläinen et al. (2018) are included in our online dictionary. In the future, their approach could be included in a dynamic fashion in our system so that whenever a new entry is added on the MediaWiki platform, a set of possible translations together with links to meaning groups in other languages could be brought as suggestions to the dictionary editor.

Semantiikan juurielementti

MG: X

MG: X

Käännökset

Kielen tunnus (esim. eng)

Käännös	Sanaluokka	Nimi	Arvo	Poista
ohukainen	N	• Nimi: mg	Arvo: 0 X	X
			<input type="button" value="Lisää arvo"/>	
levy	N	• Nimi: mg	Arvo: 1 X	X
			<input type="button" value="Lisää arvo"/>	

Figure 3: Meaning groups in the MediaWiki edit form.

Meaning groups (MGs) have editable locally unique IDs in the edit form of MediaWiki, as seen in Figure 3. Meaning groups can be added as needed. Translations in different languages are grouped together when the dictionary data is visualized for the user based on the meaning group IDs.

SemUr and SemFi (Hämäläinen, 2018) are automatically extracted semantic databases for Skolt Sami, Erzya, Moksha, Komi-Zyrian and Finnish. These databases represent corpus frequencies of co-occurrences of two words given a syntactic relation. Through this data it is possible, for instance, to see which words can act as a subject or object for a given verb. This can be a useful resource for a lexicographer especially as it reveals

information about polysemy, not to mention the number of links it introduces in between the different dictionary entries.

The graph like relation structure calls for a different visualization strategy to what is commonly used in MediaWiki. Therefore, we create our own MediaWiki extension that can be used to visualize and browse the semantic databases. This visualization can be accessed from a dictionary entry on the MediaWiki.

Figure 4 shows the interface incorporated in our MediaWiki-based dictionary for browsing the semantic data. In the example, the adjective modifiers and verbs with the subject relation are shown for the Finnish word *kirves* ‘axe’. The interface gives the possibility to focus on related words of a certain part-of-speech or syntactic relation.

Recent work using neural networks to extend cognate relations for Skolt Sami and North Sami (Hämäläinen & Rueter, 2019) is an important data point for lexicographic work. Cognates from closely related languages can further be used in a multitude of language technology applications. Cognate relations are introduced to our online dictionary by linking words sharing a cognate relation to each other. This way, a dictionary user can move from one entry to its cognate easily. The same linking functionality is also used to link compound words with their constituents.

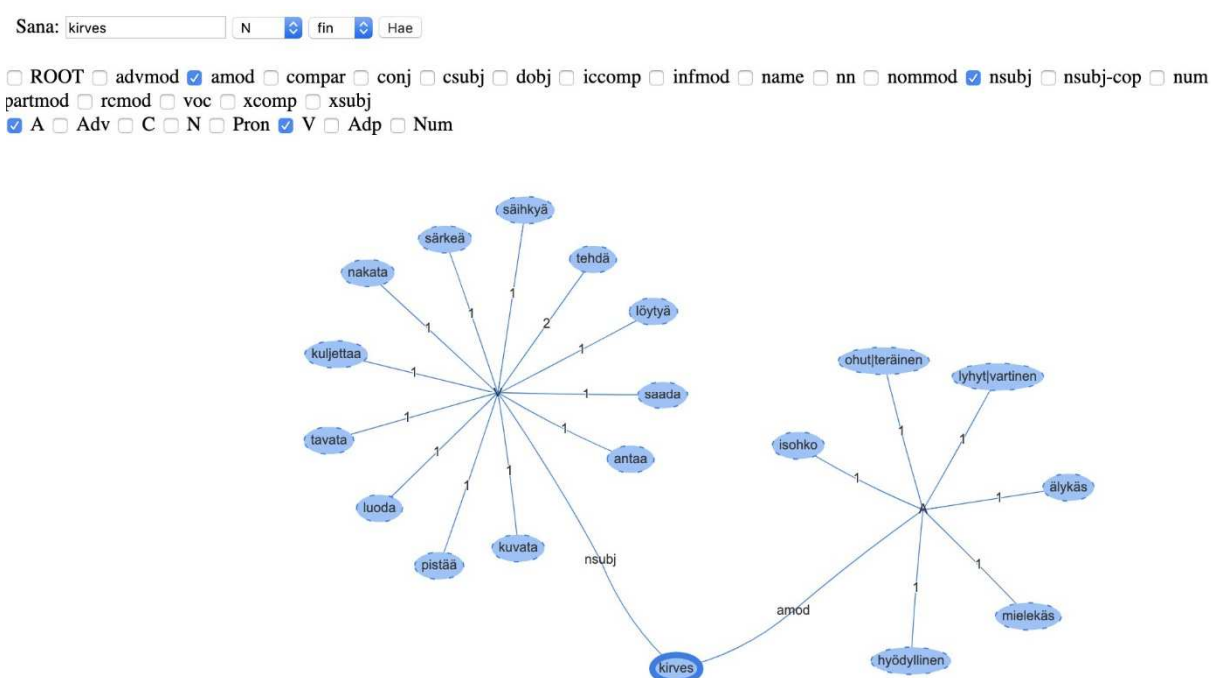


Figure 4: Interface for browsing semantic data.

č̣iõhč̣lõs (adjektiiv)

Näytä etymologia

- [čiekčalas](#) (cognate)
 - modality - plausible
 - pos - A
 - src - aku_2019-02-24
 - xml_lang - sme

Figure 5: Cognate view in the dictionary.

Cognates can be viewed by clicking on a button titled *Näytä etymologia* ‘show etymology’, as seen in Figure 5. Information is shown about the cognate word together with a link to its entry in the other language.

All of this new information introduced into the system has been made available for programmatic access over the custom API of the MediaWiki dictionary. The access to this API has been integrated into UralicNLP Hämäläinen (2019), which is an open-source Python library for processing endangered Uralic languages.

5. Crowd-sourcing

Our initial experiments in crowd-sourcing have been limited to a small number of people due to the fact that the communities speaking the endangered languages in question are not as big as they are in the case of majority languages. Nevertheless, crowd-sourcing serves for the purpose of exposing the XML structured dictionaries to non-technical linguists and community members.

Work with the Skolt Sami, Erzya and Komi-Zyrian language communities has included actual editing of MediaWiki materials that have directly augmented the dictionary database and hence enhanced the materials and tools available on the parallel Giellatekno infrastructure. During the summer of 2017, one work involving community linguists added much needed verbal derivation content in addition to example sentences from language archive materials at the Giellagas Institute in Oulu, Finland. In this two-month trial, conflicts between MediaWiki editors and XML editors were resolved. Additional input parameters that were found necessary were incorporated into the infrastructure to allow for sound-to-text alignment of archive materials in future work with Skolt materials, i.e. this was ground-breaking with regard to future work with other languages as well.

A second encounter with community collaboration was organized at the end of 2017. This time around, native and virtually native speakers were asked to evaluate automatically aligned concept translations. The alignments consisted of one source-language word with translations into several target-language words. The alignment had been facilitated using two pivot languages. In this way, new translations were shared

between dictionaries for the source languages Skolt Sami, Erzya, Moksha and Komi-Zyrian. Translation languages included English, Finnish, French, German, North Sami, Norwegian, and Russian, as well as some other minority languages. The task consisted of (i) accepting, (ii) not accepting, and if not accepting (iii) noting. Although the nature of the task was relatively straightforward, finding native speakers with adequate knowledge in three or more languages was a problem, but not entirely unsurpassable.

Crowd-sourcing introduces issues of access and tools in general. Work with language communities lacking active representatives in the Finnish academic community introduces a need for issuing non-academic usernames and access. This required the system to be moved away from using Haka credentials, which is a nation-wide authentication system for academic institutions in Finland. Levels of access must then be established that, on the one hand, allow access to language community activists and researchers and, on the other hand, ensure the integrity of the open-source multilingual lexical data synchronously maintained in Tromsø, Norway and Helsinki, Finland. Once access has been established, there is a need to maintain quality control of the data, i.e. one source of problems is that Skolt Sami has several Latin characters available only on a few open-source keyboards, the same applies to Komi-Zyrian and the Mari languages, which have letters from outside the Russian Cyrillic alphabet – should there be a virtual keyboard available.

6. Discussion and future work

Our online dictionary system represents a big leap towards the correct direction in making language resources available both for regular dictionary users and for more technically oriented users through the open API. However, as indicated by our crowd-sourcing experiments, some additional care has to go into streamlining the usability of the dictionary editing. Currently, the edit form reveals a myriad of detailed information such as continuation lexicon and stem group, which might be overwhelming for an average language speaker. This calls for more user-centric usability testing to be conducted in the future.

The combined meaning groups from Hämäläinen et al. (2018) have been introduced into the system in a static fashion. However, their method could, in the future, be integrated into our system in a more dynamic way. In practice, this would mean that a dictionary editor adding a new entry for any language in the system would get recommendations for other candidate translations to choose from. This could speed up the process of conducting lexicographic work with endangered languages.

More active engagement of the community members is needed in the future. The first step to make contributing to the dictionary as easy as possible would be localization of the interfaces used. First and foremost to Russian, as a vast majority of the endangered Uralic languages are spoken in Russia, but also localization to all the supported endangered languages.

7. References

- Beesley, K. R. & Karttunen, L. (2003). *Finite-State Morphology*. Stanford, CA: CSLI Publications, pp. 451–454.
- Hunt, B., Chen, E., Schreiner, S. L. & Schwartz, L. (2019). Community lexical access for an endangered polysynthetic language: An electronic dictionary for St. Lawrence Island Yupik. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 122–126. <https://www.aclweb.org/anthology/N19-4021>.
- Hämäläinen, M. (2018). Extracting a Semantic Database with Syntactic Relations for Finnish to Boost Resources for Endangered Uralic Languages. In *Proceedings of the Logic and Engineering of Natural Language Semantics 15 (LENLS15)*.
- Hämäläinen, M. (2019). UralicNLP: An NLP Library for Uralic Languages. *Journal of Open Source Software*, 4(37), p. 1345.
- Hämäläinen, M. & Rueter, J. (2018). Advances in Synchronized XML-MediaWiki Dictionary Development in the Context of Endangered Uralic Languages. In *Proceedings of the Eighteenth EURALEX International Congress*. pp. 967–978.
- Hämäläinen, M. & Rueter, J. (2019). Finding Sami Cognates with a Character-Based NMT Approach. In *Proceedings of the Workshop on Computational Methods for Endangered Languages*, volume 1.
- Hämäläinen, M., Tarvainen, L. L. & Rueter, J. (2018). Combining Concepts and Their Translations from Structured Dictionaries of Uralic Minority Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA).
- Laxström, N. & Kanner, A. (2015). Multilingual Semantic MediaWiki for Finno-Ugric dictionaries. In *Septentrio Conference Series*, volume 2, pp. 75–86.
- Littell, P., Pine, A. & Davis, H. (2017). Waldayu and Waldayu Mobile: Modern digital dictionary interfaces for endangered languages. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*. Honolulu: Association for Computational Linguistics, pp. 141–150. <https://www.aclweb.org/anthology/W17-0119>.
- Moshagen, S., Rueter, J., Pirinen, T., Trosterud, T. & Tyers, F. M. (2014). Open-Source Infrastructures for Collaborative Work on Under-Resourced Languages. In *The LREC 2014 Workshop “CCURL 2014 - Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era”*, pp. 71–77.
- Nasution, A.H., Murakami, Y. & Ishida, T. (2018). Designing a Collaborative Process to Create Bilingual Dictionaries of Indonesian Ethnic Languages. In *Proceedings of the 11th Language Resources and Evaluation Conference*. Miyazaki, Japan: European Language Resource Association. <https://www.aclweb.org/anthology/L18-1536>.
- Rueter, J. (2014). The Livonian-Estonian-Latvian Dictionary as a threshold to the era of language technological applications. *Eesti ja soome-ugri keeleteaduse ajakiri*.

Journal of Estonian and Finno-Ugric Linguistics, 5, p. 251.

Rueter, J. & Hämäläinen, M. (2017). Synchronized MediaWiki Based Analyzer Dictionary Development. In *Proceedings of the Third Workshop on Computational Linguistics for Uralic Languages*, pp. 1–7.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



The Semantic Network of the Spanish Dictionary During the Last Century: Structural Stability and Resilience

Camilo Garrido^{1,2}, Claudio Gutierrez^{1,2}, Guillermo Soto³

¹ Department of Computer Science, Universidad de Chile

² Millennium Institute of Foundational Research on Data

³ Department of Linguistics, Universidad de Chile

E-mail: cgarrido@dcc.uchile.cl, cgutierr@dcc.uchile.cl, gsoto@uchile.cl

Abstract

The semantic network of a dictionary is a mathematical structure that represents relationships among words of a language. In this work, we study the evolution of the semantic network of the Spanish dictionary during the last century, beginning in 1925 until 2014. We analysed the permanence and changes of its structural properties, such as size of components, average shortest path length, and degree distribution. We found that global structural properties of the Spanish dictionary network are remarkably stable. In fact, if we remove all the labels from the network, networks from different editions of the Spanish dictionary are practically indistinguishable. On the other hand, local properties change over the years offering insights about the evolution of lexicon. For instance, the neighbourhood of a single word or the shared neighbourhood between a pair of words. This paper presents preliminary evidence that dictionary networks are an interesting language tool and good proxies to study semantic clouds of words and their evolution in a given language.

Keywords: semantic networks; dictionary networks; Spanish language

1. Introduction

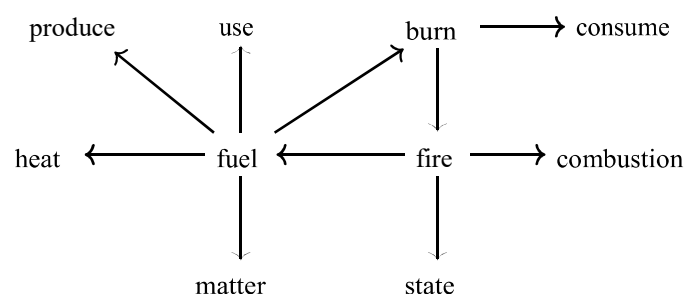
The lexicon of a language can be organized as a semantic network by considering the words as nodes and the similarities of some kind among the words as representing edges. A suitable proxy to such a network is the one obtained from a dictionary, built as follows: The nodes are the dictionary entries (properly cleaned), and for each entry define an edge from it to all the words that occur in its definition (which, when properly cleaned, occur as entries too) (see Figure 1). These *dictionary networks* are well known and have attracted linguistic interest (cf. Picard et al., 2009; Levary et al., 2012).

Until now the studies of dictionary networks have focused on static versions of dictionaries. But a dictionary evolves over time. New words are added to the lexicon, due to the introduction of a new, previously incommunicable concept, or to increase the different ways of mentioning an existing concept. Additionally, some words experience some slight changes in their meanings to adapt to new cultural trends. A few words are eliminated. Some new organizing criteria are introduced.

The evolution of a dictionary suggests studying the corresponding evolution of its associated network. Are there observable patterns in such evolution that can be of

linguistic interest? What can the evolution of such networks tell us about the evolution of the lexicon of a language? These are the types of questions that motivated this research.

In this paper we study the historical evolution along the last century of the networks associated with the most traditional Spanish dictionary. This dictionary has been issued by the Spanish Royal Academy since 1780, with regular periodicity and a rather stable philosophy and methodology.



Fire: Fuel in a state of combustion.

Fuel: Any matter used to produce heat by burning.

Burn: To consume with fire.

Figure 1: The network built from the entries fire, fuel, burn, and their definitions.

We investigate the permanence and changes of structural properties of the network of this Spanish dictionary beginning in 1925. There are two groups of network properties that we explore: global and local properties. The *global properties* are those capturing aspects as a whole and give an overall view of the network, for example, ratio of number of nodes versus edges, connectivity, centrality, etc. The *local properties* correspond to those topological properties of vicinities of nodes, such as clustering coefficient, the number of triangles in a particular location or the similarities and differences between the cloud surrounding two words in dictionary networks.

We highlight two main findings of our study. First, the structural properties of these networks are remarkably stable. Simply put: if we delete the labels of the nodes (i.e. of the words), and normalize the size of the networks, it would be very difficult, if not impossible, to tell which network corresponds to which year. The 1925 and the 2014 dictionary networks have almost the same structure. In particular, these networks are highly resilient, that is, they keep their structure in spite of the deletion of words and local perturbations. Second, the (historically) successive networks offer insights on how the semantic neighbourhood of a word evolves, that is, how relationships among words evolve. As we considered the dictionary as a suitable proxy of the lexicon of the language itself, this could shed light on the evolution over time of particular meanings and senses of concepts. One example we present is that of the noun *sex* (sexo) and adjective *sexual*. Early in the 20th century the word *sex* was in a cloud of biological

terms and almost disjoint from *sexual*, which refereed to human behaviours, the former with higher presence than *sexual*. In 2014, the cloud around word *sexual* became bigger than that of *sex* and more words directly connecting both entries appeared.

The paper is organized as follows: Section 2 presents an overview of dictionary networks. Sections 3 and 4 present the structural stability of the Spanish Dictionary network and the changes in local features. Section 5 analyses related work. Section 6 then presents our conclusions.

2. Dictionary networks: an overview

The definition of a word involves recursively new words, senses and meanings. Litkowski (1978) observed that this relation naturally forms a network that has linguistic interest. See Section 5, Related Work, for a more detailed overview of the developments of these type of networks.

2.1 Basic network model

In this work we utilize a simple (naive) model of a dictionary network that lacks any information on the type of word on nodes and edges, that is, just words pointing to other words represented in a standard form. At first sight, this simplification might seem impractical since it misses a lot of linguistic information (e.g. type, morphology, inflection, etc.) present in a dictionary. Nevertheless, several studies have shown the power of this simple model (Clark, 2003; Picard et al., 2009; Levary et al., 2012). In fact, besides facilitating the analysis of the network and its comparison with those in other fields, it captures the main features of the structure of these networks.

For this work we implemented the following procedure to build the networks:

1. *Model or design*: Consider all types of words as a single type: forget if they were nouns, verbs, adverbs, etc. Merge the entries that correspond to the same word into one definition, e.g. *Singer: A machine for sewing cloth.* and *Singer: One who, or that which, sings.* Forget the role and place of occurrence of a word, as well as its number of occurrences, inside a sentence (i.e. transform the defining text of a word in a set of words).
2. *Clean*: Remove entries that are inflected forms, e.g., *singing: from Sing*. Remove prepositions, conjunctions, and articles from entries and definitions. We consider them stopwords. They appear too often in any text and they would add noise to the graph. Lemmatize each word occurring in the definitions (transform nouns into singular; verbs into the infinitive; adjectives into their male singular form). In this work, we used Freeling (Padró & Stanilovsky, 2012) for the lemmatization of Spanish words and StanfordNLP (Qi et al., 2018) for the lemmatization of English words. Finally, remove any word that does not appear in the dictionary, e.g. prefixes and suffixes like *Ex-* and *-able*.

3. *Mathematical model of the dictionary*: Build the graph over the previous data. At this point, the dictionary D has become a universe of words W and a set of pairs $(w, \text{def}(w))$, where $w \in W$ is an entry in D and $\text{def}(w) \subseteq W$ is the set of words occurring in the definition of w .
4. *Build the network*: From the data in (3), construct a directed graph $G = (V, E)$, where the nodes are $V = \{w / (w, S) \in D\}$ and the edges $E = \{(w, w^o) / (w, S) \in D \text{ and } w^o \in \text{def}(w)\}$. For example, from the entry “*Eaglet (n.) A young eagle, or a diminutive eagle.*” we get the edges $(\text{Eaglet}, \text{young})$, $(\text{Eaglet}, \text{eagle})$ and $(\text{Eaglet}, \text{diminutive})$.

2.2 Main structural features

The network of a dictionary allows one to explore and study the global and topological properties that emerge from the network of words that cannot be captured locally (e.g. considering only isolated entries and their definitions). A classic global property is component analysis that allows finding subgroups of words according to connectivity. It shows four categories (Figure 2):

Giant Strongly Connected Component (SCC), this refers to words that recursively use themselves, which amount to about 1/3 of all words, most of them corresponding to entries never used in a definition. Bidirectional Component, words that mutually use each other in their definitions. Bidirectional Strongly Connected Component, words that mutually use each other and recursively use all other words in the category – amounting to 10% of all dictionary words. And Triangle Strongly Connected Component, triples of words that mutually need each other and recursively use all other words in the category. We will see that these components are stable parts of a dictionary.

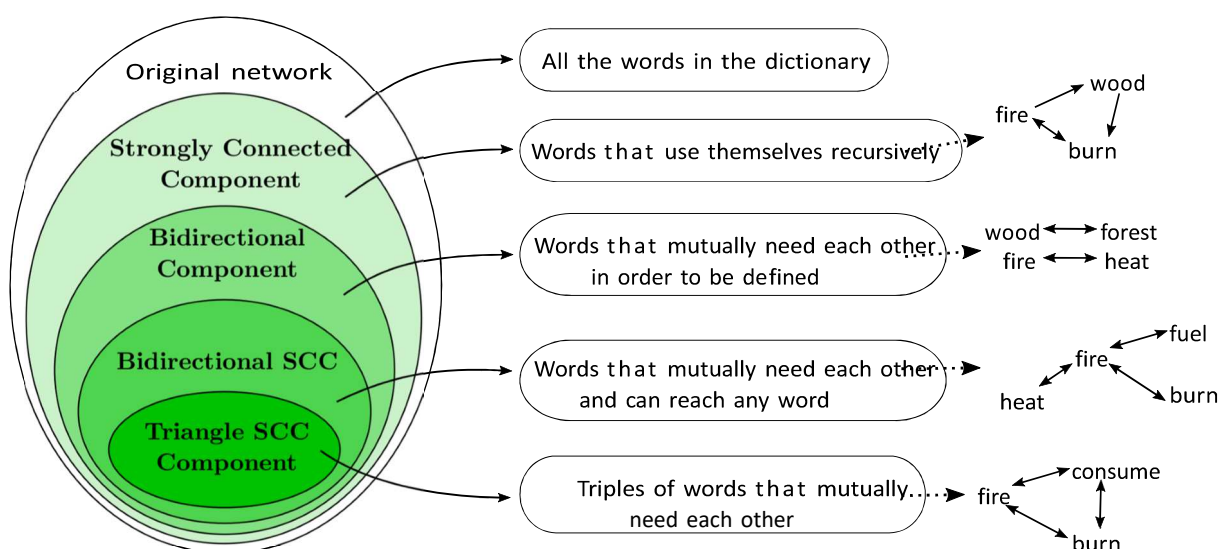


Figure 2: Structural components of a dictionary network. Examples on the right taken from the *Online Plain Text English Dictionary (OPTED)*.

3. The Spanish dictionary network: a stable and resilient structure

The Spanish Language dictionary (*Diccionario de la Lengua Española*, DLE) is a dictionary issued periodically since 1780 by the Spanish Royal Academy (currently in its 23rd edition). The new versions present updated lexicon and linguistic and editorial reorganizations¹.

In this section we study the network of the DLE and show that its basic structure remains stable and resilient over the years. We analyze three editions of the DLE: the 15th (published in 1925), the 18th (1956), and the current, 23rd edition (2014). According to the Royal Spanish Academy, the 1925 and 2014 editions are especially significant. The former (1925) incorporates attention to different Spanish-speaking territories besides Spain, and describes simpler definitions. The latter (2014), the most recent version, besides updating the lexicon, modifies its structure to facilitate searches, and incorporates other features, e.g. showing variations of entries and a consistent treatment of their male and female forms. To have an intermediate reference point, with a logarithmic interval between the extremes (30 and 60 years), we employed the 18th edition (1956). We used printed versions (none exist for the 1925 and 1956 editions) and for reasons of space we avoid the description of the tedious work and lessons obtained from scanning, cleaning and tuning the final texts.

3.1 Basic measures

A first snapshot of the evolution of dictionary networks is given by basic network measures (see Table 1) (Newman, 2003). The number of nodes (n) indicates the number of words in the dictionary. The dictionary grows about 15% every 30 years in this period. Edges (m) do not grow at the same rate, and the current dictionary has on average less edges per node (z) than previous years (meaning shorter definitions on average). Despite the changes in the number of nodes and edges, the average distance between entries (l) is not affected, staying around 4. The parameter α , the exponent of the degree distribution function ($p_k \sim k^{-\alpha}$), also remains almost unaffected over the years with the value $\alpha \approx 2.6$. The clustering coefficients over the years are also very similar, both global (c^1) and local (c^2). In dictionary networks, two entries having a common (non-frequent) word in their definitions are likely to be related. Lastly, the degree correlation coefficient (r) indicates whether the high-degree vertices in the network associate link preferentially with other high-degree vertices or not ($r = 1$ means high and $r = -1$ means low connectivity). This coefficient falls over the years. This may be caused by lexicographic decisions between editions, e.g. the removal of adverbs with the suffix *-mente* or past participles of verbs.

¹ <http://www.rae.es/diccionario-de-la-lengua-espanola/presentacion>

	n	m	z	l	α	c^1	c^2	r
DLE 1925	60,823	1,058,012	17.39	4.03	2.59	0.019	0.227	0.042
DLE 1956	69,719	1,174,912	17.49	4.03	2.58	0.017	0.225	0.039
DLE 2014	87,255	1,076,377	12.34	4.09	2.65	0.015	0.224	0.002
OPTED	95,095	979,523	20.60	4.64	3.13	0.009	0.217	-0.008
WordNet	84,967	1,134,957	26.72	2.99	2.99	0.029	0.203	-0.016

Table 1: Basic measures for the networks of the Spanish dictionary (DLE) over the years. The Online English dictionary OPTED and WordNet networks are shown for comparison. n and m are the number of nodes and edges, respectively; the other parameters are explained in Section 3.1.

3.2 Component analysis

Components are classic features when describing the topology of networks (Section 2.2). For the Spanish dictionary network (Table 2), the Giant Strongly Connected Component for all three editions remains around 30% of the whole network. The Bidirectional Component stays around 17% of all the words over the years. The Bidirectional Strongly Connected Component covers about 11% of the network. Finally, one of the strongest notions of connectivity is the subgraph induced by the strongly connected component of triangles. It represents less than the 3% of the network in each dictionary. The ratio of the size of each component is consistent over time. The words composing the components are also very consistent. Note that around 80% of the words in a component in 1925 remain in the same component in 2014 (Table 3).

3.3 Centrality measures

We tested four classic centrality measures for each DLE network: Betweenness Centrality, Closeness Centrality, Degree, and PageRank (Boldi & Vigna, 2014). The ranks are very similar if we just take into the account the top nodes/words. Here we present the recurrent words (RW) in the top 20 ranking for each measure:

Betweenness (9 RW): *acción, cosa, dar, estar, hacer, mano, parte, ser, tener.*

Closeness (10 RW): *alguno, cosa, dar, decir, estar, hacer, otro, persona, ser, tener.*

Degree (12 RW): *acción, alguno, cosa, dar, decir, estar, hacer, otro, parte, persona, ser, tener.*

PageRank (13 RW): *acción, alguno, cosa, dar, decir, efecto, estar, hacer, otro, persona, poder, ser, tener.*

Over the last century, half of the words stayed in the top 20 ranking. These are basic words that help to put together definitions and the dictionary, e.g. *Natación: acción y efecto de nadar* (Swimming: action or effect of swim).

3.4 Cliques

Cliques are sets of nodes such that any pair among them is connected by an edge. In the context of dictionary networks, cliques are a local property that allows identification of a strong dependency among words (each one occurs in the definition of all others). For example, *cosa*, *dar*, *decir*, *hacer*, *ser*, *tener*, *todo*.

In the Spanish dictionary network, the number of cliques grows from edition to edition, but the growth rate seems to slow down over the years (Figure 4). There are no bigger cliques than K_7 in any of the three editions.

3.5 Resilience of the dictionary network

Resilience refers to the vulnerability or the ability of a network to resist link or node failures. This happens to be a relevant property in dictionary networks. As a notion of resilience, we use the variation of the size of the largest component as nodes are removed from the network. We use two approaches to node removal: random choice and high in-degree nodes, the latter meaning the removal of words that occur the most in other definitions. As baseline, we compare the behaviour of dictionary networks with that of a random graph. We use the random graph model proposed by Barabási and Albert (1999) based on the idea of preferential attachment. It is frequently used for language networks comparison (Dorogovtsev & Mendes, 2001; Steyvers & Tenenbaum, 2005).

	1925	1956	2014
Original network	60,823	69,719	87,255
SCC	18,307	21,538	26,989
Bidirectional Component	10,462	12,061	16,025
Bidirectional SCC	6,125	7,429	11,308
Triangle SCC	1,033	1,318	2,359

Table 2: Component sizes of the Spanish dictionary networks in number of words.

	1925-2014	% of 1925
Original	54,235	89.1%
SCC	15,514	84.7%
Bidirectional Component	7,665	73.3%
Bidirectional SCC	4,841	79.0%
Triangle SCC	828	80.2%

Table 3: Number of words (and percentage) from 1925 that remain in the same component in 2014.

It turns out that removing random nodes produces almost no damage at all. All three dictionaries and random graphs resist the attacks well. The size of the component decreases linearly with respect to the number of nodes removed. On the other hand, dictionary networks and random graphs behave very differently when removing high in-degree nodes.

Dictionary networks resist more attacks than random graphs (Figure 3). Random graphs decline quickly. Removing just 10% of the high in-degree nodes is necessary to completely destroy and scatter the graph. That is not the case with dictionary networks. The giant component of dictionary networks decreases almost linearly until we remove about a third of the network. From that point forward, the giant component starts to decline rapidly, scattering completely when 37% of the high in-degree nodes are removed. It is important to note that the resilience of connectivity of dictionary networks does not rely on frequently used words that connect the network, but on the high connectivity among all words. One could express this by saying that it is very difficult to completely remove a cloud of close concepts; there will always remain other ways to express them. This seems to be a particular property of dictionary networks, as results for other real world networks do not show this behaviour (Jeong et al., 2001; Dunne et al., 2002; Newman et al., 2002).

4. What changes: the local features

Despite its structural stability, there are changes in the successive versions of the DLE: new entries are incorporated, some entries are removed and some definitions are enriched or modified. In this section, we focus on these changes in the dictionary.

4.1 Definitional and interchangeable entries

The entries in the DLE can be divided into two groups: *definitional* entries are words used to define other words and *interchangeable* entries correspond to words that do not occur in any definition at all. In network terms, definitional words are those that have inlinks and outlinks, while interchangeable words have only outlinks. The fact that a word has only outlinks means that in some sense is “disposable”, that is, it could be replaced by the words in its definition (Levary et al., 2012), hence the name interchangeable.

If we study how incorporations and deletions of entries from one version of the dictionary to another occur, eight possible outcomes show up (Figure 4). Definitional entries can (1) stay as a definitional entry, (2) become an interchangeable entry, (3) be removed from the dictionary. Likewise, interchangeable entries can (4) stay as an interchangeable entry, (5) become a definitional entry, or (6) be removed from the dictionary. Additionally, new entries are incorporated into the dictionary as (7) new definitional entries or (8) new interchangeable entries.

	1925	1956	2014
K_3	2,208	3,007	5,911
K_4	489	917	1,311
K_5	95	299	347
K_6	10	69	69
K_7	1	8	5

Table 4: Cliques in DLE networks.

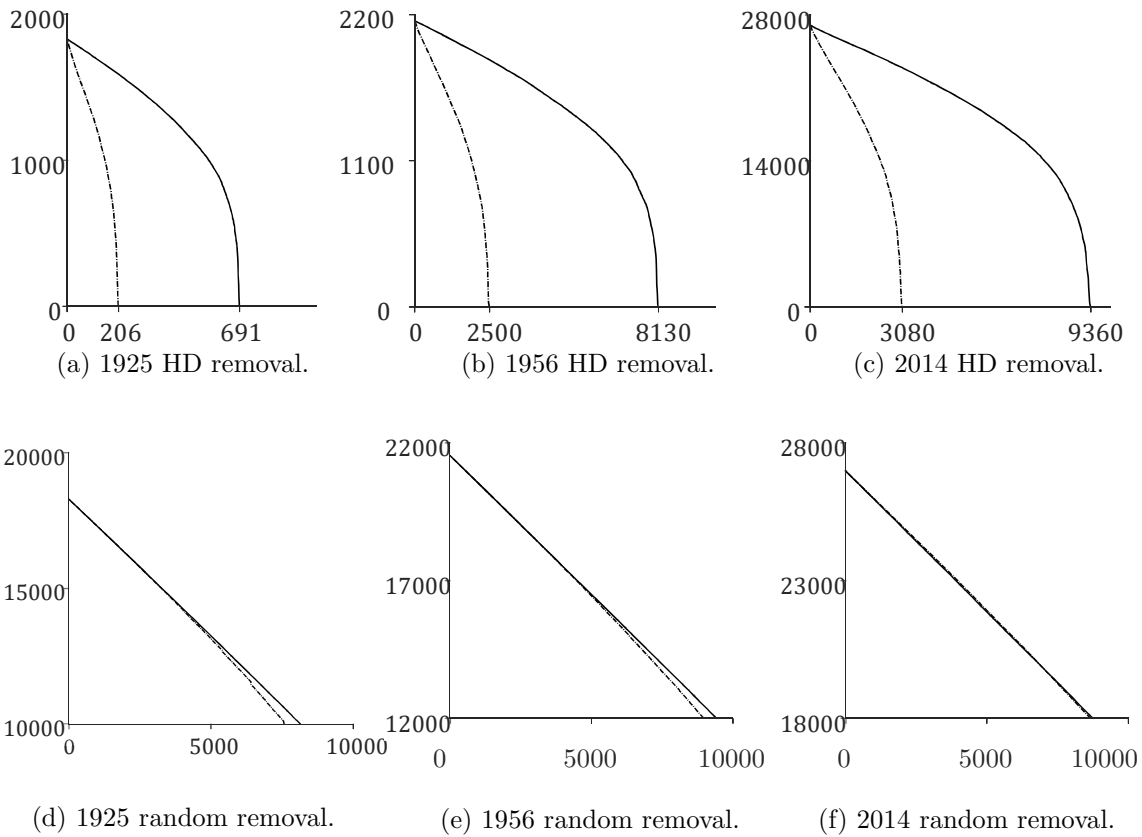


Figure 3: Sizes of the giant component as nodes are removed. On the left, high degree (HD) node removal. DLE network (solid line) keeps its structure (giant component) as compared to a random network (dotted line). On the right, random removal does not affect the size of the giant component in either DLE or a random network.

Most of the entries in a dictionary do not change their type between versions. In fact, in the DLE (with new versions approximately every 30 years) between 80%-90% of definitional entries stay as definitional, and a similar percentage of interchangeable entries stay as interchangeable (1 and 4 in Figure 4). When new words are added to the dictionary, most of them (76%-95%) enter as interchangeable (8 in Figure 4); only a few of them occur in definitions (7 in Figure 4). On the other hand, almost all of the entries that are removed from the dictionary were interchangeable entries (6 in Figure 4).

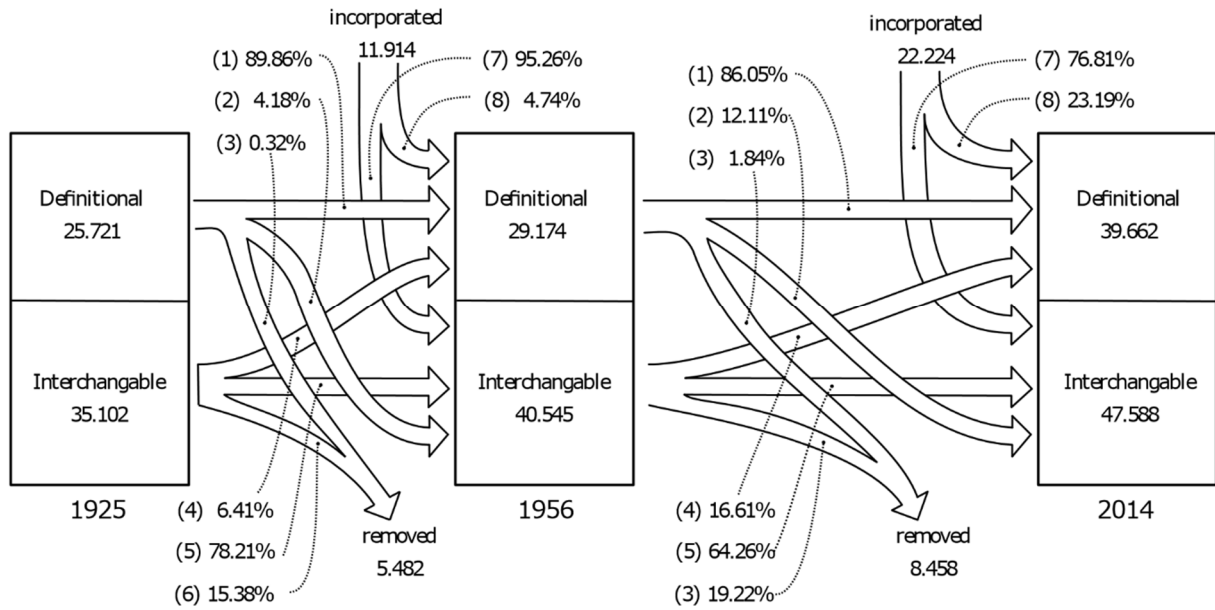


Figure 4: Changes in the entries of the DLE from 1925 to 1956 and 1956 to 2014. For a detailed explanation of the figure see Section 4.1.

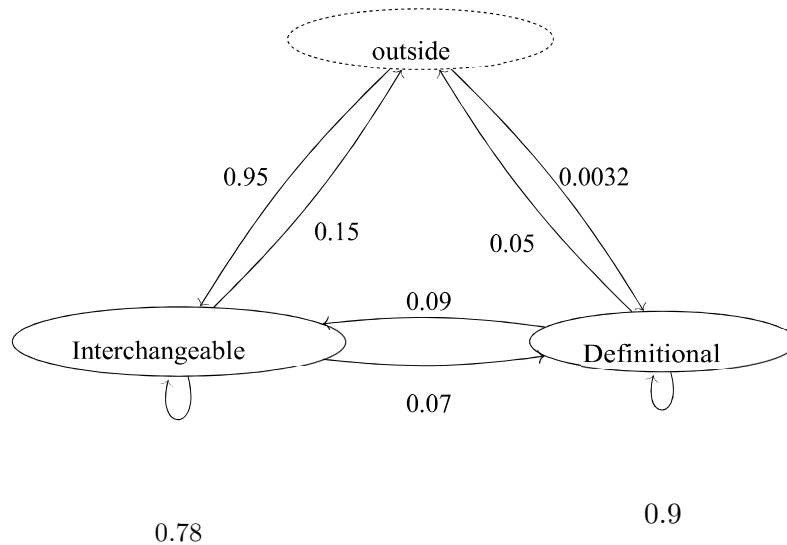


Figure 5: Markov chain that describes the probability of transitions among types of words every 30 years in the Spanish Dictionary.

In order to better describe the transitions among the types of words, we build a Markov chain using the empirical data of the transitions over the years (see Figure 5). A Markov chain is a stochastic model that describes the transitions between possible states using only its current state. It can be described as a directed graph with probabilities for edges and states for nodes. A word can be in one of three states. It can be a definitional, it can be interchangeable, or it can be “outside”. The state outside means that the word is not in the dictionary. This model allows us to estimate the probability of a word being in a state in future editions of the Spanish dictionary and the paths it is going to take. For example, a definitional word has a probability $p = 0.9$ of staying as

definitional in 30 years in the future (one iteration). If we consider a span of 90 years (three iterations), a definitional word has a probability of $p = 0.729$ (calculated as $0.9 \cdot 0.9 \cdot 0.9$) of always staying as definitional. The model allows us to calculate the probability of more complex transitions. For example, the probability of a definitional word becoming interchangeable in one iteration and then being removed from the dictionary in the next iteration is $p = 0.0135$ (calculated as $0.07 \cdot 0.15$).

4.2 Examples of simple local changes

These changes do not affect or change the overall structure of the network (as we saw in Section 3). But they impact at the local level. In fact, these changes alter the structure of the vicinity of some words (not only those whose definition explicitly changes). We will illustrate these changes through some examples in order to offer insights on how the evolution of the network structure speaks about semantic features.

First, entering and outgoing words. *Aeropuerto* (airport) is an obvious case of an entering word that was not present in the 1925 edition. In fact, airplanes and other aerial words were emerging concepts at the time. In 1956, *aeropuerto* is already incorporated as a definitional entry. Later, in 2014, *aeropuerto* is still a definitional entry being used by 17 different words in their definitions, such as airfield (*aeródromo*), checkroom (*consigna*), and tower (*tower*). On the other hand, there are words that were slowly put aside in the dictionary. These words were definitional entries in 1925. In 1956, they became interchangeable entries, as they did not appear in any definition. And in 2014, they were completely removed from the dictionary. Examples are *Adolecente* (old form of adolescent); *fecundante* (someone who impregnates or fertilizes); *escaza* (an Aragonese word referring to a certain type of pot).

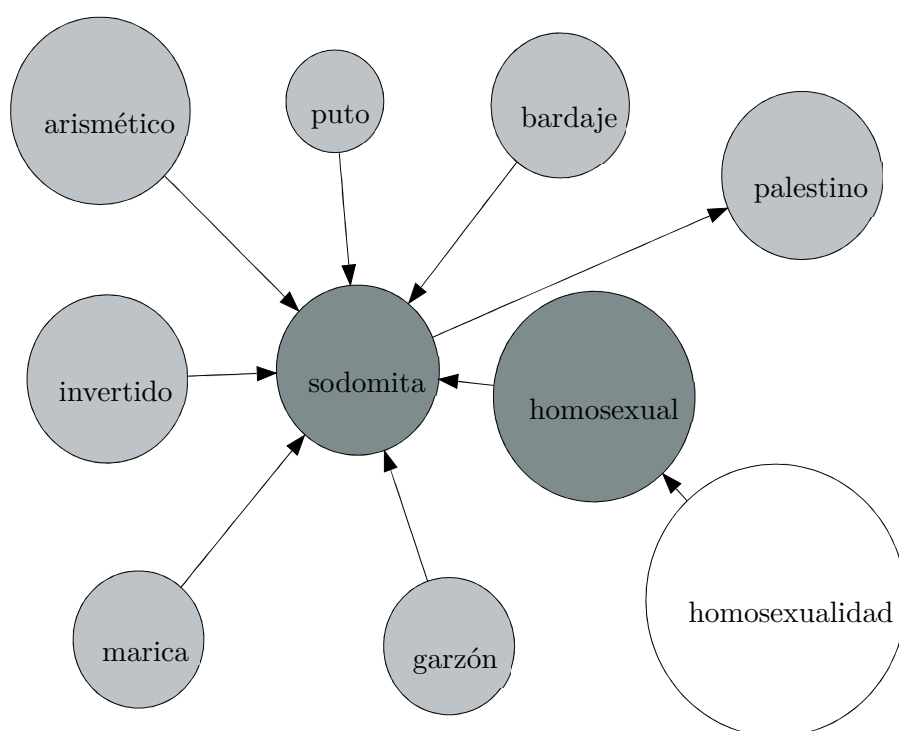
Second, words whose cloud of meaning changes. Consider the word *prostituta* (prostitute). The 1925 dictionary contains the definitional entry *prostituta* defined as *ramera* (whore). There is no definition for the male noun. However, the dictionary contains the interchangeable entry “*prostituto, ta*” (the suffix denotes it can be male or female). This entry refers to the irregular past participle of the verb *prostituir* (prostitute). In the 1956 dictionary, these entries remain with few changes. Both of them keep their definitions, but the entry *prostituta* became an interchangeable word. Most of the changes occurred in the 2014 edition. First, the entry *prostituta* was removed from the dictionary. Second, the entry “*prostituto, ta*” became a definitional entry. And third, the entry “*prostituto, ta*” no longer refers to the past participle, but to the noun, covering both the male and female forms. It also got a neutral gender and a less derogatory definition: a woman or man who engages in sexual acts for money.

4.3 More complex local changes

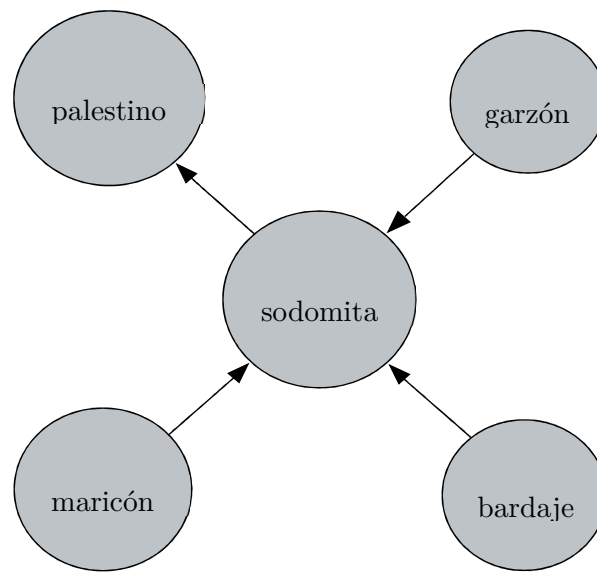
The above changes are not particularly surprising (one could guess them, although in

the network can be detected automatically!). There are more interesting cases that we think would be difficult or virtually impossible to detect without having a network, and thus, demonstrate in some sense the potentialities of the network methodology. A good example is the evolution in the relationship between the words *sexo* (sex) and *sexual* (sexual) and between *homosexual* (homosexual) and *sodomita* (sodomite).

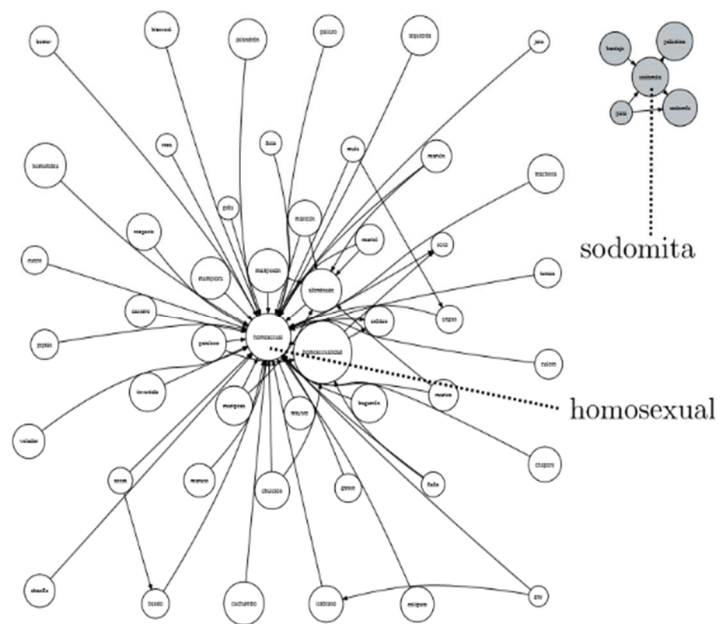
The words *sex* and *sexual* are directly related since the definition of sexual is basically “of or pertaining to sex”. However, it is interesting to observe how the relationships between their neighbourhoods change. In 1925 (Figure 7a), the neighbourhood of *sex* is noticeably larger than the neighbourhood of *sexual*; moreover, *sex* was surrounded by biological terms, such as plant, walnut, sweet potato, male, female, hermaphrodite, etc. Later in 1956 (Figure 7b), the size of the neighbourhoods became very similar as *sexual* occurs in more definitions. The neighbourhood of *sexual* expanded to a particular subject. Words such as *sperm*, *egg*, *orgasm*, incorporated *sexual* in their definitions. There are many paths between *sex* and *sexual*, but this edition is the first one to have a word that connects them directly (i.e. there is a path of length 2): *masochism* is defined using both *sex* and *sexual*. Now, in 2014, both neighbourhoods increase their size (Figure 7c), hence their semantic weight. The cloud around *sexual* becomes bigger than that of *sex* and both entries appear where more words connect directly, such as *sexuality*, *venereal*, and *transsexual*.



(a) 1925

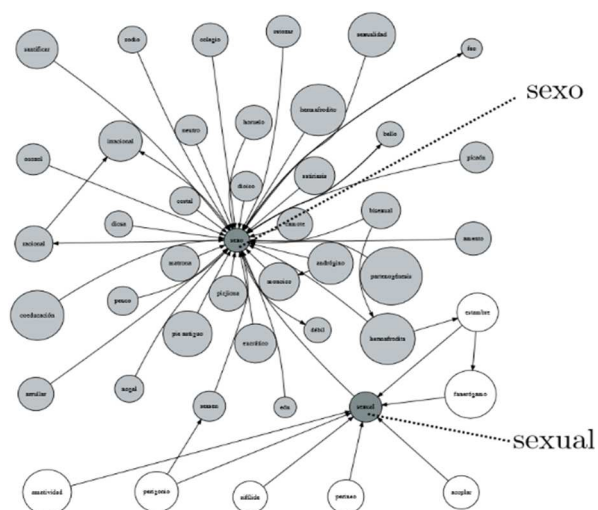


(b) 1956

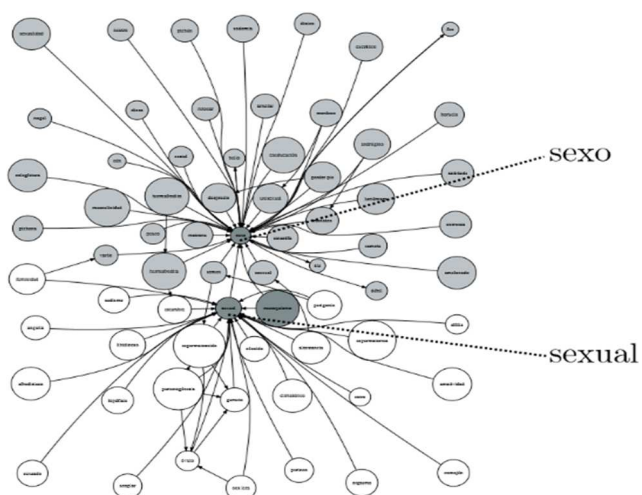


(c) 2014

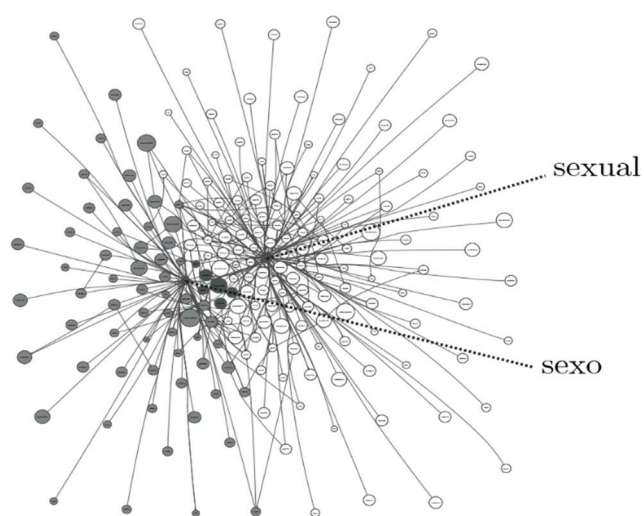
Figure 6: Sub-network around the words *homosexual*(homosexual) and *sodomita* (sodomite).



(a) 1925



(b) 1956



(c) 2014

Figure 7: Sub-network around the words *sexo* (sex) and *sexual* (sexual).

The relationship between *homosexual* (homosexual) and *sodomite* (sodomite) presents a different evolution. In 1925 (Figure 6a), *homosexual* was not defined in the dictionary, while *sodomite* occurred as definitional entry. *Sodomite* covered two concepts: a demonym of an old Palestinian city and a person who engages in sodomy. In 1956 (Figure 6b), the entry *homosexual* was incorporated into the dictionary as a definitional entry. However, it was not a proper definitional entry. It was incorporated as a synonym of *sodomite*, working as a proxy for other words like *homosexuality* to reach *sodomite*. This situation changed in 2014 (Figure 6c), when *homosexual* no longer expressed the meaning of sodomite. It is now defined using concepts such as homosexuality and sexual attraction to persons of the same sex. Its neighbourhood grew considerably; more than 50 words use it in their definitions. Lastly, both entries are not connected anymore. Their concepts diverged. *Sodomite* holds the same meaning since 1925 and *homosexual* evolves from not being in the dictionary, passing to be a synonym of *sodomite*, to become an entry with its own meaning. Last but not least, note that in this analysis the use of neighbourhoods of the network was essential.

5. Related work

Litkowski (1978) was one of the first to state the importance of studying and exploiting dictionary networks, as sources of material for natural language and to unravel the complexities of meaning. He presented three models for representing a dictionary. One based on the relationship *x is used to define y*. The second model incorporates senses of words as nodes. The final model considered the nodes as concepts, having different nodes when words in a definition have more than one meaning.

After Litkowski, there were several investigations about dictionaries and the information that could be extracted from them (Amsler, 1980, 1981; Calzolari, 1984; Chodorow et al., 1985; Calzolari & Picchi, 1988). For example, Picard et al. (2009) aimed to reduce a dictionary to its grounding kernels from which all the other words could be defined. They define a hierarchy of definitional distance and show it correlates with psycholinguistic variables. Levary et al. (2012) studied loops and self-reference in the definition of words. They observed that definitions have a great amount of short loops (length < 6). Muller et al. (2006) presented a method that exploits a directed weighted graph derived from a dictionary to compute distance between two words. The work of Steyvers and Tenenbaum (2005) presented an analysis of the large scale of three types of semantic networks: WordNet (Miller, 1995), word association norms (Nelson et al., 2004), and Roget's Thesaurus (Roget, 1911). They focused on a statistical analysis, concluding that these networks have a small-world structure, characterized by sparse connectivity, short average path lengths between words, and strong local clustering.

Less directly related to our work are lexical databases represented in the form of networks. Built from diverse sources in a manually annotated process, they cover the current use of words and their meanings. WordNet (Miller, 1995) groups words into

sets of cognitive synonyms (synsets) and FrameNet (Baker et al., 1998) annotates examples of how words are used in actual texts.

6. Conclusions

This work shows that the study of semantic networks derived from dictionaries could offer insights and tools to study the evolution of the lexicon of a language. We developed in this paper the case of the Dictionary of the Spanish Royal Academy. Among the most relevant findings, is the fact that the network is has a stable structure over the years and is highly resilient. We hypothesize that this is valid for definitional dictionaries in other languages (we tested, although did not present the results here, the case of the English OPTED dictionary). The study presents preliminary evidence that dictionary networks are interesting artefacts and good proxies to study semantic clouds of words and their evolution in a given language.

7. Acknowledgements

The research leading to these results has received funding from under grant CONICYTPCHA/Doctorado Nacional/2015-21150149.

8. References

- Amsler, R.A. (1980). The structure of the Merriam-Webster pocket dictionary.
- Amsler, R.A. (1981). A taxonomy for English nouns and verbs. In *Proceedings of the 19th annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, pp. 133–138.
- Baker, C. F., Fillmore, C. J. & Lowe, J. B. (1998). The Berkeley Framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*. Association for Computational Linguistics, pp. 86–90.
- Barabási, A. L. & Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439), pp. 509–512.
- Boldi, P. & Vigna, S. (2014). Axioms for centrality. *Internet Mathematics*, 10(3-4), pp. 222–262.
- Calzolari, N. (1977). An empirical approach to circularity in dictionary definitions. *Cahiers de Lexicologie Paris*, 31(2), pp. 118–128.
- Calzolari, N. (1984). Detecting patterns in a lexical data base. In *Proceedings of the 10th International Conference on Computational Linguistics and 22nd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, pp. 170–173.
- Calzolari, N. & Picchi, E. (1988). Acquisition of semantic information from an on-line dictionary. In *Proceedings of the 12th conference on Computational linguistics-Volume 1*. Association for Computational Linguistics, pp. 87–92.
- Chodorow, M. S., Byrd, R. J. & Heidorn, G. E. (1985). Extracting semantic hierarchies

- from a large on-line dictionary. In *Proceedings of the 23rd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, pp. 299–304.
- Clark, G. (2003). Recursion through dictionary definition space: Concrete versus abstract words. Technical report, University of Southampton Tech Report).
- Dorogovtsev, S. N. & Mendes, J. F. F. (2001). Language as an evolving word web. *Proceedings of the Royal Society of London B: Biological Sciences*, 268(1485), pp. 2603–2606.
- Dunne, J. A., Williams, R. J. & Martinez, N. D. (2002). Food-web structure and network theory: the role of connectance and size. *Proceedings of the National Academy of Sciences*, 99(20), pp. 12917–12922.
- Jeong, H., Mason, S. P., Barabási, A. L. & Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature*, 411(6833), pp. 41–42.
- Levary, D., Eckmann, J. P., Moses, E. & Tlusty, T. (2012). Loops and self-reference in the construction of dictionaries. *Physical Review X*, 2(3), p. 031018.
- Litkowski, K. C. (1978). Models of the semantic structure of dictionaries. *American Journal of Computational Linguistics*, 81, pp. 25–74.
- Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, 38(11), pp. 39–41.
- Muller, P., Hathout, N. & Gaume, B. (2006). Synonym extraction using a semantic distance on a dictionary. In *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing*. Association for Computational Linguistics, pp. 65–72.
- Nelson, D. L., McEvoy, C. L. & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, 36(3), pp. 402–407. URL <http://dx.doi.org/10.3758/BF03195588>.
- Newman, M. E. (2003). The structure and function of complex networks. *SIAM review*, 45(2), pp. 167–256.
- Newman, M. E., Forrest, S. & Balthrop, J. (2002). Email networks and the spread of computer viruses. *Physical Review E*, 66(3), p. 035101.
- Padró, L. & Stanilovsky, E. (2012). FreeLing 3.0: Towards Wider Multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*. Istanbul, Turkey: ELRA.
- Picard, O., Blondin-Massé, A., Harnad, S., Marcotte, O., Chicoisne, G. & Gargouri, Y. (2009). Hierarchies in dictionary definition space. *arXiv preprint arXiv:0911.5703*.
- Qi, P., Dozat, T., Zhang, Y. & Manning, C. D. (2018). Universal Dependency Parsing from Scratch. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Brussels, Belgium: Association for Computational Linguistics, pp. 160–170. URL <https://nlp.stanford.edu/pubs/qi2018universal.pdf>.
- Roget, P. M. (1911). Roget's Thesaurus of English Words and Phrases. <http://www.gutenberg.org/etext/10681>. Last accessed 01 July 2017.

- Sparck-Jones, K. (1967). Dictionary Circles. Technical report, System Development Corp Santa Monica California.
- Steyvers, M. & Tenenbaum, J. B. (2005). The Large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive science*, 29(1), pp. 41–78.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



Towards a Graded Dictionary of Spanish Collocations

Marcos García Salido, Marcos Garcia, Margarita Alonso-Ramos

Universidade da Coruña, CITIC, Grupo LyS, Dpto. de Letras,

Fac. de Filoloxía. 15071, A Coruña

E-mail: {marcos.garcias, marcos.garcia.gonzalez, margarita.alonso}@udc.gal

Abstract

Several recent studies have observed that texts of different quality and written by learners at different proficiency levels also vary in the lexical combinations they contain. Such variation can be operationalized by quantitatively measuring the association between the components of these lexical combinations. In particular, pointwise mutual information (MI) has proved to be a good predictor of proficiency development, as several studies on English learners' writing have shown. This paper examines whether association measures are also a good predictor for the proficiency level of texts written by learners of Spanish, with a view to using such information for grading lexical combinations in order to include them in a collocation dictionary of Spanish. The study also investigates whether the association measures that correlate with learners' proficiency level can discriminate between phraseological collocations and non-collocations. Our results show that, whereas the MI of learner texts' lexical combinations is a better predictor of author proficiency than frequency, the latter performs better in identifying phraseological collocations among the whole set of lexical combinations.

Keywords: graded collocation dictionary; CEFR proficiency level; association measures

1. Introduction

Phraseological expressions permeate discourse to a considerable extent. Erman and Warren (2000) estimate that, on average, 55% of texts is made up of prefabricated expressions. Collocations surely are a subset among those prefabricated expressions and are therefore an essential component of learning a new language. In fact, several studies have found that collocations are a challenging aspect of language learning: see Granger (1998), Nesselhauf (2004), or Vincze et al. (2016), to cite but a few.

Despite the importance of collocations in language learning, the attention given to this phenomenon in curricula or assessment materials is not always evident, as noticed by Paquot (2018). According to her, the Common European Framework of Reference for Languages—henceforth CEFR (Council of Europe, 2001)—assumes a very traditional understanding of phraseology, by obviating frequent word combinations and using the term *phrase* mostly for stock phrases and pragmatically conditioned expressions. Paquot emphasises that, by ignoring learners' phraseological competence, we are losing a valuable assessment criterion for language proficiency.

In the particular case of Spanish, Higuera García (2017) argues that, whereas research has devoted considerable attention to collocations, these combinations are still not very

well treated in Spanish Language Teaching. She also favours a flexible conception of the notion, which includes frequently used combinations, even if they are not properly the result of lexical restrictions. As for their introduction to learners, although she mentions frequency, Higuera clearly prefers other selection criteria, such as their relation with syllabus' topics and with communicative functions.

Even though, from the situation Higuera García (2017) depicts, collocations seem to be a phenomenon rather neglected by the Spanish Teaching community in general, learners of this language have some reference works at their disposal. The *Diccionario de colocaciones del español* (DiCE; Alonso-Ramos, 2004) is an online dictionary that follows the principles of the Meaning-Text Theory in the treatment of collocations. It is an ongoing project that so far incorporates lexical units related to the sentiment's lexical field and is in the process of including academic collocations. The sentiment collocations provide users with CEFR level indications. The *Diccionario combinatorio práctico del español contemporáneo* (PRÁCTICO; Bosque, 2006) is a paper dictionary based on a mostly theoretical combinatorial dictionary (REDES, Bosque, 2004). In its structure, it is more similar to other learner-targeted collocation dictionaries, such as Benson et al. (1986), than REDES. In contrast to DiCE, and in spite of being corpus-based, PRÁCTICO in general does not provide notes on collocation frequency, but occasionally indicates their semantic prosody. Finally, the *Herramienta de Ayuda a la Redacción en español* (HARenEs; Alonso Ramos et al., 2015) is a web tool that gives its users collocations directly extracted from a corpus in a more user-friendly manner than concordancers.

This paper explores the possibilities of lexical association measures in grading learners' lexical combinations and in identifying phraseological collocations among such combinations. Its final aim is to explore a method to compile a collocation dictionary that combines features offered separately by some of the reference works reviewed: firstly, by including a vast set of collocations representative of Spanish, like PRÁCTICO; and secondly, by offering notes on the CEFR level of collocations, like DiCE, providing thus guidelines to the Spanish teaching community for grading lexical contents and for assessment. In what follows, we review some related work in Section 2. Next, the method proposed is described in Section 3. Section 4 presents and discusses the results and evaluates the viability of the method for compiling a graded collocation dictionary of Spanish before moving onto the conclusions (Section 5).

2. Lexical combinations, frequency-based association measures and proficiency

When it comes to grading vocabulary, lexical frequency shows up repeatedly as a useful criterion (Nation, 2001; Alvar Ezquerro, 2005). The rationale behind this recommendation is that the most frequent vocabulary of a language covers larger

proportions of text than less frequent units (be they word-families, lemmas or word-forms). Consequently, learning this first would theoretically lead to great advances in understanding and producing texts. Frequency has been proposed as a grading criterion for multi-word vocabulary as well. Martinez (2013) suggests to give priority to lexical combinations that are both frequent and semantically opaque.

Frequency as a grading criterion has been applied to several vocabulary repertoires directed to Spanish teaching. The *Plan curricular del Instituto Cervantes* (henceforth, PCIC)¹ is a set of guidelines that adapts the recommendations of the CEFR with a greater degree of specificity. Several of its sections provide vocabulary graded by proficiency—including some collocations. This document states that vocabulary selection is based on frequency and usability as perceived by experienced professionals, among other criteria. In a similar vein, corpus frequency has been used for grading the collocations included in at least two collocation dictionaries: the DiCE (García Salido & Alonso Ramos, 2017) and the *Dizionario delle Collocazioni Italiane per Apprendenti* (Spina, 2016)—in the case of the latter, frequency has been used along with aspects such as the topic to which the vocabulary is related.²

Irrespective of whether most frequent lexical combinations are taught first, examination of learners' production seems to back the idea that such combinations are acquired and used earlier than less frequent ones. Thus, in an analysis of texts of intermediate and advanced learners of English, Granger and Bestgen (2014) observe that the first group uses a larger proportion of bi-grams with high t-score values³ than the latter. However, that is not the whole story. Granger and Bestgen (2014) also analyse their learners' bi-grams in terms of another association measure: pointwise mutual information (MI), which results from the ratio between the observed and expected frequencies of a combination. In this case, it is advanced learners who use the combinations with highest MI values more often. In a further study, Bestgen and Granger (2014) fail to observe a significant correlation between t-score and text quality as determined by professional English teachers, but they do find a significant positive correlation between quality and MI.

More recently, Paquot (2018) has studied the phraseological use of advanced learners of English (levels B2 through C2) and found that the MI of lexical combinations used in learner texts predicts teachers' ratings better than any other measure of syntactic or lexical complexity. In contrast to the earlier references, in this study Paquot focuses on combinations of two lexical units related by a syntactic dependency (namely, verb plus

¹ https://cvc.cervantes.es/ensenanza/biblioteca_ele/plan_curricular/

² Here, we limit ourselves to review collocation dictionaries somehow including CEFR level information, since offering a complete picture of vocabulary selection and grading, as undertaken in projects such as the English Vocabulary Profile (Capel, 2010, 2012), falls outside the scope of this paper.

³ A measure highly correlated with co-occurrence frequency.

object noun, verb plus modifying adverb and noun plus modifying adjective), rather than on bigrams.

In summary, whereas priming frequent phraseological combinations and introducing them first in curricula seems reasonable by virtue of its usability, MI seems a better predictor of proficiency when examining learner production.

3. Methodology

In what follows we describe the methods used to explore to what extent association measures (AMs) predict the proficiency of learners of Spanish in order to apply this information in the compilation of a graded collocation dictionary. With this aim we extracted lexical combinations from a corpus of learner texts and assigned them the AMs corresponding to those very combinations in a reference corpus of Spanish.

The learner corpus used in this study comes from CEDEL2 (Lozano & Mendikoetxea, 2013). We chose texts whose authors got a score of 50% or higher in a placement test⁴ administered to them at the time the corpus was compiled. Also, we included only texts that had a length of at least 200 words. The resulting sample consisted of 234 texts comprising 102,621 words. These were graded according to CEFR levels by three expert teachers of Spanish as a Second/Foreign Language who reached a consensus of 67% (Krippendorff's $\alpha = 0.7$). Texts were assigned the level chosen by the majority of raters. The distribution of texts across levels was unequal, as can be seen in Table 1.

CEFR level	A1	A2	B1	B2	C1	C2
No. of texts	23	66	95	39	8	4

Table 1: No. of texts by CEFR level.

There is a clear relationship between authors' median scores in the placement test and the grading of their texts by experts, with the exception of levels C1 and C2, where this correspondence is reversed, as can be seen in Figure 1. On the other hand there is a considerable overlap between the scores obtained by authors of B2 through C2 texts, on one hand, and A1 and A2 texts, on the other. Likewise, B1 texts correspond to a spread range of scores in the placement test.

This learner corpus was tokenized and lemmatized by means of LinguaKit (Garcia & Gamallo, 2015) and PoS-tagged using FreeLing (Padró & Stanilovsky, 2012). Lemmas and PoS-tags were manually revised in order to assign existent Spanish forms to possible lemmas. Only those forms that could be identified with a Spanish word beyond reasonable doubt were corrected.⁵ These data were then submitted to syntactic parsing

⁴ The test in question is a standardized level-placement test developed by the University of Wisconsin (Lozano & Mendikoetxea, 2013).

⁵ For instance, the token *siguente* was lemmatized as the canonical *siguiente* 'next', but

using UDPipe models (Straka et al., 2016).

From this corpus we extracted pairs of lemmas in the following syntactic dependencies: object-verb (obj), subject-verb (nsubj) and noun plus modifying adjective (amod). The collocation candidates extracted from the learner corpus were then assigned the association measures corresponding to these very collocations in a reference corpus. The reference corpus from which the measures were extracted was a 170-million word fragment of Mark Davies' *Corpus del español*⁶ automatically processed with the same tools used for the learner corpus—but without manual supervision this time. In spite of the variety of lexical association measures available, we chose MI and frequency given their previous use as predictors of proficiency level in several studies, namely Bestgen and Granger (2014), Granger and Bestgen (2014), and Paquot (2018),⁷ as noted in Section 2, even though some other measures might perform better in the detection of phraseological combinations (Pecina, 2010).

These data were then used to fit a generalized linear mixed model. The association measures of the combinations that reached a frequency of 3 or higher⁸ in the reference corpus were the independent variables of the model, and the dependent variable was the CEFR levels assigned by the teachers to the texts where they appeared. For this analysis we tried different solutions, namely: (i) using the AMs of the whole set of combinations of each text meeting the conditions already mentioned; (ii) assigning a mean score to each text based on the combinations of each dependency type; and (iii) calculating a unique mean score based on the three dependencies considered taken together. Only in the latter case did we obtain significant results (see Section 4 below).

Additionally, all the lexical combinations extracted from the learner corpus and in the above-mentioned syntactic dependencies (plus noun–preposition–noun) were manually revised in order to identify those that qualified as phraseological collocations. For this, the annotator followed Meaning-Text Theory's definition of *collocation* (Mel'čuk, 2012). According to it, collocations are compositional phrasemes consisting of two elements: one freely chosen by speakers (the base); the other (collocate), which predicates some meaning of the base, is selected depending on the latter: cf. Sp. *vuelta* and its English equivalent *walk*, which cannot be combined with the direct translations of their respective support verbs: *dar una vuelta* 'lit. *give a walk' vs. **tomar una vuelta* 'lit. take a walk', in spite of the sense equivalence of the two expressions. This definition

contesto used as a noun was left as was, since it seems a clear calque from English quite removed from its Spanish equivalent *concurso*, and only its tag was changed from verb to noun (it happens to coincide with the first person present of the verb *contestar* 'to answer').

⁶ <https://www.corpusdelespanol.org/>

⁷ These pieces of research use t-score rather than frequency, but the rankings yielded by both of them are strikingly similar.

⁸ This threshold was established in order to discard possible hapaxes in the reference corpus. The threshold was low in order to have as many data as possible to predict the level of each text.

encompasses quite a variety of lexical combinations, ranging from support verb constructions (e.g. **make/do the homework*), to idiosyncratic combinations where the collocate has a very restricted applicability (*leap year*). For this process our annotator had a list of candidates and could optionally check their context in the corpus by means of a link.

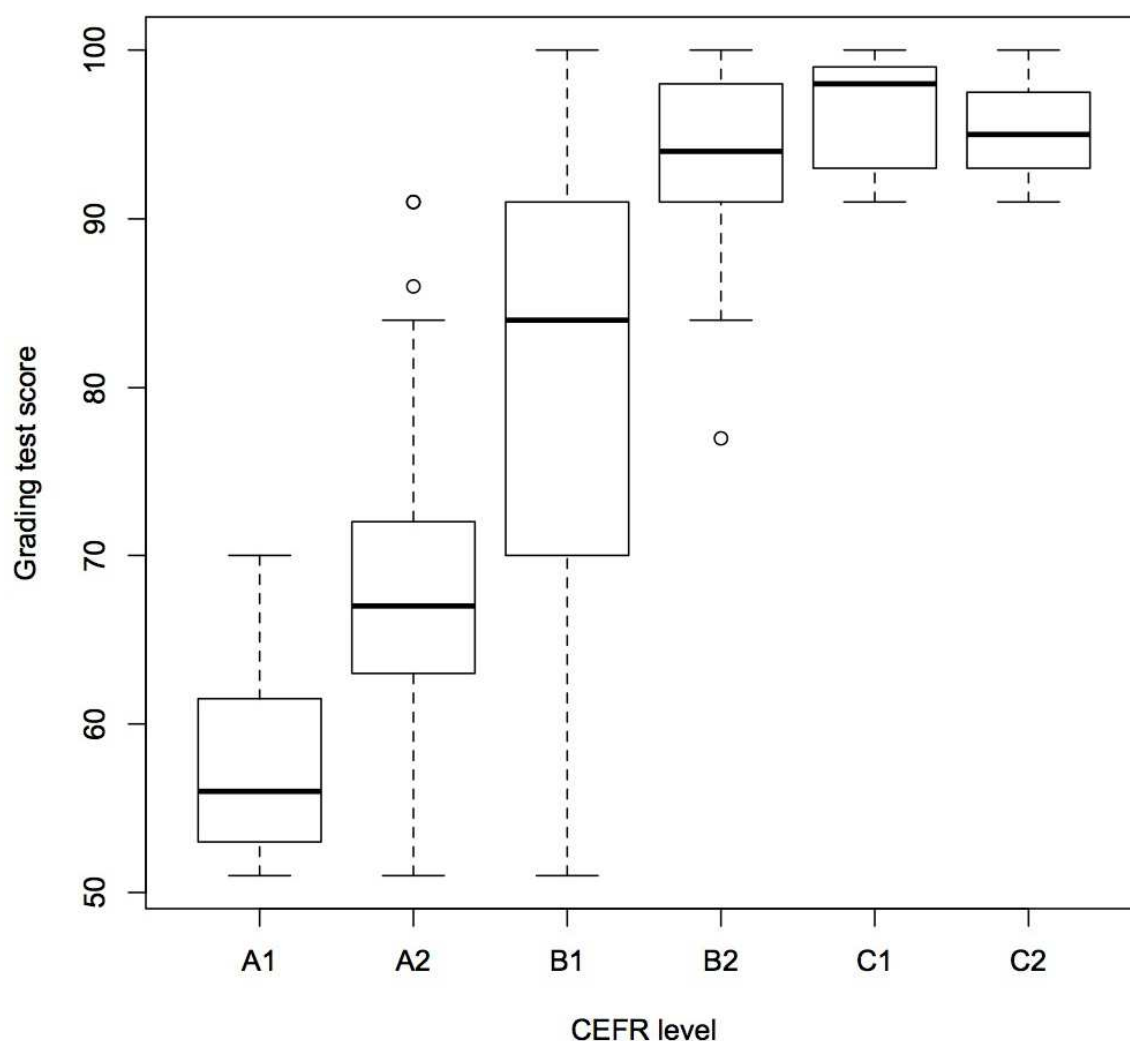


Figure 1: Correspondence between author's scores and CEFR levels of texts. Boxes represent the central 50% of data, the horizontal line in bold is the median and the segments outside boxes are the lowest and highest 25% of data. Outliers are represented by dots.

To evaluate the internal consistency of the annotation we performed an intra-annotator agreement analysis. The collocations of some texts included in the corpus had already been annotated for a previous project, although with a somewhat different procedure. On that occasion, the annotators read the entire texts and annotated their collocations in XML format. The compared samples consisted of 4,867 candidates from the 88 texts

that were annotated in both processes. The coincidence was of 87% (Fleiss' kappa = 0.7), a considerable value taking into account the differences between the two annotation processes and the time lapse between them (around five years). This agreement rate also provides indirect evidence on the syntactic parsing quality.

This annotation allowed us to establish a correspondence between the association measures of our sample and the fact that a combination was considered a phraseological collocation by a native speaker of Spanish, thus providing a further selection criterion for candidates inclusion in a dictionary.

4. Results and discussion

All the processes just described resulted in a set of collocation candidates associated to the CEFR level of the texts where they appeared and two association measures taken from their occurrences in the reference corpus.⁹ Using these data we examined whether there was any relation between proficiency level and the statistically measured association of candidates appearing in texts graded with such level.

The correspondence between candidate combinations' AMs and CEFR level can be seen in Figures 2 and 3. The first set of data correspond to the whole set of combinations, whereas in Figure 3 the data are mean scores for each text obtained from the association measures of the combinations it contains. As for the data in Figures 2, there seems to be a positive correlation between MI median and CEFR level in all three syntactic patterns, even though a considerable overlap between the different levels is apparent. It is also noticeable that MI values are rather low, particularly in the case of subject-verb combinations, where the medians in all levels fall below 3.

In the case of frequency, the correspondence between level and association scores is much less clear: it could be an inverse correlation in the case of subject-verb combinations, but in the other two syntactic patterns no such tendency emerges and the overlap for verb-object frequency values is almost total.

If we assign an average score to each text based on the AMs of candidates it contains, like in Bestgen and Granger (2014), Granger and Bestgen (2014) and Paquot (2018), a somewhat clearer picture emerges, but the general tendency is similar to that discussed above. In Figure 3 one can observe a clearer tendency for texts of higher levels to obtain higher average values of MI in all three syntactic patterns examined, particularly in adjective-noun combinations. Based on the average scores for these combinations, C1 and C2 are clearly detached from the rest. The MI values for the other two syntactic patterns are generally lower (especially in the case of subject-verb, as before), and some groups divert from the general tendency (namely, B2 in verb-object, which has a lower

⁹ In the case of frequency, we used the logarithm of base 10 of the raw frequency—cf .van Heuven et al. (2014).

median than B1, and C1 in subject-verb combinations).

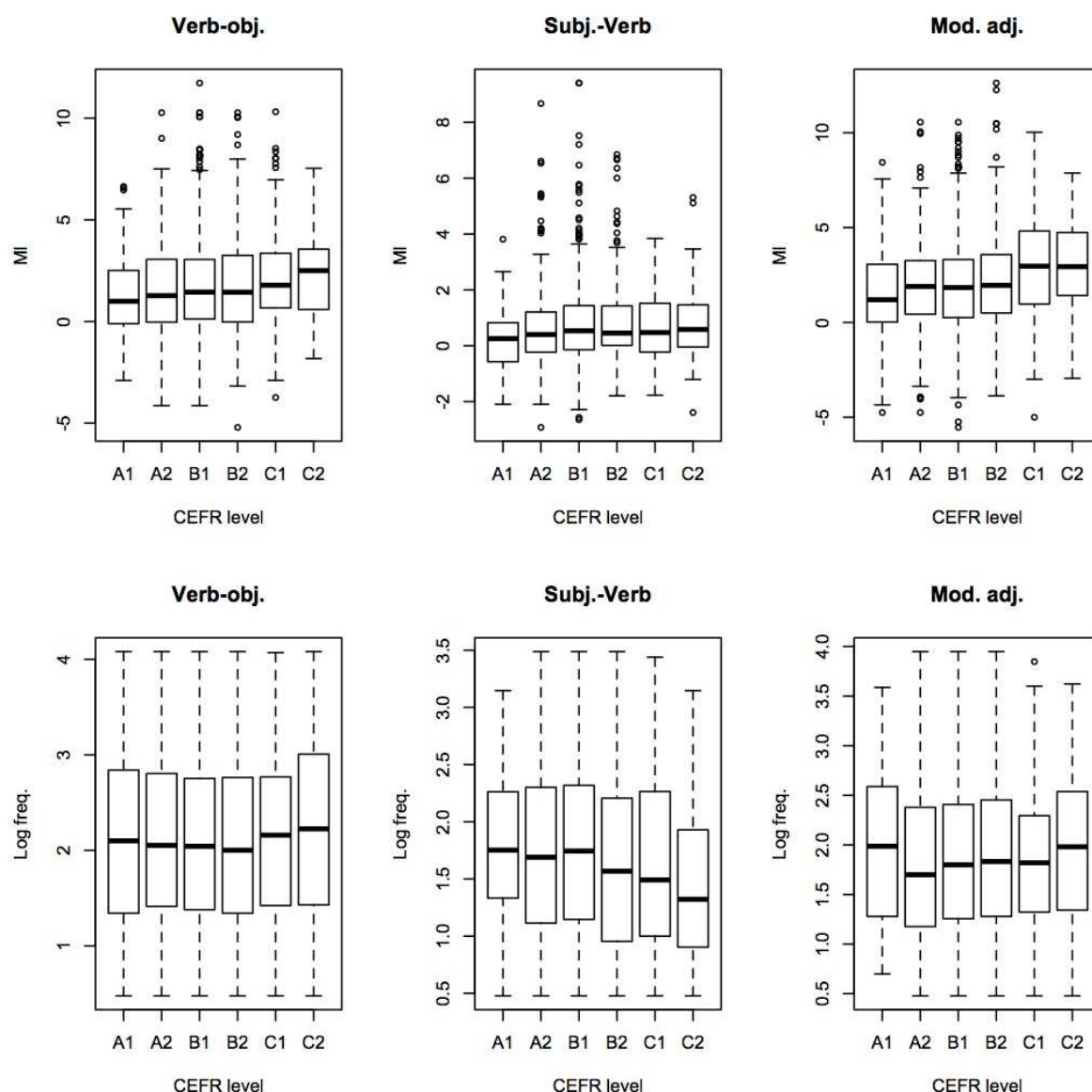


Figure 2: Association measures and CEFR level

With respect to frequency, again no linear progression in text averages can be observed in any of the syntactic patterns studied (although the median of verb-object based averages seems to draw a parabolic line).

In order to establish whether the observed tendencies reached statistical significance, the data were submitted to a generalized linear mixed model analysis.¹⁰ We treated the CEFR levels assigned to the texts as the dependent variable and the mean scores based on AMs as independent variables. Corpus texts were included into the model as random

¹⁰ For this we used R's lme4 package Bates et al. (2015): <https://cran.r-project.org/package=lme4>.

factors. Using the texts' mean scores based on the three syntactic patterns examined separately did not yield significant results. However, when the mean scores obtained from taking together the AMs of the three syntactic types of combinations were used, significant effects for mean MI and for the interaction between MI and frequency were observed, as can be seen in Table 2. The fixed effects of this model explains 39% of the variance.¹¹

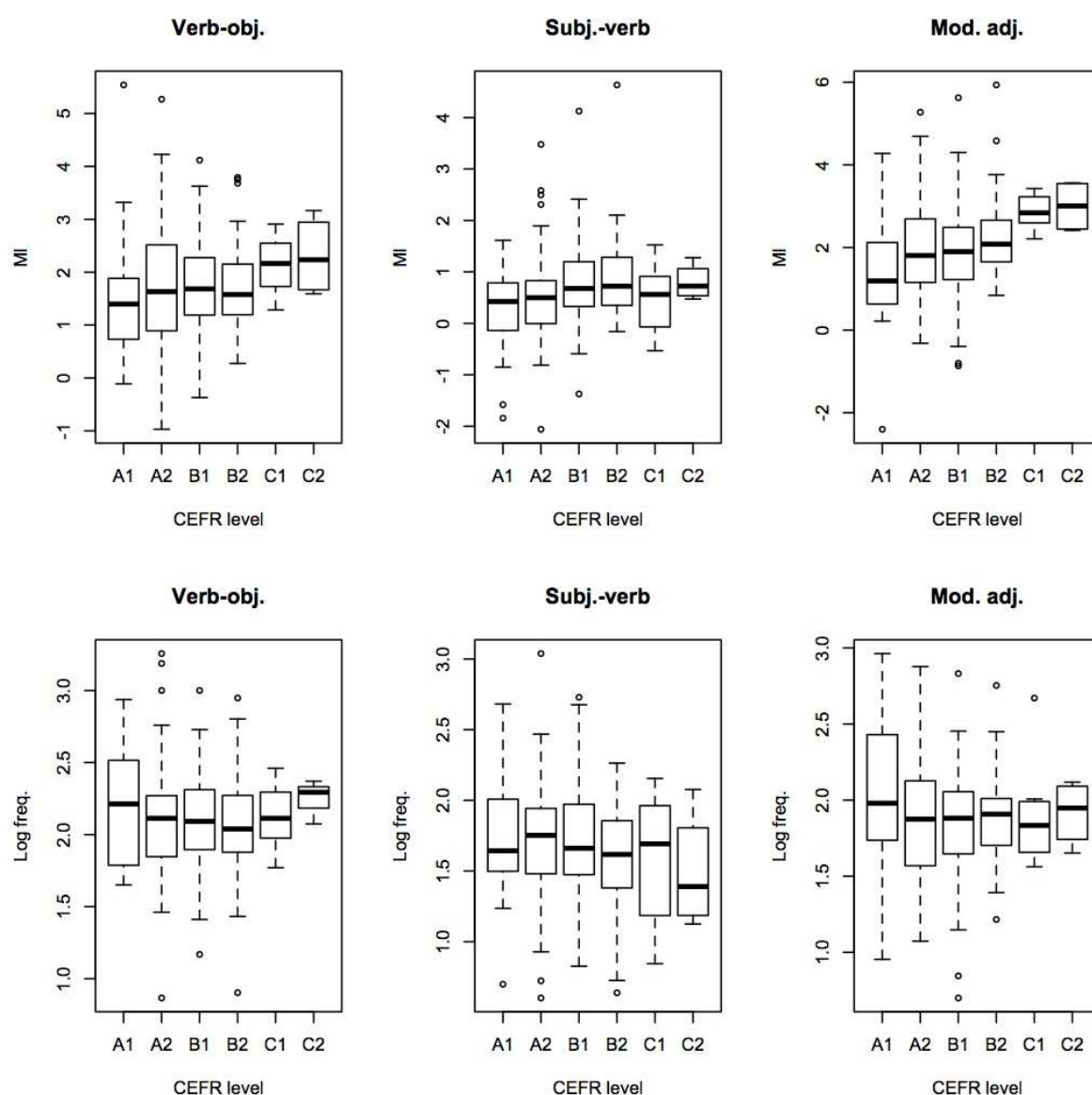


Figure 3: Association measure text means and CEFR level

These results indicate that the raters graded texts containing lexical combinations with higher MI values at more advanced levels. Frequency alone did not have an effect on

¹¹ R^2 was calculated with R's MuMIn package Barton (2019): <https://cran.r-project.org/package=MuMIn>

the level assigned to the texts, but it counteracted the effect of MI. This suggests that frequent lexical combinations, even if their members are highly associated (i.e., they have high MI scores), are not perceived as markers of advanced proficiency, at least not so clearly as less frequent combinations with equally high MI scores.

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-3.879	3.937	-0.985	0.32447
mean MI	9.302	3.269	2.845	0.00444 **
mean log freq.	2.386	2.047	1.165	0.24398
mean MI : mean log freq.	-4.028	1.602	-2.514	0.01195 *

Table 2: Fixed effects for the generalised linear mixed model predicting CEFR level. Factors marked with ** are significant at the 0.01 level and those marked with * at the 0.05 level.

All the above indicates that, when it comes to classifying a repertoire of collocations into CEFR proficiency levels, the strength of association between their constituents as measured by MI is more relevant than frequency of co-occurrence. This is in consonance with the findings of Bestgen and Granger (2014) and Paquot (2018) for English. However, given the effect of frequency in the opposite direction, those two dimensions should be somehow combined in such a classification.

The data examined so far refer to lexical combinations that only had to meet two conditions: they must have at least three occurrences in the reference corpus and be instances of one of the three syntactic dependencies referred to above. However, compiling a collocation dictionary that includes all the combinations that met these two requirements is probably not very interesting. Thus, for instance, we have seen the especially low MI values of subject-verb combinations that puts into question the phraseological status of many of the candidates belonging to this pattern. In order to refine the set of candidates, we will now review the data coming from the manual collocation annotation of the learner corpus' sample.

When examining the correlation between the human annotator's criterion and the AMs used here, frequency of co-occurrence shows up as slightly superior to MI in separating good from bad candidates, as can be seen in the precision-recall curves of Figure 4. Even using association measures, our results point to the need of human intervention in compiling a collocation dictionary. Thus, if we wanted to retrieve 80% of phraseological collocations included in the sample by using a log10 frequency value as the cut-off point (which in this case was $\log_{10}(\text{frequency}) \geq 1.53$, or 34 occurrences in raw frequency), the mean precision would be 35%, that is, 65% of candidates would have to be manually discarded.

If we extrapolate these figures to the data extracted from our reference corpus, we will end up with a set of ca. 50,000 candidates, which would eventually yield around 18,000

phraseological collocations.

To gauge what kind of collocation candidates would be extracted for each CEFR level using frequency and MI thresholds, we have used the sextiles corresponding to the MI values of the reference corpus data as cut-off points and extracted the ten best candidates for each level. In the case of A1 through B2 levels the candidates were those with the highest frequencies, whereas for C1 and C2 we extracted those with medium frequencies, in order to reflect somehow the negative interaction between frequency and MI. The results can be seen in Table 3. For the sake of clarity, we occasionally have used inflectional variants different from the lemma form.

A1	tener tiempo ‘have time’; tener cosa(s) ‘have thing(s)’; tener vida ‘have life; tener dinero ‘have money’; tener trabajo ‘have [a] job’; tener poder ‘have power’; (la) gente tiene/tenía/etc. ‘people have’; tener punto ‘have point(s)?’; tener día ‘have day’; tener nombre ‘have name’;
A2	tener idea ‘have idea’; tener relación ‘have relationship’; tener posibilidad ‘have possibility’; tener opción ‘have option’; ver cosa(s) ‘see thing(s)’; tener efecto ‘have an effect’; tener experiencia ‘have experience’; dar vida ‘give life’; persona tiene/tenía/etc. ‘people have’; tener valor ‘have value’; dar tiempo ‘give time’
B1	hacer cosa(s) ‘do thing’; tener problema(s) ‘have problem’; hacer tiempo ‘lit. make time, time ago’; tener razón ‘be right have reason’; tener sentido ‘have sense make sense’; tener derecho ‘have right’; tener suerte ‘have luck’; tener gana(s) ‘have desire’; tener oportunidad ‘have opportunity’; tener año(s) ‘have year’; tener hijo(s) ‘have children’
B2	hacer falta ‘need, lit. make lack’; mismo tiempo ‘same time’; mayor parte ‘most of’; gran parte ‘large part’; dar paso(s) ‘take step(s), lit. give step(s)’; llevar tiempo ‘take time, lit. carry time’; ver película ‘watch film’; decir cosa(s) ‘say things’; gran cantidad ‘large quantity’; dar oportunidad ‘give opportunity’; hacer daño ‘do harm’
C1	desarrollar trama ‘develop plot’; transformación profunda ‘deep transformation’; volcán alto ‘high volcano’; añadir aceite ‘add oil’; alojamiento web ‘web hosting’; provocar aparición ‘cause apparition’; artista extranjero ‘foreign artist’; respetar autor ‘respect author’; escuchar banda ‘listen (to a) band’; ganar batalla ‘win (the) battle’
C2	aumento considerable ‘considerable increase’; pedir auxilio ‘call for help’; bebida gaseosa ‘soda’; coger bici ‘take (the) bike’; beber café ‘drink coffee’; célula cerebral ‘brain cell’; certificado digital ‘digital certificate’; comida casera ‘home-cooked food’; compañía aseguradora ‘insurance company’; sintetizar concepto ‘sum up (a) concept’

Table 3: Samples of candidates by CEFR level.

Candidates with low MI and high frequency, i.e., those corresponding to A1 and A2 levels, tend to be support verb constructions with one of the most frequent verbs in

Spanish (*tener* ‘have’). This is in keeping with what the PCIC proposes. Thus, the most common verbs occurring in multiword expressions at levels A1 and A2 are *hacer* ‘do, make’ and *tener* ‘have’ (in addition to *ser* ‘be’). The sample here only includes the ten most frequent candidates, and is not very informative about other types of combinations (particularly, noun+adjective), which are less frequent and more scarce. It is at C1 and C2 levels (for which we took samples of medium frequency) where noun+adjective combinations start to appear regularly. Another issue is the presence of some free combinations (*tener cosa(s)* ‘have things’) seemingly not very interesting for learners, as well as combinations hardly recognisable out of context (*tener punto(s)* ‘have points?, have a score?’;).

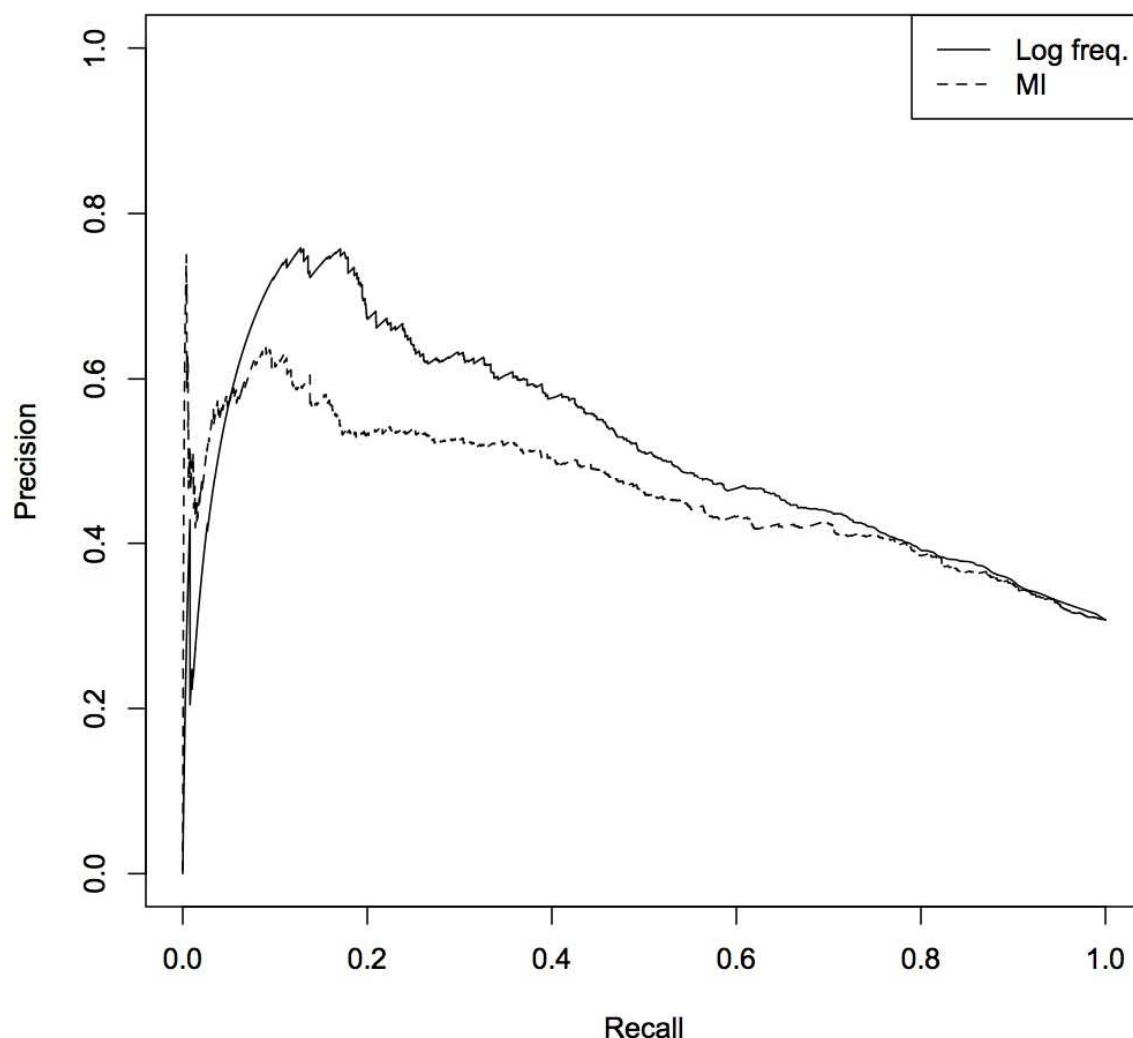


Figure 4: Precision-recall curves for candidates annotated as collocations

As for B1 and B2 levels, although candidate combinations with *tener* are also predominant here, they seem more interesting than those at lower levels. Among them,

however, there are some idioms (e.g. *hacer falta* ‘be necessary’) that require a different lexicographic treatment than collocations. Finally, the candidates in levels C1 and C2 are more variegated: there are more adjective-nouns combinations, although they also include free (e.g. *añadir aceite* ‘add oil’) and non-compositional combinations (*alojamiento web* ‘web-hosting’).

These observations call for a manual revision of candidates when compiling a dictionary: uninteresting free combinations probably should be excluded, non-compositional combinations should be distinguished from the rest in order to give them a different lexicographic treatment, etc. Notwithstanding, pre-processing using AMs seems a valuable guiding principle for selection and grading. As far as collocation selection is concerned, using a frequency cut-off point or examining only the n-best candidates ranked by frequency can alleviate the task of lexicographers, since as seen in Fig. 4 different values of AMs are associated with different precision rates. With respect to grading, we have proven that MI has an effect on the CEFR level given by raters.

5. Conclusion

This paper explored the use of association measures to extract collocations with a view to populating a dictionary of Spanish graded by CEFR levels. When it comes to grading collocations, much like in the case of single words, frequency seems in principle a reasonable criterion to determine the sequence of vocabulary presentation in curricula: giving priority to high-frequency lexical elements provides learners with valuable knowledge, both in terms of comprehension and production, given the high coverage rates of these elements.

However, the co-occurrence frequency of lexical combinations is not a good predictor of the proficiency level of learners’ texts. In this respect, MI has shown up as clearly superior. This finding is in line with what Granger and Bestgen (2014) and Paquot (2018) observe regarding the text quality of English learners. In consequence, future lexicographic ventures should take into account MI when it comes to grading lexical combinations.

Frequency, in turn, seems to perform slightly better than MI in distinguishing collocations from other types of lexical combinations (free, non-compositional) as identified by human annotators following phraseological criteria. This is at odds with some previous research (Ellis et al., 2008) and probably deserves further investigation. A possible reason is that here we used candidates in a syntactic relationship, rather than candidates within a given text span, in contrast to Ellis et al. (2008) (cf. Garcia et al., 2019, for similar results with a native sample).

Whereas the two association measures examined can ease the task of lexicographers by promoting collocational candidates (frequency) and providing a sequencing criterion (MI), they cannot guarantee a completely automated process with high quality results.

This study presented an initial approach that opens up further lines of research, starting with replications with more balanced data in terms of the representation of the different CEFR levels in the corpus—not an easy task given the difficulty to come by Spanish learner corpora of sizes comparable to those pertaining to other genres. Additionally, we have dealt with only two AMs, due to their spread use in related studies. However, a plethora of lexical AMs has been proposed (Pecina, 2010). It will be interesting, therefore, to study the correspondence between those measures and learners’ proficiency in future studies.

6. Acknowledgements

Marcos García Salido is the recipient of a post-doctoral grant from Xunta de Galicia (ED481D 2017/009), and Marcos García’s research is funded by a Juan de la Cierva incorporación grant (IJCI-2016-29598) and a 2017 Leonardo Grant for Researchers and Cultural Creators (BBVA Foundation). This research was also supported by a project funded by Ministerio de Economía, Industria y Competitividad (FFI2016-78299-P).

7. References

- Alonso Ramos, M., Roberto Carlini, J. C. F., Orol González, A., Vincze, O. & Wanner, L. (2015). Towards a learner need-oriented second language collocation writing assistant. In F. Helm, L. Bradley, M. Guarda & S. Thouësy (eds.) *Critical CALL—Proceedings of the 2015 EUROCALL Conference*, 2015. p. 16. <https://reference.research-publishing.net/publication/chapters/978-1-908416-29-2/304.pdf>.
- Alonso-Ramos, M. (2004). Diccionario de colocaciones del español. <http://www.dicesp.com/>.
- Alvar Ezquerro, M. (2005). La frecuencia léxica y su utilidad en la enseñanza del español como lengua extranjera. In M. A. C. Carballo, O. C. Moya, J. M. G. Platero & J. P. M. Gutiérrez (eds.) *Actas del XV Congreso Internacional de ASELE*. pp. 19–39. http://cvc.cervantes.es/ensenanza/biblioteca_ele/asele/pdf/15/15_0017.pdf.
- Barton, K. (2019). Package ‘MuMIn’. R Package Version 1.43.6. <https://cran.r-project.org/package=MuMIn>.
- Bates, D., Mächler, M., Bolker, B. & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), pp. 1–48.
- Benson, M., Benson, E. & Ilson, R. (1986). *The BBI combinatory dictionary of English: A guide to word combinations*. John Benjamins Publishing.
- Bestgen, Y. & Granger, S. (2014). Quantifying the development of phraseological competence in L2 English writing: An automated approach. *Journal of Second Language Writing*, 26, pp. 28–41. <http://dx.doi.org/10.1016/j.jslw.2014.09.004>.
- Bosque, I. (2004). *REDES. Diccionario combinatorio del español contemporáneo*. Madrid: SM.
- Bosque, I. (2006). *Diccionario combinatorio práctico del español contemporáneo*.

- Madrid: SM.
- Capel, A. (2010). A1 - B2 vocabulary: insights and issues arising from the English Profile Wordlists project. *English Profile Journal*, 1(01), p. e3. <http://dx.doi.org/10.1017/S2041536210000048>.
- Capel, A. (2012). Completing the English Vocabulary Profile: C1 and C2 vocabulary. *English Profile Journal*, 3(June 2012), p. e1. <http://dx.doi.org/10.1017/S2041536212000013>.
- Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Cambridge University Press.
- Ellis, N. C., Simpson-Vlach, R. & Maynard, C. (2008). Formulaic language in native and second-language speakers. *TESOL Quarterly*, 42(3), pp. 375–396.
- Erman, B. & Warren, B. (2000). The idiom principle and the open choice principle. *Text*, 20(1), pp. 29–62.
- Garcia, M. & Gamallo, P. (2015). Yet Another Suite of Multilingual NLP Tools. In J.P.L. J. L. Sierra-Rodríguez & A. Simões (eds.) *Languages, Applications and Technologies. Communications in Computer and Information Science*. Cham: Springer, pp. 65–75.
- Garcia, M., García-Salido, M. & Alonso-Ramos, M. (2019). A comparison of statistical association measures for identifying dependency-based collocations in various languages. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWEWN 2019)*. Florence: Association for Computational Linguistics.
- García Salido, M. & Alonso Ramos, M. (2017). Asignación de niveles de aprendizaje a las colocaciones del Diccionario de Colocaciones del español. *Revista Signos*, 51(97), pp. 153–174.
- Granger, S. (1998). Prefabricated patterns in advanced EFL writing: Collocations and formulae. In A. P. Cowie (ed.) *Phraseology: Theory, Analysis and Applications*. Oxford: Oxford University Press, pp. 145–160.
- Granger, S. & Bestgen, Y. (2014). The use of collocations by intermediate vs. advanced non-native writers: A bigram-based study. *IRAL - International Review of Applied Linguistics in Language Teaching*, 52(3), pp. 229–252.
- Higuera García, M. (2017). Pedagogical principles for the teaching of collocations in the foreign language classroom. In S. Torner & E. Bernal (eds.) *Collocations and other lexical combinations in Spanish: theoretical, lexicographical and applied perspectives*. London & New York: Routledge, pp. 250–266.
- Lozano, C. & Mendikoetxea, A. (2013). Learner corpora and second language acquisition. The design and collection of CEDEL2. In A. Díaz-Negrillo, P. Thompson & N. Ballier (eds.) *Automatic Treatment and Analysis of Learner Corpus Data*. Amsterdam/Philadelphia: John Benjamins, pp. 65–100.
- Martinez, R. (2013). A framework for the inclusion of multi-word expressions in ELT. *ELT Journal*, 67(April), pp. 184–198.
- Mel'čuk, I. (2012). Phraseology in the language, in the dictionary, and in the computer. *Yearbook of Phraseology*, 3, pp. 31–56.

- Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.
- Nesselhauf, N. (2004). *Collocations in a Learner Corpus*. Amsterdam/Philadelphia: John Benjamins.
- Padró, L. & Stanilovsky, E. (2012). FreeLing 3.0: Towards Wider Multilinguality. In N. Calzolari, K. Choukri, T. Declerck, M.U. Dogan, B. Maegaard, J. Mariani, J. Odiijk & S. Piperidis (eds.) *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC2012)*. European Language Resources Association (ELRA), pp. 2473–2479. <http://dblp.uni-trier.de/db/conf/lrec/lrec2012.html#PadroS12>.
- Paquot, M. (2018). Phraseological Competence: A Missing Component in University Entrance Language Tests? Insights From a Study of EFL Learners' Use of Statistical Collocations. *Language Assessment Quarterly*, 15(1), pp. 29–43.
- Pecina, P. (2010). Lexical association measures and collocation extraction. *Language Resources and Evaluation*, 44(1-2), pp. 137–158.
- Spina, S. (2016). Learner corpus research and phraseology in Italian as a second language: The case of the DICI-A, a learner dictionary of Italian collocations. In B. Sanromán Vilas (ed.) *Collocations Cross-Linguistically. Corpora, Dictionaries and Language Teaching*. Helsinki: Société Néophilologique de Helsinki, pp. 219–244.
- Straka, M., Hajic, J. & Straková, J. (2016). UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA), pp. 1659–1666.
- van Heuven, W. J., Mandera, P., Keuleers, E. & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology*, 67(6), pp. 1176–1190.
- Vincze, O., García-Salido, M., Orol, A. & Alonso-Ramos, M. (2016). A corpus study of Spanish as a Foreign Language learners' collocation production. In M. Alonso-Ramos (ed.) *Spanish Learner Corpus Research*. Amsterdam/Philadelphia: John Benjamins, pp. 299–331. <https://benjamins.com/catalog/scl.78.11vin>.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



Designing an Electronic Reverse Dictionary Based on Two Word Association Norms of English Language

Jorge Reyes-Magaña^{1,2}, Gemma Bel-Enguix¹, Gerardo Sierra¹,

Helena Gómez-Adorno³

¹ Instituto de Ingeniería, Universidad Nacional Autónoma de México, Circuito Escolar s/n, Ciudad Universitaria, Delegación Coyoacán, Ciudad de México, México

² Facultad de Matemáticas, Universidad Autónoma de Yucatán, Anillo Periférico Norte, Tablaje Cat. 13615, Colonia Chuburná Hidalgo Inn, Mérida, Yucatán, México

³ Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México, Circuito Escolar s/n, Ciudad Universitaria, Delegación Coyoacán, Ciudad de México, México

E-mail: jorge.reyes@correo.uady.mx, gbele@iingen.unam.mx, gsierram@iingen.unam.mx, helenagomez@iimas.unam.mx

Abstract

This work introduces the exploitation of some language resources, namely word association norms, for building lexical search engines. We used the Edinburgh Associative Thesaurus and the University of South Florida Free Association Norms for the construction of knowledge graphs that will let us execute algorithms over the nodes and edges in order to do a lexical search. The aim of the search is to perform an inverse dictionary search that, given the description of a concept as a query in natural language, will retrieve a target word. We evaluated two graph approaches, namely Betweenness Centrality and PageRank, using a corpus of human-definitions. The results are compared with the BM25 text-retrieval algorithm and also with an online reverse dictionary— OneLook Reverse Dictionary. The experiments show that our lexical search method is competitive with the IR models in our case study, even with a slight outperformance. This demonstrates that an inverse dictionary is possible to build with these kind of resources, no matter the language of the Word Association Norm.

Keywords: inverse dictionary; norm association words; graph theory

1. Introduction

Two types of dictionaries can be distinguished in order to link a concept with its meaning: semasiological and onomasiological. The former provide meanings, i.e. given a word, the user obtains the meaning of the word. The latter work in the opposite way, given the description of a word, the user obtains the related concept (Baldinger, 1970). The problem of building an onomasiological dictionary has been tackled in diverse ways, since in printed onomasiological dictionaries the words are not isolated, but usually arranged by shared semantic or associated features grouped under headwords (Sierra, 2000b; Sierra & McNaught, 2003). The main disadvantage in this type of search is that a really specific idea of the concept is needed in order to search in the right place of the index or headwords. Currently, an onomasiological dictionary can be thought of as

a simple internet search, thanks to the information accessible through different digital resources for almost any kind of topic. Unfortunately, the outcome of the search tends to be even more confusing, or it simply shows other results that do not correspond to the concept.

The present paper presents two algorithms that perform a lexical search over a knowledge graph in a similar way onomasiological dictionaries help to find a concept, starting from a definition or a set of clue words. We developed a model based on graph-based techniques, the *Betweenness centrality* and *PageRank*, to perform the search of a given concept on a dataset of word association norms for English, the Edinburgh Associative Thesaurus (EAT) (Kiss et al., 1973b), and the University of South Florida Free Association Norms (Nelson et al., 2004).

We used an evaluation corpus consisting of seven concepts. Although this is a small evaluation corpus, it can be considered as an illustrative example on how our method allows the building of reverse dictionaries using WAN. For each concept, 10 definitions were provided by human native speakers. In most cases, the definitions are very different from the ones found in dictionaries; they lack specialized terms and include cultural references and connotations. This allows us to design a more realistic electronic application, that will help people find a target word even with a limited knowledge of specific details. We used the 70 definitions as queries in our search model and compared the results with an information retrieval (IR) model (BM-25) and the online reverse dictionary *OneLook*¹. Our model achieved better results than the baseline IR model for this case of search scenario.

2. Related Work

2.1 Onomasiological searching

There are some specialized texts that aim to help writers who need to go from a meaning or concept to a corresponding word. These resources are gathered according to their behaviour in the following three features: a) the type of information they contain, b) the structure of the wordbook, and c) the type of search undertaken. We distinguish four different groups: *Thesauri*, *Reverse dictionaries*, *Synonymy and antonymy dictionaries* and *Pictorial dictionaries*.

The whole scenario of onomasiological searches changed with the universalization of the internet and language technologies, that allowed building online resources powered by the huge corpus the world wide web provides. In the last two decades, several online dictionaries have been designed that allow natural language searches. The users enter their own definition in natural language and the engine looks for the words that match the definition.

¹ <https://www.onelook.com/thesaurus/>

One of the first online dictionaries allowing this type of search was the one created for French by Dutoit and Nugues (2002). Another interesting contribution was introduced by Bilac et al. (2004), who designed a dictionary for Japanese. El-Kahlout and Oflazer (2004) built a similar resource for Turkish. For English, there exists an online onomasiological dictionary, OneLook Reverse Dictionary,² that retrieves acceptable results. One of the main works in Spanish is the one by Sierra (2000a), which was improved by Hernández (2012).

2.2 Free word associations

Free word associations (WA) are commonly collected by presenting a stimulus word (SW) to the participant and asking him to produce in a verbal or written form the first word that comes to his mind. The answer generated by the participant is called a response word (RW).

Compilations of WA are called Word Association Norms. Many languages have this type of resources, which are time-consuming to collect and need many volunteers.

In recent years, the web has become a natural way to get data to build such resources.

*Jeux de Mots*³ provides an example in French (Lafourcade, 2007), whereas the *Small World of Words*⁴ contained datasets in 14 languages at the time of writing. Nevertheless, the norms are only available in German. The authors (De Deyne et al., 2013) will make the other languages available as soon as they finish collecting the material. Such repositories have the problem of being collected without control over who is actually adding to the content, the linguistic proficiency of the users, and their age, gender or level of studies.

For Spanish, there exist several datasets of word associations. Algarabel et al. (1998) integrate 16,000 words, including statistical analyses of the results. Macizo et al. (2000) build norms for 58 words based on the responses of children, and Fernández et al. (2004) derived the free-association norms for the Spanish names of Snodgrass and Vanderwart pictures (Sanfeliu & Fernández, 1996).

The use of free word associations to compute relationality between words is not new. Borge-Holthoefer and Arenas (2009) describe a model (RIM) to extract semantic similarity relations from free association information. In recent years, Bel-Enguix et al. (2014) used techniques of graph analysis to calculate associations from large collections of texts. Additionally, Garimella et al. (2017) published a model of word associations

² <https://www.onelook.com/reverse-dictionary.shtml>.

³ <http://www.jeuxdemots.org/>.

⁴ <https://smallworldofwords.org/>.

that was sensitive to the demographic context.

The only resource designed and compiled for Mexican Spanish is the *Corpus de Normas de Asociación de Palabras para el Español de México*⁵ (Arias-Trejo et al., 2015).

Among the available compilations, the best-known in English are the *Edinburgh Associative Thesaurus*⁶ (EAT) (Kiss et al., 1973a) and the collection of the University of South Florida (USF) (Nelson et al., 1998)⁷. This work proposes the use of these datasets to be the basis of the design of a lexical search system that works from the clues or definitions to the concept, i.e., from the responses to the stimuli in order to build the reverse dictionary.

3. Word Association Norms datasets and graph

The EAT was mainly collected with undergraduate students from different British universities. The participants were between 17 and 22 years old, among which 64% were males and 36% were females. Every informant gave responses for 100 words, and every word was given to 100 participants. The resource was elaborated between 1968 and 1971 and published in 1973.

We used an XML version of the resource⁸, prepared by the University of Montreal, that consists of 8,211 stimulus words, and 20,445 different words including both, stimuli and responses.

The USF norms were collected with more than 6,000 participants that produced nearly three-quarters of a million responses to 5,019 stimulus words. Participants were asked to write the first word that came to mind that was meaningfully related or strongly associated with the presented word on the blank shown next to each item. The norms are distributed as plain text files separated by commas⁹ so that the document can be opened in a variety of different programs and databases. In this format, data for 5,019 normed words and their 72,176 responses can be found.

The graph representing the WAN's datasets has been elaborated with lemmatized lexical items. It is formally defined as: $G = \{V, E, \varphi\}$ where:

- $V = \{v_i / i = 1, \dots, n\}$ is a finite set of nodes of length n , $V \neq \emptyset$, that corresponds to the *stimuli* and their *associates*.

⁵ <http://www.labpsicolinguistica.psicol.unam.mx/Base/php/general.php>

⁶ <http://www.eat.rl.ac.uk/>

⁷ <http://web.usf.edu/FreeAssociation>

⁸ <http://rali.iro.umontreal.ca/rali/?q=en/Textual%20Resources/EAT>

⁹ <http://w3.usf.edu/FreeAssociation/AppendixA/index.html>

- $E = \{(v_i, v_j) | v_i, v_j \in V, 1 \leq i, j \leq n\}$, is the set of edges.
- $\varphi : E \rightarrow \mathbb{R}$, is a function over the weight of the edges.

We built separate graphs, each one is undirected so that every *stimulus* is connected to every associated word without any precedence order.

For the weight of the edges we used one of the following functions:

Frequency (F) Counts the number of occurrences of every associate to its *stimulus* in the whole dataset. For the system to work in the shortest paths, we need to calculate the *IF*, inverse frequency, that is defined in the following way: being *F* the frequency of a given associated word, and ΣF the sum of the frequencies of the words connected to the same *stimulus*, $IF = \Sigma F - F$

Association Strength (AS) Establishes a relation between the frequency (F) and the number of associations for every stimulus. It is calculated as follows: being *F* the frequency of a given associated word, and ΣF the sum of the frequencies of the words connected to the same *stimulus* (the total number of responses), the association strength (AS) of the word *W* to such *stimulus* is given by the formula:

$$AS_w = \frac{F}{\Sigma F}$$

For our experiments, we need to calculate the inverse association strength, *IAS*, in order to prepare the system to work with graph-based algorithms:

$$IAS_w = 1 - \frac{F}{\Sigma F}$$

Figure 1 depicts a subgraph of the EAT dataset, containing only four stimuli with their corresponding associates. It can be observed that there are some associate words that are common to different stimuli, even for this small subgraph. We can also find relationships between two stimuli; for example, *hamburger* and *lion*. Figure 2 depicts a subgraph of the USF dataset, containing the same four stimuli presented in Figure 1, but in this case the corresponding responses were the available in the American resource. We can observe that the associate word *food* is shared by *spaghetti* and *hamburger*.

4. Graph algorithms and the reverse dictionary

Given a definition, we search in the graph for the word that better matches it. For this purpose we considered centrality measures, because these type of algorithms identify the most important nodes in a graph; for example, the degree centrality assumes that

important nodes have many connections. The degree centrality is not suitable for our purposes because we need to find the most important nodes for a specific subset of nodes (the nodes that represent the words in a definition). In order to build the inverse dictionary we choose two algorithms, the *Betweenness Centrality* and *PageRank*, described in the following sections.

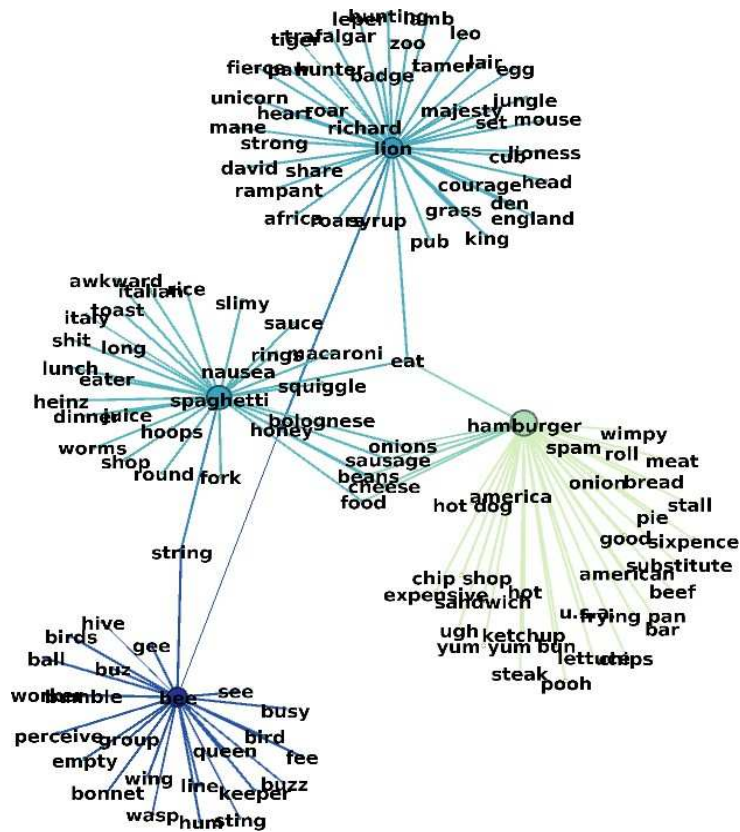


Figure 1: Subgraph based on EAT with the stimuli *bee*, *lion*, *hamburger*, and *spaghetti* with their corresponding associates.

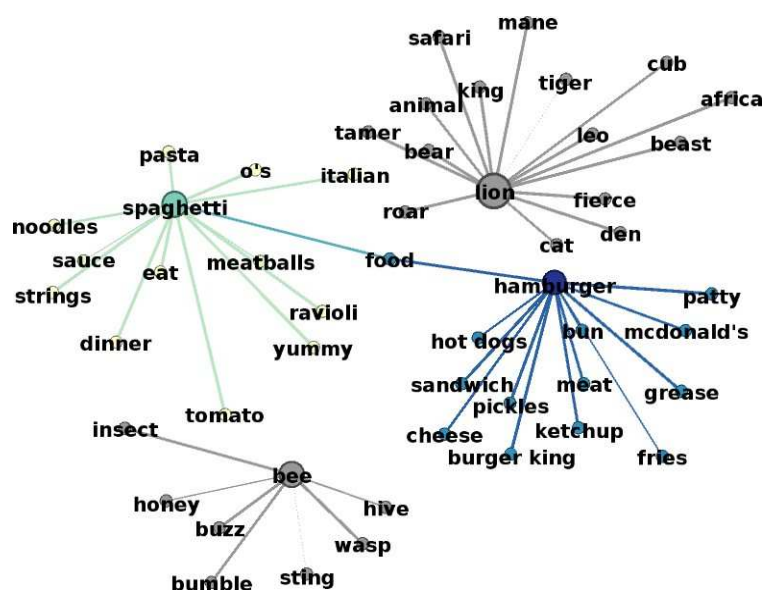


Figure 2: Subgraph based on Florida Free Association Norms with the stimuli *bee*, *lion*, *hamburger* and *spaghetti* with their corresponding associates.

4.1 Betweenness Centrality

We choose a variation of the *Betweenness Centrality* (BT) algorithm (Freeman, 1977) which instead of computing the BT of all pairs of nodes in a graph, calculates the centrality based on a sample (subset) of nodes (Brandes, 2008). The traditional betweenness algorithm assumes that important nodes connect other nodes. For a given node (v) in a graph (G), the BT is calculated as the relation between the number of shortest paths between nodes i and j that pass through v and the number of shortest paths between i and j . It is formally described as follows:

$$C_{btw}(v) = \sum_{i,j \in V} \frac{\sigma_{i,j}(v)}{\sigma_{i,j}} \quad (1)$$

where:

V = is the set of nodes, $\sigma_{i,j}$ is the number of shortest paths between i and j , and $\sigma_{i,j}(v)$ is the number of those paths that pass through some node v that is not i or j .

In a non-weighted graph, the algorithm looks for the shortest path. In a weighted graph, like the one we have built, it finds the path that minimizes the sum of the weights of the edges.

The BT algorithm was introduced based on the general idea that when a particular person in a group is strategically located on the shortest communication path connecting pairs of others, that person is in a central position (Bavelas, 2002). Noting the importance of the shortest paths, we adapted the information available in WAN, letting the most important nodes and their relations be represented as minimal values, as explained before. This is why we have adopted the weighting function based on the inverse frequency and the inverse association strength.

We employ the approximation of the BT algorithm in order to search for the concept related to a given definition. This is because it only uses a subset of nodes to find the most central ones in the graph. Our hypothesis is that, if we use a subset, the nodes of the WAN graph (WG) that represent the words of a definition as initial and final nodes in the BT algorithm, and calculate the centrality of the other nodes in WG taking these nodes as pairs, then the more central nodes will be the concept of such a definition. This approximation is formally described as follows:

$$C_{btw_approx}(v) = \sum_{i \in I, f \in F} \frac{\sigma_{i,f}(v)}{\sigma_{i,f}} \quad (2)$$

where: I is the set of initial nodes, F is the set of final nodes, $\sigma_{i,f}$ is the number of shortest paths between i and f , and $\sigma_{i,f}(v)$ is the number of those paths that pass through some node v that is not i or f .

Therefore, we define a subgraph composed by the words (nodes) of the definition. This subgraph is used as both initial and final nodes, for calculating the shortest paths from each of the nodes of the initial nodes set to each one of the nodes of the final nodes set. Finally, the nodes are ranked taking the measure of BT as a parameter for the comparison of the most important nodes found by the algorithm.

4.2 PageRank

PageRank computes a ranking of the nodes in a graph G based on the structure of the incoming links. It was originally designed as an algorithm to rank web pages. It was developed by Page et al. (1999), it is formally described as:

Let u be a web page. Then let F be the set of pages u points to and B be the set of pages that point to u . Let $N_u = |F_u|$ be the number of links from u and let c be a factor used for normalization (so that the total rank of all web pages is constant).

R represents the computation of PageRank, as follows:

$$R(u) = c \sum_{v \in B_u} \frac{R(v)}{N_v} \quad (3)$$

The rank of a page is divided among its forward links evenly to contribute to the ranks of the pages they point to. The equation is recursive but it may be computed by starting with any set of ranks and iterating the computation until it converges. In the most general and intuitive manner, PageRank corresponds to the standing probability distribution of a random walk on the graph of the Web.

Figure 3 shows Mathematical PageRanks for a simple network, expressed as percentages. Each value in the nodes represents the probability of a random walker finishing the path in it. The highest value is seen in node B, as it is the one with the most connections in the graph.

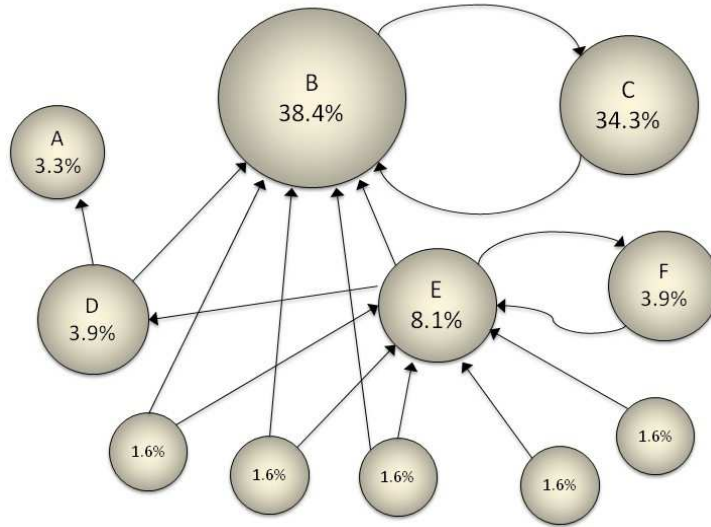


Figure 3: PageRank percentage in a simple network.

In our case, the pages described above are the words in the WAN datasets, the webpage links correspond to all the relations given by the stimuli-response between words. The hypothesis here is that the higher scores returned by the PageRank algorithm correspond to a target word being matched with a suitable definition. In this case, we didn't need the original graph to be tested with the algorithm because it will return the most relevant node of all the WAN dataset, instead, we pruned the graph considering some aspects described in the following subsection.

4.3 Algorithm description

Algorithm 1: Reverse dictionary

Data: WAN datasets, definitions to search

Result: list of ranked concepts

pre-process(WAN datasets);

pre-process(definitions to search);

GraphWAN = build-graph(WAN datasets);

GraphWAN = prune-graph(GraphWAN);

for each definition **do**

 definition = remove-StopWords(definition);

 definition = filter-WordsInWAN(definition);

 build_subgraph(definition);

 ranking_nodes_BT = BT(GraphWAN,subgraph);

 ranking_nodes_PR = PR(GraphWAN);

 ascending_order(ranking_nodes_BT);

 ascending_order(ranking_nodes_PR);

Algorithm 1 presents the overall schema of our model. The WAN datasets used here as input refer to both EAT or USF norms. First, we perform some pre-processing steps. All the *stimuli* and the *responses* are lemmatized, leaving each word as the most representative of the flexed forms. The same pre-processing is applied to the definitions to be searched by the model. This process provides us with more matches in the case when the definition contains *table*, *tables*, etc. because it will be transformed into its lemma, *table*. For this purpose, we use the lemmatization process available in *spacy*¹⁰.

Later, we built GraphWAN with the Python package Networkx (Hagberg et al., 2005). Due to various experiments carried out with the original graph we discovered that compression was needed in order to get a more compact graph to be processed, and for this purpose we prune the original graph taking all the neighbours for each word of the definition to be searched, i.e. all nodes that have a connection with the words of the definition were selected considering the original graph structure.

Then, for each definition to be searched we removed all the functional words using the stop words list available in the *NLTK* package (Bird & Loper, 2004). Next, with the list of words with lexical meaning, we kept only the ones that belong to the vocabulary in WAN. With this we built a subgraph to be the input in the Betweenness Centrality algorithm. Finally, the nodes were sorted out according to the highest centrality measure, which corresponds to the words that are closer to the ones of the definition.

5. Experiments and results

5.1 Evaluation corpus

For the experiments, a small corpus containing 10 definitions for seven concepts was used, and these definitions were taken from Sierra and McNaught (2000), originally used for evaluating their work. These definitions are reported to be gathered with a small group of twenty undergraduate students in the area of terminology. From two sets, each student was asked to take a set and write on a blank sheet of paper, similar to an onomasiological search, a concept, a definition or the ideas suggested to them by each word. After exchanging the sheets, the other students participating in the experiment wrote the word or words designating the concepts identified or written on the blank sheets by the previous student.

The selected words used for evaluating our system are: *water*, *squirrel*, *bench*, *hurricane*, *lemon*, *bucket* and *clothes*. Table 1 presents an example of 10 definitions of the same concept given by different students.

¹⁰ <https://spacy.io/>

It's a little rodent and can be red or grey, it has a big bushy tail
A small rodent living in trees with a long bushy tail
A small rodent which lives in trees, collects nuts and has a bushy tail
Animal, grey/red, bushy tail, lives in trees, buries nuts
Small animal, lives in trees, eats acorns, has a bushy tail
Animal, bushy tail, eats nuts, builds nests in trees called dreys
Small funny animal with big, bushy tail, likes nuts, likes trees
Animal that lives in trees and collects acorns, has a long tail
A small-sized animal, habitat in trees
Small grey mammal, relative to the rodent, found in both countryside and town

Table 1: Definitions of *squirrel* given by the students.

5.2 Results with the inverse dictionary and graphs

The experiments were performed taking into account weighted graphs with the two previously mentioned functions: Inverse Frequency (IF) and Inverse Association Strength (IAS). Considering separated graphs with each of the WAN datasets.

For the evaluation of the inference process, we used the technique of precision at k ($p@k$) from Manning et al. (2009). For example, $p@1$ shows that the concept associated with a given definition was ranked correctly in the first place, in $p@3$ the concept was in the first three results, and the same applies to $p@5$ and $p@10$.

The results are shown in Tables 2 and 3. As a general statement when the model searches over the graphs weighted with IAS the results are higher than when searching on the graph weighted with IF in both datasets. Psychologists agree that Association Strength (AS) is the measure that implies a cognitive relationship between two terms, and this idea is reflected in our results. Frequency is closely related to AS, but it lacks the generalization of the latter function.

Regarding the WAN datasets, the best results are achieved using USF Word Association Norms processed with Betweenness Centrality. We consider this is because this algorithm lets us create a source and target of nodes that exactly correspond to the words given by a user in the definition, compared to PageRank that analyses the graph built with the neighbourhood around this words.

Weighting function	Graph Algorithm	p@1	p@3	p@5	p@10
Inverse Frequency (IF)	Betweenness Centrality (BT)	0.152	0.186	0.220	0.237
Inverse Association Strength (IAS)	Betweenness Centrality (BT)	0.152	0.220	0.237	0.254
Inverse Frequency (IF)	PageRank (PR)	0.000	0.074	0.129	0.129
Inverse Association Strength (IAS)	PageRank (PR)	0.000	0.0740	0.129	0.129

Table 2: Results in terms of precision of our model with EAT dataset

Weighting function	Graph Algorithm	p@1	p@3	p@5	p@10
Inverse Frequency (IF)	Betweenness Centrality (BT)	0.236	0.309	0.418	0.436
Inverse Association Strength (IAS)	Betweenness Centrality (BT)	0.290	0.363	0.418	0.5272
Inverse Frequency (IF)	PageRank (PR)	0.037	0.074	0.129	0.222
Inverse Association Strength (IAS)	PageRank (PR)	0.037	0.074	0.148	0.222

Table 3: Results in terms of precision of our model with USF dataset

5.3 Results

In order to evaluate the relevance of our method, we performed experiments with other well-known IR methods.

First, we compared the performance of our method with the results of a reverse dictionary. To do that, we used the OneLook Thesaurus that allows you to describe a concept and returns a list of words and phrases related to that concept. The definitions were manually checked using the OneLook web application¹¹.

Secondly, we performed experiments with one of the most successful text-retrieval algorithms, Okapi BM25, based on probabilistic models and developed in the seventies by Stephen E. Robertson and Karen Spärck Jones (1976). The algorithm implemented following Robertson and Zaragoza (2009) is based on the bag-of-words method. Given a query, it ranks a list of documents according to their relevance for such query. We have applied it considering as a document every definition and every set of responses to a stimulus.

¹¹ <https://www.onelook.com/thesaurus/>

Method	P@1	P@3	P@5	P@10
OneLook	0.202	0.347	0.376	0.434
Reverse Dictionary with USF (IAS)	0.290	0.363	0.418	0.5272
BM25 with EAT	0.257	0.357	0.414	0.471
BM25 with USF	0.257	0.400	0.457	0.514

Table 4: Comparative precision results

The results achieved using the two baselines, OneLook and BM25, are reported in Table 4, where they are compared with the best result obtained by the inverse dictionary with our model. The BM25 algorithm showed better performance than the OneLook reverse dictionary when the search was performed over the WAN datasets. The BM25 was implemented using both WAN datasets. For each *stimuli* we built a document containing all the *responses* established in the resource. The better results are consistent with the ones seen in the reverse dictionary, USF norms show the best performance with this *IR* algorithm. It is observed that this algorithm is the most competitive against our model, but we outperformed the results in $p@1$ and $p@10$, while we unperformed in $p@3$ and $p@5$.

The system is fast, efficient and demonstrates high performance. However, the structure of the resource we have built favours the fact that two words that are not really related by association could have a short path between them because they share a connected word. This is expected to be a problem of our reverse dictionary based on WANs, although it can be minimized by performing some kind of lexical filter in the future.

6. Conclusions and future work

This paper introduces a model for onomasiological searches that has some novelties; among them the simplicity, the use of graph-based techniques and the WAN datasets the method is based on. However, we observed that the graph built with all the nodes and edges contained in the datasets tends to be not so good, due to the number of paths that lead to the wrong results. In order to solve this problem, we had to make a graph reduction keeping the most relevant nodes and their paths.

We have shown how descriptions of concepts that are made by ordinary people with non-scientific specifications can retrieve accurate results using our method. This is possible thanks to the nature of the dataset. Indeed, word association norms group words that are closely related in a cognitive way, and taking advantage of the metrics in the original resource that can be used to produce weighted edges in the graph that is built.

The success of the system with non-scientific input can drive new lines of applied research, and the implementation of different assistant writing systems especially oriented to people with a range of aphasias, like dysnomia and Alzheimer’s disease.

Our algorithm has shown competitive performance compared with other baseline systems.

7. Acknowledgments

This research has been supported by projects PAPIIT IA401219 from the Universidad Nacional Autónoma de México; and CONACYT Fronteras de la Ciencia 2016-01-2225.

8. References

- Algarabel, S., Ruíz, J. C. & Sanmartín, J. (1998). *The University of Valencia's computerized Word pool*. Instruments & Computers.: Behavior Research Methods.
- Arias-Trejo, N., Barrón-Martínez, J.B., Alderete, R.H.L. & Aguirre, F.A.R. (2015). *Corpus de normas de asociación de palabras para el español de México [NAP]*. Universidad Nacional Autónoma de México.: UNAM.
- Baldinger, K. (1970). *Teoría semántica: hacia una semántica moderna*, volume 12. Alcalá.
- Bavelas, A. (2002). A mathematical model for group structure. *Social networks: critical concepts in sociology*, New York: Routledge, 1, pp. 161–88.
- Bel-Enguix, G., Rapp, R. & Zock, M. (2014.). A Graph-Based Approach for Computing Free Word Associations. In *Proceedings of the 9th edition of the Language Resources and Evaluation Conference*. pp. 221–230.
- Bilac, S., Watanabe, W., Hashimoto, T., Tokunaga, T. & Tanaka, H. (2004). Dictionary search based on the target word description. In *Proceedings of the Tenth Annual Meeting of the Association for Natural Language Processing*. pp. 556–559.
- Bird, S. & Loper, E. (2004). NLTK: the natural language toolkit. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*. Association for Computational Linguistics, p. 31.
- Borge-Holthoefer, J. & Arenas, A. (2009). Navigating Word Association norms to Extract Semantic Information. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society*.
- Brandes, U. (2008). On variants of shortest-path betweenness centrality and their generic computation. *Social Networks*, 30(2), pp. 136–145.
- De Deyne, S., Navarro, D.J. & Storms, G. (2013). Better explanations of lexical and semantic cognition using networks derived from continued rather than single-word associations. *Behavior research methods*, 45(2), pp. 480–498.
- Dutoit, M. & Nugues, P. (2002). A Lexical Database and an Algorithm to Find Words from Definitions. In *Proceedings of the 15th European Conference on Artificial Intelligence*. pp. 450–454.
- El-Kahlout, I. & Oflazer, K. (2004). Use of Wordnet for Retrieving Words from Their Meanings. In *2nd Global WordNet Conference*.
- Fernández, A., Díez, E., Alonso, M.A. & Beato, M.S. (2004). Free-association norms form the Spanish names of the Snodgrass and Vanderwart pictures. *Behavior*

- Research Methods, Instruments & Computers*, 36, pp. 577–583.
- Freeman, L.C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, pp. 35–41.
- Garimella, A., Banea, C. & Mihalcea, R. (2017). Demographic-aware word associations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. pp. 2285–2295.
- Hagberg, A., Schult, D. & Swart, P. (2005). Networkx: Python software for the analysis of networks. *Mathematical Modeling and Analysis, Los Alamos National Laboratory*.
- Hernández, L. (2012). *Creación semi-automática de la base de datos y mejora del motor de búsqueda de un diccionario onomasiológico*. Universidad Nacional Autónoma de México.
- Kiss, G., Armstrong, C., Milroy, R. & Piper, J. (1973a). *An associative thesaurus of English and its computer analysis*. Edinburgh.: Edinburgh University Press.
- Kiss, G.R., Armstrong, C., Milroy, R. & Piper, J. (1973b). An associative thesaurus of English and its computer analysis. *The computer and literary studies*, pp. 153–165.
- Lafourcade, M. (2007). Making people play for Lexical Acquisition. In *Proceedings of the th SNLP 2007, Pattaya, Thailand*, 7, pp. 13–15.
- Macizo, P., Gómez-Ariza, C.J. & Bajo, M.T. (2000). Associative norms of 58 Spanish for children from 8 to 13 years old. *Psicológica*, 21, pp. 287–300.
- Manning, C., Raghavan, P. & Schütze, H. (2009). *Introduction to Information Retrieval*. Cambridge University Press.
- Nelson, D.L., McEvoy, C.L. & Schreiber, T.A. (1998). *Word association rhyme and word fragment norms*. The University of South Florida.
- Nelson, D.L., McEvoy, C.L. & Schreiber, T.A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, 36(3), pp. 402–407.
- Page, L., Brin, S., Motwani, R. & Winograd, T. (1999). The PageRank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.
- Robertson, S. & Sparck Jones, K. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27(3).
- Robertson, S. & Zaragoza, H. (2009). The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval*, 3(4), pp. 333–389.
- Sanfeliu, C. & Fernández, A. (1996). A set of 254 Snodgrass' Vanderwart pictures standardized for Spanish: Norms for name agreement, image agreement, familiarity, and visual complexity. *Behavior Research Methods, Instruments, & Computers*, 28, pp. 537–555.
- Sierra, G. (2000a). Design of an onomasiological search system: A concept-oriented tool for terminology. *Terminology*, 6(1), pp. 1–34.
- Sierra, G. (2000b). The onomasiological dictionary: a gap in lexicography. In *Proceedings of the Ninth Euralex International Congress*. pp. 223–235.
- Sierra, G. & McNaught, J. (2000). Extracting semantic clusters from MRDs for an

- onomasiological search dictionary. *International Journal of Lexicography*, 13(4), pp. 264–286.
- Sierra, G. & McNaught, J. (2003). Natural Language System for Terminological Information Retrieval. In A. Gelbukh (ed.) *Computational Linguistics and Intelligent Text Processing*. Berlin, Heidelberg: Springer, pp. 541–552.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



Towards Electronic Lexicography for the Kurdish Language

Sina Ahmadi¹, Hossein Hassani², John P. McCrae¹

¹ Insight Centre for Data Analytics, National University of Ireland Galway

² Department of Computer Science and Engineering, University of Kurdistan Hewlêr

E-mail: sina.ahmadi, john.mccrae@insight-centre.org, hosseinh@ukh.edu.krd

Abstract

This paper describes the development of lexicographic resources for Kurdish and provides a lexical model for this language. Kurdish is considered a less-resourced language, and currently, lacks machine-readable lexical resources. The unique potential which Linked Data and the Semantic Web offer to e-lexicography enables interoperability across lexical resources by elevating the traditional linguistic data to machine-processable semantic formats. Therefore, we present our lexicon in Ontolex-Lemon ontology as a standard model for sharing lexical information on the Semantic Web. The research covers the Sorani, Kurmanji, and Hawrami dialects of Kurdish. This research suggests that although Kurdish is a less-resourced language, in terms of documented lexicons, it has a wide range of resources, but because they are not machine-readable they could not contribute to the language processing. The outcome of this project, which is made publicly available, assists scholars in their efforts towards making Kurdish a resource-rich language.

Keywords: Kurdish; e-lexicography; less-resourced languages; machine-readable dictionary

1. Introduction

Linguistic resources are knowledge repositories which not only provide lexical and semantic descriptions of words but also reflect the culture and civilization of speakers of a language. In an era when human language is more and more frequently processed by machines, such resources are crucial components of language technology and natural language processing (NLP). Kurdish, as a less-resourced Indo-European language spoken in several dialects and written using different scripts (Forcada et al., 2019), still lacks such resources. In an attempt to remedy the lack of resources for Kurdish, we provide machine-readable dictionaries for three of the five main dialects of Kurdish, namely Kurmanji, Sorani, and Hawrami.

A machine-readable dictionary (MRD) not only provides lexicographic information in an electronic form, but is also a database which can be queried and therefore integrated in NLP tools. As the body of the research in Kurdish language processing is still scant, we believe that such resources will pave the way for further developments in the field. We also believe that lexical resources will enable researchers to address more NLP tasks which may require lexicographic resources such as word sense disambiguation (Navigli & Ponzetto, 2012) and semantic parsing (Shi & Mihalcea, 2005) and enhance the quality of the existing NLP applications.

The Semantic Web as an extension of the World Wide Web (WWW) represents an effective means of data representation and enables users and computers to retrieve and share information efficiently (Berners-Lee et al., 2001). The Resource Description Framework (RDF) is the foundational data model for the Semantic Web. Unlike traditional databases where data has to adhere to a fixed schema, RDF documents are not prescribed by a schema and can be described without additional information, making RDF data model self-describing (Klyne & Carroll, 2004). More recently, the concept of the Web of Linked Data, which makes RDF data available using the HyperText Transfer Protocol (HTTP) and Linguistic Linked Open Data (Chiarcos et al., 2013), has gained traction along with the Semantic Web, particularly in the NLP community as a standard for linguistic resource creation. Moreover, the unique potential which the Semantic Web and Linked Data offer to e-lexicography enables interoperability across lexical resources by leveraging printed or unstructured linguistic data to machine-readable semantic formats.

This paper has two major contributions:

- It provides a thorough review of the current state of Kurdish lexicography, both traditional and electronic. Such a review includes an analysis of the properties of the existing Kurdish dictionaries, such as type of dictionary (monolingual, bilingual, multilingual), script of the Kurdish text (Persian-Arabic, Latin or Cyrillic), description of the content and size of dictionaries. Although very few in comparison to printed dictionaries, terminological resources and electronic dictionaries are also covered in this paper. This review helped us to differentiate between the lack of resources and unavailability of lexicographic resources in electronic forms. We discovered that Kurdish, from the lexicographic point of view, is not as less-resourced as claimed in the literature. Instead, other issues have hindered the availability of these resources in a machine-readable form, which has resulted in the perception that the language lacks such essential assets. This is not equally true for all the Kurdish dialects, but it is obvious for the two widely spoken dialects, namely Kurmanji and Sorani.
- We present three machine-readable dictionaries based on the OntoLex-Lemon model for Kurmanji, Sorani and Hawrami. We not only included frequent headwords in the dictionaries, namely 4,172 entries for Kurmanji, 5,683 entries for Sorani and 1,184 for Hawrami, but also tried to create a prototypical resource which may be easily adapted by future Kurdish lexicographers. Despite the existence of a few electronic word lists and glossaries for Sorani and Kurmanji, our electronic Hawrami dictionary is the first one of its kind for this dialect.

For this, we consider two stages in the development of our resources. First, we collect the vocabulary for each dialect. This stage includes manual work for the extraction of entries and annotating each part of their description, such as gender, part-of-speech (PoS), sense, English translations, example and etymology. This step is followed by a semiautomatic normalization of the scripts and orthography. In the next step, the

lexicographic information is semi-automatically transformed from a tabular format into the OntoLex-Lemon model in the Resource Description Framework (RDF).

The rest of this paper is organized as follows. We first describe the Kurdish language, its various dialects and scripts in Section 2. In Section 3, we provide a survey on the history of Kurdish lexicography and available lexicons. Section 4 describes the development of our resources according to the OntoLex-Lemon standard. Following this section, insights into the developed resources are provided in Section 5. The paper is concluded in Section 6, where we provide suggestions for modern e-lexicography for Kurdish and future steps in this direction. Note that throughout this study, lexicon and dictionary are used interchangeably.

2. Kurdish language

2.1 Dialects

Kurdish is an Indo-European multi-dialect language which is spoken by about 30 million speakers (Hassani, 2018). The dialects are referred by different names, namely Kurmanji, Sorani, Hawrami and Kirmashani (Hassani, 2018). The Kurmanji speakers, as the majority of Kurdish speaking population, are located in different areas of Syria, Iraq, Turkey and Iran. Sorani is the second most popular dialect, which is mainly spoken among Kurds in Iran and Iraq. Similarly, Hawrami is primarily spoken in Iran and Iraq, but among a smaller community. Moreover, almost all Kurdish dialects are also spoken among a large Kurdish diaspora in different western countries (Hassanpour, 1992).

The debate over the concept of dialects versus languages, the attribution of different dialects to Kurdish or considering some as separate languages has been around for decades (Hassani, 2018). According to the literature (Hassanpour, 1992; Haig & Matras, 2002; Hassani, 2018), the debate expanded to how to categorize and name the dialects. However, to avoid drifting beyond our purpose, in this research we prefer to follow the common approach among the researchers in Kurdish NLP with regard to dialect attribution, their categorization, and naming style according to the way presented in the Kurdish BLARK (Hassani, 2018).

2.2 Scripts and orthographies

Kurdish poetry and prose narratives were historically transmitted orally (Kreyenbroek, 2005), therefore the language does not have a long history of written texts (Hassani & Medjedovic, 2016). While some scholars have different opinions, the dominant conclusion dates the appearance of the first written Kurdish text to circa 1600 (Hassani, 2018). Since then, the language has been written in Persian-Arabic until the beginning of the 20th century, when due to geopolitical conditions the usage of Latin, Cyrillic, and to a limited extent, Armenian scripts was started. In the 1920s, the first attempts to present a standard writing system for Kurdish began. As a result, in 1932 Jeladet Ali Bedirkhan (in Kurdish, *Celadet Elî Bedirxan*) introduced a

Latin-based orthography (also known as Bedirxan alphabet) (Bedirxan & Lescot, 1970), while a group of scholars introduced one based on the Persian-Arabic script in Iraq. These orthographies are both based on the phonetics of the language. The usage of Cyrillic and Armenian was mainly restricted to the communities in Armenia and the former Soviet countries (Hassanpour, 1992). Gradually, the Persian-Arabic and Latin-based scripts have become more dominant in various Kurdish speaking regions, although their popularity differs from region to region. The Persian-Arabic orthography is dominant in the Kurdish regions of Iraq, Iran, and Syria (Haig & Matras, 2002). On the other hand, the Latin-based orthography is used by the Kurds in Turkey. According to Hassani and Medjedovic (2016), the usage of Latin-based orthography is growing and becoming more popular in Iraq and Syria, with a greater usage by the Kurdish media, particularly in the Kurdistan Region of Iraq.

In an attempt to standardize and unify the scripts for all Kurdish dialects, the Kurdish Academy of Language has recently introduced a Unified Kurdish Alphabet, *Yekgirtû*¹, which is based on the Latin orthography. Figure 1 illustrates Kurdish phonemes in all dialects and their corresponding letters in the alphabets. The grey cases refer to non-existing characters.

2.3 Kurdish language processing

Hassani (2018) provides a summary of the Kurdish NLP situation in which the status of the available data and tools for Kurdish NLP are presented. However, we also address a few essential efforts on Kurdish NLP which are pertinent to the current research, particularly on Kurdish language processing resources and tools.

The initiative to create corpus for Kurdish dates back to 1998 (Gautier, 1998). However, efforts in creating machine-readable corpora for Kurdish are recent. The first machine-readable corpus for Kurdish is the Leipzig Corpora Collection which contains some 56,000 sentences of Sorani Kurdish constructed using different sources such as the Internet, newspapers, and Wikipedia (Biemann et al., 2007). In 2013, the Kurdish Language Processing Project created Pewan (Esmaili et al., 2013) which is composed of 115,000 Sorani and 25,000 Kurmanji news articles. KurdNet (Aliabadi et al., 2014) is the Kurdish WordNet, and currently only contains Sorani translations of the Base Concept of the English WordNet (Miller, 1995). Bianet is a parallel news corpus of Turkish, English and Kurmanji containing 3,214 articles (Ataman, 2018). In addition, researchers have created Kurdish corpora for particular NLP tasks, for example, part-of-speech (PoS) annotation (Walther & Sagot, 2010; Walther et al., 2010), dialectology (Hassani & Medjedovic, 2016; Malmasi, 2016), creating dependency treebanks (Gökırmak & Tyers, 2017), and intralanguage and interlanguage machine translation (Hassani, 2018; Ahmadi, 2019).

¹ <http://kurdishacademy.org/?p=111>

Kurdish phonemes (IPA)	Our suggestion	Latin-based	Yekgirtú	Persian-Arabic-based			
				Initial	Middle	Final	Single
[a:]	A a	A a	A a	ا	ا	ا	ا
[b]	B b	B b	B b	ب	ب	ب	ب
[t̪]	Ç ç	Ç ç	C c	چ	چ	چ	چ
[d̪]	C c	C c	J j	ج	ج	ج	ج
[d]	D d	D d	D d	د	د	د	د
[æ]	E e	E e	E e	ه	ه	ه	ه
[e:]	Ê ê	Ê ê	É é	ئ	ئ	ئ	ئ
[f]	F f	F f	F f	ف	ف	ف	ف
[g]	G g	G g	G g	گ	گ	گ	گ
[h]	H h	H h	H h	ه	ه	ه	ه
[I]	I i	I i	I i				
[i:]	Î î	Î î	Í í	ئ	ئ	ئ	ئ
[ʒ]	J j	J j	Jh jh	ژ	ژ	ژ	ژ
[k]	K k	K k	K k	ک	ک	ک	ک
[l]	L l	L l	L l	ل	ل	ل	ل
[ʎ]	Ł ł	Ll ll	Ll ll	ل	ل	ل	ل
[m]	M m	M m	M m	م	م	م	م
[n]	N n	N n	N n	ن	ن	ن	ن
[o:]	O o	O o	O o	و	و	و	و
[p]	P p	P p	P p	پ	پ	پ	پ
[q]	Q q	Q q	Q q	ق	ق	ق	ق
[r]	R r	R r	R r	ر	ر	ر	ر
[r̥]	Ř ř	Rr rr	Rr rr	ر	ر	ر	ر
[s]	S s	S s	S s	س	س	س	س
[ʃ]	Ş ş	Ş ş	Sh sh	ش	ش	ش	ش
[t̪]	T t	T t	T t	ت	ت	ت	ت
[ʊ]	U u	U u	U u	و	و	و	و
[u:]	Û û	Û û	Ú ú	وو	وو	وو	وو
[v]	V v	V v	V v	ف	ف	ف	ف
[w]	W w	W w	W w	و	و	و	و
[x]	X x	X x	X x	خ	خ	خ	خ
[j]	Y y	Y y	Y y	ی	ی	ی	ی
[z]	Z z	Z z	Z z	ز	ز	ز	ز
[h̥]	Ĥ ĥ		H', h'	ح	ح	ح	ح
[ç]	Ĭ ĭ		'	ع	ع	ع	ع
[ɣ]	Ǧ ǧ		X', x'	غ	غ	غ	غ
[ʁ:]	Û û		Û ù	ق	ق	ق	ق
[ɣ]	Ď ě			ڤ	ڤ	ڤ	ڤ
[ʔ]	'			ئ	ئ	ئ	ئ
[ʁ]	Ǧ ǧ						

Figure 1: A comparison of the alphabets used for Kurdish writing.
A unified script for all dialects is suggested.

Kurdish NLP is a young sector in the realm of worldwide NLP. Particularly, to be able to prepare the underlying resources to leverage its language processing capacity, it needs a wide range of tools such as Optical Character Recognition (OCR), thesauri, treebanks, machine-readable lexicons, a variety of language models, and transliterators

for its various scripts, to name a few. However, currently, most of these tools either do not exist or they are in their infancy. The situation and the requirements have been addressed by several researchers (Hassani, 2018; Yaseen & Hassani, 2018; Ahmadi, 2019). The current research is an attempt to improve this situation.

3. Kurdish Lexicography

Since poems have historically had a special place in Kurdish literature, the earliest works in Kurdish lexical studies were in verse. *Nûbihara Biçûkan* (*The Kids' Spring*) which dates back to 1683, is considered the first Kurdish dictionary and the first Kurdish work in children's literature (Yıldırım, 2008). This resource contains 1,000 Kurdish-Arabic pairs which were taught for years at Kurdish elementary schools to teach Arabic for Koranic studies (Hassanpour, 1992). Poetic resources have been historically used among the Kurds for educational purposes as the translations are provided in rhythm. Bolelli and Ertekin (2017) count eight poetic resources for various Kurdish dialects, which mostly provide Arabic translations. Recently, Ertekin (2017) presented a Turkish-Kurmanji dictionary in verse. The following is an example from the Nodeyî (1936) which was created according to Yıldırım (2008) in verse in Sorani Kurdish:

(أَيْنَ) لَهكۆی (مَنْ) كَییه (أَيَّانَ) كَهی (أَيْنَ) eger (مَا) قِ (مَتَى) key (سَمَ) jar	(أَيْنَ) lekwe (مَنْ) kê ye (أَيَّانَ) key (أَيْنَ) if (مَا) what (مَتَى) when (سَمَ) poison
---	---

Figure 2: A couplet from (Nodeyî, 1936) Arabic-Sorani Kurdish work (original on the left, transliterated to Latin in the middle, translated into English on the right). Kurdish words appear in parentheses immediately after the source words in Arabic.

Despite the historical popularity of poetic resources in traditional Kurdish schools, such resources can hardly be categorized as dictionaries due to the superficial representation of lexical information and the poetic structure. Moreover, these resources cannot be consulted, and therefore it is impossible to systematically retrieve data from them.

In this section, we describe some of the existing lexicographic resources for Kurdish which have played an essential role in forming Kurdish lexicography. A complete list of Kurdish lexicographic resources is provided in Appendix A. We have not considered word lists and glossaries which appear as part of other works in linguistics and literature (e.g. MacKenzie, 1966; Kahn, 1974; Cano & Şêrgo, 1991; Paul, 1998; Thackston, 2006a,b). The list in Appendix A also presents various characteristics of the dictionaries, such as target dialects, script, and entry description.

3.1 Before the 20th century

Three major lexicographic resources were published before the 20th century:

Garzoni's Kurdish Grammar and Vocabulary Book (Garzoni, 1787). This dictionary is a part of the earliest scientific European studies on the Kurdish language and civilization which dates back to the late 18th century. The research carried out by various Christian missionaries (Yarshater, 1982). (Garzoni, 1787) collected materials for his *Grammatica e vocabolario della lingua Kurda* (*Grammar and Vocabulary for the Kurdish Language*) (Garzoni, 1787) in Amedi (Amadyia), which is now located in the Kurdistan region of Iraq. This book is an Italian-Kurmanji dictionary and grammar guide which was written to enable missionaries to converse with Kurmanji speakers.

Jaba's Kurmanji Kurdish-French Dictionary (Jaba, 1879). This dictionary presents its entries in both Arabic (Ottoman Turkish script) and Latin orthographies. The latter is used for phonological purposes and therefore can be considered as the pronunciation of the entry. Although definitions and etymological information are mostly provided alongside the entries, the PoS and the gender of the nouns are less frequently present in the dictionary.

Maqdisi's Kurmanji Kurdish-Arabic Dictionary (Mokri, 1987). This dictionary was published in 1892 based on the dialect of Bitlis, now located in Turkey, by a Palestinian Arab Ottoman official. Although neither the PoS nor the gender of nouns is indicated, the present stem of verbs is regularly included. Another version of the dictionary was published with Turkish translations rather than the original Arabic in 1978 (Paşa & Bozarslan, 1978).

3.2 After the 20th century

Kurdish lexicography flourished in the 20th century through the efforts of Kurdish native scholars and orientalists, particularly by the researchers of the former Soviet Union (Leezenberg et al., 2011). This section describes a number of these dictionaries under bilingual, monolingual, and multilingual categories which were published during the mentioned period. These dictionaries are selected based on their contribution significance to Kurdish lexicography.

3.2.1 Bilingual dictionaries

Bakaev's Kurmanji Kurdish-Russian Dictionary (Bakaev, 1957). This dictionary was one of the first linguistic works in the former Soviet Union. The author was a native Russian speaker whose mother tongue was Kurdish. The combination of the author's philological background and his practical knowledge of Kurdish enabled him to produce a standard dictionary. The vocabulary is Kurmanji based on the language of the Kurdish community in the former Armenian Soviet Socialist Republic (SSR) and the former Georgian SSR.

Bakaev collected the dictionary data from various sources, such as folklore texts published mainly during era of the former Soviet Union, the works of the folklorists affiliated with the institutes of the Academy of Sciences of the Armenian SSR and Yerevan State University, the literary work translated into or originally written in Kurdish published in Armenian SSR, and translated textbooks from Russian and Armenian into Kurdish. This explains the presence of many words which were not common in Kurdish daily life (Chyet, 1998).

Kurdoev's Kurmanji Kurdish-Russian Dictionary (Kurdoev, 1960). The author of the dictionary set himself the task to most fully reflect the vocabulary fund of the modern Kurdish language. The vocabulary includes household, agricultural and modern literary language and the press. The dictionary is based on the vocabulary used in a Kurmanji speaking area in Soran, which is currently located in the Kurdistan Region of Iraq. Although the dictionary presents a more diverse vocabulary in comparison to Bakaev's work, the reliability of its data and also its scientific approach have been questioned by some scholars (Chyet, 1998).

Wahbi and Edmonds's Sorani Kurdish-English Dictionary (Wahby & Edmonds, 1966). This dictionary comprises the lexical material of the "standard language of belles-lettres, journalism, official and private correspondence and formal speech as it has been developed, on the basis of the Southern-Kurmanji dialect of Sulaymaniyah in Iraq since 1918" (Mokri, 1987). Moreover, the dictionary contains words unique to the sub-dialects spoken in Erbil, Kirkuk, and Sanandaj. The dictionary does not provide bibliographic information about its lexicographic resources. However, according to Bodrogligeti (1967), Sheikh Mihammadi Khal's *Ferhenî Xal by Xal* (1960), work by MacKenzie (1961, 1962, 1966), and McCarus (1958) perhaps contributed to the compilation of this dictionary.

Kurdoev and Yusupova's Sorani Kurdish-Russian Dictionary (Kurdoev & Yusupova, 1983). This dictionary is the first Sorani Kurdish-Russian based the Sulaimani sub-dialect of Sorani. The authors compiled the dictionary based on the translations of the entries of dictionaries by Kurdoev (1960) and Mukryani (1950). The information provided for the entries in this dictionary includes pronunciation, PoS, idioms, and expressions.

Chyet's Kurdish-English Dictionary (Chyet & Schwartz, 2003). Chyet's dictionary is a seminal work in Kurdish lexicography containing all the main Kurdish dialects. The entries in Kurmanji Kurdish are in Latin and Arabic orthographies, followed by the PoS, numbered definitions in English, synonyms and variant forms. Moreover, the dictionary contains etymological and linguistic remarks along with expressions and examples with translations in English. Interestingly, relevant forms of a word in Early, Middle and Modern Iranian followed by Sorani, Zaza and Gorani-Hawrami equivalents are provided. Chyet used several dictionaries to compile this resource.

Hakem's Sorani Kurdish-French Dictionary (Hakem, 2012). This dictionary contains

around 22,000 entries, 3,000 variants corresponding to entries, nearly 2,000 sub-entries (compound verbs) and more than 1,000 expressions. There are radicals of each simple verb or that of the compounds when the simple verb is no longer used in the spoken or written language. The grammatical category of each entry is indicated, as well as the language level whenever it seems necessary. The entries are written in Arabic characters and in Latin transcription. In some cases, expressions are also provided for entries. This dictionary focuses on contemporary language in the different registers of writing and speaking, both in the Kurdistan of Iraq and the Kurdistan of Iran.

University of Kurdistan Dictionaries (M. Rohani, 2012, 2018). These two dictionaries were compiled at the University of Kurdistan in Sanandaj based on *Henbane Borîne* (Sharafkandi, 1991) with enriched details added such as pronunciation, etymology, definition, synonyms, translations and variant forms. Moreover, they include neologisms for technical terms. M. Rohani (2012) addressed all Kurdish dialects, which makes it distinctive among bilingual dictionaries.

3.2.2 Monolingual dictionaries

Bedirxan and Keskin (2009) published the first Kurdish-Kurdish dictionary in Kurmanji and later, the *Xal Dictionary* (Xal, 1960) was published as the first Sorani Kurdish monolingual dictionary. In addition, there have been efforts to create dictionaries within Kurdish dialects, such as Habiballah's (Bedar) (2010) dictionary in Hawrami with Sorani translations, Izadpanah's (1978) dictionary in two Southern Kurdish dialects, Laki and Lori, with Sorani translations and Sohrabi and Sreshabadi's (2012) dictionary of the Garusi sub-dialect of Southern Kurdish with Sorani translations. Other monolingual dictionaries which are mostly in Kurmanji Kurdish are Botî (2006), Demîrhan (2007) and Mukryani (2007).

3.2.3 Multilingual dictionaries

Blau's Kurmanji Kurdish-English-French dictionary (Blau, 1965) is the first multilingual Kurdish dictionary which was created based on newspaper articles published in the 1930's and 1940's based on Kurmanji journals. However, the English translations provided in this resource have been questioned by scholars (Chyet, 1998; M., 1966). Several years later, the author published a book consisting of a linguistic analysis, Sorani glossaries and folkloric texts with French translations (Blau, 1975). The two glossaries contain richer descriptions, including gender, part-of-speech, present stem of verbs and oral example texts.

Henbane Borîne (Sharafkandi, 1991) is a Kurdish-Persian dictionary that incorporated all Kurdish dialects in its compilation. In addition to the Persian translations, this resource provides synonyms and senses in Kurdish, including all dialects. Therefore, it sets a foundation for the unification of the dialects. Many dictionaries, which have been compiled following *Henbane Borîne*, have referred to it as one of their essential resources.

3.3 Terminological resources

Various terminological resources exist in Kurdish, such as a glossary of the names of animals (Justi, 1878), glossary of plants (Kasimoğlu, 2013), glossary of law (Talbani, 2006), and engineering (Soğancı, 2014). A valuable terminological resource for Kurdish is *Kurmancî*, which is a biannual linguistic magazine published by the Kurdish Institute of Paris since 1987. The aim of the magazine is to spread the results of the Institute’s linguistic seminars on problems of terminology and standardization of the Kurdish language. The periodical contains headwords in Kurmanji Kurdish and translations in French, English, and Turkish. A database containing the words published in all issues of this periodical is available online².

3.4 Electronic dictionaries

Gautier (1996) was the pioneer in creating the first electronic dictionaries for Kurdish in his *Dirêjî Kurdî* (Kurdish Dimension) project. This project aimed at developing a lexicographic software environment specifically for Kurdish to deal with various early-age technical issues such as character representation. FreeDict, as a project which provides open-source bilingual dictionaries for most languages, also has dictionaries in Kurmanji to Turkish, German and English, as well as Sorani to Kurmanji. The dictionaries are publicly available in the TEI XML format. However, the sources of the resources are not clear in all cases. Moreover, there are a few collaboratively created dictionaries, or Wiktionaries³(available for Kurmanji and Sorani), which provide electronic content. There have been a few efforts in creating electronic lexicons on the Web based on a printed dictionary, but as none of them are documented, we cannot cite them here.

4. Methodology

In order to create our dictionaries, we follow the pipeline illustrated in Figure 3:

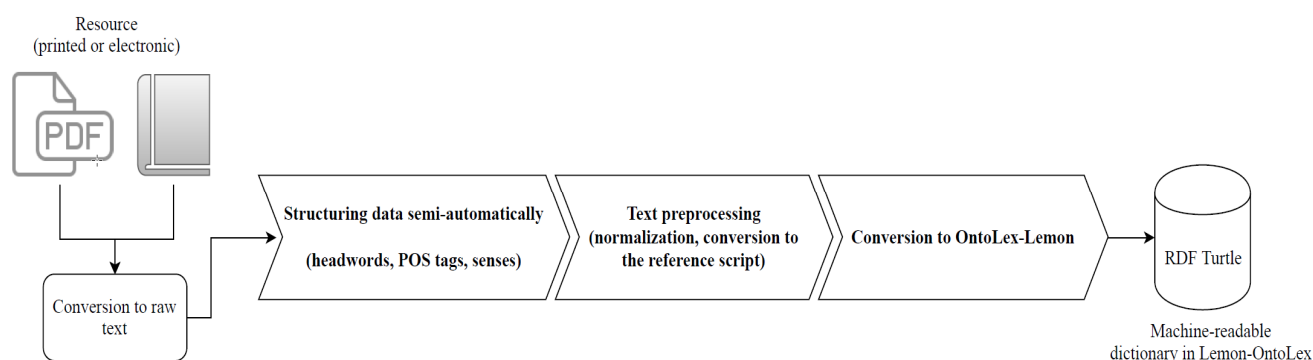


Figure 3: Our resource creation pipeline for creating dictionaries in OntoLex-Lemon from PDF documents.

² <https://www.institutkurde.org/en/publications/kurmanci/>

³ <https://www.wiktionary.org/>

4.1 Data collection

As a wide range of published Kurdish dictionaries are available, we selected three dictionaries for our experiment following three selection criteria: i) the number of entries to be manageable in a research project, ii) the availability of the resource and iii) the copyright situation of the resource. Eventually, we selected the word lists provided in three grammar books of Kurmanji (Thackston, 2006a), Sorani (Thackston, 2006b) and Hawrami (MacKenzie, 1966). In addition to reliability, these resources provide a workable sample of a few thousand frequent entries in those dialects. We were not able to find a similar resource for the Kirmashani (Southern Kurdish) dialect.

The Kurmanji and Sorani word lists were available in searchable Portable Document Format (PDF), hence we extracted information into an unstructured text semi-automatically. Because this semi-automatic extraction created some noise, improper transformation to text and misplaced portions of texts, we manually cleaned the text by removing noise and recreating the micro- and macro-structure of the lexicon using tabulations. In the case of the Hawrami lexicon, we had to re-type the word list manually as only the printed book was available.

Moreover, we modified a few traditional lexicographic norms in the resources, such as replacing ~ by headword and placing relevant lexemes of an entry as new entries if with different PoS or etymological roots. Figure 5, on the left, illustrates the Kurdish entry “*bend*” (bond in English) in the Kurmanji-English dictionary where “~ *kirin*” (to arrest, to fetter) and “*man di ~a*” (to wait for) are respectively replaced by “*bend kirin*” and “*man di benda*” as new entries. Similarly, we modified the English translations, particularly in cases where two synonym verbs are provided, the preposition “to” is only provided for the first verb.

Following the data extraction, we unified the orthography and the scripts of the resources. The word lists were originally written in orthographies suggested by the authors and used for teaching purposes. Having various scripts for writing in Kurdish causes a burden for the computation process (Ahmadi, 2019). Moreover, none of the current Kurdish scripts can be used for all Kurdish dialects. Therefore, we suggest a new character setup, illustrated in Figure 1, based on Latin orthography and the phonetics of the language to deal with the missing characters and to accommodate computation needs. The suggested script introduces a single character for the phonemes in all Kurdish dialects, such as ğ and ^d used in the Zaza and Hawrami dialects, respectively. As the orthographies were based on the phonetics of the language (in Latin), we could automatically transliterate the original text into our suggested orthography. We ignored the Persian-Arabic equivalent of Sorani lexicon at this stage.

4.2 Conversion to OntoLex-Lemon

In recent years there have been efforts to create specific data models providing support

for representing linguistic data on the Semantic Web. The OntoLex-Lemon (McCrae et al., 2017) is a model based on the Lexicon Model for Ontologies (lemon) which provides rich linguistic grounding for ontologies, such as representation of morphological and syntactic properties of lexical entries. This model draws heavily on previous lexical data models, particularly LexInfo (Cimiano et al., 2011), LIR (Montiel-Ponsoda et al., 2008) and LMF (Francopoulo et al., 2006), with improvements such as being RDF-native, descriptive and modular justifying its promise of adaptability in linguistic resource management. The core vocabulary of Lemon is the Ontology Lexicon (Ontolex), known as OntoLex-Lemon, which is illustrated in Figure 4.

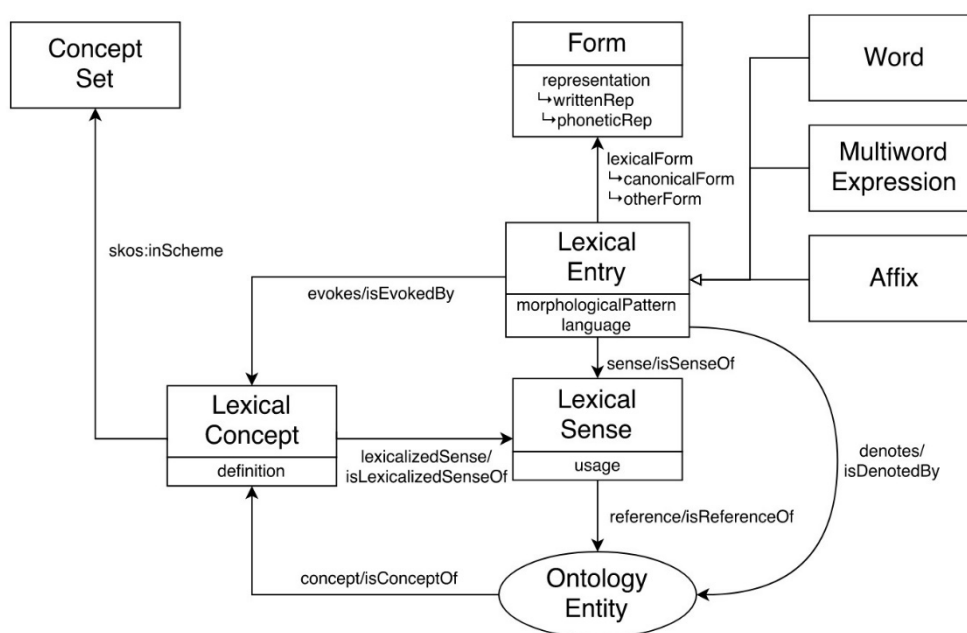


Figure 4: Lemon-OntoLex Core (McCrae et al., 2017).

The previous step yielded a tabular format of the lexicographic information, making it possible to convert the data semi-automatically into RDF triples in OntoLex-Lemon.

Figure 5 illustrates the equivalent of the entry “*bend*” in the Kurmanji-English dictionary in RDF Turtle in Ontolex-Lemon. We have used language tags according to ISO 639-3⁴, `kmr` for Kurmanji, `ckb` for Sorani and `hac` for Hawrami (registered as Gorani). As there are many scripts for Kurdish writing, we also include a subtag expressing script following the language tag. For instance, `kmr-latn` shows that the literal is in Kurmanji Kurdish and written in the Latin script. The script code `Arab` can be used for Arabic script as well.

⁴ https://iso639-3.sil.org/code_tables/639/

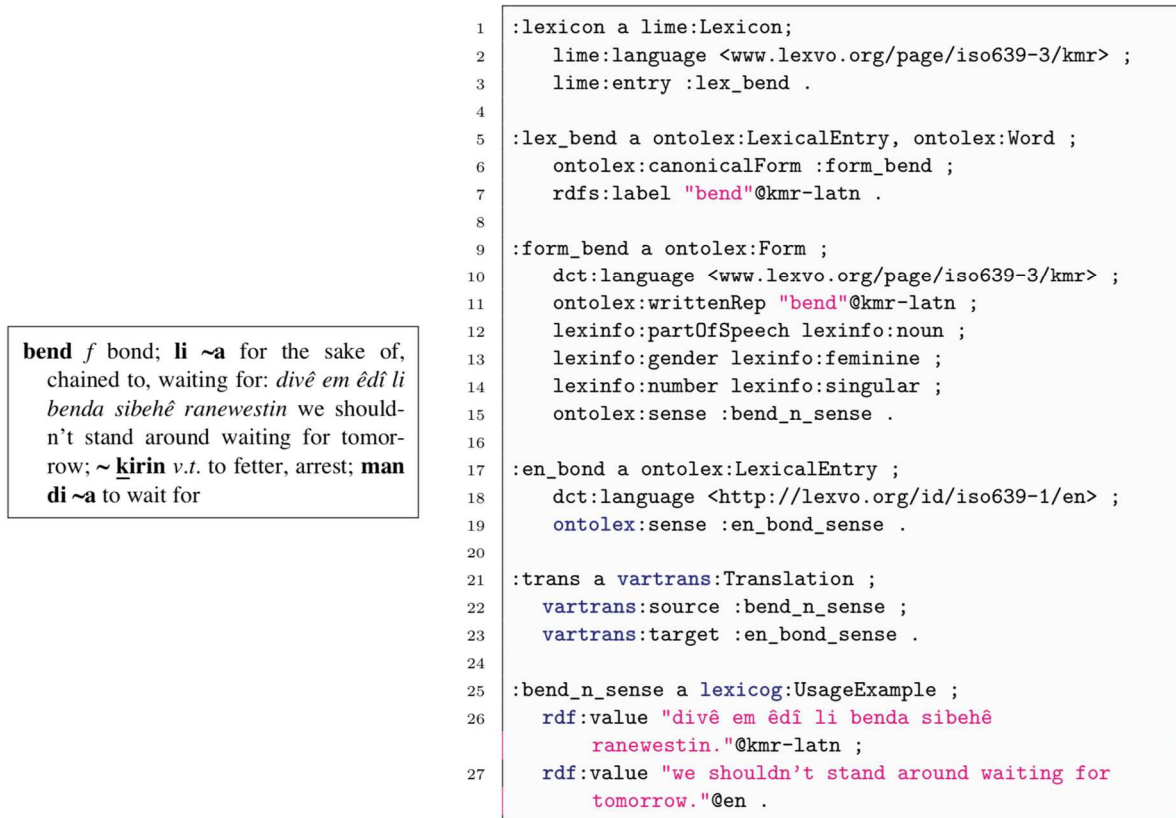


Figure 5: An example entry from our Kurmanji-English dictionary. The original printed entry on the left and the equivalent in RDF Turtle based on the OntoLex-Lemon model.

In addition to OntoLex-Lemon core, we used the following modules:

- Linguistic Metadata (lime) allows to describe metadata at the level of the lexicon-ontology interface with information such as lexical entries and language (lines 1 to 3 in Figure 5).
- Syntax and Semantics (synsem) enables us to describes syntactic behaviour. We use syntactic frames to relate a lexical entry to one of its various syntactic roles, such as the canonical form of the word *bend* described in lines 5 to 7 in Figure 5).
- Lexinfo (lexinfo) (Cimiano et al., 2011) for describing relevant linguistic categories and properties, particularly part-of-speech, gender and number (lines 9 to 15 in Figure 5).
- Variation and translation (vartrans) is used to describe relations between lexical entries, particularly translations. As our resources are not currently connected to any external English resource, we also create entries for English words as shown in lines 17 to 19 in Figure 5.
- Lexicography module (lexicog) (Bosque-Gil et al., 2017) represents information,

structures and annotations commonly found in lexicography. The Lexicographic Resource class in this module is used to represent the original printed entry structures. In addition, we used the UsageExample class for representing examples of the usage of a sense (lines 25 to 27 in Figure 5).

Multi-word expressions (MWEs) are lexical units which are semantically unique, greater than a word, and can bear both idiomatic and compositional meanings (Masini, 2005). Therefore, we create new entries for MWEs using `ontolex:MultiwordExpression`. Regarding Kurdish MWEs, we could not find any writing standard. In both orthographies, Persian-Arabic-based and Latin-based, words in MWEs are written either with spaces or without. For instance, “*toz-û-telaz*” (dust) can be found as “*tozûtelaz*”, “*tozwtelaz*” or “*toz û telaz*” in the literature. Hence, we followed the English norm of using hyphens, i.e. -, for Kurdish MWEs. Furthermore, regarding the idioms, we create new entries as they are semantically different from the canonical forms as well.

5. Analysis

Table 1 provides a statistical analysis of various characteristics of our lexicographic resources. # Entries refers to the number of entries in the electronic dictionary. Put in other terms, it refers to the number of triples with `lime:entry` properties. This feature does not have the same value as the printed original resources, as idioms and MWEs are presented as new entries in the electronic resources while they are presented in the description of the entries in the printed resources. Furthermore, statistics regarding attributes such as gender, PoS tag, etymological roots, example sentences and idioms are provided. The Sorani and Hawrami dictionaries have the highest number of Gender and PoS tags and etymological roots, respectively.

Resource	Number of entries		Attributes				Polysemy degree
	Word	MWE	Gender & POS	Etymology	# idioms	Examples	
Kurmanji	4172	122	3420 (76.64%)	213 (4.96%)	340	265 (6.35%)	1.03%
Sorani	5683	160	5348 (91.37%)	111 (1.89%)	82	543 (9.55%)	1.06%
Hawrami	1184	165	1184 (87.76%)	242 (17.93%)	123	10 (0.008%)	1.01%

Table 1: Lexicographic resources statistics

We define polysemy degree as the number of unique senses divided by the number of entries. This measure varies in the range of 1.01% and 1.06%, indicating that a small proportion of less than 1% of the entries are polysemous, and for the rest there is only one sense available.

6. Conclusion and future work

In this paper, we provided a review on the current state of Kurdish lexicography and described the development of dictionaries for three out of five main dialects of Kurdish, namely Sorani, Kurmanji and Hawrami. Having more than 60 printed dictionaries and terminological resources, we demonstrate that Kurdish is fairly rich in printed resources, although this is not the case with respect to electronic and machine-readable resources. The lack of such resources makes Kurdish a less-resourced language.

Our lexicographic resources are created using the word lists provided in three grammar books of Kurmanji (Thackston, 2006a), Sorani (Thackston, 2006b) and Hawrami (MacKenzie, 1966) and according to the OntoLex-Lemon model. As Kurdish is written in more than one script and some of the dialectal phonemes do not have a character in those scripts, we suggest a few characters based on the Latin script which can lead to a unification of the scripts. The resources are publicly available for non-commercial use under the CC BY-NC-SA 4.0 license ⁵ at <https://github.com/KurdishBLARK/KurdishLex>.

The current study aims at paving the way for Kurdish e-lexicography by developing prototypical resources. Enriching our dictionaries using additional resources and scripts and, linking the dictionaries across dialects and resources, such as KurdNet (Aliabadi et al., 2014), may be addressed in the future work. Creating specific standards for Kurdish, particularly regarding the scripts, will also be suggested as future work. We would also like to highlight solutions to tackle some of the current challenges in Kurdish lexicography such as the following:

- Lexicographic infrastructure: as our findings suggest, more than half of Kurdish dictionaries were created before 2000. In order to create machine-readable version of these resources, retrodigitization tools, such as Optical Character Recognition, are required. On the other hand, tools for creating and maintaining dictionaries are needed.
- Raising awareness: we believe that the lexicography community should be aware of the current computer-based solutions for creating resources and collecting data.
- Creating basic Kurdish text processing tools such as lemmatizer, spell-checker (Salavati & Ahmadi, 2017) and name entity recognizer.
- Copyright issues: the majority of the dictionaries cited in this paper were available online in scanned version or searchable PDF. This is against the copyrights and creator licences, which leads to discouragement in the lexicographers' community.

⁵ <https://creativecommons.org/licenses/by-nc-sa/4.0/>

7. Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

8. References

- Abdollahpour, H. (2008). *Ferhengê Hejîr: Farsî-Kurdî (Hejîr dictionary: Persian-Kurdish) (Sorani)*, volume 2. Mukryani Publishing House.
- Ahmadi, S. (2019). A Rule-based Kurdish Text Transliteration System. *Asian and LowResource Language Information Processing (TALLIP)*, 18(2), pp. 18:1–18:8.
- Aliabadi, P., Ahmadi, M. S., Salavati, S. & Esmaili, K. S. (2014). Towards building Kurdnet, the Kurdish Wordnet. In *Proceedings of the Seventh Global Wordnet Conference*. pp. 1–6.
- Amin, P. (2003). *Yad dictionary: English-Arabic-Kurdish*. Erbil.
- Anter, M. (1967). *Ferhenga Kurdî-Tirkî (Kurdish-Turkish dictionary)*. Istanbul: Yeni Matbaa.
- Arif, H. K. (2006). *Govend-Zinar: Ferhengê Farsî-Kurdî (Govend-Zinar Persian-Kurdish Dictionary) (Sorani)*, volume 2. Mukryani Publishing House.
- Ataman, D. (2018). Bianet: A Parallel News Corpus in Turkish, Kurdish and English. *arXiv preprint arXiv:1805.05095*.
- Bakaev, C. K. (1957). *Kurdish-Russian dictionary*. State Publishing House of Foreign and National Dictionaries, Moscow.
- Bedirxan, C. A. & Keskin, A. (2009). *Ferheng: Kurdî, Kurdî (Kurdish-Kurdish dictionary) (Kurmanji)*, volume 2. Avesta.
- Bedirxan, C. A. & Lescot, R. (1970). *Grammaire kurde (Kurdish Grammar)*. Librairie d'Amérique et d'Orient.
- Berners-Lee, T., Hendler, J., Lassila, O. et al. (2001). The Semantic Web. *Scientific american*, 284(5), pp. 28–37.
- Biemann, C., Heyer, G., Quasthoff, U. & Richter, M. (2007). The Leipzig Corpora Collection-monolingual corpora of standard size. *Proceedings of Corpus Linguistic*, 2007.
- Blau, J. (1965). *Dictionnaire kurde-français-anglais (Kurdish-French-English dictionary)*, volume 9. Centre pour l'étude des problèmes du monde musulman contemporain.
- Blau, J. (1975). *Le Kurde de Amadiya et de Djabal Sindjar: analyse linguistique, textes folkloriques, glossaires*. Librairie C. Klincksieck.
- Bodrogligeti, A. (1967). A Kurdish-English dictionary. By Taufiq Wahby and C. J. Edmonds, pp. xii, 179. Oxford, Clarendon Press, 1966. *Journal of the Royal Asiatic Society of Great Britain & Ireland*, 99(2), pp. 152–155.
- Bolelli, N. & Ertekin, N. (2017). Ferhengên Menzûm Di Edebîyata Kurdî De. *Bingöl Üniversitesi Yaşayan Diller Enstitüsü Dergisi*, 3(5), pp. 21–44. (in Kurdish).
- Bosque-Gil, J., Gracia, J. & Montiel-Ponsoda, E. (2017). Towards a module for

- lexicography in OntoLex. *DICTIONARY News*, 7.
- Botî, K. (2006). *Ferhenga Kamêran: Kurdî-Kurdî (Kamêran dictionary: Kurdish-Kurdish) (Kurmanji)*. Spîrêz Press & Publisher. <https://books.google.ie/books?id=AVErAQAAIAAJ>.
- Cano, D. & Şêrgo, M. (1991). *Ferheng Erebi-Kurdî Zaraveyê Kurdmancî (Arabic-Kurdish dictionary) (Kurmanji)*, volume 2. Beirut, Lebanon.
- Chiarcos, C., McCrae, J., Cimiano, P. & Fellbaum, C. (2013). Towards open data for linguistics: Linguistic linked data. In *New Trends of Research in Ontologies and Lexical Resources*. Springer, pp. 7–25.
- Chyet, M. (1998). Kurdish Lexicography a Survey and Discussion. *Iran and the Caucasus*, 2(1), pp. 109–118.
- Chyet, M. L. & Schwartz, M. (2003). *Kurdish-English Dictionary (Kurmanji)*. Yale University Press.
- Cimiano, P., Buitelaar, P., McCrae, J. & Sintek, M. (2011). LexInfo: A declarative model for the lexicon-ontology interface. *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(1), pp. 29–51.
- Darvishian, A. (1997). *Kermanshahi Kurdish Dictionary*. Sahand Publication House.
- Demîrhan, U. (2007). *Ferhenga Destî-Kurdî bi Kûrdî (Kurdish-Kurdish pocket dictionary)*.
- Ebrahimpour, T. (2008). *Ferhengê Kurdî-Îngilîsî (Kurdish-English Dictionary) (Sorani)*. Saha Publications, Tehran.
- Ertekin, Z. (2017). İlk Manzum Türkçe-Kürtçe Sözlük: Nûbihara Mezinan. *e-Şarkiyat İlmî Araştırmaları Dergisi/Journal of Oriental Scientific Research (JOSR)*, 9(1), pp. 89–105.
- Esmaili, K. S., Eliassi, D., Salavati, S., Aliabadi, P., Mohammadi, A., Yosefi, S. & Hakimi, S. (2013). Building a test collection for Sorani Kurdish. In *Computer Systems and Applications (AICCSA), 2013 ACS International Conference on*. IEEE, pp. 1–7.
- Farizov, I. (1957). *Russian-Kurdish dictionary (Kurmanji)*. State Publishing House of Foreign and National Dictionaries, Moscow.
- Forcada, M. L., Esplà-Gomis, M., Pérez-Ortiz, J. A., Sánchez-Cartagena, V. M. & SánchezMartínez, F. (2019). D1.1 – Survey of relevant low-resource languages. Technical report, Global Under-Resourced MEdia Translation (GoURMET).
- Francopoulo, G., George, M., Calzolari, N., Monachini, M., Bel, N., Pet, M. & Soria, C. (2006). Lexical markup framework (LMF). In *International Conference on Language Resources and Evaluation-LREC 2006*, p. 5.
- Garzoni, M. (1787). *Grammatica e vocabolario della lingua kurda composti dal p. Maurizio Garzoni de'predicatori ex-missionario apostolico*. nella Stamperia della Sacra Congregazione di Propaganda Fide.
- Gautier, G. (1996). Dirêjî Kurdî: a lexicographic environment for Kurdish language using 4th Dimension R . In *5th International Conference and Exhibition on Multilingual Computing (ICEMCO)*, volume 5. pp. Session-of.
- Gautier, G. (1998). Building a Kurdish Language Corpus: An Overview of the

- Technical Problems. *Proceedings of ICEMCO*.
- Gewranî, A. S. A. (1985). *Ferhenga Kurdî Nûjen: Kurdî-Erebî (Nûjen Kurdish dictionary: Kurdish-Arabic) (Kurmanji)*. Amman: A.S.A.
<https://books.google.ie/books?id=kJncNQAACAAJ>.
- Gharib, K. (1975). *Arabic-English-Kurdish Dictionary*. Alajial, Baghdad.
- Gökirmak, M. & Tyers, F. M. (2017). A dependency treebank for Kurmanji Kurdish. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pp. 64–72.
- Habiballah (Bedar), J. (2010). *Wişename (Lexicon)*. Aras Publishing and Printing House.
- Haig, G. & Matras, Y. (2002). Kurdish linguistics: a brief overview. *STUF-Language Typology and Universals*, 55(1), pp. 3–14.
- Hakem, H. (2012). *Dictionnaire kurde-français: Sorani (Kurdish-French dictionary: Sorani)*. L'Asiathèque, Maison des langues du monde.
- Hassani, H. & Medjedovic, D. (2016). Automatic Kurdish dialects identification. *Computer Science & Information Technology*, 6(2), pp. 61–78.
- Hassani, H. (2018). BLARK for multi-dialect languages: towards the Kurdish BLARK. *Language Resources and Evaluation*, 52(2), pp. 625–644.
- Hassanpour, A. (1992). *Nationalism and language in Kurdistan, 1918-1985*. Edwin Mellen Press.
- Ismail Hassan, S. (2019). *Kurdish-English Dictionary*. Tafseer Office.
- Izadpanah, H. (1978). *Lak and Lor dictionary*. Kurdish Scientific Council, Baghdad.
- Izoli, D. (1992). *Ferheng: Kurdî-Tirki, Türkçe-Kürtçe (Dictionary: Kurdish-Turkish, Turkish-Kurdish) (Kurmanji)*. Istanbul: Deng Yayinlari.
- Jaba, A. (1879). *Dictionnaire kurde-français (Kurdish-French dictionary) (Kurmanji)*. Commissionaire de l'Académie Impériale des Sciences.
- Jalilian, A. (2010). *Ferhengî Başûr (Başûr Dictionary)*. Aras Publishing and Printing House.
- Jalilian, A. A. (2009). *Ferhengî Başûr (Başûr Dictionary) (Southern Kurdish)*. Enstîtûy Kelepûrî Kurdî.
- Justi, F. (1878). *Les noms d'animaux en kurde (name of animals in Kurdish) (Kurmanji)*. Imprimerie nationale.
- Kahn, M. (1974). *Kurmanji-English, English-Kurmanji Lexicon*. Ann Arbor: The University of Michigan. Unpublished manuscript.
- Karadaghi, R. (2006). *The Azadi: English-Kurdish Dictionary*. Ehsan Publication House.
- Kasimoğlu, A. (2013). *Ferhenga Naven Nebatan A Kurdi (Dictionary of Plants in Kurdish, Turkish and Latin) (Kurmanji)*. İmaj Matbaacılık Sanayi.
<https://books.google.ie/books?id=4p0JugEACAAJ>.
- Keidane, K., Mukriani, K. & Mitrokhina, V. (1977). *Educational Russian-Kurdish Dictionary*. Moscow, Publisher "Russian Language".
- Khalidgul, M. (2002). *Ferhenga Gulî: Farsî-Kurdî (Gulî dictionary: Persian-Kurdish) (Kurmanji)*. Spîrêz Press & Publisher.

- Kiani Kolivand, K. (2011). *Kian dictionary: Laki lexicon (in Persian)*, volume 2. Sifa, Khorramabad.
- Klyne, G. & Carroll, J. (2004). Resource Description Framework (RDF): Concepts and Abstract Syntax. <https://www.w3.org/TR/rdf-concepts/>.
- Kreyenbroek, P.G. (2005). Kurdish written literature. *Encyclopædia Iranica*, p. 2.
- Kurdoev, K. K. & Yusupova, Z. (1983). *Kurdish-Russian Dictionary (Sorani)*. Moscow, Publisher "Russian Language".
- Kurdoev, K. K. (1960). *Kurdish-Russian dictionary (Kurmanji)*. State Publishing House of Foreign and National Dictionaries, Moscow.
- Leezenberg, M. et al. (2011). Soviet Kurdology and Kurdish Orientalism. *The Heritage of Soviet Oriental Studies*, pp. 86–102.
- MacKenzie, D. N. (1961). The origins of Kurdish. *Transactions of the Philological Society*, 60(1), pp. 68–86.
- MacKenzie, D. N. (1962). *Kurdish dialect: studies*, volume 2. Oxford University Press.
- MacKenzie, D. N. (1966). Joyce Blau: Kurdish-French-English dictionary. *Bulletin of the School of Oriental and African Studies*, 29(3), p. 672–673.
- MacKenzie, D. N. (1966). *The dialect of Awroman (Hawraman-i Luhon) Grammatical sketch, texts, and vocabulary*. Copenhagen: Munksgaard.
- Malmasi, S. (2016). Subdialectal differences in Sorani Kurdish. In *Proceedings of the third workshop on nlp for similar languages, varieties and dialects (vardial3)*, pp. 89–96.
- Masini, F. (2005). Multi-word expressions between syntax and the lexicon: the case of Italian verb-particle constructions. *SKY Journal of Linguistics*, 18(2005), pp. 145–173.
- Mayi, T. M. (2009). *Ferhenga Mayî: Kurdî-Erebî (Mayî Dictionary: Kurdish-Arabic) (Kurmanji)*. Spîrêz Press & Publisher.
- McCarus, E. N. (1958). *A Kurdish Grammar: descriptive analysis of the Kurdish of Sulaimaniya, Iraq*. 10. American Council of Learned Societies. <http://files.eric.ed.gov/fulltext/ED089545.pdf>.
- McCrae, J. P., Bosque-Gil, J., Gracia, J., Buitelaar, P. & Cimiano, P. (2017). The OntoLexLemon model: development and applications. In I. Kosem et al. (eds.) *Proceedings of eLex 2017 conference, Leiden, Netherlands*. Brno: Lexical Computing, pp. 19–21.
- Miller, G.A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, 38(11), pp. 39–41.
- Mokri, M. (1987). *Dictionnaire Kurde-Arabe de Dia'Ad-din Pacha Al-Khalidi*. Kanz almutun wa-al-dirasat al-madhhabiyah, al-lughawiyah wa-al-ijtima'iyah, al-hadarah al-Islamiyah, al-lisan wa-al-thaqafah al-Iraniyah. Librairie du Liban. <https://books.google.ie/books?id=JrYpDgAAQBAJ>.
- Montiel-Ponsoda, E., De Cea, G. A., Gómez-Pérez, A. & Peters, W. (2008). Modelling multilinguality in ontologies. *Coling 2008: Companion volume: Posters*, pp. 67–70.
- Mukryani, G. (1950). *Rêber dictionary: Arabic-Kurdish (Sorani)*. Erbil.

- Mukryani, G. (1961). *Mahabad dictionary: Kurdish-Arabic (Sorani)*. Erbil.
- Mukryani, G. (1966). *Pellke-zêrrîne dictionary: Kurdish-Arabic-Persian-French-English (Sorani)*. Erbil.
- Mukryani, G. (2005). *Nobere dictionary: Arabic-Kurdish dictionary for Educational Puropses (Sorani)*. Aras Publishing House, Erbil.
- Mukryani, K. (2007). *Haraşan dictionary: Kurdish-Kurdish (Sorani)*. Kurdistan region, Ministry of Culture.
- Nahid, M. (2011). *Ferhengê Nahîd: Kurdî-Kurdî-Farsî (Nahid dictionary: Kurdish-KurdishPersian) (Sorani)*. Akadîmyay Kurdî, Erbil.
- Nanwazade, A. (2005). *Ferhengê Kurdîy Herman (Herman Kurdish Dictionary) (Sorani Kurdish)*, volume 2. Mukryani Publishing House.
- Navigli, R. & Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193, pp. 217–250.
- Nawkhosh, S. (2012). *Kurdish-Arabic-English Dictionary*.
- Nizameddin, F. (2003). *Ferhengê Estêregeşe: Kurdî-Erebî (Estêregeşe dictionary: KurdishArabic) (Sorani)*. Aras, Erbil.
- Nodeyî, M. (1936). *Ahmadi dictionary: Arabic-Kurdish*. Jyan Publishing house, Sulaymaniyah.
- Özcan, M. (1997). *Zazaca-Türkçe Sözlük (Zaza-Turkish dictionary)*. Kaynak Yayınları.
- Paşa, Y. Z. & Bozarslan, M. E. (1978). *Kürtçe-Türkçe sözlük (Kurdish-Turkish dictionary)*. Çıra Yayınları.
- Paul, L. (1998). *Zazaki: Grammatik und Versuch einer Dialektologie (Grammar and dialectology experiment)*. Reichert Verlag.
- Qazzaz, S. (2000). *The Sharezoor Kurdish-English dictionary: Farhang-i Sharazur kurdiinglizi*. Aras Publishing and Printing House.
- Rizgar, B. (1993). *Kurdish-English, English-Kurdish Dictionary (Kurmanji)*. MF Onen, London.
- Rohani, M. (2012). *University of Kurdistan Dictionary: Persian-Kurdish*, volume 3. University of Kurdistan, Sanandaj Iran.
- Rohani, M. (2018). *University of Kurdistan Dictionary: Kurdish-Kurdish-Persian*, volume 4. University of Kurdistan, Sanandaj Iran.
- Saadalla, S. (1998). *Saladin's English-Kurdish Dictionary (Kurmanji)*. Sweden. (Arabic script).
- Saadalla, S. (2000). *Saladin's English-Kurdish Dictionary (Kurmanji)*. Avesta. (Latin script).
- Salavati, S. & Ahmadi, S. (2017). Building a Lemmatizer and a Spell-checker for Sorani Kurdish. *LTC'17 The 8th Language & Technology Conference*.
- Selma Abdallah, K. A. (2006). *English-Kurdish Kurdish-English Dictionary*. Star Publications (P) Ltd., India.
- Sharafkandi, A. H. (1991). *Hanbana Borina: Kurdish-Persian dictionary (Sorani)*, volume 2. Soroush, Tehran.
- Shi, L. & Mihalcea, R. (2005). Putting pieces together: Combining FrameNet, VerbNet

- and WordNet for robust semantic parsing. In *International conference on intelligent text processing and computational linguistics*. Springer, pp. 100–111.
- Siabandov, S. & Châchân, A. (1957). *Armenian-Kurdish Dictionary*. State Press of Armenia (HayPetHrat), Yerevan.
- Soğanci, M. (2014). *Ferhenga Zaravên Teknîki: Kurdî-Türkçe-English (Dictionary of Technical Terms: Kurdish-Turkish-English) (Kurmanji)*. Türkiye Mühendis ve Mimar Odaları Birliği.
- Sohrabi, R. & Sreshabadi, J. (2012). *Farhang-e Garus (Garus Dictionary) (Southern Kurdish)*. University of Kurdistan.
- Talbani, N. (2006). *Legal dictionary: Arabic-Kurdish-French-English (Sorani)*. Hoshiyari Publication House.
- Thackston, W. M. (2006a). *Kurmanji Kurdish:-A Reference Grammar with Selected Readings*. Harvard University.
- Thackston, W. M. (2006b). *Sorani Kurdish-A Reference Grammar with Selected Readings*. Harvard University.
- Turgut, H. (2001). *Zazaca-türkçe sözlük (Zaza-Turkish dictionary)*. Wêjiayîşê Tiji-Tij Yayınları.
- Turgut, H. (2008). *Türkçe-Zazaca Sözlük (Turkish-Zaza dictionary)*. Do Yayınları.
- Ulumaskan, A. (2016). *Ferheng - Wörterbuch: Kurdî - Almanî, Kurdisch - Deutsch & Deutsch - Kurdisch, Almanî - Kurdî (Kurdish-German, German-Kurdish dictionary) (Kurmanji)*. Mezopotamien Verlag und Vertrieb GmbH. <https://books.google.ie/books?id=LK9YnQAACAAJ>.
- Wahby, T. & Edmonds, C. J. (1966). *A Kurdish-English Dictionary*. Clarendon Press.
- Walther, G. & Sagot, B. (2010). Developing a large-scale lexicon for a less-resourced language: General methodology and preliminary experiments on Sorani Kurdish. In *Proceedings of the 7th SaLTMiL Workshop on Creation and use of basic lexical resources for less-resourced languages (LREC 2010 Workshop)*, p. 8.
- Walther, G., Sagot, B. & Fort, K. (2010). Fast Development of Basic NLP Tools: Towards a Lexicon and a POS Tagger for Kurmanji Kurdish. In *International conference on lexis and grammar*, p. 0.
- Xal, M. (1960). *Ferhengî Xall (Xall dictionary) (Sorani)*. Aras Publishing House, Erbil.
- Yarshater, E. (1982). *Encyclopaedia Iranica*, volume 2. Routledge & Kegan Paul.
- Yaseen, R. & Hassani, H. (2018). Kurdish Optical Character Recognition. *UKH Journal of Science and Engineering*, 2(1), pp. 18–27.
- Yıldırım, K. (2008). *Nubihara Biçûkan, Arapça-Kürtçe-Arapça Sözlük*. Avesta publishing house.
- Zîlan, R. (1989). *Ferhenga Swêdî-Kurdi (Kurmancî) (Swedish-Kurdish dictionary) (Kurmanji)*. SIL, Statens Institut för Läromedel. <https://books.google.ie/books?id=WHX2AQAACAAJ>.

A. Appendix

There are reportedly more than 71 dictionaries and terminological resources available for Kurdish (Jalilian, 2010). In the following list, however, we only provide those to which we could have access to. In order to save space, we used the following symbols: † to refer to an unsystematic script which is not based on the known orthographies, ‡ to denote Ottoman Turkish Arabic script, * to show our estimation based on the number of the pages and density of entries per page. Furthermore, Sor., Kur., SK, HK are used to respectively refer to Sorani, Kurmanji, Southern Kurdish (Kirmashani) and Hawrami dialects. In the Script column, P-A (Persian-Arabic), L (Latin) and C (Cyrillic) are used to show the scripts in which the Kurdish entries or lexemes are written. → and ↔ are used to show translation directions from source language (in the left) to the target language (in the right). In cases of uncertainty, we use ?.

Table 2: The list of Kurdish dictionaries in chronological order based on which the Kurdish lexicography review in Section 3 is carried out in this paper

No	Author	Type	Year	Languages	Entries	Script	Description
1	(Garzoni, 1787)	bilingual	1787	Italian →Kur.	5,250*	L†	translation
2	(Jaba, 1879)	bilingual	1879	Kur. →French	14,340*	P-A‡ and L†	translation, example sentences
3	(Mokri, 1987)	bilingual	1892	Kur. →Arabic	7,200*	P-A‡	translations, present stem of verbs
4	(Mukryani, 1950)	bilingual	1950	Arabic →Sor.	15,000	P-A	translations
5	(Bakaev, 1957)	bilingual	1957	Kur. →Russian	14,000	C	translations, gender, expressions, variant forms
6	(Farizov, 1957)	bilingual	1957	Russian →Kur.	30,000	L	translations, gender, expressions, variant forms
7	(Siabandov & Châchân, 1957)	bilingual	1957	Armenian →Kur.	23,000	C	translation
8	(Kurdoev, 1960)	bilingual	1960	Kur. →Russian	34,000	C	detailed translations with polysemy, gender, expressions, variant forms

9	(Xal, 1960)	monolingual	1960	Sor. →Sor.	22,000*	P-A	definition, synonyms
10	(Mukryani, 1961)	bilingual	1961	Sor. →Arabic	30,000	P-A	translations
11	(Bedirxan & Keskin, 2009)	monolingual	1962	Kur. →Kur.	15,000*	L	synonyms
12	(Blau, 1965)	multilingual	1965	Kur. →(French-English)	6,000*	L	translations, gender, present stem of verbs, limited PoS
13	(Wahby & Edmonds, 1966)	bilingual	1966	Sor. →English	6,500*	L ^t , P-A	PoS, synonyms and variant forms, rich description
14	(Mukryani, 1966)	multilingual	1966	Sor. →(PersianArabic-English-French)	4,000*	P-A	translations
15	(MacKenzie, 1966)	bilingual	1966	HK →English	1,000	L _t	translation, PoS, gender, idioms, example sentences, variant forms
16	(Anter, 1967)	bilingual	1967	Kur. →Turkish	?	L	simple translations
17	(Blau, 1975)	bilingual	1975	(Kur.Sor.) →French and English	2,000*	L	translation, gender, PoS
18	(Gharib, 1975)	multilingual	1975	Arabic →Sor. and English,	1,000*	P-A and L	illustrations
19	(Keidane et al., 1977)	bilingual	1977	Sor. →Russian	2,100	P-A	translation
20	(Paşa & Bozarlan, 1978)	bilingual	1978	Kur. →Turkish	7,200*	L	simple translations, present stem of verbs
21	(Izadpanah, 1978)	multilingual	1978	(Lori- Laki) ↔ (Sor.-Persian)	4,800*	P-A	translations

22	(Kurdoev & Yusupova, 1983)	bilingual	1983	Sor. →Russian	25,000	P-A	translations, PoS, gender, expressions, variant forms
23	(Gewranî, 1985)	bilingual	1985	Kur. →Turkish	25,000*	L	translations
24	(Zîlan, 1989)	bilingual	1989	Swedish →Kur.	5,000	L	translations, synonyms, illustrations
25	(Sharafkandi, 1991)	multilingual	1991	Sor. →(Sor.-Persian)	60,000	P-A [†]	translations, synonyms
26	(Izoli, 1992)	bilingual	1992	Turkish ↔Kur.	25,000-30,000	L	translations, definitions, gender, PoS
27	(Rizgar, 1993)	bilingual	1993	Kur. ↔English	15,000	L	translations, PoS, gender, synonyms, expressions, variant forms
28	(Darvishian, 1997)	bilingual	1997	SK →Persian	?	P-A	translation, pronunciation
29	(Özcan, 1997)	bilingual	1997	Zaza →Turkish		L	
30	(Saadalla, 1998)	bilingual	1998	English →Kur.	72,000	P-A	translation, gender
31	(Qazzaz, 2000)	bilingual	2000	Sor. →English	10,000*	P-A and L	translation, synonyms, PoS, idioms, proverbs
32	(Saadalla, 2000)	bilingual	2000	English →Kur.	72,000	L	translation, PoS
33	(Turgut, 2001)	bilingual	2001	Zaza →Turkish	?	L	?
34	(Khalidgul, 2002)	bilingual	2002	Persian →Kur.	4,000	P-A	translations
35	(Chyet & Schwartz, 2003)	bilingual	2003	Kur. →English	59,360*	P-A and L	translations, PoS, gender, expressions, synonyms, variant forms, etymology, example sentence

36	(Demîrhan, 2007)	monolingual	2003	Kur. →Kur.	19,680*	L	?
37	(Nizameddin, 2003)	bilingual	2003	Sor. →Arabic	13,650*	P-A	translations, synonyms
38	(M. Amin, 2003)	multilingual	2003	English →(Sor.-P-A)	35,000*	P-A	translation
39	(Mukryani, 2005)	bilingual	2005	Arabic →Sor.	25,000	P-A	translations
40	(Nanwazade, 2005)	monolingual	2005	Sor. →Sor.	10,000*	P-A	definitions, etymology, idioms
41	(Botî, 2006)	monolingual	2006	Kur. →Kur.	15,000*	L	gender, definition, synonyms
42	(Arif, 2006)	bilingual	2006	Persian →Sor.	36,300	P-A and L	translations
43	(Selma Abdallah, 2006)	bilingual	2006	Sor. ↔English.	3,300*	P-A and L	translations
44	(Karadaghi, 2006)	bilingual	2006	English →Sor.	44,000	L and P-A	translations
45	(Mukryani, 2007)	monolingual	2007	Sor. →Sor.	3,000*	P-A	?
46	(Abdollahpour, 2008)	bilingual	2008	Persian →Sor.	28,000	P-A	translations, synonyms
47	(Ebrahimpour, 2008)	bilingual	2008	Sor. →English	40,800*	P-A and L	translation
48	(Turgut, 2008)	bilingual	2008	Turkish →Zaza	?	L	?
49	(Jalilian, 2009)	multilingual	2009	SK →(Sor.-Persian)	31,600	P-A and L	translations, example sentences
50	(Habiballah Bedar), 2010)	monolingual	2009	HK →Sor.	56,000*	P-A	synonyms
51	(Mayi, 2009)	bilingual	2009	Kur. →Arabic	20,700*	L and P-A	translation, definitions

52	(Ismail Hassan, 2019)	bilingual	2009	Sor. →English	35,000*	P-A and L	synonyms, PoS, pronunciation
53	(Kiani Kolivand, 2011)	bilingual	2011	Laki →Persian	30,000	?	translations, etymology
54	(Nahid, 2011)	multilingual	2011	Sor. →(Sor.-Persian)	21,000*	P-A	translation
55	(Hakem, 2012)	bilingual	2012	Sor. →French	?	P-A	?
56	(Nawkhosh, 2012)	bilingual	2012	Sor. →Arabic and English	3,000*	P-A and L	synonyms
57	(M. Rohani, 2012)	bilingual	2012	Persian →Sor.	18,500	P-A	translation, pronunciation, PoS, idioms, example sentences, variant forms, synonyms
58	(Sohrabi & Sreshabadi, 2012)	multilingual	2012	Garusi →(Sor.-Persian)	8,000	P-A	translation, pronunciation
59	(Ulumaskan, 2016)	bilingual	2016	Kur. ↔German	25,000	L	?
60	(M. Rohani, 2018)	multilingual	2018	Sor. →(Sor.-Persian)	93,000	P-A	translation, PoS, idioms, example sentences, variant forms

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



Introducing Kosh, a Framework for Creating and Maintaining APIs for Lexical Data

Francisco Mondaca¹, Philip Schildkamp², Felix Rau²

¹ Cologne Center for eHumanities, University of Cologne

² Data Center for the Humanities, University of Cologne

E-mail: f.mondaca@uni-koeln.de, philip.schildkamp@uni-koeln.de, f.rau@uni-koeln.de

Abstract

In recent years, the use of application programming interfaces (APIs) throughout the Internet has increased significantly. The main reason for this growth is the multiplicity of scenarios where APIs can be employed. In the case of APIs for lexical data, their use varies from applications for mobile devices, desktop applications to natural language processing (NLP) applications, among others. While some publishers offer their data via APIs, for most small or medium size publishers implementing and providing an API is still an obstacle due to the costs and technical expertise required for their development and maintenance. Against this background, we have developed Kosh, an open-source framework for creating and maintaining APIs for lexical data. Kosh has been conceived to minimize the technical expertise required for its use, while offering generic, flexible and efficient data management. In this article, we present the methodology employed in Kosh's development and describe its architecture and functionality.

Keywords: API; Elasticsearch; framework; GraphQL; REST

1. Introduction

The development of digital lexicography over the past decades has been focused on the production of lexical data, either by digitizing printed works or by creating born-digital lexical data from scratch. Therefore, software production within this field of expertise has been directed towards the development of tools for compiling lexical data. Lexical data access has been confined mainly to the development of web applications, which are the heirs of printed dictionaries. The emergence of NLP applications and mobile devices, among other use cases, has increased the necessity to focus on the development of efficient ways of accessing lexical data. APIs satisfy this requirement, as a single API instance can provide data for multiple applications at the same time.

Although the use of APIs seems to ease several aspects of data access, there are no software solutions focused on API development and maintenance. While it is possible for large publishers to develop their own APIs, the main problem faced by small or medium sized publishers is the absence of technical expertise in-house and expensive external solutions. Against this background we created an easy-to-use framework to serve lexical data via APIs in order to lower this technical and financial hurdle.

The structure of this article is as follows: In Section 2 the motivation and decisions made about data format, data control, data persistence and efficient data access are explained. In Section 3 the architecture and functioning of Kosh are described. Section 4 concludes with a summary of the presented work and future development of the framework. Referenced publications are listed in Section 5.

2. Development Methodology

2.1 Background and motivation

Kosh has been conceived to provide API access to any XML¹-encoded lexical dataset, and its name *Kosh* derives from the Hindi word for dictionary or lexicon, कोश *koś* or *kosh*, which in turn derives from Sanskrit कोश *kośa* with the same meaning. Kosh's origin is related to multilateral API development for TEI²-encoded Sanskrit dictionaries at the University of Cologne, where the most noted digital collection of Sanskrit dictionaries worldwide is hosted.

Using the Cologne Digital Sanskrit Dictionaries web portal³, users can query all of the 37 dictionaries available through various web applications and even download each dataset in XML format. The underlying digitization project started in 1996, when XML and Unicode were not available, while in 2003 the dictionaries had been converted into XML. During the Lazarus project⁴ (2013-2015) three dictionaries were encoded in TEI-P5⁵, among them the two with the most complex structure of the entire collection (Böhtlingk & Roth, 1855-1875; Monier-Williams, 1899). Those were chosen to develop a custom schema⁶ to be employed for all future TEI-P5 dictionaries in the collection.

The first iteration of Kosh were the C-SALT APIs for Sanskrit Dictionaries (Mondaca, 2018), a proof-of-concept developed within the context of the currently running VedaWeb project⁷. One of this project's most important outcomes is to link each word of the Rigveda, the oldest text of the Indo-Aryan language family, to a dictionary specifically compiled for this text (Grassman, 1873). And in order to provide API access to this TEI-P5-encoded dictionary to the VedaWeb web application and other possible use cases, the C-SALT APIs for Sanskrit Dictionaries were implemented and have been

¹ Extensible Markup Language, <https://www.w3.org/XML>

² Text Encoding Initiative, <https://tei-c.org>

³ Cologne Digital Sanskrit Dictionaries, <https://www.sanskrit-lexicon.uni-koeln.de>

⁴ Cologne Center for eHumanities, Lazarus project, <https://cceb.uni-koeln.de/lazarus>

⁵ Text Encoding Initiative, P5 Encoding Guidelines, <https://tei-c.org/guidelines/P5>

⁶ C-SALT Dictionary Schema, https://github.com/cceb/c-salt_dicts_schema

⁷ VedaWeb, <https://vedaweb.uni-koeln.de>

transformed into a data module⁸ served by Kosh.

The guiding principle of both iterations is and has been to provide efficient access to the underlying lexical data through means of open-source software. But unlike the first iteration, the C-SALT APIs for Sanskrit Dictionaries, which were hard-coded to only serve their one designated dataset, Kosh is a *generic* solution for XML-encoded dictionaries, i.e. how each dictionary is structured is not relevant, and any XML-encoded dictionary can be indexed and accessed through Kosh's APIs.

2.2 Modular rather than monolithic

The early-stage development of Kosh consisted partly of researching software with similar features, and we noticed a lack of tools that focus on providing API access to lexical data. Most of the dictionary writing systems (DWS), commercial as well as open-source, are focused on compiling lexical data, but bear no means of providing API access to the generated data. This is reflected in a recent survey among lexicographers (Kallas et al., 2019: 33), asking respondents to identify their wishes or needs to be solved in the next 10-15 years; API access was one of the mentioned topics.

An exception in this respect is the open-source DWS Jibiki, which provides access to lexical data contained within the platform through an API (Mangeot & Enguehard, 2018: 29). But to use this API, its clients must previously register with the system. While for many publishers this might be a desired feature, as it gives them an extra layer of control and is integrated into the DWS, we opted for a different approach to Kosh's software architecture: Modularity.

When following a modular approach to software development, resolving errors or scaling up/down specific aspects of a system is usually less complex than in the case of monolithic applications, the prime architecture in traditional software development. For example, if an API module exhibits undefined behaviour (an error), this should not affect or propagate to the whole DWS, but should be contained within the erroneous module. This is one reason why the microservices architecture, essentially modular, has reached such a high level of popularity throughout the software industry.

The task of a DWS should be focused on creating and compiling new lexical data and if required accessing external sources via standardized APIs. As is the case with Lexonomy⁹, a cloud-based DWS that can access data from Sketch Engine¹⁰, a corpus manager tool, via an API. When keeping it modular, lexical data generated with this or another DWS is published by a different software component than the DWS itself, such as Kosh.

⁸ C-SALT APIs for Sanskrit Dictionaries, https://cceh.github.io/c-salt_sanskrit_data

⁹ Lexonomy, <https://www.lexonomy.eu>

¹⁰ Sketch Engine, <https://www.sketchengine.eu>

2.3 Input Data Format

In order to keep the complexity of Kosh as minimal as possible, we decided to support only the most common serialization format in lexicography: XML. At scholarly level, the use of XML-based models such as the TEI is well-known, especially in the digitization of printed dictionaries. DWS such as TLex Suite¹¹ or the Dictionary Production System¹² also output XML data. Other popular formats employed in dictionary compilation such as Toolbox¹³, prevalent in language documentation, can be transformed into XML with open access tools¹⁴, as is also the case for most of other formats such as JSON¹⁵ or YAML¹⁶. XML is widely used for representing dictionaries modelled as trees, but it is also employed to serialize graph-based models such as RDF¹⁷, although other serialization formats for graph-based models such as Turtle¹⁸ or JSON-LD¹⁹ have gained more popularity.

Kosh can handle any XML-encoded lexical dataset. We believe that developing a generic framework for APIs means that the framework should be agnostic towards the structure of the dictionaries involved: Searchable fields vary between dictionaries and they have to be defined by the publisher. Kosh can handle all types of structures as long as they are serialized in XML: Graphs, trees or graph-augmented trees, a tree-like structure that allows elements to have more than one parent (Měchura, 2016, 2018). The only limitation of our generic approach is the requirement to specify only one single XPath expression to represent an entry of the respective dictionary.

When indexing RDF datasets with Kosh, the problem that arises is to choose which nodes will be indexed, as lexical data is normally to be found in different nodes, unlike in tree-based models. If data has been encoded in OntoLex-Lemon (McCrae et al., 2017), one of the most employed graph-based models for lexical data, we would ideally index a top level node such as `LexicalEntry`. The problem then is that most of the lexicographic information such as forms and senses is normally to be found in other nodes i.e. `Form` or `LexicalSense`. So, in this case, we would need three indexes to access these nodes. For indexing the English WordNet²⁰, only two indexes are required

¹¹ TLex Suite, <https://tshwanedje.com/tshwanelex>

¹² Dictionary Production System, <http://dps.cw.idm.fr>

¹³ Toolbox, <https://software.sil.org/toolbox>

¹⁴ Natural Language Toolkit, Toolbox Reader, https://www.nltk.org/_modules/nltk/toolbox.html

¹⁵ JavaScript Object Notation, <https://www.json.org>

¹⁶ YAML Ain't Markup Language, <https://yaml.org>

¹⁷ Resource Description Framework, <https://www.w3.org/RDF>

¹⁸ Turtle, <https://www.w3.org/TR/turtle>

¹⁹ JSON for Linking Data, <https://json-ld.org>

²⁰ English WordNet, <https://en-word.net>

for the types of nodes available: `LexicalEntry` and `Synset`²¹.

2.4 Simply generic

As mentioned above, our starting point, conceptually and technically, was the C-SALT APIs for Sanskrit Dictionaries. Therefore, decisions such as which web framework, which search engine and which API paradigms to use were already made. The main issue we had to tackle was to conceive a generic method for any XML-encoded dictionary to be parsed and indexed. For this purpose, we set two goals: i) Make the configuration of this process as human-friendly as possible, and ii) from a software development perspective as elegant as possible.

Another question was: Which notation system should be used to determine the location of the nodes to be indexed? As we are parsing XML files, a rational alternative was to choose XPath²², a query language designed for selecting nodes in an XML document. As Kosh relies on lxml²³, a library for manipulating XML documents, which supports XPath 1.0 but not XPath 2.0, all XPath notations must conform to XPath 1.0.

Regarding the human interaction required to configure Kosh, one must specify which nodes of which XML documents contain lexical entries and which subnodes contain fields to be indexed. Elasticsearch²⁴ indices can be configured by external JSON files (see Section 3.2); such a file is used by Elasticsearch to setup an index and its fields with their respective data types, which are specified under the property `properties`. Following this pattern, we employ the `_meta` property to store Kosh-specific data without integrating it with the respective Elasticsearch index. In conclusion, by enriching the standard Elasticsearch JSON index definition with all required Kosh-specific data, we are able to drastically minimize human configuration effort.

2.5 Searching lexical data

A crucial decision in developing Kosh has been to employ a search engine, Elasticsearch, instead of a database, relational or not, for searching through and retrieving lexical data. We abstained from using a database management system (DBMS) with a mounted search engine on top of it as our primary data storage, as this solution seemed to add a level of complexity that is too cumbersome for a framework that should deal with different datasets and update them automatically when modified. The central question here is, why would a database be useful for this purpose?

²¹ English WordNet Kosh data module,
https://github.com/cceh/kosh_data/tree/master/wordnet_en

²² XML Path Language, <https://www.w3.org/TR/1999/REC-xpath-19991116>

²³ lxml, XML and HTML with Python, <https://lxml.de>

²⁴ Elasticsearch, <https://www.elastic.co/products/elasticsearch>

Databases were conceived and are employed for storing and managing data. Some of them (e.g. PostgreSQL²⁵) allow full-text searches, and most of the search scenarios required by the average dictionary consumer might be covered by this functionality, but DBMS in general are not tailored to automatically hash fields to minimize response latencies nor to provide different means of fuzzy query logic as search engines are. Search engines are thus the best performing systems, and Elasticsearch is one of the most used and best documented search engines servers available, so we chose to employ it.

2.6 Tracking data changes

An ideal scenario to collaboratively edit dictionaries and track changes would be to place all the datasets on a git²⁶ repository. One of the main features of git is versioning, and if the modules are on a cloud repository then all users with access can track changes and contribute. This aspect is particularly useful if a dataset contains errors or is open to modifications, and as dictionaries are continuously being edited and extended, versioning is a major improvement in their compilation process.

While not being part of Kosh's core, any publisher using Kosh can easily setup data synchronization pipelines by e.g. hooking into GitHub events²⁷, and as soon as Kosh notices the changes being propagated to its local data modules (i.e. filesystem watches are triggered), the respective search indexes get updated.

2.7 Choosing API paradigms

Authors like Tarp (2015: 34) have pointed out that one of the central features of a dictionary is to retrieve information in an easy and efficient way. Since we second this perspective, Kosh provides access to indexed lexical data not only via a single API paradigm, but the two most popular among the request-response APIs: REST (Fielding, 2000) and GraphQL (Shevat et al., 2018: 224). Besides these two main API paradigms, there are some less-employed technologies available, e.g. XQuery²⁸, which we thought of implementing but refrained from at this early stage of development.

REST has been the most popular API paradigm in the last decade, but GraphQL has risen in popularity considerably during the last few years. The reduced data load that GraphQL offers towards mobile applications is an attractive factor for its implementation in such environments (see Section 3.4). And as our goal is to satisfy as

²⁵ PostgreSQL, <https://www.postgresql.org>

²⁶ git Source Control Management, <https://git-scm.com>

²⁷ GitHub Developer Guide, <https://developer.github.com/webhooks>

²⁸ XML Query Language, <https://www.w3.org/TR/xquery-31>

many consuming and publishing use cases as possible with this framework, serving endpoints for both APIs per dataset offers the highest compatibility and therefore coverage.

While Kosh’s lexical input data has to be in XML format, both APIs return data in JSON format. The reason for this decision lies in the fact that parsing JSON is less cumbersome than parsing XML. This statement might be misleading, as Kosh by default indexes the whole entry in XML format, independently of the searchable fields defined by the publisher. If the client needs information that is not available through these fields, it must parse the full XML entry returned by Kosh’s APIs.

2.8 Open-Source Licensing

Kosh is an open-source framework and relies extensively and exclusively on open-source technologies. It runs natively on Unix-based systems, in particular Linux (Torvalds, 1997), the operating system prevalent in server environments. Elasticsearch, the search engine server, is also open-source, as is Python, the programming language that Kosh is written in. Both API paradigms offered by Kosh, REST and GraphQL, are also open-source, as is Docker²⁹, which may be used to deploy Kosh (see Section 3.5). In terms of licensing, Kosh is available under the MIT Licence³⁰.

3. Architecture and Functioning

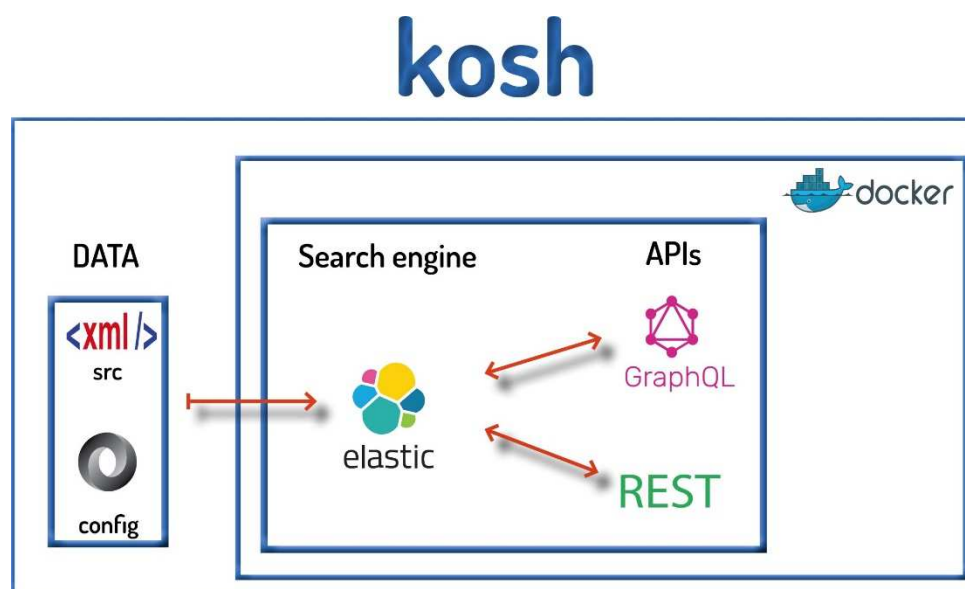


Figure 1: Overview of Kosh’s Architecture.

²⁹ Docker, <https://docs.docker.com>

³⁰ MIT License, <https://opensource.org/licenses/MIT>

3.1 Overview

Kosh's core relies on the search and analytics engine Elasticsearch and access to data indexed by this search engine is provided by GraphQL and REST APIs. While currently only two API paradigms are implemented, Kosh's application structure is designed to be modular, wherefore implementing new API paradigms to provide access to the underlying lexical data is part of our vision. A Kosh data module (input data) consists of:

1. A dataset in XML format containing lexical data
2. A JSON file containing information about the elements and their data types to be extracted from the XML file(s) in XPath 1.0 notation. This information is used by the XML parser, by Elasticsearch and by the API components
3. A kosh dotfile (.kosh) providing the following information:
 - The data module(s) name(s)
 - Filesystem path(s) to the XML file(s) containing lexical data (see 1.)
 - Filesystem path(s) to the aforementioned JSON file(s) (see 2.)

Kosh is written in Python and can be deployed in Unix-based systems. XML parsing is done with lxml, the library elasticsearch-dsl³¹ is employed for communicating with Elasticsearch, and Flask³² is used as Kosh's web application framework. Kosh's core can be downloaded as a Docker image from Docker Hub³³ or accessed directly on GitHub³⁴.

3.2 Data and metadata

Kosh processes lexical data in XML format and datasets might be split into multiple files (see e.g. de_alcedo³⁵). Further, a single Kosh instance can serve multiple data modules, while each data module is accessible through its own API endpoints. But Kosh's main innovation lies in the possibility to define the searchable fields, their respective data types and thus the perspective on each individual dataset. The only constraint is that for each index only one top-level node, i.e. entry, is allowed, but it is possible to create multiple indexes for a single XML file (see Section 2.3).

³¹ Elasticsearch DSL, <https://elasticsearch-dsl.readthedocs.io>

³² Flask, <http://flask.pocoo.org>

³³ Kosh Docker image, <https://hub.docker.com/r/cceh/kosh>

³⁴ Kosh GitHub repository, <https://github.com/cceh/kosh>

³⁵ De Alcedo Kosh data module, https://github.com/cceh/kosh_data/tree/master/de_alcedo

A lexical entry may contain different substructures, e.g. headword, part-of-speech (PoS), senses, etc., but Kosh is agnostic in this respect. The only information required for parsing and indexing a lexical entry is its XPath within the XML file(s). If no further fields (and their XPaths), e.g. headword or PoS, are specified, users can search in the whole entry but not in specific substructures, as the whole entry is indexed per default and analysed without its markup. This might be relevant for some use cases, especially when a dataset cannot be encoded in a more fine-grained manner.

```
{
  "mappings": {
    "_meta": {
      "_xpaths": {
        "id": "./@id",
        "root": "//entry",
        "fields": {
          "lemma": "./form/orth",
          "[sense_def]": "./sense/def",
          "[sense_pos]": "./sense/gramGrp/pos/q",
          "[dicteg]": "./sense/dicteg/q"
        }
      }
    },
    "properties": {
      "lemma": {
        "type": "keyword"
      },
      "sense_def": {
        "type": "text"
      },
      "sense_pos": {
        "type": "text"
      },
      "dicteg": {
        "type": "text"
      }
    }
  }
}
```

Figure 2: JSON configuration file for the Basque dictionary *Hiztegi Batu Oinarrituna*³⁶(HBO)

The JSON file seen in Figure 2 is used to configure the underlying Elasticsearch index. Relevant for Elasticsearch is the `mappings` property. It must contain the `properties` key, which specifies the fields to be indexed and their respective data types. For handling strings, the data types `keyword` and `text` may be chosen. The difference between them is that the latter is analysed by the standard analyser, which tokenizes the input string based on the Unicode Text Segmentation algorithm, while the former does not analyse or modify the input string. This should be taken into consideration when indexing headwords, because if they are indexed as `text` the analyser converts the input strings to lowercase and splits them if they contain spaces. In some cases this could render exact matches (term queries in Elasticsearch terminology) impossible.

Kosh-specific configuration values, e.g. information relevant for XML parsing, are

³⁶ HBO Kosh data module, https://github.com/cceh/kosh_data/tree/master/hiztegitatua

stored within the `_meta` property. It contains the XPath to lexical entries within the mandatory property `_xpath.root` and any additional fields to be extracted within the property `_xpath.fields`. Usually every lexical entry in parsed XML files contains a unique ID, which is also required by Kosh. This applies to the dataset as seen in Figure 2, but some datasets might not contain unique entry IDs (see e.g. `freedicts`³⁷). In such cases, Kosh generates IDs by calculating SHA1 hashes from a normalized form of each entry, so those IDs only change when the respective entry changes and therefore are reproducible.

Data modules are identified by Kosh through the existence of a `.kosh` dotfile. Such a `.kosh` file acts as an entry point for the data module by specifying its names, file system paths to XML files that contain the lexical data to be indexed, and to the JSON metadata files containing the previously described definitions for the respective data module.

Finally, when running Kosh on an operating system capable of notifying³⁸ file changes, Kosh automatically updates the respective Elasticsearch index and re-binds all API components to reflect changes made to the data module definitions or its lexical data.

3.3 Elasticsearch engine

Kosh employs Elasticsearch as its search engine server. Currently, the supported query types are those available for unique fields with the properties `keyword` or `text` in the configuration file, e.g. `prefix`, `term` and `match`. Query types on multiple fields, e.g. `multimatch` and `bool`, have not yet been implemented but are being actively developed. String based queries can be classified as full-text or term-level, and clients can perform both types of queries on all indexed fields. Queries might return different results when using a `term` query (exact matching), if the queried field has been indexed as `text` instead of `keyword`, because `text` fields are analysed, i.e. they are tokenized and lowercased. For example, if a dictionary has uppercased lemmas which have been indexed as `text`, any uppercased term-level query on the respective field will not deliver results.

By default, Elasticsearch (and thus Kosh) returns ten elements per query, but a client can request more results by providing a specific integer value in the `size` query field. Further, Kosh's default configuration offers two term-level query types that expose all the indexed entries at once: `regexp` and `wildcard`. And the `prefix` query type can return all entries in a couple of requests. If the publisher wishes to restrict access to his lexical data, i.e. only offer queries that return a subset of the data, these query types have to be disabled in Kosh's source code.

³⁷ Freedict Kosh data module, https://github.com/cceh/kosh_data/tree/master/freedict

³⁸ inotify Linux Manpage, <http://man7.org/linux/man-pages/man7/inotify.7.html>

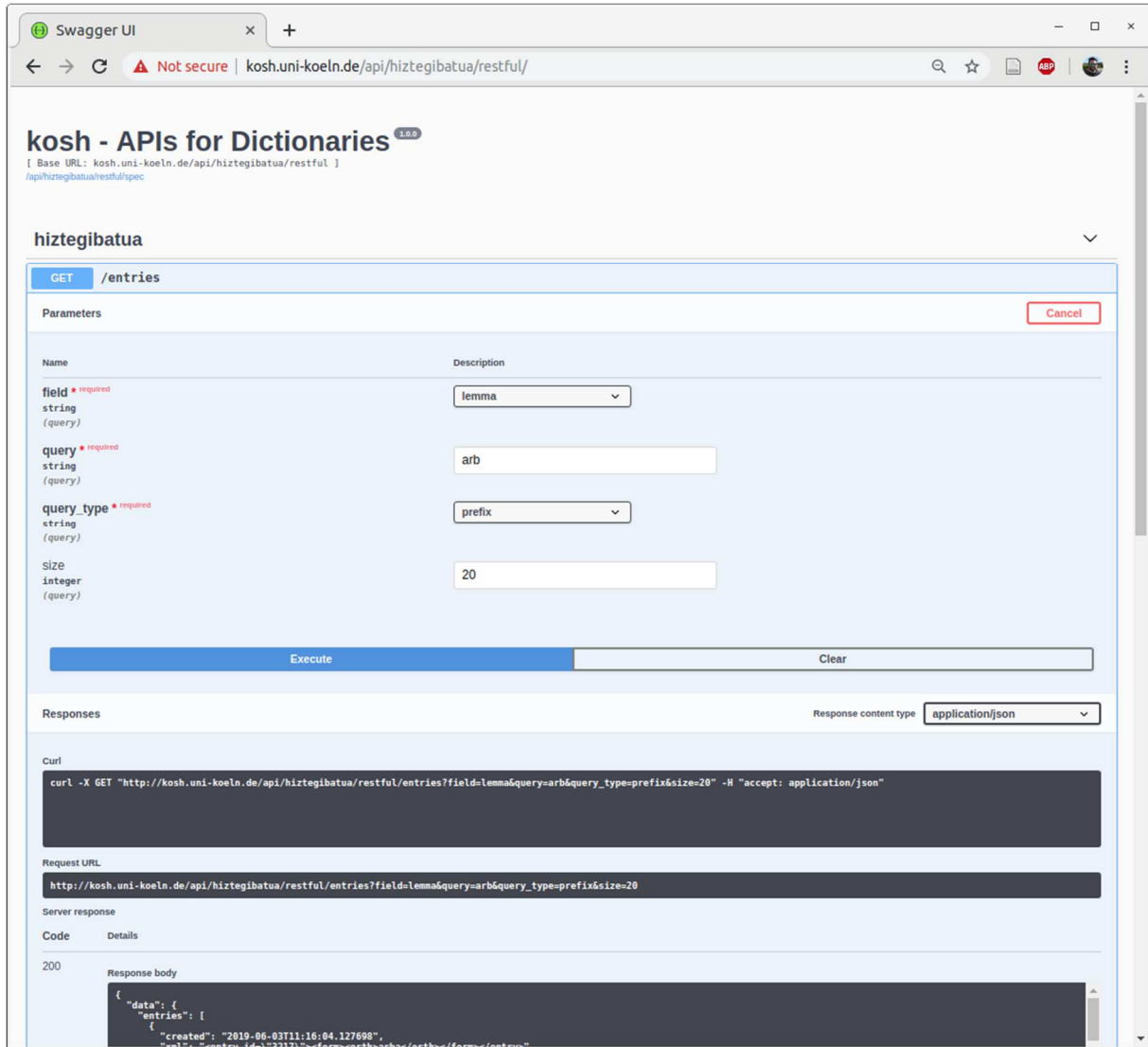


Figure 3: Swagger UI³⁹, prefix query for ‘arb’ in the HBO, max. 20 results.

3.4 API access

While all the previously described layers of Kosh are crucial for its functioning, the APIs represent the visible and most relevant layer for clients. Kosh offers the two most popular API paradigms of the last decade for each indexed dataset: REST and GraphQL, and both return data in JSON format. The main differences between them are that GraphQL has a single endpoint, is typed, and that in a GraphQL query, unlike when using REST, the fields to be returned need to be explicitly specified. While this function can also be implemented in a REST API via sparse fieldsets⁴⁰, it is not a constraint on its implementation. For example, when using GraphQL, a client may

³⁹ Swagger UI, <https://swagger.io/tools/swagger-ui>

⁴⁰ JSON API, Sparse Fieldsets, <https://jsonapi.org/format#fetching-sparse-fieldsets>

retrieve only the headword field of all entries matching a specific query, e.g. all headwords that have the prefix ‘arb’, while a RESTful query would retrieve all the available fields related to the matched entries. Thus, one of the main advantages of GraphQL over REST is reduced data load, which can be relevant for mobile applications in areas with connectivity problems.

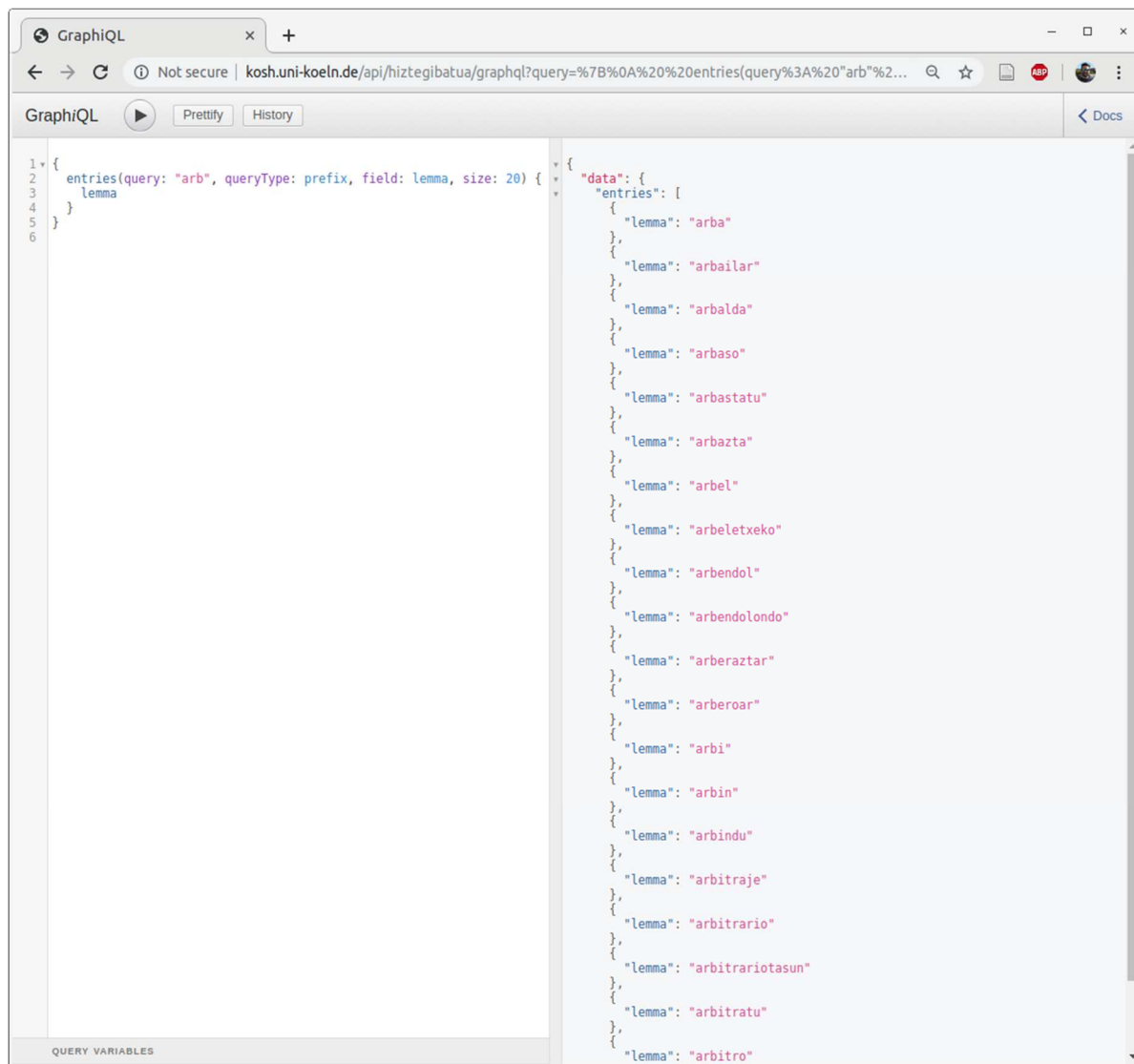


Figure 4: GraphQL - Prefix query for ‘arb’ in the HBO, fetching only related lemmas, max. 20 results

The framework provides user interfaces for all APIs (including documentation of available endpoints, queries, and typings). This way, all those interested in accessing the lexical data provided by the APIs can easily test and integrate them. For each data module Kosh serves an instance of Swagger UI (see Figure 3), running against all RESTful endpoints, and a GraphQL⁴¹ instance (see Figure 4), to allow running all available queries.

⁴¹ GraphQL GitHub repository, <https://github.com/graphql/graphql>

Kosh offers two RESTful endpoints per data module: `entries` and `ids`. Using the `entries` endpoint, a client may search within the default available `xml` field as well as within any field defined by the respective data module. The `ids` endpoint fetches entries by specifying one or more entry IDs. For each GraphQL API endpoint the same two types of queries are available: `entries` and `ids`. All those API endpoints only offer consumption of lexical data, no modifications can be made to the underlying dataset, i.e. only HTTP GET requests are allowed.

3.5 Deployment

Kosh can be deployed in two different ways: Either natively or via Docker. The first option requires a Unix-based system with Python 3.6+ and Elasticsearch installed and running. This can be achieved by simply running the included `Makefile`, which installs Kosh and all required Python libraries, and providing a suitable configuration either on Kosh's command line or via a configuration file.

The second deployment option requires Docker and is the easiest method to deploy and maintain a Kosh instance. Docker is an operating-system-level virtualization tool which is popular among developers and administrators due to the possibility of distributing software packages as containers, i.e. isolated from each other. At the same time it offers clear and effective ways of bundling them together. To orchestrate containers and integrate them as services, Docker provides `docker-compose`⁴², which in this case is employed to bundle an Elasticsearch and a Kosh instance together.

When using the included `docker-compose.yml` and `docker-compose.local.yml`, Kosh can be easily setup without the need to install any additional software. Docker will pull the Elasticsearch and Kosh images from Docker Hub, where they are both built automatically, i.e. the images always contain the latest versions of Kosh and Elasticsearch.

Kosh's source code is available on GitHub. For demonstration purposes, we also provide another GitHub repository, Kosh Data⁴³, that contains different data modules, so that users may transfer the structure of Kosh data modules onto their own datasets.

4. Conclusions and further development

In this article, we have presented Kosh and its main goal: To provide efficient and easy-to-configure access to lexical data. For this purpose, we have described the various theoretical considerations and technical decisions that have been made: i) Choosing XML as the data input format, ii) selecting Elasticsearch as Kosh's storage layer, and

⁴² Docker Compose, <https://docs.docker.com/compose>

⁴³ Kosh Data GitHub repository, https://github.com/cceh/kosh_data

iii) adopting REST and GraphQL as its default API paradigms.

Kosh is a stable and high performing microservice that offers cutting-edge technologies with a relatively low learning curve for users without strong technical skills. Still, if it is used in production then aspects such as deploying a web server or user analytics should ideally be addressed by technical staff. Currently, only one field may be queried via the APIs, while the underlying search engine offers a much more fine-grained query logic. We plan to expose more of this functionality through Kosh's APIs in the future. We also envision the implementation of further API paradigms to enrich Kosh with more possibilities of serving lexical data. Besides such long-term goals, we are also committed to accomplish short-term development milestones, including continuous support in form of upstream library updates and bug fixes.

5. References

- Böhtlingk, O. & Roth, R. (1855-1875). *Sanskrit-Wörterbuch*. St. Petersburg: Kaiserliche Akademie der Wissenschaften.
- Fielding, R. T. (2000). *Architectural Styles and the Design of Network-based Software Architectures*. Ph.D. thesis, University of California, Irvine. URL https://www.ics.uci.edu/~fielding/pubs/dissertation/fielding_dissertation.pdf.
- Grassman, H. G. (1873). *Worterbuch zum Rig-veda*. Wiesbaden: O. Harrassowitz.
- Kallas, J., Koeva, S., Kosem, I., Langemets, M. & Tiberius, C. (2019). D1.1 Lexicographic Practices in Europe: A Survey of User Needs. Deliverable D1.1, Elexis. European Lexicographic Infrastructure. https://elex.is/wp-content/uploads/2019/02/ELEXIS_D1_1_Lexicographic_Practices_in_Europe_A_Survey_of_User_Needs.pdf.
- Mangeot, M. & Enguehard, C. (2018). Dictionaries for Under-Resourced Languages: from Published Files to Standardized Resources Available on the Web. Research Report, Laboratoire d'informatique de Grenoble. URL <https://hal.archives-ouvertes.fr/hal-02056905>.
- McCrae, J. P., Bosque-Gil, J., Gracia, J., Buitelaar, P. & Cimiano, P. (2017). The OntoLexLemon Model: Development and Applications. In I. Kosem et al. (eds.) *Proceedings of the 5th Biennial Conference on Electronic Lexicography (eLex 2017)*. Leiden, the Netherlands, pp. 587– 597. <https://elex.link/elex2017/wp-content/uploads/2017/09/paper36.pdf>.
- Mondaca, F. (2018). C-SALT APIs for Sanskrit Dictionaries: A Novel Approach for Accessing Digital Lexical Resources Online. Workshop on eLexicography: Between Digital Humanities and Artificial Intelligence. Co-located with EADH 2018 - Data in Digital Humanities. December 19, 2018. Galway, Ireland. https://lexdhai.insight-centre.org/Lex_DH_AI_2018_paper_7.pdf.
- Monier-Williams, M. (1899). *A Sanskrit-English dictionary: Etymologically and philologically arranged with special reference to Cognate indo-european languages*. Oxford: The Clarendon Press.
- Měchura, M. (2016). Data Structures in Lexicography: from Trees to Graphs. In *The*

- 10th Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2016, Karlova Studanka, Czech Republic, December 2-4, 2016.* pp. 97–104. URL <http://nlp.fi.muni.cz/raslan/2016/paper04-Mechura.pdf>.
- Měchura, M. (2018). Shareable Subentries in Lexonomy as a Solution to the Problem of Multiword Item Placement. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (eds.) *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*. Ljubljana University Press, Faculty of Arts, pp. 223–232. <http://euralex.org/wp-content/themes/euralex/proceedings/Euralex%202018/118-4-2964-1-10-20180820.pdf>.
- Shevat, A., Sahni, S. & Jin, B. (2018). *Designing Web APIs*. Sebastopol: O'Reilly Media.
- Tarp, S. (2015). La teoría funcional en pocas palabras. *Estudios de Lexicografía*, 4, pp. 31–42.
- Torvalds, L. (1997). *Linux: a Portable Operating System*. Master of Science Thesis, University of Helsinki. https://www.cs.helsinki.fi/u/kutvonen/index_files/linus.pdf.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



Enriching an Explanatory Dictionary with FrameNet and PropBank Corpus Examples

**Pēteris Paikens¹, Normunds Grūzītis², Laura Rituma²,
Gunta Nešpore², Viktors Lipskis², Lauma Pretkalniņa²,
Andrejs Spektors²**

¹Latvian Information Agency LETA, Marijas street 2, Riga, Latvia

²Institute of Mathematics and Computer Science, University of Latvia,
Raina blvd. 29, Riga, Latvia

E-mail: peteris.paikens@leta.lv, normunds.gruzitis@ailab.lv

Abstract

This paper describes ongoing work to extend an online dictionary of Latvian – Tezaurs.lv – with representative semantically annotated corpus examples according to the FrameNet and PropBank methodologies and word sense inventories. Tezaurs.lv is one of the largest open lexical resources for Latvian, combining information from more than 300 legacy dictionaries and other sources. The corpus examples are extracted from Latvian FrameNet and PropBank corpora, which are manually annotated parallel subsets of a balanced text corpus of contemporary Latvian. The proposed approach augments traditional lexicographic information with modern cross-lingually interpretable information and enables analysis of word senses from the perspective of frame semantics, which is substantially different from (complementary to) the traditional approach applied in Latvian lexicography. In cases where FrameNet and PropBank corpus evidence aligns well with the word sense split in legacy dictionaries, the frame-semantically annotated corpus examples supplement the word sense information with clarifying usage examples and commonly used semantic valence patterns. However, the annotated corpus examples often provide evidence of a different sense split, which is often more coarse-grained and, thus, may help dictionary users to cluster and comprehend a fine-grained sense split suggested by the legacy sources. This is particularly relevant in case of frequently used polysemous verbs.

Keywords: explanatory dictionary; FrameNet; PropBank; semantic annotation; Latvian

1. Introduction

A major function of an explanatory dictionary is to describe the word senses and illustrate their usage with examples. The separation of word senses is usually done by a lexicographer, based on linguistic intuition and corpus evidence. For less-resourced languages, however, modern corpus-based dictionaries are often missing or works in progress, and the established dictionaries and their senses are not based on corpus evidence. As a consequence, the word sense split is often too fine-grained, which can make it difficult even for a native speaker to grasp the difference, while certain contemporary word senses tend to be missing.

These issues are particularly salient when working on semantic resources for the needs of computational linguistics. Word sense inventories used for automatic word sense disambiguation and semantic parsing tasks need to be formal, well-defined and exhaustive, while the existing dictionaries leave much to the reader’s interpretation and rely on illustrative examples of various word usages.

The current work is aimed to extend Tezaurs.lv,¹ the largest Latvian online reference dictionary (Spektors et al., 2016). Tezaurs.lv is structured as an explanatory dictionary which has been compiled from approximately 300 dictionaries and other sources, and contains more than 310,000 entries. In addition to common dictionary content, Tezaurs.lv has been extended with structured data for various natural language processing needs – inflectional paradigm and inflection tables, phonetic transcriptions, domains of usage, stylistic markers and usage restrictions.

Currently the dictionary entries contain usage examples – citations automatically selected from a balanced text corpus of modern Latvian (Levane-Petrova, 2019). These corpus examples tend to illustrate the most common senses and not represent the whole variety of word usage.

However, semantically annotated corpora have sufficient information to separate substantially different uses of the same word, and thus provide examples for each such subsense. In this work we describe the process and results of adding this information to Tezaurs.lv. Section 2 describes the semantically annotated datasets used for this task, Section 3 contains the implementation details, and Section 4 illustrates the resulting changes to the online dictionary.

2. Semantically annotated Latvian corpora

A dataset of semantically annotated Latvian text units is being created within a larger research project “Full Stack of Language Resources for Natural Language Understanding and Generation in Latvian” (Gruzitis et al., 2018b). The goal of that project is to build a balanced multilayer corpus based on cross-lingually oriented syntactic and semantic representations: Universal Dependencies (Nivre et al., 2016), FrameNet (Fillmore et al., 2003), PropBank (Palmer et al., 2005), Abstract Meaning Representation (Banarescu et al., 2013), as well as auxiliary layers for named entity and coreference annotation.²

The data is selected to provide a balanced and representative medium-sized corpus of Latvian: around 13,000 sentences annotated at all the above mentioned layers, including FrameNet. To ensure that the corpus is balanced not only in terms of text genres and writing styles but also in terms of lexical units, the text unit of this corpus is an isolated

¹ Open access at www.tezaurs.lv

² Available at <https://github.com/LUMII-AILab/FullStack>.

paragraph. Paragraphs are manually selected from a balanced 10-million-word text corpus (Levane-Petrova, 2019): 60% news, 20% fiction, 7% academic texts, 6% legal texts, 5% spoken language, 2% miscellaneous. The corpus is considered a representative selection of contemporary literary Latvian, including diverse sources starting from year 1991 but excluding translations and genres such as user-generated comments and chat.

The paragraph selection is done with the goal to ensure good coverage for the 1,000 most frequently used Latvian verbs and each of their coarse-grained word senses. We assume that the corpus will prove to be balanced also with respect to nominal lexical units, as the source data is well balanced in terms of genres and frequency distribution. The corpus is not large but has good coverage of the most frequently used verbs, which also tend to be the most ambiguous ones, and there is ongoing work to increase this corpus.

2.1 FrameNet annotations

The annotation of the general-purpose Latvian FrameNet is based on the latest Berkeley FrameNet (BFN) frame inventory (v1.7). The choice to rely on the English BFN frames was made in order to reuse the BFN frame hierarchy and other inter-frame relations, as well as semantic types of frame elements (FE), and the definitions of frames and FEs in general. Another reason for BFN compatibility is to facilitate use cases that require cross-lingual semantic parsing.

In annotating the Latvian FrameNet a concordance approach was followed: frame instances are annotated separately for each target word instead of going through all documents and sentences. Such an approach increases the annotation consistency. In the current annotations only core FEs (which characterize and define the frame) and two non-core FEs (Time and Place) are systematically annotated.

The annotations follow a corpus-driven approach: lexical units in Latvian FrameNet are created only based on the annotated corpus examples. Moreover, the FrameNet annotation is done on the top of the underlying Universal Dependency treebank layer (Pretkalnina et al., 2018), so the annotation of frames and frame elements is thus guided by the dependency structure of a sentence. The currently annotated dataset contains approximately 1,600 distinct lexical units.

2.2 PropBank annotations

The Latvian PropBank corpus is derived from the Latvian FrameNet corpus, thus, this is a parallel dataset. The same original sentences are used, however, the annotations at times are substantially different. The initial draft configuration is automatically generated using the suggested mapping alternatives between English FrameNet and English PropBank. This was followed by linguists mapping the lexical units from

Latvian FrameNet annotation to the semantic frames of English PropBank, and verification of the mapping between FrameNet and PropBank semantic roles, which generally depends on the underlying sentence syntax.

The reason for integrating both FrameNet and PropBank corpus examples into Tezaurs.lv entries is that PropBank tends to provide even more robust and fine-grained sense splits. The semantic roles of the PropBank semantic predicates follow the syntactic argument structure of the target verb, while FrameNet frame elements are often annotated beyond the syntactic argument structure of the target verb. The totality of PropBank annotations for a particular verb essentially constitute a valency dictionary, describing the syntactic relations possible (and used in corpus) for every semantic argument of that verb. Another benefit is that PropBank predicates are lexical compared to the highly abstract FrameNet frames. Therefore both representations are complementary from the Tezaurs.lv user perspective.

A particular source of difficulty is the alignment of Latvian verbs with the English PropBank – unlike some other languages (Haverinen et al., 2015; Xue, 2008), the annotation project chose to use the English PropBank sense inventory instead of native Latvian senses so that the results are immediately aligned and usable for multilingual processing tools. This requires upfront work with translation dictionaries to appropriately map the intended meaning of each verb to its English equivalent. If multiple PropBank verbs match the intended meaning, then extra attention is paid to verb argument structures, however in some cases the choice between multiple options is mostly subjective.

It's worth noting that the sentences are not fully annotated with PropBank roles – only the verbs expressing the FrameNet annotation are targeted, and only the arguments matching the FrameNet core roles are annotated.

3. Technical implementation

For a given lexical entry of Tezaurs.lv, illustrative annotated examples from the Latvian FrameNet and PropBank corpora are selected and visualized as follows.

From the Latvian FrameNet dataset (Section 2.1), we first select all annotation sets where the headword is the target word. Each annotation set represents a single frame, together with its core elements, evoked by the target word. If the same sentence contains more than one frame instance, each instance is encoded in a separate annotation set.

Latvian FrameNet annotation sets are encoded in an extended CoNLL-U format,³ since

³ <https://universaldependencies.org/format.html>

FrameNet annotations are added on top of dependency trees of Latvian UD Treebank⁴ (Gruzitis et al., 2018a). The extension follows the CoNLL-2009 format.⁵ Figure 1 illustrates an annotation set from the Latvian FrameNet corpus for the sentence “as soon as Sophie had closed [the] garden gate behind her she opened [the] envelope” with the Closure frame evoked by the verb ‘to close’ (“*aizvērt*”), and its elements (semantic roles) Agent and Container_portal filled by the subject (nsubj – “*Sofija*”/‘Sophie’) and object (obj – “*vārtin, us*”/‘gate’) arguments of the verb respectively.

```
# sent_id      = a-d199-p12s1
# text         = Tiklīdz Sofija bija aizvērusi aiz sevis dārza vārtin,us, viņa atvēra aploksni.
# word-by-word = As-soon-as Sophie had closed behind her garden gate      , she opened envelope.
1 Tiklīdz      tiklīdz      SCONJ cs          _ 4 mark      _ _ _ _ _
2 Sofija       Sofija       PROPN npfsn4      _ 4 nsubj   _ _ _ _ _ Agent
3 bija         būt          AUX    vcnisii30an     _ 4 aux     _ _ _ _ _
4 aizvērusi    aizvērt      VERB   vmnpdfsna      _ 11 advcl  _ _ Y Closure _
5 aiz         aiz          ADP    spsg          _ 6 case    _ _ _ _ _
6 sevis        sevis        PRON   px000gn        _ 4 obl     _ _ _ _ _
7 dārza        dārzs        NOUN   ncmsg1         _ 8 nmod    _ _ _ _ _
8 vārtin,us    vārtin,     NOUN   ncmpa1         _ 4 obj     _ _ _ _ _ Container_portal
9 ,            ,            PUNCT  zc             _ 4 punct   _ _ _ _ _
10 viņa        viņa         PRON   pp3fsmn         _ 11 nsubj  _ _ _ _ _
11 atvēra      atvērt      VERB   vmnist130an     _ 0 root    _ _ _ _ _
12 aploksni    aploksne    NOUN   ncfsa5          _ 11 obj    _ _ _ _ _
13 . . .      PUNCT  zs       _ 11 punct   _ _ _ _ _
```

Figure 1: Sample FrameNet annotation set. Fields 1–10 correspond to the CoNLL-U fields: ID, FORM, LEMMA, UPOS, XPOS, FEATS, HEAD, DEPREL, DEPS, MISC; fields 11–13 correspond to the CoNLL-2009 fields: FILLPRED, PRED, APRED. To save space, values of FEATS, DEPS and MISC are excluded from the sample. The word-by-word English translation is added for clarity.

Since the Latvian PropBank corpus is derived from the Latvian FrameNet corpus (Section 2.2), PropBank annotation sets are available as parallel data in the same extended CoNLL-U format (see Figure 2). The initial CONLL-U columns of both datasets are identical, containing the Universal Dependencies (UD) syntactic representation, but the final columns contain the relevant semantic annotation.

For each lexical unit in Latvian FrameNet and Latvian PropBank, there are seven annotation sets on average. To automatically select concise sets of annotated examples to be included in Tezaurs.lv entries of the corresponding verbs, the following selection criteria are applied (in this order):

⁴ https://github.com/UniversalDependencies/UD_Latvian-LVTB

⁵ <http://ufal.mff.cuni.cz/conll2009-st/task-description.html>

1. The annotation sets corresponding to each separate frame using this word are selected.
2. If an annotation set is a subset of another annotation set in terms of the evoked frame and its frame elements, it is excluded from the selection, i.e. example sentences representing more frame elements are preferred over example sentences representing less frame elements for the same frame.
3. For each group of so far selected example sentences, shorter examples (containing less tokens) are preferred over longer examples.

```
# sent_id      = a-d199-p12s1
# text         = Tiklīdz Sofija bija aizvērusi aiz sevis dārza vārtiņus, viņa atvēra aploksni.
# word-by-word = As-soon-as Sophie had closed behind her garden gate , she opened envelope.
1 Tiklīdz      tiklīdz      SCONJ cs          _ 4 mark      _ _ _ _ _
2 Sofija       Sofija       PROPN npfsn4      _ 4 nsubj    _ _ _ _ _ ARG0-PAG
3 bija         būt          AUX      vcnisii30an      _ 4 aux      _ _ _ _ _
4 aizvērusi    aizvērt      VERB      vmnpdfsnsn4      _ 11 advcl   _ _ Y close.01 _
5 aiz         aiz          ADP      spsg          _ 6 case     _ _ _ _ _
6 sevis        sevis        PRON      px000gn          _ 4 obl      _ _ _ _ _
7 dārza        dārzs        NOUN      ncmsg1          _ 8 nmod     _ _ _ _ _
8 vārtiņus     vārtiņi     NOUN      ncmpa1          _ 4 obj      _ _ _ _ _ ARG1-PPT
9 ,            ,            PUNCT     zc             _ 4 punct    _ _ _ _ _
10 viņa        viņa         PRON      pp3fsnn          _ 11 nsubj   _ _ _ _ _
11 atvēra      atvērt      VERB      vmnist130an      _ 0 root     _ _ _ _ _
12 aploksni    aploksne    NOUN      ncfsa5          _ 11 obj     _ _ _ _ _
13 . . .      PUNCT     zs             _ 11 punct    _ _ _ _ _
```

Figure 2: Sample PropBank annotation set – a complementary semantic annotation to FrameNet (cf. Figure 1).

Additionally, frequency counts are summarized for each lexical unit and are used to sort the selected FrameNet- and PropBank-annotated example sentences (for each Tezaurs.lv entry). In the Tezaurs.lv user interface, the selected annotated examples are visualized using the *brat* JavaScript library⁶ (Stenetorp et al., 2012). To generate annotation visualizations in SVG and PNG formats, two kinds of data structures (JSON objects) are generated from the FrameNet- and PropBank-annotated corpus examples.

First, a common stylesheet object is generated (as illustrated in Figure 3) from the FrameNet and PropBank frame inventories, listing all frames (predicates) and frame elements (semantic roles) and their visualization properties. Second, a *brat* annotation object (Figure 4) is generated from the corresponding FrameNet annotation set (Figure 3) for each selected corpus example. Similarly, a *brat* annotation object is generated from the corresponding PropBank annotation set. Note that frame elements (semantic roles) in the Latvian FrameNet and PropBank corpora are added to the root nodes of

⁶ <http://brat.nlplab.org>

the respective syntactic subtree, instead of whole text spans (syntactic phrases). The text spans are calculated while generating the brat annotation objects, based on the dependency links encoded in the underlying UD annotations (the HEAD column in the CoNLL-U data structures; see Figure 1).

```
{
  "entity_types": [{"type": "FE", "bgColor": "yellow", "borderColor": "darken"}],
  "event_types": [
    {...},
    {"type": "Closure", "bgColor": "lightgreen", "borderColor": "darken", "arcs": [
      {"type": "Agent", "color": "blue"},
      {"type": "Time", "color": "blue"},
      {"type": "Place", "color": "blue"},
      {"type": "Containing_object", "color": "blue"},
      {"type": "Result", "color": "blue"},
      {"type": "Container_portal", "color": "blue"}
    ]},
    {...},
    {"type": "Body_movement", "bgColor": "lightgreen", "borderColor": "darken", "arcs": [
      {"type": "Agent", "color": "blue"},
      {"type": "Place", "color": "blue"},
      {"type": "Path", "color": "blue"},
      {"type": "Body_part", "color": "blue"},
      {"type": "Addressee", "color": "blue"}
    ]},
    {...}
  ]
}
```

Figure 3: An incomplete example stylesheet for the FrameNet frames and frame elements. A similar *brat* stylesheet is generated also for PropBank predicates and semantic roles.

```
{
  "text": "Tiklīdz Sofija bija aizvērusi aiz sevis dārza vārtiņus, viņa atvēra aploksni.",
  "triggers": [{"T0", "Closure", [[20, 29]]}],
  "events": [{"E1", "T0", [{"Agent", "T1"}, {"Container_portal", "T2"}]],
  "entities": [
    ["T1", "FE", [[8, 14]]],
    ["T2", "FE", [[40, 54]]]
  ]
}
```

Figure 4: Example sentence with the *brat* annotation, corresponding to the FrameNet annotation given in Figure 1. A similar annotation object is also generated for the corresponding PropBank-annotated corpus example.

Finally, an SVG or a PNG image is generated for each FrameNet and PropBank corpus example (as illustrated in Figure 5) from the common *brat* stylesheet object and the example-specific *brat* annotation objects.



Figure 5: A corpus example (“as soon as Sophie had closed the garden gate behind her [she opened the envelope]”) with parallel FrameNet and PropBank annotation, illustrating the sense and use of the headword “aizvērt” (‘to close’).

4. Enriched online dictionary

The currently intended use case for the FrameNet- and PropBank-annotated corpus examples is to provide separate yet complementary information to the relevant dictionary entries. A set of concise and representative annotated corpus examples is shown to the dictionary user.

Figure 6 illustrates how such frame-semantic information would be displayed in the Tezaurs.lv interface. The original Tezaurs.lv entry contains:

1. the headword: “aizvērt” (‘to close’);
2. shorthand grammatical information in the Latvian lexicographic tradition, in this particular case showing some key inflectional forms and indicating that the verb is transitive: “-veru, -ver, -ver, pag. (‘past’) -vēru; trans.”;
3. definitions of word senses: (1) “verot aizdarīt” ~ ‘to become closed, shut’, (2) “verotaizvirzīt aiz kā, kam cauri” ~ ‘to move behind something, through something’ (the marker “apv.” indicates that this sense is used only in some regions);
4. definitions of subsenses: e.g. the first sense has a subsense for closing body parts like eyes and lips – “aizdarīt (acis, plakstus, lūpas, muti)”;
5. idioms (“frazelogismi”): collapsed in this example;
6. references to source dictionaries (“avoti”);
7. inflection table (“morfologija”) automatically provided by a complementary web-service: collapsed in this example;

8. plain-text corpus examples (“korpusa piemēri”) automatically selected by a complementary web-service: it is not certain that the provided corpus examples cover all common senses of the headword, and the examples are selected by the lemma, without explicitly linking them to word senses.

aizvērt -veru, -ver, -ver, pag. -vēru; trans.

1. Verot aizdarīt.

// imperf. Vērt ciet; aiztaisīt.

// Aiztaisīt, uzliekot vāku (parasti, virās iestiprinātu).

// Aizdarīt (acis, plakstus, lūpas, muti).

2. apv. Verot aizvirzīt aiz kā, kam cauri.

FRAZEOLOĢISMI: +

AVOTI: LLVV, ĒiV

MORFOLOĢIJA: darbības vārds, 1. konjugācija +

KORPUSA PIEMĒRI: —

«Baidās tik ļoti, ka nespēj aizvērt acis.» (Guntis Berelis, *Mīnotaura medības*. Rīga, Atēna, 1999.)

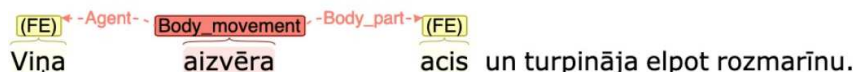
«Sveta notrīs, aizver logu un apsēžas pie galda.» (Elita Franciska Cimare, *Sarkanie ūdeņi*. Rīga, AGB, 2001.)

«Vairākas iepriekšējās nakts nebija pat uz mirkli aizvēris acis.» (Egils Lukjanskis, *Kam neskanēs zvans*. Rīga, Zvaigzne ABC, 2006.)

Piemēri ir atlasīti automātiski un var būt neprecīzi. [Vairāk piemēru...](#)

FRAMENET PIEMĒRI: —







[Vairāk piemēru...](#)

Figure 6: Tezaurs.lv entry for the verb ‘aizvērt’ (to close): <https://tezaurs.lv/#/sv/aizvērt>. The original entry, consolidated from two source dictionaries (LLVV and EiV), is enriched with automatically extracted usage examples from (i) a balanced text corpus (‘Korpusa piemēri’), and (ii) a FrameNet-annotated corpus (‘FrameNet piemēri’). FrameNet annotations can be switched to the parallel PropBank annotations.

In the supplementary section of FrameNet corpus examples (“FrameNet piemēri”), a concise annotated example is given for each of the different semantic frames evoked by the headword to illustrate its sense split and semantic valency according to FrameNet.

In the above example, two of the FrameNet frames – Closure and Body_movement – align with the first sense (and its third subsense) of the headword, and it is debatable whether Body_movement is a subsense of Closure or not (for this particular verb). However, the third FrameNet example which evokes Locale_closure, illustrates a distinct meaning of the verb *‘aizvērt’*, which is missing in the original Tezaurs.lv entry despite being a commonly used word sense for already a long time. Also note that the second word sense provided by Tezaurs.lv is rare and possibly obsolete, and therefore is not represented in the balanced FrameNet-annotated corpus.

5. Conclusions and future work

In summary, we propose to extend online dictionaries by adding frame-semantically annotated corpus examples. Such examples enable complementary analysis of word senses and word valence patterns from the perspective of frame semantics, which is substantially different from the traditional lexicographic approach.

In our opinion, the major benefit of the suggested approach for everyday dictionary users is the following: it often provides an alternative and more coarse-grained split of word senses based on semantically annotated corpus evidence according to the robust FrameNet and PropBank methodologies.

Since Latvian FrameNet uses the abstract frame inventory of Berkeley FrameNet and the more concrete semantic predicate inventory of English PropBank, it also makes it easier for language learners to understand the differences between particular word senses, assuming that they know English better than Latvian.

Another benefit is the modernization of legacy dictionaries. A large portion of Tezaurs.lv entries and word sense splits originate from Latvian dictionaries of 1970s, but the semantically annotated corpus represents contemporary usage of the language. Because of this, corpus examples illustrate usage and sense split of words in more contemporary contexts, some of which were not identified in the earlier dictionaries.

The differences in sense splitting between legacy dictionaries and examples from the semantically annotated corpora illustrate the need for future work on updating the Latvian word sense inventory based on corpus evidence, either as part of the traditional lexicographic workflow or as a separate lexical resource in the likeness of WordNet (Miller, 1995; Bond & Foster, 2013).

Another direction of future work is the handling of multi-word expressions (MWEs) such as phrasal verbs. For example, the verb *‘aiziet’* (‘to go away’) has distinct senses invoked by *‘aiziet bojā’* (‘to perish’), *‘aiziet mūžībā’* (‘to die’). Such MWEs are explicitly annotated in the Latvian FrameNet dataset, but are currently not included in the CoNLL-style output format and, thus, are not included in the FrameNet example visualizations.

6. Acknowledgements

This work has received financial support from the European Regional Development Fund under grant agreement No 1.1.1.1/16/A/219 and from the Latvian State research program “Latvian language” (VPP-IZM-2018/2-0002).

7. References

- Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M. & Schneider, N. (2013). Abstract Meaning Representation for Sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*. Sofia, Bulgaria, pp. 178–186.
- Bond, F. & Foster, R. (2013). Linking and Extending an Open Multilingual Wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Sofia, Bulgaria: Association for Computational Linguistics, pp. 1352–1362.
<https://www.aclweb.org/anthology/P13-1133>.
- Fillmore, C. J., Johnson, C. R. & Petruck, M. R. (2003). Background to FrameNet. *International Journal of Lexicography*, 16(3), pp. 235–250.
- Gruzitis, N., Nespore-Berzkalne, G. & Saulite, B. (2018a). Creation of Latvian FrameNet based on Universal Dependencies. In *Proceedings of the International FrameNet Workshop (IFNW)*. Miyazaki, Japan, pp. 23–27.
- Gruzitis, N., Pretkalnina, L., Saulite, B., Rituma, L., Nespore-Berzkalne, G., Znotins, A. & Paikens, P. (2018b). Creation of a Balanced State-of-the-Art Multilayer Corpus for NLU. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC)*. Miyazaki, Japan, pp. 4506–4513.
<http://www.lrec-conf.org/proceedings/lrec2018/pdf/935.pdf>.
- Haverinen, K., Kanerva, J., Kohonen, S., Missilä, A., Ojala, S., Viljanen, T., Laippala, V. & Ginter, F. (2015). The Finnish Proposition Bank. *Language Resources and Evaluation*, 49(4), pp. 907–926. <https://doi.org/10.1007/s10579-015-9310-y>.
- Levane-Petrova, K. (2019). LVK2018: Līdzsvarotais mušdienu latviešu valodas tekstu korpuss, tā nozīme gramatikas pētījumos. *Language: Meaning and Form*, 10.
- Miller, G.A. (1995). WordNet: A Lexical Database for English. *Commun. ACM*, 38(11), pp. 39–41. URL <http://doi.acm.org/10.1145/219717.219748>.
- Nivre, J., de Marneffe, M. C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R. & Zeman, D. (2016). Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*. pp. 1659–1666.
- Palmer, M., Gildea, D. & Kingsbury, P. (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1), pp. 71–106.
- Pretkalnina, L., Rituma, L. & Saulite, B. (2018). Deriving Enhanced Universal Dependencies from a Hybrid Dependency-Constituency Treebank. In *Text*,

- Speech, and Dialogue*, volume 11107. Springer, pp. 95–105.
<https://www.researchgate.net/publication/327520269>.
- Spektors, A., Auzina, I., Dargis, R., Gruzitis, N., Paikens, P., Pretkalnina, L., Rituma, L. & Saulite, B. (2016). Tezaurs.lv: the largest open lexical database for Latvian. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*. Portoroz, Slovenia, pp. 2568–2571.
- Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S. & Tsujii, J. (2012). brat: a Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Avignon, France: Association for Computational Linguistics, pp. 102–107. <https://www.aclweb.org/anthology/E12-2021>.
- Xue, N. (2008). Labeling Chinese Predicates with Semantic Roles. *Comput. Linguist.*, 34(2), pp. 225–255. <http://dx.doi.org/10.1162/coli.2008.34.2.225>.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



Karst Exploration: Extracting Terms and Definitions from Karst Domain Corpus

Senja Pollak^{1,2}, Andraž Repar¹, Matej Martinc¹, Vid Podpečan¹

¹Jožef Stefan Institute, Ljubljana, Slovenia

²Usher Institute of Population Health Sciences and Informatics,
Edinburgh Medical School, Edinburgh, UK

E-mail: senja.pollak@ijs.si, repar.andraz@gmail.com, matej.martinc@ijs.si,
vid.podpecan@ijs.si

Abstract

In this paper, we present the extraction of specialized knowledge from a corpus of karstology literature. Domain terms are extracted by comparing the domain corpus to a reference corpus, and several heuristics to improve the extraction process are proposed (filtering based on nested terms, stopwords and fuzzy matching). We also use a word embedding model to extend the list of terms, and evaluate the potential of the approach from a term extraction perspective, as well as in terms of semantic relatedness. This step is followed by an automated term alignment and analysis of the Slovene and English karst terminology in terms of cognates. Finally, the corpus is used for extracting domain definitions, as well as triplets, where the latter can be considered as a potential resource for complementary knowledge-rich context extraction and visualization.

Keywords: karstology; term extraction; term embeddings; term alignment; definition extraction; triplets; specialized corpora

1. Introduction

The totality of means of expression in a language can be divided into general language and specialized language. Even if there is no distinct boundary between the two, it can be said that general language defines the sum of the means of linguistic expression encountered by most speakers of a given language, whereas specialized language goes beyond the general vocabulary based on the socio-linguistic or the subject-related aspect. The latter arises as a consequence of constant development and specialization in the fields of science, technology, and sociology (Svensen, 1993: 48-49). Similar to the definition of technical language by Svensen, in the context of terminology, specialized language, also called language for special purposes, is defined as a “language used in a subject field and characterized by the use of specific linguistic means of expression” (ISO 1087-1:2000).

If lexicologists and lexicographers mainly focus on words or lexemes, terminologists focus on terms, i.e., the words with a protected status when used in special subject domains (Pearson, 1998: 7). In contemporary approaches, the dichotomy ‘word-term’ no longer exists. For Kageura (2002) terms are functional variants of words. Cabré

Castellví (2003: 189) claims that all terms are words by nature and notes that “we recognize the terminological units from their meaning in a subject field, their internal structure and their lexical meaning”. According to Myking (2007: 86), the traditional terminology is concept-based and the new directions are lexeme-based.

A definition is a characterization of the meaning of the lexeme (Jackson, 2002: 93). It is “a representation of a concept by a descriptive statement which serves to differentiate it from related concepts” (ISO 12620:2009). The concept to be defined is called a *definiendum*, the part defining its meaning *definiens*, and the optional element (usually a verb) connecting the two parts in a sentence is called a hinge.

Granger (2012) highlights the six most significant innovations of electronic lexicography in comparison to the traditional methods: a) corpus integration, meaning the inclusion of authentic texts in the dictionaries; b) more and better data, since there are no more space limitations and one has the possibility to add multimedia data; c) efficiency of access (quick search and different possibilities of database organization); d) customization, meaning that the content can be adapted to the user’s needs; e) hybridization, denoting that the limits between different types of language resources—e.g., dictionaries, encyclopaedias, term banks, lexical databases, translation tools—are breaking down; and f) user input, since collaborative or community-based input is integrated. Similar can be claimed for terminological work, where recent approaches in terminology science consider knowledge (represented in texts) as conceptually dynamic and linguistically varied (Cabr , 1999; Kageura, 2002), and where novel methods in data acquisition, organization and representation, are being constantly developed. Knowledge can be extracted from specialized resources automatically, benefiting from the advances in the field of natural language processing. Moreover, attempts in dynamic, visual representation of domain knowledge have been proposed in recent years, e.g., EcoLexicon¹ (Faber et al., 2016).

In this work, we present the extraction of specialized knowledge from a corpus of karstology, i.e. an interdisciplinary domain at the intersection of geology, hydrology, and speleology. The domain is of high interest, as karst is possibly the most prominent geographical feature of Slovenia (with karst formations being some of popular tourist and natural attractions in the country). It is also an interesting example of how terminology is dynamically evolving in a cross-linguistic context. The literature published in English contains many local Slovenian scientific terms and toponyms for typical geomorphological karst structures, which makes it appropriate for research and identification of cognates, as well as homonym terms, with possible differences in meaning across cultures.

¹ <http://ecolexicon.ugr.es/en/index.htm>

Within the TermFrame² project, we focus on the specialized knowledge of karst science, and plan to develop methods that allow for context- and language-dependent investigation into a domain, relying on semi-automated tools. In this paper, we apply some of the methods that we have previously developed to a new domain, resulting in a repository of karst term and definition candidates in Slovene and English, contributing to the karstology terminological science. Next, we propose a word embedding based term list extension and triplet extraction method that can be used for visualization. These are novel components, contributing to terminological domain modelling.

This paper is structured as follows. After presenting the related work in automated specialized knowledge extraction in Section 2, we present the resources used (Section 3), methods (Section 4), results (Section 5) and conclude the paper with a discussion and plans for future work (Section 6).

2. Related work

Terminological work has undergone a significant change with the emergence of computational approaches resulting in semi-automated extraction of terms, definitions and other knowledge structures from raw text. Automatic terminology extraction has been implemented for various languages, including English (e.g., Sclano & Velardi, 2007; Frantzi & Ananiadou, 1999; Drouin, 2003) and Slovene (e.g., Vintar, 2010; Pollak et al., 2012), which are the languages in our corpus. In the last few years, word embeddings (Mikolov et al., 2013) have become a very popular natural language processing technique, and several attempts have already been made to utilize word embeddings for terminology extraction (e.g., Amjadian et al., 2016; Zhang et al., 2017). We use word embeddings techniques for extending term lists.

Numerous approaches have also been proposed in bilingual term extraction and alignment, including Gaussier (1998), Kupiec (1993), Lefever et al. (2009), Vintar (2010), Baisa et al. (2015), as well as Aker et al. (2013), who treat bilingual term alignment as a binary classification task. The modified version of the latter approach described in Repar et al. (2018), is also used in this paper.

Automated definition extraction approaches have been developed for several languages, including English (e.g., Navigli & Velardi, 2010), Slovene (e.g., Fišer et al., 2010) and multilingual methods (e.g., Faralli & Navigli, 2013). In our work we use a pattern-based definition extraction method for English and Slovene (Pollak et al., 2012).

In addition to definitions, authors have focused on extracting different types of semantic relations. Pattern-based approaches (Hearst, 1992; Roller et al., 2018), and machine learning techniques have also been proposed (cf. Nastase et al., 2013). In contrast to

² <http://termframe.ff.uni-lj.si/>

extracting predefined semantic relations, the Open Information Extraction (OIE) paradigm considers relations as expressed by parts of speech (Fader et al., 2011), paths in a syntactic parse tree (Ciaramita et al., 2005), or sequences of high-frequency words (Davidov & Rappoport, 2006). In our experiments we use the ReVerb triplet extractor by Etzioni et al. (2011).

This study presents the knowledge extraction steps within the TermFrame project, complementing previous work in karstology modelling presented in Vintar and Grčić-Simeunović (2017), and contributing to the emerging karstology knowledge base. The extracted knowledge was used in the frame-based annotation approach, identifying the semantic categories, relations and relation definitors in definitions of karst concepts, as presented in Vintar et al. (2019), as well as in topic modelling using term co-occurrence network presented in Miljković et al. (2019). The work is also closely related to Faber et al. (2016), a multilingual visual thesaurus of environmental science, which was developed following a frame-based, cognitively-oriented approach to terminology.

3. Resources

The corpus of karstology was constructed within the TermFrame project; it consists of Slovene, Croatian and English texts. We focus on the Slovene and English parts of the TermFrame corpus (v1.0). The English subcorpus contains cca. 1.6 M words and the Slovene one cca. 1 M words (see Table 1 for details).

	English	Slovene
Vocabulary size	64,079	73,813
Documents	24	60
Sentences	103,322	57,575
Words	1,673,132	1,041,475
Tokens	1,972,320	1,231,039
Type-to-token ratio	0.032	0.060

Table 1: Statistics for English and Slovenian subcorpora.

In addition, we are using a short gold standard list of Karst domain terms, called the QUIKK term base³. The QUIKK term base consists of terms in four languages, but for the purposes of our experiments, the Slovene and English term lists are used, containing 57 and 185 terms, respectively.

³ <http://islovar.ff.uni-lj.si/karst>

4. Methods

4.1 Term candidate extraction

First, we present the procedure of extracting terms by comparing the words in the noun phrases in the domain and reference corpora, and next we present a method using word embeddings to extend the list of terms.

4.1.1 Statistical term extraction

For extracting domain terms we use the LUIZ-CF term extractor (Pollak et al., 2012), which is a variant of LUIZ (Vintar, 2010) refined with scoring and ranking functions. The term extraction uses part-of-speech patterns for detecting noun phrases and compares the frequencies of words (lemmas) in the noun phrase in the domain corpus of karstology and a reference corpus.

The output is a list of term candidates in Slovene and English, above a selected frequency⁴ and/or termhood threshold. In addition, we applied the following filtering and term merging procedures:

- *Nested term filtering*: Nested terms are the terms that appear within other longer terms and may or may not appear by themselves in the corpus (Frantzi et al., 2000). As in Repar et al. (2019), the difference between a term and its nested term is defined by a frequency difference threshold: if a term in a corpus appears predominantly within a longer string, only the longer term is returned. If not (if a shorter term appears independently of a longer term more frequently than the set parameter), both terms are included in the final output.⁵
- *Stop word filtering*: If a term candidate is found on the stop word list, the term is excluded from the final list.⁶
- *Term merging by fuzzy matching*: Frequently, we can find terms that are extracted as separate terms but are in fact duplicates because they are written in different variants. This can be due to spelling variations (e.g., British and American English, using hyphenation or not), typos (which are relatively

⁴ We set minimum frequency to 15.

⁵ In our experiments, the parameter is set to 15 to match minimum frequency.

⁶ General stop words are not problematic, as they are frequent also in a reference corpus, and therefore not identified as terms by LUIZ-CF. However, the words specific to the academic discourse, are not frequent in general language and therefore often appear as extracted term candidates. To exclude them, we use the following short stop word list: *example, use, source, method, approach, table, figure, percentage, et, al., km.*

frequent when we deal with large text collections), errors due to pdf-to-text conversions etc. The proposed term merging is based on Levenshtein edit distance (Levenshtein, 1966): if two terms are nearly identical (default threshold is 95%), they will be merged and mapped to a common identifier. In addition, a rule which handles the case when two terms have a different prefix but the same tail and should not be recognized as duplicates can be applied.

4.1.2 Extending term lists with word embeddings

Word embeddings are vector representations of words, where each word is assigned a multidimensional vector of real numbers, characterizing the word based on the lexical context in which it appears. When vectors are computed on very large corpora, and especially with recent advances in models using neural networks, these representations have seen a huge success within various natural language processing tasks.

The embeddings capture certain degree of semantics, as words that are similar or semantically related are closer together in the vector space. Previous research conducted by Diaz et al. (2016) showed that embeddings can be successfully used for expanding queries on topic specific texts. In this research, we test if word embeddings can be used for a similar task of extending the gold standard term lists to find more domain terms. According to the research conducted by Diaz et al. (2016), embeddings trained only on small topic specific corpora outperform non-topic specific general embeddings trained on very large general corpora for the task of query expansion due to strong language use variation in specialized corpora. Therefore, we use the same approach for extending the term list and train custom embeddings on the specialized corpus instead of using pretrained embeddings.

In our experiments, we have trained FastText embeddings (Bojanowski et al., 2017) on the Slovenian and English karst subcorpora and use them to find the twenty closest words (according to cosine distance between embeddings) for the first fifty terms in the QUIKK term base⁷. These related words are sorted according to their proximity to the term and the first, second, tenth and twentieth ranked words are used in manual evaluation. Embeddings for multi-word terms are generated by averaging the word embeddings for each word in the term.⁸

⁷ To be exact, 50 English terms, and 47 Slovene terms, since only 47 Slovenian terms from the QUIKK term base appear in the Slovenian corpus.

⁸ There are several possible multi-word term aggregation approaches, such as summation of component word vectors, averaging of component word vectors, creating multi-word term vectors, etc. As comparing different techniques is beyond the scope of this study, we decided for the simple averaging technique, as previous research on this topic conducted on the medical domain (Henry et al., 2018) found no statistically significant difference between any multi-word term aggregation method.

4.2 Cognates detection and term alignment

English terms are mapped to Slovene equivalents using a data mining approach by Aker et al. (2013) reimplemented in Repar et al. (2018). Bilingual term alignment is treated as a binary classification, with a support vector machine classifier trained on various dictionary and cognate-based features that express correspondences between the words (composing a term) in the target and source languages. The first take advantage of dictionaries (Giza++) created from large parallel corpora, and the latter exploit string-based word similarity between languages (cf. Gaizauskas et al., 2012). In addition, the cognate-based features (see Table 2) allow users to identify cognate term pairs, which are interesting as karst terms in different languages clearly share their origin, but there exist also well-known examples of non-equivalent cognates (e.g., Slovene “dolina” vs. English “doline”).

Feature	Description
Longest Common Subsequence Ratio	Measures the longest common non-consecutive sequence of characters between two strings
Longest Common Substring Ratio	Measures the longest common consecutive string (LCST) of characters that two strings have in common
Dice similarity	$2 * \text{LCST} / (\text{len}(\text{source}) + \text{len}(\text{target}))$
Normalized Levensthein distance (LD)	$1 - \text{LD} / \max(\text{len}(\text{source}), \text{len}(\text{target}))$

Table 2: Cognate-based features used for term alignment.

4.3 Definition candidates extraction

We use the pattern-based module of the definition extractor (Pollak et al., 2012), which is available online.⁹ The soft pattern matching is used to extract sentences of forms NP is NP, NP refers to NP, NP denotes NP, etc., and the parameters contain language (EN, SL), as well as the position of the term in Slovene (if the term must be at the beginning of the sentence, after a larger set of predefined start patterns (our choice) or anywhere in a sentence).

4.4 Triplet extraction

As predefined definition patterns (cf. Section 4.3) were designed for extracting specific knowledge contexts, we complement the approach by open-relation extraction (this experiment is conducted only for English, as for Slovene the tools are not available).

⁹ <http://clowdflows.org/workflow/8165/>

We use ReVerb (Fader et al., 2011), which extracts relation phrases and their arguments and results in triplets of form:

<argument1, relation phrase, argument2>

We believe that in the case that argument1 and argument2 match domain terms, the triplets can be exploited as a method for extraction of knowledge-rich contexts (an alternative to definitions). They are also a useful input for visualization of terminological knowledge and can meet the needs of frame-based terminology, aiming at facilitating user knowledge acquisition through different types of multimodal and contextualized information, in order to respond to cognitive, communicative, and linguistic needs (Gil-Berrozpe et al., 2017). Previously, triplets have been used in other domains, e.g., in systems biology for building networks from domain literature (Miljković et al., 2012).

5. Evaluation setting and results

5.1 Term candidate extraction

5.1.1 Statistical term extraction

We extracted 4,397 English term candidates and 2,946 Slovene term candidates. A domain expert and a linguist specialized in terminology with high domain understanding manually evaluated all term candidates for Slovene and the top 1,823 (above a selected threshold)¹⁰ term candidates for English. The following categories were used:

- Not a term (label: 0)
- Karst term (label: 1)
- Broader domain terms (label: 2)
- Named entity (label: 3)

To distinguish between karst and broader domain terms, the following criterion is used. While karstology is in itself an interdisciplinary field, in TermFrame the focus is on karst geomorphology entailing surface and underground landforms, and karst hydrology

¹⁰ The reason for the discrepancy in the number of evaluated terms is that the evaluation for Slovene yielded a much lower number of terms (categories 1 or 2) in Slovene than in English. Since we need a large number of terms for additional steps, i.e. term alignment, we instructed the evaluators to process the full list of term candidates for Slovene. If we took the same number of top terms for Slovene as for English (top 1,823), we get the following results (cf. Table 3): Not a term: 1,187, Karst term: 140, Domain term: 174, Named entity: 220, Precision: 0.293.

with its typical forms and processes. Terms from neighbouring domains (geography, biology, geochemistry, etc.) which are not exclusive to karst are considered broader domain terms. In case of disagreement, the two annotators achieved consensus on the final category. As presented in Table 3, the resulting list of terms contains 351 karst terms for English and 158 for Slovene. The newly extracted karst terms, such as *cave*, *uvala*, *doline*, *denudation* describing landforms, processes, environment, etc., can serve for the extension of the manual QUIKK karstology term base, while for example the term candidate *karst region* is not considered a term because it is too generic and compositional, denoting a different underlying semantic relation (a region which contains karst).

The precision of term extraction is 0.516 for English and 0.235 for Slovene. For examples of terms in each category, see Table 4, while top terms sorted by termhood score for English and Slovene are presented in Tables 5 and 6, respectively.

Lang	Evaluated terms	Not a term	Karst term	Broader domain term	Named entity	Precision
Slovene	2,946	2,228	158	194	341	0.235
English	1,823	882	351	434	156	0.516

Table 3: Term extraction results. Precision is calculated as the sum of all three positive categories (1, 2, 3) divided by the number of evaluated terms.

In addition, we evaluate our filtering methods. All nested terms (306 for English, 105 for Slovene) removed by the nested term filtering are correctly eliminated, the stop word filter did not detect any terms which should not be removed, and all near duplicates (11 for English, 22 for Slovene) detected with the fuzzy match filter are also correct (e.g., “ground-water” was detected as a duplicate of “ground water”).

Lang	Not a term	Karst term	Broader domain term	Named entity
Slovene	dinarska smer	slepa dolina	naplavna ravnica	Planinsko polje
	ilovnat material	udornica	ravnovesna meja	Podgorski kras
	kataster jam	kalcijev karbonat	mehansko preperevanje	Gorski kotar
English	deepest cave	karst aquifer	sea level	Southeast Asia
	world heritage	subterranean water	carbonic acid	Castleguard Cave
	largest spring	phreatic cave	cave habitat	Central America

Table 4: Examples of term extraction evaluation categories.

Rank	Frequency	Term	Categorization
1	19269	cave	1
2	451	karst aquifer	1
3	522	karst area	1
4	459	cave system	1
5	314	dinaric karst	3
6	414	carbonate rock	1
7	348	cave passage	1
8	218	crna reka	3
9	271	karst system	1
10	209	karst feature	1
11	192	karst terrain	1
12	201	karst landscape	1
13	203	karst region	0
14	192	karst spring	1
15	564	united state	3
16	146	troglobitic specie	2
17	187	cave entrance	1
18	227	lava tube	2
19	169	cave sediment	1
20	164	karst rock	1

Table 5: Top 20 English karst term candidates with frequencies and categorization to karst terminology (1), broader domain terminology (2), named entity (3) or non-term (0).

5.1.2 Extending term lists with word embeddings

The method was tested on 47 English and 50 Slovene source terms (i.e. the terms from the gold standard list), for which out of the 20 most related words (according to the cosine distance between the source term and the related word), four per each source term were selected for evaluation (first, second, tenth and twentieth ranked words), resulting in 200 term-word pairs for English and 188 for Slovene.¹¹ Examples of ranked related words for five English and five Slovene terms are presented in Table 7.

¹¹ In this section, we intentionally name related words as words and not as terms, to contrast them to the gold standard list of terms to which they are compared. As shown in the evaluation, they can be evaluated as terms or not in the next step.

Rank	Frequency	Term	Categorization
1	1,966	nadmorska višina	0
2	9,543	jama	1
3	4,472	kras	1
4	6,359	voda	0
5	713	slepa dolina	1
6	4,481	dolina	0
7	405	brezstropa jama	1
8	2,948	apnenec	1
9	623	Pivška kotlina	3
10	2,573	sediment	0
11	3,418	dno	0
12	425	erozijski jarek	2
13	3,608	polje	1
14	2,770	rov	1
15	728	kraško polje	1
16	2,049	udornica	1
17	4,619	del	0
18	2,564	kamnina	2
19	507	suha dolina	1
20	3,882	oblika	0

Table 6: Top 20 Slovene karst term candidates with frequencies and categorization to karst terminology (1), broader domain terminology (2), named entity (3) or non-term (0).

Term	R1	R2	R10	R20
sinkhole	shakehole	suburban	sinkpoint	dump
aggressive water	aggressively	aggressiveness	qc	coldwater
epikarst zone	epikarstic	subcutaneous	cutaneous	epiphreatic
caprock sinkhole	sinkpoint	overbank	suburb	evacuation
seacave	seacoast	sealevel	vrulja	caveand
udornica	udornina	zapornica	koliševka	kamojstrnik
agresivna voda	sposoben	mehurček	skozi	preniči
epikras	epikraški	prenikujoč	epr	vadozen
vrtača	vrtačast	mikrovrtača	globel	neizravn
rečna jama	reža	narečen	mohoričev	vodokazen

Table 7: Examples of ranked related words for five English (upper five examples) and five Slovene (lower five examples) terms.

The two human evaluators evaluated the related words according to two criteria:

- Is the word a term?
- Semantic similarity to the term

The first criterion is measured on a scale with four nominal classes (see Section 5.1.1), while the second criterion uses a numerical scale from zero to ten, following the evaluation procedure of Finkelstein et al. (2002), where zero suggests no semantic similarity and ten suggests very close semantic relation (fractional scores were also allowed). The inter-annotator agreement between the two evaluators (according to the Cohen's kappa coefficient) is 0.689 for the first criterion and 0.513 for the second criterion for English, and 0.594 for the first criterion and 0.389 for the second criterion for the Slovene evaluation.

Table 8 presents the results for the evaluation of embeddings-based term extension. Out of 200 English term-word pairs, 112 were manually labelled as term-term pairs by at least one evaluator, which suggests that, at least for English, embeddings can be used for extending the term list. Out of these 112 related terms, 52 were labelled as karst specific terms by at least one evaluator. For Slovenian, the results are worse, since out of 188 term-word pairs only 69 were labelled as term-term pairs, and out of these only 36 are karst specific.

Out of 112 English term-term pairs, 62 were ranked first and second and 50 were ranked tenth and twentieth according to the cosine distance similarity. Out of 69 Slovenian term-term pairs, 39 were ranked first or second and 30 were ranked as tenth or twentieth. This suggests that words that have most similar embeddings to terms according to the cosine distance (rank 1 and rank 2) are also more likely to be terms themselves than words that have less similar embeddings (rank 10 and rank 20). Similar reasoning applies to karst specific term-term pairs, where for English 30 were ranked first or second and 22 were ranked tenth or twentieth. For Slovenian, 24 out of 36 were ranked first or second and 12 were ranked tenth or twentieth.

When it comes to semantic similarity, unsurprisingly better ranked related words were manually evaluated as semantically more similar. For example, the first ranked (most similar to terms according to the cosine distance) English related words got an average semantic similarity score¹² of 4.040 out of ten, and the first ranked Slovenian related words got an average semantic similarity score of 4.468. These are larger than the semantic similarity score averages of 2.610 and 3.064 for English and Slovenian related words ranked as twentieth, respectively. Another interesting observation is the fact that the average semantic similarity score is the highest for English karst specific term-terms pairs (5.702) and much lower if all the term-word pairs are considered (3.325). If we

¹² The semantic similarity score for each related word is calculated as an average between the two semantic similarity scores given by two evaluators.

consider all term-term pairs, the average semantic similarity score is 4.710. The same applies for Slovenian term-word pairs, with semantic similarity score average rising from 3.859 when all term-words pairs are considered, to 5.536 when only term-term pairs are considered, and up to 6.722 when only karst specific term-term pairs are considered.

We also measure the correlation between cosine distances and the semantic similarity scores for term-word pairs using Pearson and Spearman correlation coefficients. The correlation is generally low, the highest being measured for Slovenian Karst specific term-term pairs where the Pearson correlation reached the value of 0.341 and Spearman the value of 0.208. There was no correlation measured on Slovene term-term pairs and surprisingly, a small negative Pearson correlation was measured on Slovenian karst specific term-term pairs and a small negative Spearman correlation was measured on English pairs which were labelled as terms.

5.2 Cognate detection and term alignment

We evaluate the approach first on the QUIKK gold standard, where 100% precision and recall above 40% were obtained. Next, we also add to the QUIKK gold standard the terms extracted using the statistical method and term embeddings that were positively evaluated. The total list of 908 English terms and 391 Slovene terms were input to the term alignment algorithm. The resulting list of 93 aligned term pairs was manually evaluated. In this experiment, the precision was 77.42% (72 term alignments out of 93 were correct), while the recall could not be calculated, as the gold standard alignment was not available.

	English				Slovene			
All words	200				188			
Avg. sem. score	3.325				3.859			
Avg. cos. dist.	0.747				0.760			
Pearson corr.	0.181				0.231			
Spearman corr.	0.136				0.194			
	R1	R2	R10	R20	R1	R2	R10	R20
Distribution	50	50	50	50	47	47	47	47
Avg. sem. score	4.040	3.540	3.110	2.610	4.872	4.468	3.032	3.064
Terms	112				69			

Avg. sem. score	4.710	5.536
Avg. cos. dist.	0.757	0.771
Pearson corr.	0.176	-0.018
Spearman corr.	0.160	-0.016
	R1 R2 R10 R20	R1 R2 R10 R20
Distribution	32 30 29 21	17 22 15 15
Karst terms	52	36
Avg. sem. score	5.702	6.722
Avg. cos. dist.	0.761	0.780
Pearson corr.	0.151	-0.152
Spearman corr.	0.070	-0.067
	R1 R2 R10 R20	R1 R2 R10 R20
Distribution	16 14 15 7	12 12 5 7
Not Terms	88	119
Avg. sem. score	1.563	2.887
Avg. cos. dist.	0.734	0.753
Pearson corr.	-0.010	0.341
Spearman corr.	-0.110	0.208
	R1 R2 R10 R20	R1 R2 R10 R20
Distribution	18 20 21 29	30 25 32 32

Table 8: English and Slovenian embeddings evaluation according to two criteria described in Section 4.1.2. *Avg. sem. score* stands for the average of manually prescribed semantic similarity scores for each term-word pair, *Avg. cos. dist* stands for the average cosine distance, *Pearson corr.* is a Pearson correlation coefficient between the semantic similarity score and cosine distance values and *Spearman corr.* is a Spearman correlation coefficient between the semantic similarity score and cosine distance values.

As described in Section 4.2, karst terminology contains a considerable amount of cognates. See Table 9 for cognate values for Longest Common Substring Ratio, Longest Common Subsequence Ratio, Dice Similarity, and Normalized Levensthein Distance).

5.3 Definition candidate extraction

In total, 1,320 definition candidates were extracted for English, and 1,218 for Slovene. Definition candidates were manually validated by domain experts following two criteria: whether the sentence defines the concept, and whether the concept belongs to the domain of karstology. To distinguish between definitions and non-definitions the experts checked whether the sentence explains what the concept is, either by specifying its hypernym and a set of distinguishing features (analytical), or by listing its hyponyms (extensional), or by using another explanatory strategy (e.g., functional definitions). The definition candidates were then assigned one of the following three categories:

- Definitions of karst terms (Example: *Aggressiveness is an attribute of groundwater that corresponds to a chemical potential for mobilization of a dissolved matter from the rock.*)
- Definitions of broader domain terms (biology, geology etc.). (Example: *Exploration geophysics is the science of seeing into the earth without digging or drilling.*)
- Non-definitions (Example: *The oldest rocks are the sandstones of Permian age, which are only locally present.*)

English term	Slovene term	LCSTR	LCSSR	Dice	NormLD
mineralization	mineralizacija	0.71	0.79	0.71	0.79
salinization	salinizacija	0.67	0.75	0.67	0.75
nitrification	nitrifikacija	0.54	0.69	0.54	0.69
aggressive water	agresivna voda	0.25	0.63	0.27	0.50
karst plateau	kraška planota	0.27	0.60	0.29	0.40
karst	kras	0.20	0.60	0.22	0.40
marble	marmor	0.50	0.50	0.50	0.50
karst drainage	kraška drenaža	0.19	0.50	0.20	0.38
karst phenomena	kraški pojav	0.13	0.47	0.14	0.20
linear stream cave	linearna epifreatična jama	0.22	0.44	0.27	0.44

Table 9: Cognate scores for a sample of Slovene and English term pairs

As presented in Table 10, for English, out of 1,320 definition candidates 218 were evaluated as karst definitions, and an additional 187 as broader domain definitions. The precision of the definition extraction on karst domain is thus 0.16 for strictly karst domain definitions, and 0.31 for broader domain definitions (incl. karst definitions). For Slovene, there are 1218 definition candidates, out of which 260 are karst definitions and 166 are from broader domain. The precision for definition extraction for Slovene is thus 0.21 for strictly karst domain, and 0.35 for karst and broader domain.

	English	Slovene
Karst definitions	218	260
Broader domain definitions	187	166
Non definitions	915	792
All definition candidates	1320	1218

Table 10: Number of extracted definition candidates, evaluated as karst definitions, broader domain definitions and non-definitions.

The karst definitions were then used by domain experts and linguists in the scope of the TermFrame project for a fine-grained, annotation process, following frame-based terminology principles (Faber, 2015). The annotation principles and results are presented in Vintar et al. (2019), where several annotation layers are proposed: definition element layers (definiendum, definator and genus); semantic categories (top level concepts are landforms, processes, geomes, entities, instruments/methods) and relations (16 relations, such as `has_form`, `has_cause`).

5.4 Triplet extraction

The English subcorpus yielded 80,564 triplets. Below we list selected examples of relevant triplets that are closely related to the karst domain:

- <Karst areas, commonly lack, surface water>
- <Karst areas, have, numerous stream beds that are dry except during periods of high runoff>
- <Sinkholes located miles away from rivers, can flood, homes and businesses>
- <Karst areas, offer, important resources>
- <Some collapse sinkholes, develop, where collapse of the cave roof reaches the surface of the Earth>

The extracted triplets are analysed according to the most common relation patterns, to estimate their potential for extending predefined definition patterns. From the relation phrase part of the triplet, the verb is identified, showing the most frequent verb structures. We remove all stopwords from the relation phrase using a general list of 174 English stopwords. Table 11 lists 20 most frequent verb structures found in the processed 24 documents. The results show that many karst-specific relations can be detected (e.g., verbs related to different geological processes, such as *occur*, *develop* and *form*) but still many general verbs are also frequent. The frequent relations from triplets will be discussed in relation to the predefined set of relations used in definition frames annotation (cf. Vintar et al., 2019).

	verb	count		verb	count
1	found	1451	11	appear	336
2	occur	1347	12	consist	323
3	use	878	13	represent	321
4	form	811	14	locate	313
5	develop	787	15	include	312
6	know	646	16	contain	310
7	provide	528	17	made	306
8	show	428	18	result	295
9	take	397	19	depend	273
10	describe	337	20	extend	272

Table 11: 20 most frequent verb structures compiled from 80,564 triplets. Note that stopwords were removed from verb structures.

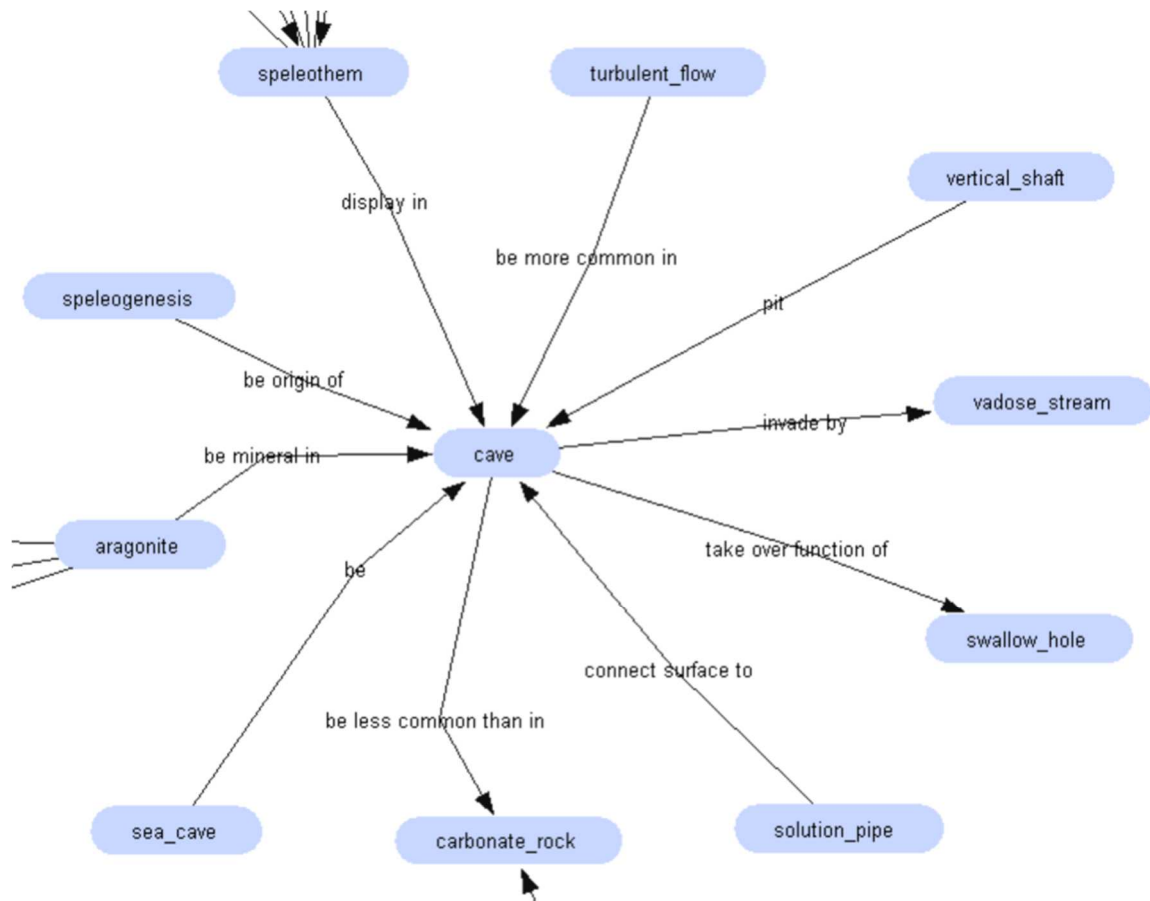


Figure 1: Visualization of a part of the triplet network. Prior to the visualization, relation phrases were lemmatized and the triplets were filtered according to the short gold standard list of Karst domain extended with an additional evaluated list of terms.

For visualization, after filtering the triplets by keeping only the ones where in a triplet <argument1, relation phrase, argument2> the two arguments are karst terms¹³, we construct a network where arguments are used as nodes and relation phrases as arcs. A visualization of a part of the triplet network obtained using Biomine network visualization tool (Eronen & Toivonen, 2012) is shown in Figure 1.

6. Conclusion and further work

We model domain knowledge utilizing a range of natural language processing techniques, including term extraction (using statistical methods, filtering and word embeddings), term alignment and cognates detection, definition extraction and triplet extraction. The proposed techniques form a pipeline for contemporary terminological work, relying on semiautomated processes for knowledge extraction from specialized domain corpora. Several modules in the pipeline rely on existing techniques, which were refined for the purposes of this work (e.g., term extraction), while we believe that the use of embeddings and triplets has not yet been sufficiently explored in the context of lexicography and terminography. The hypothesis was that embeddings offer not only a possibility of extending a list of terms, but also of grouping them to semantically related concepts, which can be of great value in the organization of domain knowledge (in term bases and similar resources), and also in contemporary lexicography resources.

We apply the proposed pipeline to a corpus of karst specialized texts. The main value of the evaluation steps of term and definition extraction is to obtain new gold standard karst knowledge resources that will be used in the scope of the TermFrame project for fine grained analysis and novel visual representation corresponding to the cognitive shifts in recent terminology science approaches. On the other hand, we believe that the evaluation of word embeddings opens new perspectives to e-lexicography and terminography, as it shows that popular techniques from natural language processing are relatively successful for automatically extending the gold standard term lists (cca. half of English and one third of Slovene terms being valid terms). The evaluation also shows that the semantic similarity score is higher for the closest matching words (considering cosine similarity between embeddings) than for the lower ranked words, which suggests that embeddings do in fact manage to capture some semantic relations despite a relatively small training corpus. On the other hand, the correlation between cosine similarity and manual similarity score is weak, which might indicate high variance in cosine similarity for related words for different terms. We believe that semantic information has a huge potential for contributing to the organization of term bases and visually interesting knowledge maps. In the same line, we illustrate how triplet extraction in combination with term matching can serve as a knowledge representation module used for visualization.

¹³ QUIKK terms and manually evaluated terms from Section 5.1.1.

In future work, we will consider extending the corpus by using web-crawling techniques. Next, our aim is to merge the pipeline to a set of services to support users in a knowledge extraction process, for populating term bases, as well as in knowledge visualization. We believe that such tools will contribute to better understanding of similarities and differences in terminological expression between languages, and support representations reflecting dynamic culture and language specific knowledge.

7. Acknowledgements

The work was supported by the Slovenian Research Agency through the core research programme (P2-0103) and research project Terminology and knowledge frames across languages (J6-9372). This work was supported also by the EU Horizon 2020 research and innovation programme, Grant No. 825153, EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media). The results of this publication reflect only the authors' views and the EC is not responsible for any use that may be made of the information it contains. We would also like to thank Š. Vintar, U. Stepišnik, D. Miljković and other members of the TermFrame project for their collaboration.

8. References

- Aker, A., Paramita, M. & Gaizauskas, R. (2013). Extracting bilingual terminologies from comparable corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pp. 402–411.
- Amjadian, E., Inkpen, D., Paribakht, T. & Faez, F. (2016). Local-Global Vectors to Improve Unigram Terminology Extraction. In *Proceedings of the 5th International Workshop on Computational Terminology (Computerm2016)*, pp. 2–11.
- Baisa, V., Ulipová, B. & Cukr, M. (2015). Bilingual terminology extraction in Sketch Engine. In *9th Workshop on Recent Advances in Slavonic Natural Language Processing*, pp. 61–67.
- Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5, pp. 135–146.
- Cabré, M.T. (1999). *Terminology: Theory, Methods, and Application*. Amsterdam, The Netherlands and Philadelphia, USA: John Benjamins Publishing.
- Cabré Castellví, M. T. (2003). Theories of Terminology: Their Description, Prescription and Explanation. *Terminology* 9 (2), p. 163–199.
- Ciaramita, M., Gangemi, A., Ratsch, E., Šaric, J. & Rojas, I. (2005). Unsupervised Learning of Semantic Relations Between Concepts of a Molecular Biology Ontology. In *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence (IJCAI'05)*, pp. 659–664.
- Davidov, D. & Rappoport, A. (2006). Efficient Unsupervised Discovery of Word

- Categories Using Symmetric Patterns and High Frequency Words. In *Proceedings of the 21th International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*. Sydney, Australia, pp. 297–304.
- Diaz, F., Mitra, B. & Craswell, N. (2016). Query expansion with locally-trained word embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Berlin, Germany, p. 367–377.
- Drouin, P. (2003). Term extraction using non-technical corpora as a point of leverage. *Terminology*, 9(1), pp. 99–117.
- Eronen, L. & Toivonen, H. (2012). Biomine: predicting links between biological entities using network models of heterogeneous databases. *BMC Bioinformatics*, 13(1), pp. 1–21.
- Etzioni, O., Fader, A., Christensen, J., Soderland, S. & Mausam (2011). Open Information Extraction: The Second Generation. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume One (IJCAI'11)*. Barcelona, Catalonia, Spain, pp. 3–10.
- Faber, P. (2015). Frames as a framework for terminology. In H. Kockaert & F. Steurs (eds.) *Handbook of Terminology*. John Benjamins, p. 14–33.
- Faber, P., León-Araúz, P. & Reimerink, A. (2016). EcoLexicon: new features and challenges. In *Proceedings of GLOBALEX 2016: Lexicographic Resources for Human Language Technology in conjunction with the 10th edition of the Language Resources and Evaluation Conference*, pp. 73–80.
- Fader, A., Soderland, S. & Etzioni, O. (2011). Identifying Relations for Open Information Extraction. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP'11)*. Edinburgh, United Kingdom: Association for Computational Linguistics, pp. 1535–1545.
- Faralli, S. & Navigli, R. (2013). A Java Framework for Multilingual Definition and Hypernym Extraction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Sofia, Bulgaria: Association for Computational Linguistics, pp. 103–108. <https://www.aclweb.org/anthology/P13-4018>.
- Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G. & Ruppín, E. (2002). Placing search in context: The concept revisited. *ACM Transactions on information systems*, 20(1), pp. 116–131.
- Fišer, D., Pollak, S. & Vintar, Š. (2010). Learning to Mine Definitions from Slovene Structured and Unstructured Knowledge-Rich Resources. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valletta, Malta, pp. 2932–2936.
- Frantzi, K., Ananiadou, S. & Mima, H. (2000). Automatic recognition of multi-word terms: the C-value/NC-value method. *International Journal on Digital Libraries*, 3(2), pp. 115–130.
- Frantzi, K. T. & Ananiadou, S. (1999). The C-Value/NC-Value Domain Independent Method for Multi-Word Term Extraction. *Journal of Natural Language*

- Processing*, 6(3), pp. 145–179.
- Gaizauskas, R., Aker, A. & Yang Feng, R. (2012). Automatic bilingual phrase extraction from comparable corpora. In *24th International Conference on Computational Linguistics*, p. 23.
- Gaussier, E. (1998). Flow Network Models for Word Alignment and Terminology Extraction From Bilingual Corpora. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (Coling-ACL)*, pp. 444–450.
- Gil-Berrozpe, J., León-Araúz, P. & Faber, P. (2017). Specifying Hyponymy Subtypes and Knowledge Patterns: A Corpus-based Study. In I. Kosem et al. (eds.) *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference*. Brno: Lexical Computing, pp. 63–92.
- Granger, S. (2012). Electronic Lexicography-from Challenge to Opportunity. In S. Granger & M. Pacqot (eds.) *Electronic Lexicography*, chapter Introduction. Oxford University Press, p. 1–15.
- Hearst, M. A. (1992). Automatic Acquisition of Hyponyms from Large Text Corpora. In *Proceedings of the 14th Conference on Computational Linguistics - Volume 2 (COLING'92)*, pp. 539–545.
- Henry, S., Cuffy, C. & McInnes, B. T. (2018). Vector representations of multi-word terms for semantic relatedness. *Journal of biomedical informatics*, 77, pp. 111–119.
- ISO 1087-1:2000 (2000). International Standard: Terminology Work — Vocabulary — Part 1: Theory and Application. Standard cited from the Glossary of Terminology Management of DG TRAD – Terminology Coordination Unit of European Parliament (Last accessed June 17, 2019). Standard. <http://termcoord.wordpress.com/glossaries/glossary-of-terminology-management/>.
- ISO 12620:2009 (2009). International Standard. Terminology and Other Language and Content Resources — Specification of Data Categories and Management of a Data Category Registry for Language Resources. Standard cited from ISOCat Web Interface (Last accessed December 1, 2013). Standard. <https://catalog.clarin.eu/isocat/interface/index.html>.
- Jackson, H. (2002). *Lexicography: An Introduction*. Routledge.
- Kageura, K. (2002). *The Dynamics of Terminology: A Descriptive Theory of Term Formation and Terminological Growth*. John Benjamins Publishing.
- Kupiec, J. (1993). An Algorithm for Finding Noun Phrase Correspondences in Bilingual Corpora. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL)*. Columbus, OH.
- Lefever, E., Macken, L. & Hoste, V. (2009). Language-Independent Bilingual Terminology Extraction from a Multilingual Parallel Corpus. In *Proceedings of the 12th Conference of the European Chapter of the ACL*, pp. 496–504.
- Levenshtein, V. I. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10, p. 707.

- Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. In *Proceedings to The International Conference on Learning Representations 2013*.
- Miljković, D., Kralj, J., Stepišnik, U. & Pollak, S. (2019). Communities of related terms in Karst terminology co-occurrence network. In I. Kosem et al. (eds.) *Proceedings of eLex 2019*, pp. 357-373.
- Miljković, D., Stare, T., Mozetič, I., Podpečan, V., Petek, M., Witek, K., Dermastia, M., Lavrač, N. & Gruden, K. (2012). Signalling Network Construction for Modelling Plant Defence Response. *PLOS ONE*, 7(12), pp. 1-18. <https://doi.org/10.1371/journal.pone.0051822>.
- Myking, J. (2007). No Fixed Boundaries. In A. Bassey (ed.) *Indeterminacy in Terminology and LSP: Studies in Honour of Heribert Picht*, chapter 6. Amsterdam, The Netherlands and Philadelphia, USA: John Benjamins Publishing, pp. 73-91.
- Nastase, V., Nakov, P., Séaghdha, D. Ó. & Szpakowicz, S. (2013). Semantic Relations Between Nominals. In G. Hirst (ed.) *Synthesis Lectures on Human Language Technologies*. London: Morgan & Claypool Publishers, pp. 1-119.
- Navigli, R. & Velardi, P. (2010). Learning Word-Class Lattices for Definition and Hypernym Extraction. In *Proceedings of the Forty-Eighth Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden, pp. 1318-1327.
- Pearson, J. (1998). Terms in Context. In E. Tognini-Bonelli & W. Teubert (eds.) *SCL Series, Vol. 1*. Amsterdam, The Netherlands and Philadelphia, USA: John Benjamins Publishing.
- Pollak, S., Vavpetič, A., Kranjc, J., Lavrač, N. & Špela Vintar (2012). NLP workflow for on-line definition extraction from English and Slovene text corpora. In J. Jancsary (ed.) *Proceedings of KONVENS 2012*. ÖGAI, pp. 53-60. Main track: oral presentations.
- Repar, A., Martinc, M. & Pollak, S. (2018). Machine Learning Approach to Bilingual Terminology Alignment: Reimplementation and Adaptation. In A. Branco, N. Calzolari & K. Choukri (eds.) *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Paris, France: European Language Resources Association (ELRA).
- Repar, A., Podpečan, V., Vavpetič, A., Lavrač, N. & Pollak, S. (2019). TermEnsembler: An ensemble learning approach to bilingual term extraction and alignment. *Terminology*, 25(1).
- Roller, S., Kiela, D. & Nickel, M. (2018). Hearst Patterns Revisited: Automatic Hypernym Detection from Large Text Corpora. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 358-363. URL <https://www.aclweb.org/anthology/P18-2057>.
- Sclano, F. & Velardi, P. (2007). TermExtractor: a Web Application to Learn the Common Terminology of Interest Groups and Research Communities. In *Proceedings of the 9th Conf on Terminology and Artificial Intelligence TIA 2007*,

- pp. 8–9.
- Svensen, B. (1993). *Practical Lexicography: Principles and Methods Of Dictionary Making*. Oxford University Press.
- Vintar, Š. (2010). Bilingual term recognition revisited: The bag-of-equivalents term alignment approach and its evaluation. *Terminology*, 16(2), pp. 141–158.
- Vintar, Š. & Grčić-Simeunović, L. (2017). Definition frames as language-dependent models of knowledge transfer. *Fachsprache: internationale Zeitschrift für Fachsprachenforschung, -didaktik und Terminologie*, 39(1/2), pp. 43–58.
- Vintar, Š., Saksida, A., Stepišnik, U. & Vrtovec, K. (2019). Knowledge frames in karstology: the TermFrame approach to extract knowledge structures from definitions. In I. Kosem et al. (eds.) *Proceedings of eLex 2019*, pp. 305–318.
- Zhang, Z., Gao, J. & Ciravegna, F. (2017). SemRe-Rank: Incorporating Semantic Relatedness to Improve Automatic Term Extraction Using Personalized PageRank. *arXiv preprint arXiv:1711.03373*.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



LexiCorp: Corpus Approach to Presentation of Lexicographic Data

Vladimír Benko

Slovak Academy of Sciences, L. Štúr Institute of Linguistics

Panská 26, 811 01 Bratislava, Slovakia

E-mail: vladimir.benko@juls.savba.sk

Abstract

We present an experiment aimed at integrating XML-encoded dictionary data with corpus processing tools. Tokenized, lemmatized and PoS-tagged, the dictionary data can be processed by a traditional corpus manager such as NoSketch Engine (NoSkE), with the main benefit being the availability of ad-hoc full-text queries, as well as queries restricted to certain structure elements, without having to know too much about the internals of the respective XML encoding. Loaded with data from several Slovak dictionaries, the beta version of the dictionary portal (referred to as LexiCorp) is already used by our lexicographers.

We demonstrate the LexiCorp operation in the “Simple Query” mode and the use of “Zone” attribute in queries. However, having in mind that all NoSkE functionalities are available, we can say that users of LexiCorp can now receive a powerful working tool.

As NoSkE is an open-source system and implementation of LexiCorp requires just a minor modification of dictionary data and NoSkE’s CSS style(s), this approach is applicable to similar lexicographic projects as well. Though not intended to be a replacement of a fully-fledged Dictionary Writing System, it can be conveniently used to supplement functionalities that may be missing there, such as the use of regular expressions, statistics based on XML attributes, and queries related to morphological forms of search expressions.

Keywords: Dictionary writing system; corpus manager; full-text querying; *NoSketch Engine*

1. Introduction

Two types of software systems are typically employed in compilation of dictionary entries. Dictionary Writing Systems (*DWSs*), such as *TLex*¹, *iLex*² or *Lexonomy*³, are used to define the respective entry structures and to fill them with the necessary data. Corpus managers, e.g., *CQPWeb*⁴ or *(No)Sketch Engine*^{5,6}, are needed to query corpora and to analyse, aggregate and process lexical evidence gathered out of them, especially if the corpora are really large. These two types of tools can cooperate to a certain extent to provide for partial automation of certain tasks, e.g., extracting suitable

¹ <https://tshwanedje.com/tshwanelex/>

² <http://groupbanker.dk/generic-en/index.htm>

³ <https://www.lexonomy.eu/>

⁴ <http://cwb.sourceforge.net/cqpweb.php>

⁵ <https://nlp.fi.muni.cz/trac/noske>

⁶ <https://www.sketchengine.eu/>

collocations or example sentences by means of the *TickBox Lexicography*⁷.

Our paper presents a different type of co-operation between dictionary data and a corpus manager, and describes an experiment in the framework of which we use corpus tools for the presentation of data of the *Dictionary of Contemporary Slovak Language*⁸ (*DCSL*, Jarošová & Benko, 2012) that is currently being compiled at our Institute.

2. The *DCSL* Project

Dictionary compilation is a rather time-consuming process. Producing a single-volume dictionary typically takes several years, and projects of multi-volume academic dictionaries may take even several decades to complete. This was also the case of the *DCSL*, whose preparatory phase was initiated already in mid-1990s, while the actual compilation of its first volume started in early 2000s. As of 2019, three *DCSL* volumes have been published (SSSJ1, 2016; SSSJ2, 2010; SSSJ3, 2016), two more volumes are currently in preparation, with the fourth volume being scheduled to be published in the end of the next year. The whole set is planned to consist of eight to nine volumes, which is most likely to occupy our lexicographic team for (at least) the next decade.

Partly due to historical reasons, our authors and editors do not work with the dictionary text in a “fully structured” format encoded in a generalized markup language, such as *SGML* or *XML*, and they instead use a light-weight markup language *LLML* (Benko, 2018). This is also one of the reasons why no “real” dictionary writing system (*DWS*) has been used yet for compilation of the *DCSL*.⁹

During the early “MS-DOS times” authors could prepare the text of the dictionary entries with any simple text editor, even with the built-in “*F4 Editor*” of Norton Commander¹⁰. With the advent of MS Windows, the most convenient editing environment has been provided by the popular *Notepad++* program¹¹ featuring user-definable syntax highlighting that could be easily adapted to our *LLML* syntax. Two sample entries as seen on the *Notepad++* screen are shown in Figure 1.

⁷ <https://www.sketchengine.eu/user-guide/user-manual/tickbox-lexicography/>

⁸ http://www.juls.savba.sk/pub_ssj.html

⁹ The *LLML* approach has been used for all lexicographic projects carried out by our Institute since early 1990s, with the advantage being the high level of compatibility of all the lexicographic data, as well as the associated custom software tools.

¹⁰ https://en.wikipedia.org/wiki/Norton_Commander

¹¹ <https://notepad-plus-plus.org/>

```

4175 |14960
4176 "lexikón" -nu/-na |pl. N| -ny |*m.| <gr.>
4177 {1} súhrnný zoznam slov z určitého odboru spracovaný
4178 encyklopedicky, výkladový náučný slovník: 'biografický
4179 l.; spoločenský l.' o etikete; 'l. slovenských dejín,
4180 obcí; detský obrázkový l.; výstavba hesla v lexikóne;
4181 vydávať encyklopédie a lexikóny'; |pren.| '65-ročného
4182 učiteľa pokladajú za živý lexikón.' [*NP 1982]
4183 vzdelaného, múdreho človeka
4184 {2} slovná zásoba jazyka, lexika, ktorou disponujú
4185 jeho používatelia, zásobáreň lexikálnych jednotiek:
4186 'jednotky lexikónu'
4187
4188 |14970
4189 "lexikónový" -vá -vé |*príd.|
4190 {0} vzťahujúci sa na lexikón, náučný slovník; typický
4191 pre lexikón: 'lexikónové diela; lexikónová definícia;
4192 l. spôsob výkladu'

```

Figure 1: Two *DCSL* entries with *LLML* markup as displayed by *Notepad++*.

It has been said that XML has *not* been used by the dictionary authors. It has been, however, used as an intermediate format during transformation of the dictionary text to the final printed and/or electronic form. The respective XML tags in this case represent typographical parameters, and can be easily mapped to typefaces, point sizes, colours, etc. Figure 2 shows an example of such XML code.

```

1 <en id="l01_lo_w1_014960" hword="lexikón">
2 <p class="main"><b1><h0><Sk>lexikón</Sk></h0></b1> <i3>-nu/-na</i3> <t0>pl.
  N</t0> <i3>-ny</i3> <t0>m.</t0> &lang;<t5>gr.</t5>&rang;
  <b8>1.</b8>&nbsp;&rtrif;&nbsp;<Sk>súhrnný zoznam slov</Sk>
  z&nbsp;<Sk>určitého odboru spracovaný encyklopedicky</Sk>, <Sk>výkladový
  náučný slovník</Sk>: <i0><Sk>biografický</Sk> l.; <Sk>spoločenský</Sk>
  l.</i0> o&nbsp;<Sk>etikete</Sk>; <i0>l. <Sk>slovenských dejín</Sk>,
  <Sk>obcí</Sk>; <Sk>detský obrázkový</Sk> l.; <Sk>výstavba hesla</Sk>
  v&nbsp;<Sk>lexikóne</Sk>; <Sk>vydávať encyklopédie</Sk> a
  <Sk>lexikóny</Sk></i0>; <t0>pren.</t0> <i0><Sk>65-ročného učiteľa pokladajú
  za živý</Sk> lexikón.</i0> <t4>[NP 1982]</t4> <Sk>vzdelaného</Sk>,
  <Sk>múdreho človeka</Sk></p>
3 <p class="sense"><b8>2.</b8>&nbsp;&rtrif;&nbsp;<Sk>slovná zásoba jazyka</Sk>,
  <Sk>lexika</Sk>, <Sk>ktorou disponujú jeho používatelia</Sk>, <Sk>zásobáreň
  lexikálnych jednotiek</Sk>: <i0><Sk>jednotky lexikónu</Sk></i0></p>
4 </en>
5
6 <en id="l01_lo_w1_014970" hword="lexikónový">
7 <p class="main"><b1><h0><Sk>lexikónový</Sk></h0></b1> <i3>-vá -vé</i3>
  <t0>príd.</t0> &rtrif;&nbsp;<Sk>vzťahujúci sa na lexikón</Sk>, <Sk>náučný
  slovník</Sk>; <Sk>typický pre lexikón</Sk>: <i0><Sk>lexikónové diela</Sk>;
  <Sk>lexikónová definícia</Sk>; l. <Sk>spôsob výkladu</Sk></i0></p>
8 </en>

```

Figure2: *DCSL* entries in “typographically motivated” XML notation.

3. Dictionary as a corpus

An XML-encoded dictionary is usually much more structured than a typical corpus. On the other hand, it *can* be treated as if it is a corpus. If processed by a standard tokenization and tagging pipeline for the respective language(s), it can be incorporated into a corpus manager without *too many* modifications needed.

The basic idea of our experiment is straightforward: as the procedures necessary to build and annotate (Slovak¹²) corpora not only do exist but they have been fine-tuned already, we just need to find a way to “force” the corpus manager to display the dictionary structure in a format the lexicographers are accustomed to, i.e., structured by entries and highlighting the respective entry elements by means of typographical devices (such as point size, bold, italics, and colour).

3.1 Why *NoSketch Engine*

Our decision has been motivated by several factors. Firstly, as heavy users of the *Sketch Engine* (Kilgarriff et al., 2014), our lexicographers are also reasonably familiar with the environment of *NoSketch Engine* (*NoSkE*, Rychlý, 2007), and no additional training is expected. Secondly, the user interface provides for complex types of queries by means of the *Corpus Query Language* (*CQL*), yet it also offers “structure-agnostic” full-text querying in the *Simple query* mode. And lastly, the *NoSkE* client allows a simple way to customize the formatting of the output though mapping the respective user-defined *XML* structures into suitable *CSS* styles. Moreover, as *NoSkE* is available under the open-source licence, we will be able to share our solution with other lexicographic projects.

The customized version of *NoSkE* containing the processed data as installed at our dictionary portal is further referred to as *LexiCorp*.

3.2 Preparing the data

Any XML-encoded dictionary data can be easily incorporated into *NoSkE*, after being converted to a compatible “vertical” format and subsequently processed by a standard corpus-processing pipeline. This contains the following steps:

- Tokenization by the *unitok*¹³ (Michelfeit et al., 2014) tool using a custom parameter file (to take into consideration the dictionary-specific abbreviations and tokens starting and ending with hyphens used to indicate suffixes and prefixes in inflected

¹² This applies, more or less, to any language with a morphosyntactic tagger available.

¹³ <http://corpus.tools/wiki/Unitok>

headword forms and elsewhere).

- Tagging by *TreeTagger*¹⁴ (Schmid, 1994) using a standard Slovak language model (Benko, 2016).
- Post-processing – fixing lemmatization and tagging issues for dictionary-specific out-of-vocabulary (*OOV*) tokens.
- Mapping native tags to a universal tagset¹⁵.
- Mapping the suitable corpus structure elements into <doc>, <p> and <s> structures used by default by the corpus manager (all other structures are preserved).
- Mapping dictionary structures into additional corpus attributes (to simplify certain types of queries).
- Indexing (“compilation”) by *NoSkE*.

3.3 Controlling the display

The standard *NoSkE* device for controlling the format of the richly structured corpora is the *DISPLAYCLASS* parameter that can be defined for each corpus structure contained in the corpus configuration file¹⁶. To make it operational, the appropriate *CSS* style has to be defined in the *view.css* file used by *NoSkE*. In a typical case, the respective dictionary *XML* structures have to be associated by a set of typographical parameters, such as typeface, point size and colour, which is fairly straightforward. Some *CSS* wizardry is needed only if some special effects (such as injections of newlines) are required.

4. First impressions

At the time of writing this paper (June 2019), the beta version of our *LexiCorp* installation contains data of all already published contemporary Slovak dictionaries produced by our Institute, as follows:

- Three volumes the *Dictionary of Contemporary Slovak Language* (SSSJ1, 2006; SSSJ2, 2010; SSSJ3, 2015)
- Live database of the *Orthographic-Grammatical Dictionary* (OGS, 2019)
- *Concise Dictionary of Slovak Language* (KSSJ, 4th Edition, 2003)
- Dictionary part of the *Rules of Slovak Orthography* (PSP, 4th Edition, 2013)
- Six volumes of the *Dictionary of Slovak Language* (SSJ, 1959–1968)
- Two volumes of the *Dictionary of Slovak Dialects* (SSN1 & SSN2, 1994; 2006).

¹⁴ <http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>

¹⁵ http://unesco.uniba.sk/aranea_about/aut.html

¹⁶ <https://www.sketchengine.eu/corpus-configuration-file-all-features/>

Besides that, *LexiCorp* also contains data of two volumes of *DCSL* (SSSJ4, SSSJ5) that are currently being in preparation, as well as merged data of all dictionaries (less the dialectal ones). The *LexiCorp* home page¹⁷ is shown in Figure 3.



Slovenská akadémia vied | Jazykovedný ústav Ľ. Štúra
Oddelenie súčasnej lexikológie a lexikografie

Lexikografický portál s podporou NoSketch Engine

SSSJ	Id	Zdrojové dáta	Veľkosť	i	Q
Lexicon Linguae Slovacae Contemporalis I ad V	1c–5c	SSSJ I až V	5,38 M	i	Q
Lexicon Linguae Slovacae Contemporalis I ad III	1c–3c	SSSJ I až III	3,71 M	i	Q
Lexicon Linguae Slovacae Contemporalis IV et V	4c–5c	SSSJ IV a V	1,66 M	i	Q
Iné slovníky					
Lexicon Orthographico-Grammaticum	og	OGS	963 K	i	Q
Lexicon Breve Linguae Slovacae	b4	KSSJ (4. vydanie)	1,07 M	i	Q
Lexicon Praeceptae Orthographiae Slovacae	p3	PSP (3. vydanie)	298 K	i	Q
Lexicon (Aborigineum) Linguae Slovacae I ad VI	1a–6a	SSJ I až VI	3,73 M	i	Q
Lexicon Dialectorum Slovacarum I et II	1d–2d	SSN I a II	2,43 M	i	Q
Spojené slovníky					
Omnia Lexica Slovaca		SSSJ + OGS + KSSJ + PSP + SSJ	11,4 M	i	Q

Užitočné odkazy

[Dotazovací jazyk CQL \(En\)](#)

[Dokumentácia Sketch Engine \(En\)](#)

Figure 3: The *LexiCorp* home page

To demonstrate the basic functionality of the system, we will show some examples.

The easiest way to work with *LexiCorp* is to use the *Simple query* mode of *NoSkE* that is suitable for most “structure-agnostic” searches. For example, if we want to find all entries containing a certain phrase, we could do it like this (see Figure 4):

Corpus: Lexicon Linguae Slovacae Contemporale I ad V (01jun19) 5.38 M ▼

Simple query: majúci veľký Make Concordance

[Query types](#)
[Context](#)
[Text types](#)
?

Figure 4: Simple query

Part of the first result screen can be seen in Figure 5.

¹⁷ The *LexiCorp* portal containing data of the dictionaries currently being in preparation is not accessible to the general public, a *LexiCorp* demo site, however, containing the GNU Collaborative International Dictionary of English (*GCIDE*, <http://gcide.gnu.org.ua/>) is already available at: <http://lexicorp.juls.savba.sk/guest>.

The screenshot shows the Lexicon Linguae Slovacae Contemporane I ad V (01jun19) 5.38 M interface. The search bar contains 'majúci veľký'. The left sidebar shows navigation options: Home, Search, Word list, Corpus info, My jobs, User guide, Save, Make subcorpus, View options, KWIC, Sentence, Sort, Left, Right, Node, References, Shuffle, Sample, Filter, Sub-hits, 1st hit in doc, Frequency, Node tags, Node forms, Doc IDs, Collocations, Visualize, and Menu position. The main content area displays search results for 'majúci veľký' (84 results, 15.62 per million). The results are listed in a table with columns for the dictionary ID (1c), the headword, and the definition. The definitions are in Slovak and include examples and references. The search expression 'majúci veľký' is highlighted in red in the original image.

Query	Results
majúci veľký 84 (15.62 per million)	
Page 1 of 5	Go Next Last
1c bachratý	bachratý -tá -té 2. st. -tejší príd. expr. 1. ▶ (o človeku, o zvierati) majúci veľké brucho, tučný; syn. bruchatý: b. chlap; bachratá žena; tvoj kapor je bachratejší ako môj; kravy bachraté, biele, s ozrutnými rohami [L. Ballek]
1c bajúzatý	bajúzatý -tá -té príd. hovor. expr. ▶ majúci veľké fúzy, fúzatý: b. doktor; bajúzatý, na slivku zosušený ujo [P. Vilikovský]; Po tmavohnedých tapetách sa strašidelne ťahal vinič a zazerali olejové portréty bajúzatých pánov. [J. Blažková]
1c bezodný	2. expr. ▶ majúci veľkú intenzitu, mieru, neohraničený rozsah, bezhraničný, nekonečný: bezodná fantázia, túžba; bezodné sklamanie, zúfalstvo; Prepadol sa kamsi do bezodnej prachovej búrky, čo zúrila naokolo. [J. Puškáš]; Aká je tu zábava? Iba bezodná nuda. [V. Mináč]
1c bláznivý	4. expr. ▶ majúci veľkú intenzitu, veľmi silný, prudký: b. strach; b. pracovný kolotoč; b. život; bláznivá odvaha, radosť, zamilovanosť; bláznivé tempo; Chcem opäť získať ten bláznivý pocit, že aj duše dvoch ľudí aspoň tak zapadajú do seba ako ich telá. [I. Hudec]
1c bohatý ¹	bohatý ¹ -tá -té 2. st. -tší príd. 1. ▶ majúci veľký majetok, dostatok materiálnych prostriedkov; syn. zámožný, majetný; op. chudobný: b. človek; b. štát; bohatá obec; pochádza z bohatej rodiny; Sníva o bohatom manželovi. [Inet 2003]
1c bruchatý	bruchatý -tá -té príd. 1. ▶ majúci veľké brucho: b. dedko; bruchatá postava, figúra; bruchatí otcovia rodín; Primáš, nie starý, ale bruchatý s príjemným altovým hlasom predspevuje. [L. Ťažký]
1c bruškatý	bruškatý -tá -té príd. 1. expr. ▶ majúci väčšie brucho: bruškatá postava; Notár bol bruškatý päťdesiatnik. [L. Ondrejov]; Na planine hrajú bruškati páni futbal. [Vč 1983]
1c ceekatý	ceekatý, cecnatý -tá -té príd. 1. hovor. expr. ▶ (o zvieratách) majúci veľké cecky: ceekatá, cecnatá koza, krava
1c ceekatý	2. pejor. ▶ (o žene) majúca veľké prsia; prsnatá: ceekatá matróna; obrázky ceekatých mladých báb
1c cenný	cenný -ná -né 2. st. -nejší príd. 1. ▶ majúci veľkú materiálnu, najmä peňažnú hodnotu; syn. drahocenný, hodnotný: c. šperk; cenné starožitnosti, ozdoby; cenné kožušiny, obrazy; cenné minerály; cenné suroviny; cenné stroje; dostať od niekoho c. dar; práv., ekon. c. papier dokument, z ktorého vyplýva právo al. majetkový nárok vlastníka voči osobe, ktorá ho vydala (akcie, dlhopisy, kupóny, vkladové listy a pod.); pošt. c. list, balík s udanou cenou

Figure 5: **Majúci veľký** (“having large”)

We can notice here several things. The “Short reference” on the left part of the display contains the *Id* of the dictionary (“1c” meaning the first volume of SSSJ), and the respective headword. The display mode was set to “Sentence”, which has been mapped to one sense in this particular dictionary.

As the dictionary text has been lemmatized (and also morphosyntactically tagged), *LexiCorp* can find the respective expression in *all* morphological forms – this is something a traditional *DWS* is typically not capable of.

The search expression is a phrase typically contained in dictionary definitions, and is hard to find elsewhere – we, therefore, do not have to bother about the dictionary structure while querying.

The entry is structured by means of typography, leaving *NoSkE* to highlight search expression by the default red colour.

Similarly, it is quite easy to make a query based on an abbreviation (See Figure 6).

Page 1 of 2 Go Next Last	
1c albatros ¹	albatros¹ -sa pl. N a -sy m. (angl. < špan., port. < arab.) ▶ veľký morský svetlý vták podobný čajke, s dlhými a tenkými krídlami, ktoré umožňujú vynikajúco plachtiť: <i>tokajúce albatrosy; nádherný let albatrosov</i> ; zool. a. <i>sťahovavý Diomedea exulans</i>
1c albatros ²	albatros² -sa pl. N -sy m. (angl. < špan., port. < arab.) 1. ▶ športové lietadlo: <i>lietať na albatrosoch</i> ; <i>technická prehliadka albatrosov</i>
1c ananás	ananás -su pl. N -sy m. (port. < indián.) 1. ▶ tropická rastlina s pichľavými mečovitými listami v listovej ružici poskytujúca veľké chutné šťavnaté plody: <i>pestovať a</i> ; bot. a. <i>pestovaný Ananas sativus</i>
1c autodafé	autodafé neskl. s. (port.) 1. hist. ▶ (v Španielsku a Portugalsku) verejné vyhlásenie inkvizitného rozsudku, po ktorom nasledovalo odvolanie bludu al. odsúdenie heretika na smrť (obyč. upálením): <i>veľkolepé a</i> ; <i>Bosorka, ktorú práve upaľujú. Autodafé</i> . [N. Tanská]
1c bajadéra	bajadéra [-d-] -ry -dér ž. (fr. < port.) ▶ indická chrámová tanečnica: <i>bajadéry ovešané zlatom a diamantmi</i> ; <i>obdivovať umenie bajadér</i> ⚙ <i>bajadérka -ky -rok ž. zdrob.</i>
1c banán	banán -na/-nu pl. N -ny m. (port., špan. < afr.) 1. ▶ žltý podlhovastý jedlý dužinatý plod banánovníka: <i>zrelý, nezrelý b.</i> ; <i>vôňa banánov</i> ; <i>pochutnať si na banánoch</i> ; <i>Máte rada flambované banány?</i> [H. Zelinová]
1c banánovníkovité	banánovníkovité -tých pl. spodst. s. (port., špan. < afr.) bot. ▶ čeľaď jednoklíčnolistových bylín pochádzajúcich z tropických oblastí s nepravým kmeňom tvoreným listovými pošvami, s obrovskými listami a voskovožltými kvetmi (Musaceae)
1c barok	barok -ka m. (fr. < port.) 1. ▶ európsky umelecký sloh v 17. a 18. stor. uplatňujúci sa najmä v architektúre a vo výtvarnom umení, vyznačujúci sa veľkoleposťou výzdoby, vyumelkovanosťou tvarov; epocha tohto slohu: <i>včasný, vrcholný, neskorý b.</i> ; <i>klasicizujúci b.</i> ; <i>b. v strednej Európe</i> ; <i>monumentalita a nádhera baroka</i> ; <i>Toto dielo tematicky čerpá z obdobia európskeho baroka a jeho historických udalostí</i> . [LT 1998]
1c betel	betel [-t-] -lu L -li pl. N -ly m. (port. < drávid.) 1. ▶ tropický popínavý krík, ktorého plody slúžia ako korenie s dráždivým účinkom, bot. piepor betelový <i>Piper betle</i>
1c bonz ¹	bonz¹ -za pl. N -zovia m. (port. < jap.) ▶ budhistický mních: <i>hlavný b.</i> ; <i>bonzovia s vyholenými hlavami, oblečení v oranžových tógach</i> ; <i>V štyridsiatke som teda odišiel do pagody a odvtedy som bonzom</i> . [L. Moneol]

Figure 6: **Port.** (Words of Portuguese origin)

Or, just a combination of metalanguage elements (see Figure 7).

pl. N -ci

Lexicon Linguae Slovacae Contemporane I ad V (01jun19) 5.38 M

vladob

Query **pl\., N, -ci** 58 (10.78 per million)

Page 1 of 3 Go Next Last

1c|besedujúci **besedujúci** -ceho pl. N -ci m. ▶ kto beseduje, kto sa zúčastňuje na besede; syn. besedník: *hlavný b. bol minister školstva*; *viaceri besedujúci mali rovnaký názor*; *besedujúci už neboli schopní vecne debatovať*; *Z náhodného besedujúceho sa vykľúče laický aktivista*. [Sme 1998] ⚙ **besedujúca** -cej pl. N -ce G -cich ž.

1c|budúci² **budúci²** -ceho pl. N -ci m. hovor. ▶ budúci manžel; syn. nastávajúci; op. bývalý: *to je môj b.*; *prišla aj so svojím budúcim*; *Berte si svojho budúceho, slečinka Zacharovie*. [M. Krno]

1c|cestujúci² **cestujúci²** -ceho pl. N -ci m. ▶ kto vykonáva cestu dopravným prostriedkom; syn. pasažier: *platiaci, neplatiaci c.*; *čakáreň pre cestujúcich*; *vyzvať cestujúcich na nástup, výstup*; *pripútať niektorých cestujúcich zapnúť im bezpečnostné pásy*; *starať sa o cestujúcich*; *obchodný c. zástupca, agent firmy* ⚙ **cestujúca** -cej pl. N -ce G -cich ž.

1c|cvok² **cvok²** -ka pl. N -ci /-kovia G -kov m. (nem.) subšt. pejor. ▶ pomätený, nepričetný človek; syn. magor, mešuge: *on je tak trochu c.*; *urobil si zo mňa totálneho cvoka*; *A ja mu na to, že je cvok, a on sa mi smial, že pri tebe scvokatiem aj ja*. [P. Andruška]

1c|čakajúci **čakajúci** -ceho pl. N -ci m. ▶ kto práve na niečo čaká: *č. na autobus, vlak*; *postaviť sa medzi čakajúcich*; *V hĺbke úzkej chodby sa vlnil živý had čakajúcich*. [G. Rothmayerová]; *Aha, medzi čakajúcimi v rade je aj ústredná postava nášho príbehu*. [V. Bednár] ⚙ **čakajúca** -cej pl. N -ce G -cich ž.

1c|ďalejslúžiaci **ďalejslúžiaci** -ceho pl. N -ci m. ▶ (prv) kto ostal slúžiť v armáde aj po skončení základnej vojenskej služby: *dobrovoľne ď.*; *Včera ste sa vôbec neženirovali, keď ste si dali dupľu omáčky a tri čaje – ako ďalejslúžiaci*. [R. Fabry]; *Zamrzol na vojenčine ako ďalejslúžiaci, hoci bol sedliaciško každým nervom*. [P. Karvaš]

1c|debatujúci **debatujúci** [d-] -ceho pl. N -ci m. ▶ kto sa zúčastňuje na debate; syn. debatér, diskutér: *názory debatujúcich sa rôznia*; *počúvať, prerušiť debatujúcich*; *zišlo sa tam niekoľko debatujúcich*; *v sále postávali hlúčky debatujúcich*; *Keďže hudobný automat je v predsieni za rohom, debatujúci nevidia, čo sa pri ňom robí*. [V. Bednár] ⚙ **debatujúca** -cej pl. N -ce G -cich ž.

Figure 7: **Pl. N -ci** (Words with a particular form in the plural nominative case)

5. The second round

Though users *could* use the *CLQ* mode of *NoSkE* to look up expressions and strings within the various dictionary structure fields, such as headword, definition, example, etc., this would not be a good solution in our situation as our lexicographers are rather reluctant to learn anything “too abstract”.

We therefore decided to employ the *part-of-speech* (*PoS*) *filter* of *NoSkE* that can be set for *Lemma* and *Word form* queries. (See Figure 8).

Figure 8: PoS filter

The *PoS filter* is based on mapping morphological tags provided by tagger into “readable” names of PoS defined in the corpus configuration file.

As *NoSkE* “does not care” about the actual values assigned to PoS, this functionality can be used to filter any attribute attached to the respective token(s), if appropriate mappings are supplied. In our case, the mappings were based on entry structure elements, such as headword, definition, example, etc.

So that the user would not be confused, we changed the “*PoS*” string in the menu to “*Zone*”, which was, in fact, the only modification of *NoSkE* source code necessary (see Figure 9).

Figure 9: Query within the *heslo* (“headword”) zone

Using this functionality, the user does not need to know the names of the respective XML elements that encode the particular “zones”, which makes the system more

accessible also for linguists not directly involved in the dictionary compilation.

In our example, the regex functionality of *NoSkE* is used to look up for all headwords related to lexicography in all dictionaries stored in *LexiCorp*, and the “1st hit in doc” filter is applied to get rid of multiple occurrences of entries caused by run-on headwords. The result is shown in Figure 10.

Query lexikog.* , b[1-6] 24 > Filter all but first hit in document 12 (1.05 per million) ⓘ	
2c lexikograf	lexikograf -fa pl. N -fi m. ▶ odborník v lexicografii, v tvorbe slovníkov: <i>lexikografi dvojazyčných slovníkov; teoreticky fundovaní lexikografi; práca lexikografov; lexikografi sú najčastejšími používateľmi korpusov</i> ⓘ lexikografka -ky -fiek ž.: <i>slovník pripravil tím skúsených lexikografiek</i>
2c lexikografia	lexikografia -ie ž. (gr.) ▶ vedecká disciplína jazykovedy zaoberajúca sa teóriou a tvorbou slovníkov, spracovaním slovnej zásoby v podobe slovníka; tvorba slovníkov: <i>dvojazyčná, viacjazyčná l.; terminologická l. terminografia; súčasná lexikológia a l.; dejiny slovenskej lexikografie; počítačová l. spracovanie slovnej zásoby jazyka pomocou počítačových nástrojov</i>
2c lexikografický	lexikografický prisl. ▶ z hľadiska lexicografie, slovníkovej tvorby; lexikografickým, slovníkovým spôsobom; <i>syn. slovníkovo: l. opísať slovnú zásobu slovenčiny; dobre l. spracovaný slovník; zachytiť nové slovo l.</i>
2c lexikografický	lexikografický -ká -ké príd. ▶ súvisiaci s lexicografiou, tvorbou slovníkov, s lexicografmi; charakteristický pre lexicografiu: <i>l. výskum; lexikografická práca; lexikografické diela, príručky; l. kolektív; l. výklad slov; lexikografické spracovanie nárečia; Lexikografické riešenie nemusí byť jediné, ale musí pravdivo odrážať jazykovú realitu.</i> [KS 1994]
og lexikograf	lexikograf -fa pl. N -fi m.
og lexikografia	lexikografia -ie ž.
og lexikografický	lexikografický prisl.
og lexikografický	lexikografický -ká -ké príd.
og lexikografka	lexikografka -ky -fiek ž.
b4 lexikografia	lexikografia -ie ž. lingv. odbor zaoberajúci sa spracovaním slov v slovníkoch, slovníkárstvo ⓘ lexikograf -a mn. -i m. odborník v lexicografii, slovníkár; lexikografka -y -fiek ž.; lexikografický príd.: <i>l-é dielo; lexikografický prisl.</i>
p3 lexikografia	lexikografia -ie ž. ⓘ lexikograf -a mn. -i/-ovia m.; lexikografka -y -fiek ž.; lexikografický ; lexikografický prisl.
1a lexikografia	lexikografia , -ie ž. 1. zostavovanie slovníkov, slovníkárská práca; náuka o zostavovaní slovníkov; 2. vydané slovníky, slovníková literatúra ⓘ lexikograf , -a, mn. č. -i/ovia m. slovníkár; lexikografický príd.: <i>l-á práca, l-á štúdia, l-á prax; lexikografický prisl.</i>

Figure 10: Lexicography-related headwords in all current *LexiCorp* dictionaries.

6. “Bells and whistles”

The beta version of *LexiCorp* turned to be a success and was “warmly welcomed”, not only by the lexicographic team members but by also by the other researchers at our Institute. This was probably the reason why no large-scale modification has been attempted since. Here are some small points to mention.

6.1 Merged dictionary data

After the unification of structures of our dictionaries, we managed to merge all data into one resource that can be conveniently looked up with a single query as shown in the previous chapter. Due to the unified format used to represent our dictionaries (Benko, op. cit.), this operation was relatively easy to perform. We must admit, however, that this needs not be the case if new dictionaries with more richly structured entries are to be incorporated into *LexiCorp*.

6.2 Typography

The graphical representation is very important when dictionary data are displayed on a computer screen. We made a series of experiments aimed at improving the legibility of the output. As a consequence we decided to change of the default sans-serif typeface used by *NoSkE* for displaying the concordances (i.e., the dictionary entries) to a serif one that better distinguishes between Roman and italicized text within the entries. As all our users work on Microsoft Windows machines, we opted for a standard Windows *Georgia*¹⁸ font that is known to have been designed with screen readability in mind.

Paper versions of our dictionaries use several special characters (custom created by a font editor) to introduce special sections of entry, such as lexicalized expressions, idioms, run-ons, etc. Some of these characters do not even have a similarly looking Unicode equivalent. To make the problem of displaying these characters easier to solve, we decided to substitute them for different ones (sometimes not even resembling the original glyphs) selected from the *Font Awesome*¹⁹ icon collection, that is used internally by *NoSkE* and therefore already installed in the system.

The text colours of the respective dictionary zones were chosen to be compatible with those used within the dictionary production environment (Benko, 2018), i.e., so that the lexicographers would see them as familiar.

A *LexiCorp* logo and a favicon have also been designed, so that the Portal had a unified “look”.

6.3 Dictionary names

Similarly to naming convention within the Aranea web corpora project (Benko, 2014), the respective dictionaries were assigned “language neutral” (Latin) names²⁰, as well as two-character *Ids* that are displayed along with the headwords in the “short reference” zone at the left side of the output screen.

7. Conclusion and further work

The experiment presented in this work proved the feasibility of our approach. The server component of *NoSkE* proved to be more than adequate for the task. The problem of the client is that is “too good”, i.e., contains too many features not necessary for typical dictionary look-ups that may confuse (especially inexperienced) users. It could

¹⁸ [https://en.wikipedia.org/wiki/Georgia_\(typeface\)](https://en.wikipedia.org/wiki/Georgia_(typeface))

¹⁹ <https://fontawesome.com/>

²⁰ It may be interesting to note that in the territory of today’s Slovakia Latin was used as an official language until the middle of the 19th century.

be, however, a good start for building a specialized client – this is, however, beyond our capacity. We are willing, however, to provide our know-how and data structures to anyone interested.

Readers may be wondering what could be the advantages of using *LexiCorp* instead of a full-fledged DWS. We are, however, not arguing in favour of using it *instead*, but rather *in parallel*. We hope that the main advantages have been addressed in the previous text.

As the compilation of *LexiCorp* out of the source dictionary data at our site is now fully automated and lasts less than 20 minutes, it can be performed regularly, theoretically even on the daily basis so that the lexicographers can work with fresh data every day. At the present stage, however, we have found that once a week is fully sufficient.

8. Acknowledgement

This work has been, in part, funded by the VEGA Grant Agency, Project No. 2/0017/17.

9. References

- Benko, V. (2016). Feeding the “Brno Pipeline”: The Case of Araneum Slovacum. In *RASLAN: Recent Advances in Slavonic Natural Languages Processing, The Tenth Workshop*, Brno: Tribun, 2016, vol. 10, pp. 19–27. ISBN 978-80-263-1340-3. ISSN 2336-4289.
- Benko, V. (2018). In Praise of Simplicity: Lexicographic Lightweight Markup Language. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (eds.) *The XVIII EURALEX International Congress Lexicography in Global Contexts. The book of abstracts*. Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani, pp. 118.
- Benko, V. Aranea: Yet another Family of (Comparable) Web Corpora. (2014). In P. Sojka et al. (eds.) *Text, Speech, and Dialogue. 17th International Conference, TSD 2014 Brno, Czech Republic, September 8–12, 2014, Proceedings*. Cham – Heidelberg – New York – Dordrecht – London: Springer. ISBN 978-3-319-10816-2.
- Jarošová, A. & Benko, V. (2012). The Dictionary of the Contemporary Slovak: A Product of Tradition and Innovation. Vladimír Benko. In R. V. Fjeld & J. M. Torjusen (eds.) *Proceedings of the 15th EURALEX International Congress 7–11 August, 2012 Oslo*. Oslo: Department of Linguistics and Scandinavian Studies, University of Oslo, 2012, pp. 257–261. ISBN 978-82-303-2095-2.
- Kilgarrieff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, 1, pp. 7–36.

- Michelfeit, J., Pomikálek, J. & Suchomel. (2014). V Text Tokenisation Using unitok. In *8th Workshop on Recent Advances in Slavonic Natural Language Processing*, Brno, Tribun EU, pp. 71–75. 2014.
- Rychlý, P. (2007). Manatee/Bonito - A Modular Corpus Manager. In *1st Workshop on Recent Advances in Slavonic Natural Language Processing*. Brno: Masaryk University, 2007, pp. 65–70.
- Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In: *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.

Dictionaries:

- KSSJ: *Krátky slovník slovenského jazyka, 4th Edition*. (2003). Eds. J. Kačala, M. Pisárčiková & M. Považaj. Bratislava: Veda, vydavateľstvo Slovenskej akadémie vied. ISBN 80-224-0750-X.
- OGS: *Ortograficko-gramatický slovník.A – Ž (používateľská verzia Slovníka súčasného slovenského jazyka)*. (2019). Eds. M. Sokolová & A. Jarošová. Available only online at <http://lex.juls.savba.sk/>
- PSP: *Pravidlá slovenského pravopisu, 4th Edition*. (2013). Ed. M. Považaj. Bratislava: Veda, vydavateľstvo Slovenskej akadémie vied. ISBN 978-80-224-1331-2.
- SSJ: *Slovník slovenského jazyka I–VI*. (1959–1968). Ed. Š. Peciar. Bratislava: Vydavateľstvo SAV.
- SSN1: *Slovník slovenských nářečí A–K*. (1994), Ed. I. Ripka. Bratislava: Veda, vydavateľstvo Slovenskej akadémie vied. ISBN 80-224-0183-8.
- SSN2: *Slovník slovenských nářečí L–P (povzchádzať)*. Eds. A. Ferencíková – I. Ripka. Bratislava: Veda, vydavateľstvo Slovenskej akadémie vied. ISBN 80-244-0900-6.
- SSSJ1: *Slovník súčasného slovenského jazyka. A–G*. (2006). Eds. K. Buzássyová & A. Jarošová. Bratislava: Veda, vydavateľstvo Slovenskej akadémie vied. ISBN 978-80-224-0932-4
- SSSJ2: *Slovník súčasného slovenského jazyka. H–L*. (2011). Eds. A. Jarošová & K. Buzássyová. Bratislava: Veda, vydavateľstvo Slovenskej akadémie vied. ISBN 978-80-224-1172-1
- SSSJ3: *Slovník súčasného slovenského jazyka. M–N*. (2015). Ed. A. Jarošová, Bratislava: Veda, vydavateľstvo SAV. ISBN 978-80-224-1485-2.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



Porting a Crowd-Sourced German Lexical Semantics Resource to OntoLex-Lemon

Thierry Declerck^{1,2}, Melanie Siegel³

¹ German Research Center for Artificial Intelligence, Stuhlsatzenhausweg 3,
66123 Saarbrücken, Germany

² Austrian Centre for Digital Humanities, Sonnenfelsgasse 19, 1010 Vienna, Austria

³ Darmstadt University of Applied Science, Max-Planck-Str. 2, 64807 Dieburg, Germany
E-mail: declerck@dfki.de, melanie.siegel@h-da.de

Abstract

In this paper we present our work consisting of mapping the recently created open source German lexical semantics resource “Open-de-WordNet” (OdeNet) into the OntoLex-Lemon format. OdeNet was originally created in order to be integrated in the Open Multilingual Wordnet initiative. One motivation for porting OdeNet to OntoLex-Lemon is to publish in the Linguistic Linked Open Data cloud this new WordNet-compliant resource for German. At the same time we can with the help of OntoLex-Lemon link the lemmas of OdeNet to full lexical descriptions and so extend the linguistic coverage of this new WordNet resource, as we did for French, Italian and Spanish wordnets included in the Open Multilingual Wordnet collection. As a side effect, the porting of OdeNet to OntoLex-Lemon helped in discovering some issues in the original data.

Keywords: Open Multilingual Wordnet; OntoLex-Lemon; OdeNet; Lexical Semantics

1. Introduction

Wordnets are well-established lexical resources with a wide range of applications in various Natural Language Processing (NLP) fields, like Machine Translation, Information Retrieval, Query Expansion, Document Classification, etc. (Morato et al., 2004). For more than twenty years they have been elaborately set up and maintained by hand, especially the original Princeton WordNet of English (PWN) (Fellbaum, 1998). In recent years, there have been increasing activities in which open wordnets for different languages have been automatically extracted from other resources and enriched with lexical semantics information, building the so-called Open Multilingual Wordnet (OMW) (Bond & Paik, 2012), which is merging more than 35 open wordnets that are linked through the Collaborative Interlingual Index (CILI) (Bond & Foster, 2013; Bond et al., 2016). The resources in OMW are of different coverage and do not always contain the same amount of information, as for example many resources are lacking definitions (or “glosses”), contrary to the PWN resource, or example sentences.

Recently we made some experiments to enrich OMW resources with morphological resources. The resources we were dealing with are “WOLF (Wordnet Libre du Français)” for French, “ItalWordNet” for Italian and “Multilingual Central Repository” for

Spanish (this resource also contains wordnets for the Catalan, Basque and Galician languages).¹ In order to link those OWM resources to full lexical and morphological descriptions we first map them onto the OntoLex-Lemon model (Cimiano et al., 2016), which is a de facto standard for the representation of lexical data in the Web (McCrae et al., 2017), especially in the Linguistic Linked Open Data cloud.²

Up until very recently no German resources were included in the OMW collection, which requires the data to be equipped with an open and free licence. This condition is probably the reason why GermaNet is not included in OMW. GermaNet is a manually well-designed WordNet resource for German (Hamp & Feldweg, 1997).³ But GermaNet is not equipped with the type of license required by OMW.

In this context, a new German lexical semantics resource with the name “Open German WordNet” (OdeNet)⁴ has been developed with the aim to be included as the first open German WordNet into the Open Multilingual Wordnet.⁵

This paper is organised as follows. In Section 2 we present the OntoLex-Lemon model. In Section 3 we give some more details on the OMW resources we mapped to OntoLex-Lemon in order to link them to corresponding morphological resources. The result of this mapping is shown in Section 4. The OdeNet resource is described in some detail in Section 5. We describe in Section 6 the current state of the representation of OdeNet data in OntoLex-Lemon, and the issues in the original data we discovered through this mapping exercise.

2. OntoLex-Lemon

The OntoLex-Lemon model was originally developed with the aim to provide a rich linguistic grounding for ontologies, meaning that the natural language expressions used in the description of ontological elements are equipped with an extensive linguistic description.⁶

This rich linguistic grounding includes the representation of morphological and syntactic properties of lexical entries as well as the syntax-semantics interface, i.e. the meaning of these lexical entries with respect to an ontology or to specialized vocabularies. The main organizing unit for those linguistic descriptions is the lexical

¹ See Sagot and Fišer (2008), Pianta et al. (2002), Toral et al. (2010) and Gonzalez-Agirre et al. (2012), respectively.

² See <http://linguistic-lod.org/> and also Chiarcos et al. (2012).

³ See also <http://www.sfs.uni-tuebingen.de/GermaNet/> for more details.

⁴ See <https://github.com/hdaSprachtechnologie/odenet> for more details.

⁵ See http://compling.hss.ntu.edu.sg/omw20/omw_wns for more details.

⁶ See McCrae et al. (2012), Cimiano et al. (2016) and also https://www.w3.org/community/ontolex/wiki/Final_Model_Specification.

entry, which enables the representation of morphological patterns for each entry (a MWE, a word or an affix). The connection of a lexical entry to an ontological entity is marked mainly by the denotes property or is mediated by the LexicalSense or the LexicalConcept properties, as represented in Figure 1, which displays the core module of the model.

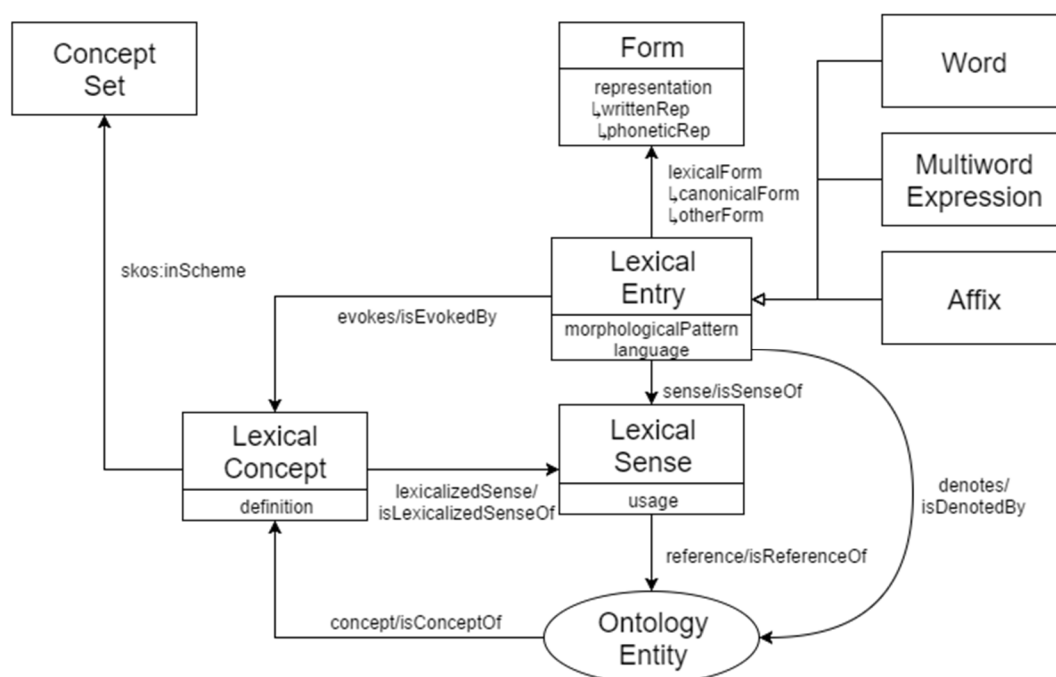


Figure 1: The core module of OntoLex-Lemon: Ontology Lexicon Interface. Graphic taken from <https://www.w3.org/2016/05/ontolex/>.

OntoLex-Lemon builds on and extends the *lemon* model (McCrae et al. (2012)). A major difference is that OntoLex-Lemon includes an explicit way to encode conceptual hierarchies, using the SKOS standard.⁷ As can be seen in Figure 1, lexical entries can be linked, via the `ontolex:evokes` property, to such SKOS concepts, which can represent WordNet synsets. This structure is paralleling the relation between lexical entries and ontological resources, which is implemented either directly by the `ontolex:reference` property or mediated by the instances of the `ontolex:LexicalSense` class.⁸ The “sets of

⁷ SKOS stands for “Simple Knowledge Organization System”. SKOS provides “a model for expressing the basic structure and content of concept schemes such as thesauri, classification schemes, subject heading lists, taxonomies, folksonomies, and other similar types of controlled vocabulary” (<https://www.w3.org/TR/skos-primer/>).

⁸ Quoting from Section 3.6 “Lexical Concept” <https://www.w3.org/2016/05/ontolex/>: “We [...] capture the fact that a certain lexical entry can be used to denote a certain ontological predicate. We capture this by saying that the lexical entry denotes the class or ontology element in question. However, sometimes we would like to express the fact that a certain lexical entry evokes a certain mental concept rather than that it refers to a class with a formal interpretation in some model. Thus, in *lemon* we introduce the class *Lexical Concept* that represents a mental abstraction, concept or unit of thought that can be lexicalized by a given collection of senses. A lexical concept is thus a subclass of *skos:Concept*.”

cognitive synonyms (synsets)”⁹, that Princeton WordNet (PWN) describes, seems to be best modelled by the `ontolex:LexicalConcept` class, while the `ontolex:LexicalSense` class is meant to represent the bridge between lexical entries and ontological entities (which do not necessarily have semantic relations between them).

3. Open Multilingual WordNet

The three Open Multilingual Wordnet resources (for French, Italian and Spanish) we were dealing with are available at the Open Multilingual Wordnet (OMW) page.¹⁰ OMW is an initiative that brings together wordnets in different languages, which are linked through the Collaborative Interlingual Index (CILI). As stated on the web page of OMW, those wordnets are of different quality, and some of those were in fact extracted from different types of language resources. OMW provided for some corrections and for an harmonization of such resources, and published them in a uniform tabular format, which is displayed below, exemplified here by entries from the Italian OMW resource:

```
08388207-n ita:lemma nobiltà
08388207-n ita:lemma aristocrazia
08388207-n ita:lemma patriziato
08388207-n ita:def_0 l'insieme degli aristocratici
08388207-n ita:def_1 l'insieme dei nobili
...
14842992-n ita:lemma terra
14842992-n ita:lemma terreno
14842992-n ita:lemma suolo
14842992-n ita:def_0 parte superficiale della
    crosta terrestre sulla quale si sta o si
    cammina
14842992-n ita:exe_0 si piegò con fatica per
    raccogliere da terra i sacchetti, pronta a
    salire sull'autobus
14842992-n ita:exe_1 il tizio comincio' a rotolarsi
    per terra in preda a dolori lancinanti
```

In the two examples displayed above, the uniform tabular format of OMW delivers information on the synset IDs (08388207-n and 14842992-n), which include the part-of-speech (“n”) of the associated lemma(s). The nominal lemmas associated with the synset-ID 08388207-n are “nobiltà” (nobility), “aristocrazia” (nobility, aristocracy) and “patriziato” (aristocracy). The nominal lemmas associated with the synset-ID

⁹ Quoted from <https://wordnet.princeton.edu/>.

¹⁰ See <http://compiling.hss.ntu.edu.sg/omw/>. For more details see also Bond and Paik (2012).

14842992-n are “terra” (earth, land, soil), “terreno” (ground, terrain, soil) and “suolo” (land, earth, ground). If available, definitions (“glosses”) are provided (marked with the feature “ita:def”), as well as examples (marked with the feature “ita:exe”).¹¹

This tabular format is used for all the OMW data sets. This makes it easier to map OMW data to a formal representation that supports the interoperability and interlinking of language resources. The next section shows the result of the mapping of OMW resources to OntoLex-Lemon.

4. Mapping the OMW Resources to OntoLex-Lemon

As mentioned earlier, the format generated by the OMW initiative is very convenient with regard to mapping onto more complex representation frameworks. A Python script was implemented for porting the OMW data sets to OntoLex-Lemon.

A design decision was to extract only the synset information and to encode the synsets as instances of the `LexicalConcept` class of OntoLex-Lemon. As we expect to have the lemmas present in already existing lexicons, we will just link the synsets to those lemmas, which are encoded as instances of the OntoLex-Lemon `LexicalEntry` class. This way we achieve a higher level of modularity. Since the synsets are now encoded as instances of the `LexicalConcept` class, each synset-ID gets a Unique Resource Identifier (URI), and does not have to be repeated for each lemma it is associated with, but can just link to those via the OntoLex-Lemon property `isEvokedBy`, as seen in Figure 1. This way we have also a more compact (graph-based) representation as in the original representation of the OMW data.

We have now 38,512 such instances of `LexicalConcept` for Spanish, 15,553 for Italian, and 59,091 for French.¹²

In Listing 1.1 we show examples of the OntoLex-Lemon encoding of two synsets for Spanish. The lemmas associated with these synsets are “cura”. In Section 2, we explain how in OntoLex-Lemon the synsets are linked to the lemmas, which are differentiated in the OntoLex-Lemon representation,¹³ which we add here, but not in the original OMW file, as in OMW the lemmas are just literals and not real lexical entries, associated with more complex linguistic information, additionally to PoS.

¹¹ We observe that using this type of text format for representing the data, one has to repeat the relevant information (for example the synset-ID) for each line introducing a lemma associated with the synset.

¹² The lower number for the Italian resource is due to the fact that we consider only the subset of *ItalWordNet* that has been curated by OMW.

¹³ Depending on the view on the word “cura” (meaning *cure* or *priest*, if the gender of the word is feminine or masculine) we can have either one lexical entry or two. Taking into consideration the distinct genders and etymologies for “cure”, we decided to have two entries.

```

: synset_spawn-13491616-n
  rdf : type ontalex : LexicalConcept ;
  ontalex : isEvokedBy : lex_cura -13491616-n ;
  skos : inScheme : spawnnet ;
.

: synset_spawn-10470779-n
  rdf : type ontalex : LexicalConcept ;
  ontalex : isEvokedBy : lex_cura -10470779-n ;
  skos : inScheme : spawnnet ;
.

: lex_cura -13491616-n a ontalex : LexicalEntry ;
  lexinfo : gender lexinfo : masc ;
  lexinfo : partOfSpeech lexinfo : noun ;
  ontalex : evokes : synset_spawn-13491616-n ;
  ontalex : canonicalForm : form_cura ;
  ontalex : otherForm : form_cura_plural .

: lex_cura -10470779-n a ontalex : LexicalEntry ;
  lexinfo : gender lexinfo : fem ;
  lexinfo : partOfSpeech lexinfo : noun ;
  ontalex : evokes : synset_spawn-10470779-n ;
  ontalex : canonicalForm : form_cura ;
  ontalex : otherForm : form_cura_plural .

```

Listing 1.1: The OntoLex-Lemon representation of two Spanish synsets with the corresponding lemmas

Current work is dedicated in enriching the three wordnets encoded in OntoLex-Lemon with further morphological semantic information. For this we already mapped the French, Italian and Spanish morphological resources included in the MMmorph data sets (Petitpierre & Russell, 1995) into OntoLex-Lemon,¹⁴ and we are bridging the two types of data sources.

5. The Open-de-WordNet (OdeNet)

The “Open-de-WordNet” (OdeNet)¹⁵ initiative is intended as a contribution to the Open Multilingual Wordnet Initiative. It is a WordNet for the German language under an

¹⁴ This mapping is described in Declerck and Racioppa (2019).

¹⁵ <https://github.com/hdaSprachtechnologie/odenet>.

open license (CC BY-SA 4.0). The main source for the synset entries is the OpenThesaurus German synonym lexicon.¹⁶ OpenThesaurus compiled approximately 120,000 entries in a crowd sourcing procedure. OdeNet transferred those data to synsets in the Global WordNet format.¹⁷ Subsequently, the resulting synsets were enriched with part-of-speech (PoS) information, semantic identifiers from OMW were identified and hierarchy relations were added.

As mentioned above, PoS information is associated with the synsets. We observe that only four PoS categories are used: Adjectives, Nouns, Verbs and “p”, which seems to be attributed to all synset/lemma combinations not being one of the three other categories. This strategy is not satisfying, and we are working on mapping all the “p” tagged lemmas to existing entries in a German lexicon in order to further specify their PoS. We also observe that phrasal multi-word units are also equipped with one of those PoS tags. In most cases this is sensible and could be accepted, as with “in Rechnung stellen” (*to bill*) or “Abschied nehmen” (*say goodbye*),¹⁸ but led to errors with idioms, as with “das geht auf keine Kuhhaut” (*this is impossible*), which cannot be marked as a verb (or as a verb phrase).

A difficulty related with the presence of such multi-word units (MWUs) for the lemmas associated with the synsets is the fact that very few morphological and lexical data sets have such MWUs as their lemmas or headwords, so that it can be hard to automatically map a lemma of OdeNet to a German lexical or morphological resources and therefore some manual work will be needed to encode such multi-word units in the OntoLex-Lemon representation. A segmentation algorithm can be helpful in this case, relating the basic components of a MWU to existing headwords in a lexicon.

Another issue with the OdeNet data is the fact that a high number of definitions associated with the synsets are only in English, as they have been first imported from the Princeton WordNet. Those definitions still need to be translated or adapted to German, preferably by a human expert.

The lemmas are also translated into English and so mapped to PWN via the semantic multilingual identifier (ili). For example “Flügel;Tragfläche;Flugzeugflügel” is translated with “wing”, which is annotated in PWN with the multilingual semantic ID “i61201”. This feature is important as it can ensure the cross-linking of OdeNet to other wordnets in OMW.

For the example “Flügel;Tragfläche;Flugzeugflügel” (*wing*) we have in the OdeNet

¹⁶ <https://www.openthesaurus.de/> and the Open Multilingual WordNet English¹⁷ resource. OpenThesaurus is a large resource, generated and updated by the crowd.

¹⁷ See <http://globalwordnet.github.io/schemas/>.

¹⁸ But in fact we would prefer to categorize those expressions as being verb phrases.

format the following lexical entries and the corresponding entry for the synset:

```

<LexicalEntry id="w3226">
  <Lemma writtenForm="Flügel" partOfSpeech="n"/>
  <Sense id="w3226\_648-n" synset="odenet-648-n"/>
  <Sense id="w3226\_4974-n" synset="odenet-4974-n"/>
  <Sense id="w3226\_8657-n" synset="odenet-8657-n"/>
  <Sense id="w3226\_9783-n" synset="odenet-9783-n"/>
  <Sense id="w3226\_10207-n" synset="odenet-10207-n"/>
  <Sense id="w3226\_11256-n" synset="odenet-11256-n"/> </LexicalEntry>
<LexicalEntry id="w39183">
  <Lemma writtenForm="Tragfläche" partOfSpeech="n"/>
  <Sense id="w39183\_9783-n" synset="odenet-9783-n"/>
</LexicalEntry>

<LexicalEntry id="w39184">\\
  <Lemma writtenForm="Flugzeugflügel" partOfSpeech="n"/>\\
  <Sense id="w39184\_9783-n" synset="odenet-9783-n"/>\\
</LexicalEntry>

<Synset id="odenet-9783-n" ili="i61201" partOfSpeech="n" dc:description="one of
the horizontal airfoils on either side of the fuselage of an airplane">
  <SynsetRelation target='odenet-3131-n' relType='holo\_ part'/>
  <SynsetRelation target='odenet-18647-n' relType='hyponym'/> </Synset>

```

From the 36,000 OdeNet synsets, about 20,000 contain links to OMW. Approximately 10,000 hyponymy relations and 2,650 antonymy relations are inserted.

In a first evaluation 7% of the PoS entries and 18% of the ili entries were not correct. There is also a need to add more relations and to correct existing ones. With the porting to OntoLex-Lemon we hope, among other things, to discover other issues for OdeNet entries that need correction.

6. Porting OdeNet to OntoLex-Lemon

In order to make OdeNet available in the Linguistic Linked Open Data cloud¹⁹ we need to transform its encoding format (compliant to the GWA²⁰ WordNet XML DTD²¹) to

¹⁹ <http://linguistic-lod.org/llod-cloud>, see also Chiarcos et al. (2012).

²⁰ “GWA” stands for Global WordNet Association. See <http://globalwordnet.org/>.

²¹ <http://globalwordnet.github.io/schemas/WN-LMF-1.0.dtd>.

an RDF²² representation. As the target representation framework we have chosen the OntoLex-Lemon model,²³ the core module of which is depicted in Figure 1.

This model is not only the de-facto standard for representing lexical data in the Linked Data framework, but it also includes a property called `ontolex:lexicalConcept`, which is very important for representing the relation between WordNet synsets and lexical data.²⁴ A key issue we had to handle with the original crowd-sourced data was that additional textual information was added to the headword, and our script for transforming the OdeNet data to OntoLex-Lemon had to clean the headword field and encode the additional information in a “comment” field. A second issue is related to the improper use of part-of-speech (PoS) information, as soon as the data was not about a noun, a verb or an adjective (the main part-of-speech information in WordNet dictionaries). We filtered out all the entries marked with PoS “p” and will link the entries to well-established German lexical data in the Linguistic Linked Data cloud in order to extract the correct PoS information. We also mapped some OdeNet codes into the LexInfo vocabulary for PoS and semantic relations.²⁵

As for now, we have in the OntoLex-Lemon encoding of OdeNet 120,012 lexical entries, the same number of lexical senses and 36,192 synsets, which are encoded as *ontolex:LexicalConcepts* and included in an SKOS²⁶ based conceptual hierarchy, supporting also the description of lexical semantic relations between synsets, like synonymy, hyponymy, etc.

It is interesting to notice that 44,506 entries contain a blank and can therefore be considered as Multi Word Expressions (MWEs). And if we add to this figure all the 14,080 compound entries²⁷ we note that approximately half of the lexical entries in the OntoLexLemon representation can be considered as segmentable lemmas.

We give now some details on the OntoLex-Lemon encoding of the first entry in OdeNet, which is “Kernspaltung” (*nuclear fission*). This example is a compound word, which we need to segment in order to be able to represent its components. This representation is supported by the Decomp module of OntoLex-Lemon, which is displayed in Figure 2. First we display the original OdeNet XML representation for “Kernspaltung”:

²² RDF stands for “Resource Description Framework”, see also <https://www.w3.org/RDF/>.

²³ See Cimiano et al. (2016) and <https://www.w3.org/2016/05/ontolex/>.

²⁴ See the section “Lexical Linkset” in https://www.w3.org/community/ontolex/wiki/Final_Model_Specification.

²⁵ See <https://www.lexinfo.net/ontology/2.0/lexinfo> and also Cimiano et al. (2011).

²⁶ See <https://www.w3.org/2004/02/skos/> for more details.

²⁷ This figure was computed merely by comparison with the list of split nominal compounds offered by the GermaNet project on its web page: http://www.sfs.uni-tuebingen.de/GermaNet/documents/compounds/split_compounds_from_GermaNet13.0.txt, We expect to have a larger number of compounds by applying a decomposition algorithm, not only to nominal entries.

```

<LexicalEntry id="w1">
  <Lemma writtenForm="Kernspaltung"
    partOfSpeech="n"/>
  <Sense id="w1_1-n" synset="odenet-1-n"/>
</LexicalEntry>
<LexicalEntry id="w2">
  <Lemma writtenForm="Kernfission"
    partOfSpeech="n"/>
  <Sense id="w2_1-n" synset="odenet-1-n"/>
</LexicalEntry>

```

Lexical senses are grouped in synsets, i.e., groups of word senses with the same meaning. Hierarchical relations are introduced as synset relations:

```

<Synset id="odenet-1-n" ili="i107577"
  partOfSpeech="n" dc:description="a
nuclear reaction in which a massive
nucleus splits into smaller nuclei with
the simultaneous release of energy">
<SynsetRelation target='odenet-5437-
n' relType='hypernym'/>
</Synset>

```

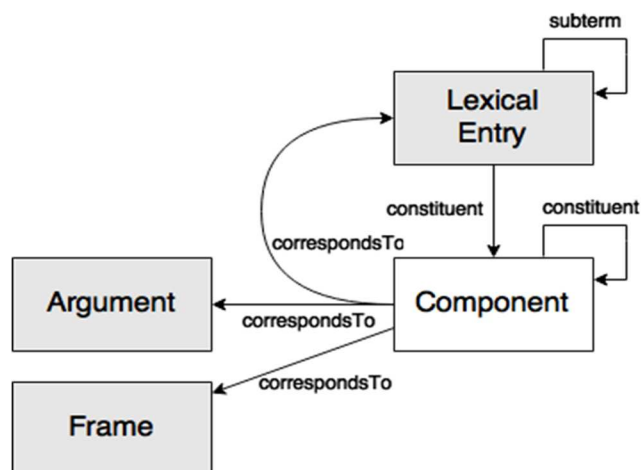


Figure 2: The Decomposition module of OntoLex-Lemon. Graphic taken from <https://www.w3.org/2016/05/ontolex/>.

In the following Listings we show the Ontolex-Lemon representation of “Kernspaltung”.

```
:entry_w1 rdf:type ontolex:LexicalEntry ;
  decomp : constituent :Kern_comp ;
  rdf:_1 :Kern_comp ;
  decomp : subterm : entry_w3542 ;
  decomp : constituent : spaltung_comp ;
  rdf:_2 : spaltung_comp ;
decomp : subterm: entry_w23527 ;
  lexinfo : hypernym : synset_odenet -5437 -n ;
  lexinfo : partOfSpeech lexinfo : noun ;
  ontolex : canonicalForm :form_w1 ;
  ontolex : sense : sense_w1_1 -n ;
  ontolex : evokes : synset_odenet -1 -n ;
.
```

Listing 1.2: The lexical entry for *Kernspaltung*

In Listing 1.2 we display the full OntoLex-Lemon entry. One aspect that can be immediately noted is the possibility to represent the components of the compound word. This demonstrates one of the benefits of linking synsets to the (complex) representation of lexical entries, as we can state (see below) the semantic relations between synsets associated with the components of a compound word and its own synset.

Listing 1.3 below shows the form information associated to the w1 entry in Listing 1.2.

```
:form_w1 rdf:type ontolex:Form ;
  ontolex : writtenRep " Kernspaltung "@de ;
.
```

Listing 1.3: The ontolex:Form *Kernspaltung*

Listing 1.4 shows the conversion of the original OdeNet sense information into an instance of the ontolex:LexicalSense class.

```
:sense_w1_1 -n rdf:type ontolex:LexicalSense ;
  ontolex : isLexicalizedSenseOf
    : synset_odenet -1 -n ;
  ontolex : isSenseOf : entry_w1 ;
  ontolex : reference
    https://www.wikidata.org/wiki/Q11429 ;
.
```

Listing 1.4: The LexicalSense associated to the entry for *Kernspaltung*

In this code we see how the property ontolex:isLexicalizedSenseOf is linking a sense to

a synset, while the entry itself can be linked to the synset via the property `ontolex:evokes`, as shown in Listing 1.1. The property (`ontolex:reference`) also links the sense to an ontological entity, here in the form of a Wikidata entry.

Listing 1.5 shows the representation of the synset associated with both the `w1` lexical entry and the `w1_1-n` sense. There we can also see that this lexical concept (synset) is also “evoked” by other entries/senses. For example by the entries for “Kernfission” or “Atomspaltung”, which are synonyms of “Kernspaltung”. The `lexinfo:hypernym` property provides information on the semantic relation this synset has to another synset.

```
: synset__odenet-1-n
  rdf : type ontolex : LexicalConcept ;
  skos : inScheme : ODEnet ;
  skos : definition "a nuclear reaction
  in which a massive nucleus splits
  into smaller nuclei with the
  simultaneous release of energy " ;
  wn: i l i i l i : i107577;
  ontolex : isEvokedBy : entry__w1 ;
  ontolex : isEvokedBy : entry__w2 ;
  ontolex : isEvokedBy : entry__w3 ;
  ontolex : isEvokedBy : entry__w4 ;
  ontolex : lexicalizedSense : sense__w1_1-n ;
  ontolex : lexicalizedSense : sense__w2_1-n ;
  ontolex : lexicalizedSense : sense__w3_1-n ;
  ontolex : lexicalizedSense : sense__w4_1-n ;
  lexinfo : hypernym : synset__odenet -5437-n ;
.
```

Listing 1.5: The LexicalConcept (synset) associated with the entry for *Kernspaltung*

Finally, in Listing 1.6 we show the “entries” for the components of the compound word “Kernspaltung”. Those components are pointing to the lexical entries they are related to. The entry `:entry_w23527` is, for example, the one corresponding to the noun “Spaltung” (*split, fission, separation, cleavage*, etc.), which has again its own senses and associated synsets. We can here disambiguate the meaning of “Spaltung” as used in the compound, as being the one of “fission”. And the whole compound can then be considered as an hyponym of the synset for “fission”.

```

:Kern_comp
  rdf : type decomp : Component ;
  decomp : correspondsTo      : entry__w3542 ;
.
:spaltung_comp
  rdf : type decomp : Component ;
  decomp : correspondsTo      : entry__w23527 ;
.

```

Listing 1.6: The two components of the entry *Kernspaltung*

In Listing 1.2 above, we can see the information on the ordering those components have in this entry, marked with the “rdf:_1” and “rdf:_2” constructs. For sure, those component “entries” can be re-used separately for other compounds, such as “Atomspaltung”. So that we can collect all the corresponding meanings of a word, even when they are used in compounds, as well as depending on their position in the compounds. Details on the decomposition module of OntoLex-Lemon are shown in Figure 2.

The porting of OdeNet to OntoLex made evident that the introduced senses in OdeNet are not really playing a role. We will in the near future replace the OdeNet senses with lexical senses established in other resources. We will also link the synsets to ontological resources, whereas the BabelNet resource from Navigli and Ponzetto (2012) can be very helpful here. We also see that there is no need to associate a PoS with a synset, as this information is present with the associated lemmas. This way we are reaching a higher level of modularity with the OntoLex-Lemon representation.

7. Current Work

We are currently linking the newly created data in the OntoLex-Lemon representation with the already existing UBY-OmegaWiki lemon-based encoding for German²⁸, which at the time of its creation (2014) could not make use of the *ontolex:LexicalConcepts* property. This work will result in the merging of two large lexical semantics German resources in OntoLex-Lemon, and make this resource accessible in the Linguistic Linked Data cloud.

8. Acknowledgements

The work presented in this paper has been partially supported by the H2020 project “Prêt-àLLOD” with Grant Agreement number 825182. Contributions by Thierry Declerck have been additionally supported in part by the H2020 project “ELEXIS”

²⁸ See https://lemon-model.net/lexica/ubyow_deu/.

with Grant Agreement number 731015. We would like to thank the anonymous reviewers for their very valuable comments, which we tried to adequately address in the final version of the paper.

9. References

- Bond, F. & Foster, R. (2013). Linking and Extending an Open Multilingual Wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. Sofia, pp. 1352–1362. <http://aclweb.org/anthology/P13-1133>.
- Bond, F. & Paik, K. (2012). A survey of wordnets and their licenses. *Small*, 8(4), p. 5.
- Bond, F., Vossen, P., McCrae, J. P. & Fellbaum, C. (2016). CILI: the Collaborative Interlingual Index. In *Proceedings of the Global WordNet Conference*, volume 2016.
- Chiarcos, C., Nordhoff, S. & Hellmann, S. (eds.) (2012). *Linked Data in Linguistics Representing and Connecting Language Data and Language Metadata*. Springer. <https://doi.org/10.1007/978-3-642-28249-2>.
- Cimiano, P., Buitelaar, P., McCrae, J. P. & Sintek, M. (2011). LexInfo: A declarative model for the lexicon-ontology interface. *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*, 9(1), pp. 29–51.
- Cimiano, P., McCrae, J. P. & Buitelaar, P. (2016). Lexicon Model for Ontologies: Community Report.
- Declerck, T. & Racioppa, S. (2019). Porting Multilingual Morphological Resources to OntoLex-Lemon. In R. Mitkov & G. Angelova (eds.) *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2019*. INCOMA Ltd.
- Fellbaum, C. (ed.) (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.
- Gonzalez-Agirre, A., Laparra, E. & Rigau, G. (2012). Multilingual Central Repository version 3.0: upgrading a very large lexical knowledge base. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*. Matsue.
- Hamp, B. & Feldweg, H. (1997). GermaNet - a Lexical-Semantic Net for German. In *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*. Madrid.
- McCrae, J., Aguado-de Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., Gomez-Perez, A., Garcia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D. & Wunner, T. (2012). Interchanging Lexical Resources on the Semantic Web. *Language Resources and Evaluation*, 46(4), pp. 701–719.
- McCrae, J. P., Buitelaar, P. & Cimiano, P. (2017). The OntoLex-Lemon Model: Development and Applications. In I. Kosem, J. Kallas, C. Tiberius, S. Krek, M. Jakubíček & V. Baisa (eds.) *Proceedings of eLex 2017*. Brno: Lexical Computing CZ s.r.o., pp. 587–597.
- Morato, J., Marzal, M., Lloréns, J. & Moreiro, J. (2004). WordNet Applications. pp. 270–278. <http://www.fi.muni.cz/gwc2004/proc/105.pdf>.

- Navigli, R. & Ponzetto, S. P. (2012). BabelNet: The Automatic Construction, Evaluation and Application of a Wide-coverage Multilingual Semantic Network. *Artif. Intell.*, 193, pp. 217–250. <http://dx.doi.org/10.1016/j.artint.2012.07.001>.
- Petitpierre, D. & Russell, G. (1995). MMORPH: The Multext Morphology Program. Multext deliverable 2.3.1, ISSCO, University of Geneva. URL <http://www.issco.unige.ch/downloads/multext/mmorph.doc.ps.tar.gz>.
- Pianta, E., Bentivogli, L. & Girardi, C. (2002). MultiWordNet: Developing an Aligned Multilingual Database. In *In Proceedings of the First International Conference on Global WordNet*. Mysore, India, pp. 293–302.
- Sagot, B. & Fišer, D. (2008). Building a free French wordnet from multilingualresources. In E.L.R.A. (ELRA) (ed.) *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco.
- Toral, A., Bracale, S., Monachini, M. & Soria, C. (2010). Rejuvenating the Italian WordNet: upgrading, standardising, extending. In *Proceedings of the 5th International Conference of the Global WordNet Association (GWC-2010)*. Mumbai.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>





ORGANIZERS

Univerza v Ljubljani



SPONSORS



A. S. Hornby Educational Trust



elex.link/elex2019