Electronic lexicography in the 21st century (eLex 2023)

Book of abstracts

edited by Marek Medveď
Michal Měchura
Carole Tiberius
Iztok Kosem
Jelena Kallas
Miloš Jakubíček
Simon Krek

27-29 June 2023

elex.link/elex2023

Electronic lexicography in the 21st century (eLex 2023)

Book of abstracts

Edited by
Marek Medveď
Michal Měchura
Carole Tiberius
Iztok Kosem
Jelena Kallas
Miloš Jakubíček
Simon Krek

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

June 2023, Brno, Czech Republic
elex.link/elex2023

eLex 2023

ORGANIZERS

LEXICAL COMPUTING

Univerza v Ljubljani

cjvt Centre for Language Resources and Technologies

EESTI KEELE INSTITUUT

/instituut voor de Nederlandse taal/

SPONSORS

Hornby

A. S. Hornby Educational Trust

elex.link/elex2023

Organizing Committee

Miloš Jakubíček
Jelena Kallas
Iztok Kosem
Simon Krek
Carole Tiberius
Ondřej Matuška
Tereza Olšanová
Michal Cukr
Vlasta Ohlídalová

Scientific committee

Aleksandra Marković
Aleš Horák
Alexander Geyken
Amália Mendes
Andrea Abel
Arvi Tavast
Anila Çepani
Annette Klosa-Kückelhaus
Carole Tiberius
Edward Finegan
Egon Stemle
Emma Sköldberg
Henrik Lorentzen
Hindrik Sijens
Ilan Kernerman
Iztok Kosem
Jelena Kallas
Kris Heylen
Kristian Blensenius
Kristina Koppel
Kristina Strkalj Despot
Lars Trap-Jensen
Laurent Romary
Lionel Nicolas
Lothar Lemnitzer
Lut Colman
Maarten Janssen
Margit Langemets
María José Domínguez Vázquez
Michal Kren
Michal Měchura
Miloš Jakubíček
Mojca Kompara Lukančič
Nicolai Hartvig Sørensen
Patrick Drouin
Paul Cook
Pilar León Araúz
Polona Gantar
Radovan Garabík
Ranka Stanković
Rizki Gayatri
Robert Lew
Said Zohairy
Sara Moze
Simon Krek
Stella Markantonatou
Sussi Olsen
Svetla Koeva
Špela Arhar Holdt
Tamás Varadi
Tanara Zingano Kuhn
Thierry Fontenelle
Vincent Ooi
Vít Suchomel
Vojtěch Kovář
Voula Giouli
Yongwei Gao
Yukio Tono
Zoe Gavriilidou

Contents

Book of Abstracts

KEYNOTE: An automatic observatory of anglicism usage in the Spanish press.

Elena Álvarez Mellado
UNED University
E-mail: ealvarezmellado@gmail.com

Anglicisms are words from English that are borrowed into another language. Anglicisms are a common source of new words in Spanish, which makes them an interesting phenomenon to observe for linguists and lexicographers. In this session we will present Observatorio Lázaro, a machine learning pipeline that monitors the Spanish press of the day and detects new anglicisms automatically.

KEYNOTE: Invisible lexicographers, AI, and the future of the dictionary

Wendalyn Nichols
Cambridge Dictionary
E-mail: wendalyn.nichols@cambridgeassessment.org.uk

Artificial intelligence is seen as an existential threat by publishers of nonfiction, most particularly the producers of reference content. What does it mean for content to be authoritative if anyone can type a question into a search box and get an answer from an AI chatbot without ever visiting a publisher's website or buying a publisher's books? Has the ubiquitous use of huge, widely-available sets of lexical data to train AI algorithms hastened the end of original lexicography, and therefore of lexicographers? AI is already disrupting the market and is not going away, but both its strengths and shortcomings can be exploited by dictionary producers to turn the threat to our advantage.

KEYNOTE: Interoperable Words. Interlinking Lexical (and Textual) Resources for Latin in the LiLa Knowledge Base

Marco C. Passaritti

Università Cattolica del Sacro Cuore (Milan, Italy)
E-mail: marco.passarotti@unicatt.it

In this talk, I will discuss the issue of interoperability between linguistic resources and how to address it by applying the principles of the Linked Data paradigm to describe several kinds of (meta)data provided by resources published on the web. In particular, I will focus on lexical resources, presenting how a few dictionaries and lexica for Latin interact with each other (and with textual corpora, too) in the LiLa Knowledge Base, i.e., a collection of multifarious resources made interoperable by adopting the same vocabulary for knowledge description, through common data categories and ontologies widely used in the Linguistic Linked Open Data community.

KEYNOTE: Using Register Analysis to Establish Style Markers in Dictionaries

Václav Cvrček
Institute of the Czech National Corpus
E-mail: vaclav.cvrcek@ff.cuni.cz

This talk investigates the potential of register analysis of text corpora for defining style labels or usage markers in dictionaries. In contemporary lexicography, large language corpora are commonly utilized to extract data on lexemes, their meanings, and frequencies, employing advanced methods such as collocation extraction or sophisticated frequency measurement that account for dispersion etc. However, when it comes to style markers, we mostly rely on mere presence (or absence) of a word in a particular genre or text type, neglecting the potential offered by corpus linguistics methods dealing with variability of texts and their functional classification.

To address this issue, this talk proposes the use of the multi-dimensional analysis (MDA) method developed by Douglas Biber (1988, 1995) for register classification. MDA is known for effectively charting the space of variation by identifying major dimensions and delimiting registers within language. By exploring the associations between words and dimensions of variability or text registers in Czech, this talk will attempt to establish style markers that are at the same time practical for the dictionary user, empirically sound, and allow for semi-automatic extraction.

The role of the invisible lexicographer in the compilation of the Slovene dictionary of abbreviations

Mojca Kompara Lukancic
University of Maribor
E-mail: mojca.kompara@gmail.com

Dictionaries of abbreviations are a type of dictionary that can be compiled almost entirely automatically (Kompara Lukančič, 2011, 2017, 2018) and automatic creation of dictionary content is possible with the development of an algorithm for automatic recognition of abbreviations and expansions in texts (Kompara Lukančič, 2011, 2017). In the article we present how invisible lexicography plays a crucial role in the compilation proces of the Slovene dictionary of abbreviations (Kompara Lukančič, 2023). The overall compilation process, starting with manual gathering of abbreviation-expansion pairs and continuing with automatic extraction and compilation of dictionary articles (Kompara Lukančič, 2011) is presented in the article. The compilation process presented is a mixture of automatic and semi-automatic approaches. The automatic recognition of abbreviations takes place at the lexical level by observing the qualities of abbreviations and their expansions we built an algorithm that recognizes abbreviations based on recognition principles, and it seeks their expansions in context. The algorithm considers different types of abbreviations and expansions; for example, overlapping abbreviations, abbreviations without conjunctions and prepositions, abbreviations made from initial letters, and abbreviations with conjunctions and prepositions and always considers the context in recognizing expansions. In the article we present the position of abbreviations in Slovene language, where we mention two dictionaries, namely Slovarček krajšav (Little Dictionary of Abbreviations; Kompara Lukančič, 2006) and the automatically generated Slovar krajšav (Dictionary of Abbreviations; Kompara Lukančič, 2011). The article draws attention to the need for a modern dictionary of abbreviations and presents the automatic creation of dictionary content for the Slovene dictionary of abbreviations (Kompara Lukančič, 2023), the micro- and macrostructure of the dictionary of abbreviations, the acquisition and selection of headwords, and the form and structure of the dictionary entries. It also addresses the issue of automatic lemmatisation, filed qualifiers, cross-references, and encyclopaedic data.

References

- Kompara Lukančič, M. (2023). Slovenski slovar krajšav. 1. izd. Maribor: Univerzitetna založba Univerze, (in print)
- Kompara Lukančič, M. (2018). Sinhrono-diahroni pregled krajšav v slovenskem prostoru in sestava slovarja krajšav. 1. izd. Maribor: Univerzitetna založba Univerze.
- Kompara Lukančič, M. (2017). Zasnova novega slovarja krajšav. Jezikoslovni zapiski : zbornik Inštituta za slovenski jezik Frana Ramovša. [Tiskana izd.]. 23, št. 1, str. 77-92.
- Kompara Lukančič, M. (2011). Razvoj algoritma za samodejno prepoznavanje krajšav in krajšavnih razvezav v elektronskih besedilih. Jezikoslovni zapiski : zbornik Inštituta za slovenski jezik Frana Ramovša. [Tiskana izd.]. 17, št. 2, str. 107-122

- Kompara Lukančič, M. (2011). Slovar krajšav. Kamnik: Amebis, Zbirka Termania.
- Kompara Lukančič, M. (2006). Slovarček krajšav. Ljubljana

Evaluation of the Cross-lingual Embedding Models from the Lexicographic Perspective

Michaela Denisová, Pavel Rychlý
Faculty of Informatics, Masaryk University
E-mail: michaeladenisova@gmail.com, pavel.rychly@sketchengine.eu

Keywords: cross-lingual embedding models; bilingual lexicon induction task; retrieving translation equivalents; evaluation

Over the years, the cross-lingual embedding models (CEMs) have drawn much attraction due to their ability to transfer lexical knowledge across languages. They facilitate the alignment of word vector representations of two or more languages into one shared space where similar words obtain similar vectors (Ruder et al., 2019).

These models are appealing for lexicography for multiple reasons. Firstly, the translation equivalents candidates can be extracted from the aligned space. Secondly, in contrast to parallel-data-based methods for finding translation equivalents candidates, they require only comparable data, i.e., comparable corpora. Comparable corpora are often available for low-resource languages or rare language combinations and are balanced in the texts they consist of. Finally, CEMs are an active research area, expected to develop and improve constantly.

In this field, finding translation equivalents candidates is referred to as bilingual lexicon induction (BLI) task. In the BLI task, the target words are induced from aligned space through the nearest neighbour search for a source word. Afterwards, they are run against a gold-standard dictionary to measure the quality of the model (Ruder et al., 2019).

The BLI task is a popular way among researchers to evaluate their models (Artetxe et al., 2016; Glavaš and Vulić, 2020; Tian et al., 2022; etc.). However, the evaluation is often inconsistent and differs from paper to paper, using various metrics and gold-standard dictionaries from multiple sources (Ren et al., 2020; Woller et al., 2021; Severini et al., 2022; etc.). This impedes our ability to correctly interpret the results and make models comparable to each other.

Moreover, many currently used gold-standard dictionaries are generated automatically (Conneau et al., 2018; Glavaš et al., 2019; Vulić et al., 2019; etc.). Therefore, they are prone to contain mistakes. For example, the most widely used gold-standard dictionaries, MUSE (Conneau et al., 2018), are criticised for occurring errors and disproportional part-of-speech distribution (Kementchedjhieva et al., 2019; Denisová and Rychlý, 2021).

On top of that, articles dealing with cross-lingual embedding models and the BLI task do not consider the utilisation in the lexicography field. They focus on the computational side of the problem and simple word-to-word extraction without reflecting on various aspects of translation.

In this paper, we investigate factors that influence the training of CEMs. We propose the most suitable parameters for the evaluation dataset based on these aspects while considering a lexicography perspective. We show that having a strong evaluation dataset and a clear evaluation process is crucial for setting appropriate training parameters. We evaluate the most common benchmark models on a distant language pair, Estonian-Slovak, a close language pair, Czech-Slovak, and language pair with different scripts, English-Korean, in various settings.

Our motivation is to determine important aspects when evaluating CEMs on the BLI task and construct a reliable, reproducible, and unifying evaluation dataset that addresses the

above-stated issues. We aim for our evaluation dataset to mirror the models' performance as accurately and transparently as possible. Moreover, we involve the lexicography point of view in the evaluation process and make CEMs more accessible for lexicographers.

References

- Artetxe, M., Labaka, G., Agirre, E. (2016). Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. Conference on Empirical Methods in Natural Language Processing.
- Conneau, A., Lample, G., Ranzato, M., Denoyer, L., J'egou, H. (2018). Word Translation Without Parallel Data. ArXiv.
- Denisová, M., Rychlý, P. (2021). When Word Pairs Matter. RASLAN.
- Glavaš, G., Litschko, R., Ruder, S., Vulić, I. (2019). How to (Properly) Evaluate Cross-Lingual Word Embeddings: On Strong Baselines, Comparative Analyses, and Some Misconceptions. Annual Meeting of the Association for Computational Linguistics.
- Glavaš, G., Vulić, I. (2020). Non-Linear Instance-Based Cross-Lingual Mapping for Non-Isomorphic Embedding Spaces. ACL.
- Kementchedjhieva, Y., Hartmann, M., Søgaard, A. (2019). Lost in Evaluation: Misleading Benchmarks for Bilingual Dictionary Induction. ArXiv.
- Ren, S., Liu, S., Zhou, M., Ma, S. (2020). A Graph-based Coarse-to-fine Method for Unsupervised Bilingual Lexicon Induction. ACL.
- Ruder, S., Vulić, I., Søgaard, A. (2019). A Survey of Cross-lingual Word Embedding Models. The Journal of Artificial Intelligence Research, 65, 569-631.
- Severini, S., Hangya, V., Jalili Sabet, M., Fraser, A.M., Schütze, H. (2022). Don't Forget Cheap Training Signals Before Building Unsupervised Bilingual Word Embeddings. ArXiv.
- Tian, Z., Li, C., Ren, S., Zuo, Z., Wen, Z., Hu, X., Han, X., Huang, H., Deng, D., Zhang, Q., Xie, X. (2022). RAPO: An Adaptive Ranking Paradigm for Bilingual Lexicon Induction. ArXiv.
- Vulić, I., Glavaš, G., Reichart, R., Korhonen, A. (2019). Do We Really Need Fully Unsupervised Cross-Lingual Embeddings? ArXiv.
- Woller, L., Hangya, V., Fraser, A. (2021). Do not neglect related languages: The case of low-resource Occitan cross-lingual word embeddings. Proceedings of the 1st Workshop on Multilingual Representation Learning.

The Dark Side of the Dictionary

Robert Lew[1], Sascha Wolfer[2]

[1]Adam Mickiewicz University Poznań, [2]Leibniz Institute for the German Language (IDS)
E-mail: rlew@amu.edu.pl, wolfer@ids-mannheim.de

Keywords: Wiktionary; English; user logs; consultation frequency; multi-word; entry inclusion

Dictionary-writing has been an extremely laborious activity: it can take as much as a century for a large team of lexicographers to produce a comprehensive dictionary (De Schryver, 2005; Gilliver, 2016). Although current developments in the automation of lexicographic work (not the least stimulated at and around eLex conferences) promise significant savings in this respect, in most cases humans are still involved. Since human labour is expensive, it should ideally not be wasted on work that serves very few or none. It is a waste of resources to create entries that (nearly) no-one is likely to consult. If we could know in advance which entries would be expendable, we could use the expertise and time towards improving more useful entries.

To this end, and in line with the theme of this year's conference, we describe an attempt at identifying dictionary entries that exist, but that no-one chooses to look at over a period of three years (2019-2021). This is the dark side of the dictionary, which is there, but no-one has seen it, much like the dark side of the Moon. We chose the English Wiktionary because it provides comprehensive logs of user visits over the years (Wikimedia, 2023) and it is a widely-used lexicographic resource, not only in English speaking countries. Using these logs from 2019 to 2021 (Wikimedia, 2022), we extract data on user visits to specific Wiktionary entries. Since the server logs do not provide any negative evidence—i.e. explicit information on the Wiktionary pages NOT visited— we also downloaded a complete list of English Wiktionary entries marked as English words and automatically cross-checked this list against page visit logs. Our dataset comprises 1.1 billion views on 545,014 entries.

It turns out that the typical Wiktionary entry is consulted ten times a month (by median). Only a handful of 101 entries have remained completely unconsulted over the three-year period. However, a much larger set of over 3,000 entries have been consulted very infrequently: less than a dozen times over the three-year period. In addition, if we look at the number of whole months in which certain articles are not consulted at all, we see that there are approximately 17,000 articles (3.1% of all entries) that had not been consulted in 24 months or more within the 36-month period. In our presentation, we discuss different operationalizations of the 'dark side' of a dictionary and try to identify patterns in rarely consulted entries.

We also look at the relationship between usage and entry age in years (the oldest entries are 20 years old), as well as simplex versus multi-word items. Perhaps surprisingly, multi-words are not amongst the least consulted entries.

References

- Atkins, S.B.T., Rundell, M. (2008). The Oxford Guide to Practical Lexicography. Oxford: Oxford University Press.

- De Schryver, G.-M. (2005): Concurrent Over- and Under-treatment in Dictionaries – The Woordeboek van die Afrikaanse Taal as a Case in Point*. In: International Journal of Lexicography, 18 (1), pp. 47–75.
- Gilliver, P. (2016): The making of the Oxford English dictionary. First edition. Oxford ; New York, NY.
- Wikimedia (2022): Wikimedia Analytics Datasets: Pageviews. URL: `https://dumps.wikimedia.org/other/pageviews/`.
- Wikimedia (2023): Pageviews Readme. URL: `https://dumps.wikimedia.org/other/pageviews/readme.html`.

The Open Dictionary Project

Tyler Nickerson
Linguistic Inc.
E-mail: tyler@linguistic.io

Keywords: dictionary authoring; dictionary format; dictionary compiler

Open access to digital, structured lexical data is increasingly prevalent in today's digital age. Despite this, no modern standard for authoring such data exists. Current open-source dictionaries are written using a plethora of formats, including (but not limited to) plain-text, JSON, CSV, and TEI. These formats offer no built-in methods of indexing or retrieving entries, so dictionaries must then be parsed and loaded into either a third-party application, like StarDict, or a portable database format, such as SQLite. Furthermore, applications such as StarDict typically leverage their own custom formats to store data in unstructured, arbitrary HTML fragments, making the process of porting or extracting structured data near impossible.

In this presentation, we introduce The Open Dictionary Project (ODict), a file format specification and compiler built specifically for authoring, assembling, and distributing digital dictionaries. ODict dictionaries are written in a human-readable XML format and compiled to '.odict' binaries with a robust command-line interface (CLI). ODict files can store thousands of entries in only a few megabytes, enabling multiple ad-hoc entry lookups in well under one second. The CLI allows users to perform a wide range of operations on the source data, such as merging, fuzzy searching, and outputting the original XML. We have built a number of language-specific bindings designed for programmatic dictionary access, currently supporting Node, Python, and the JVM.

The goal of The Open Dictionary Project is to facilitate access to the world's best dictionaries and promote the free distribution of lexical data. ODict is used extensively in the Linguistic platform to power rapid word lookups for dozens of words in parallel, as well as position us for offline dictionary support on both mobile and web (using WebAssembly). We hope that our presentation will welcome feedback on the project and allow us to further explore how ODict may be able to serve the lexicographical community.

Annotating corpora for language learning and lexicography with the Crowdsourcing for Language Learning (CrowLL) game

Tanara Zingano Kuhn[1], Kristina Koppel[2], Špela Arhar Holdt[3,4], Carole Tiberius[5], Rina Zviel-Girshin[6], Iztok Kosem[3,7]

[1]Research Centre for General and Applied Linguistics (CELGA-ILTEC), University of Coimbra, [2]Institute of the Estonian Language, [3]Faculty of Arts, University of Ljubljana, [4]Faculty of Computer and Information Science, University of Ljubljana, [5]Instituut voor de Nederlandse Taal, [6]Ruppin Academic Center, [7]Jozef Stefan Institute
E-mail: tanarazingano@outlook.com, kristina.koppel@eki.ee, arharhs@ff.uni-lj.si, carole.tiberius@ivdnt.org, rinazg@ruppin.ac.il, iztok.kosem@cjvt.si

Keywords: crowdsourcing; examples; GWAP; pedagogical corpora

In this demo, we will introduce the Crowdsourcing for Language Learning (CrowLL) game, created so far for Brazilian Portuguese, Dutch, Estonian, and Slovene. Even though computational technology has contributed extensively to identification of good examples for dictionaries and pedagogical purposes in general (e.g., Kilgarriff et al., 2008; Kosem et al., 2019; Pilán et al., 2013, 2014; Pilán, Vajjala, Volodina, 2016; Stanković et al., 2019), detection of offensiveness, sensitive content and even structural problems in sentences is the type of issue with which machines still have trouble dealing. This means that, in such cases, human verification is required. Thus, in order to streamline this task, we have developed CrowLL. The main objective of CrowLL is to have the crowd contribute to the annotation of automatically extracted sentences from corpora in order to create corpora of examples that can be used for language learning purposes, including the development of learners' dictionaries, autonomous language learning lexical resources such as SKELL (Baisa & Suchomel, 2014), and teaching materials. CrowLL is currently available as a single-player mode and has three levels, which can be played separately or combined. In level 1, the player is presented with two sentences and the question "Which sentence would you choose for teaching Portuguese (Dutch, Estonian, Slovene)? They then need to choose one, both or none of them. In level 2, the player is prompted to choose a category (or categories) of problem to which the sentence(s) that has/have not been selected in level 1 belong(s) (offensive, vulgar, sensitive content, grammar/spelling problems, incomprehensible/lack of context). In level 3, the player is asked to tap or click on the part(s) of the sentence that is/are problematic. The output of the game will provide the labels for the examples, i.e., a certain sentence can be non-problematic or problematic, and if problematic, there will be labels showing the type of problem and the problematic items will be marked. These annotated examples will compose pedagogical corpora that can be filtered by lexicographers, material developers and teachers according to their needs. It should be highlighted that not only additional languages can be added to the game, but the game itself can be adapted for other lexicographical purposes that can benefit from the use of crowdsourcing techniques, such as the collection of judgments on whether a certain example is good to illustrate the use of a certain word. In the next stage of this project, we will use these annotated sentences to prepare machine learning algorithms to automatically identify problematic sentences in corpora, thus contributing to the further growth of these pedagogical corpora.

References

- Baisa, V., Suchomel, V. (2014). SkELL: Web Interface for English Language Learning. In A. Horák, P. Rychlý (eds) Proceedings of the Eighth Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2014. Brno: Tribun EU, pp. 63-70.
- Kilgarriff, A., Husák, M., McAdam, K., Rundell, M., Rychlý, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus. Proceedings of the XIII EURALEX international congress (Vol. 1), 425–432.
- Kosem, I., Koppel, K., Kuhn, T. Z., Michelfeit, J., Tiberius, C. (2019). Identification and automatic extraction of good dictionary examples: the case(s) of GDEX. International Journal of Lexicography, 32 (2), 119–137.
- Pilán, I., Vajjala, S., Volodina, E. (2016). A readable read: Automatic assessment of language learning materials based on linguistic complexity. ArXiv.
- Pilán, I., Volodina, E., Johansson, R. (2013). Automatic selection of suitable sentences for language learning exercises. 20 Years of EUROCALL: Learning from the past, looking to the future: 2013 EUROCALL Conference Proceedings, 218–225.
- Pilán, I., Volodina, E., Johansson, R. (2014). Rule-based and machine learning approaches for second language sentence-level readability. Proceedings of the ninth workshop on innovative use of NLP for building educational applications, 174–184.
- Stanković, R., Šandrih, B., Stijović, R., Krstev, C., Vitas, D., Marković, A. (2019). SASA dictionary as the gold standard for good dictionary examples for Serbian. Proceedings of the eLex 2019 conference, 248–269.

The Perceptions of Using KBBI Online as a Speaking Guide by Advanced Learners of Bahasa Indonesia

Rizki Gayatri[1], Zamzam Hariro[1], Siti Rahajeng N.H[2]
[1]Kantor Bahasa Provinsi NTB, [2]Universitas Indonesia
E-mail: rizki.gayatri@kemdikbud.go.id, hariro.zam@gmail.com, sitirahajengnh@gmail.com

Keywords: user experience; speaking guide; KBBI Daring; BIPA Learners.

This study is aimed at describing perceptions of the use of the Big Indonesian Online Dictionary (KBBI Online) as a guide for students in learning Indonesian for Advanced Foreign Speakers (BIPA) in speaking Indonesian. KBBI Online is a website-based monolingual Indonesian dictionary available online. This dictionary is published by the Language Development and Fostering Agency, Ministry of Education, Culture, Research and Technology of the Republic of Indonesia. As of October 2022, KBBI Online has published 118,021 entries and has been accessed 191,812,129 times. The features of KBBI Online include definitions, word classes, phonemic transcripts, and word meanings. The method used in this study is face-to-face conversation using questionnaires and recording techniques (Zaim, 2014). Through the questionnaire provided, eleven respondents spread across six countries were asked to provide their perceptions of the online KBBI features. Apart from that, they are also asked to record ten basic Indonesian vocabulary words. The respondents of this research come from Mongolia, Japan, Thailand, South Korea, China, and Malaysia. The result of the study shows that as much as 90.91% of respondents agree that KBBI Online helps them pronounce Indonesian words. Furthermore, as many as 18.18% of respondents say that they are not familiar with the phonemic transcription presented in the KBBI Online, while in the phonetic transcription, all respondents are familiar with the transcription. Overall, 54.55% respondents show that they are quite satisfied with the existing features. The lack features of online KBBI are the absence of phonetic transcription and audio samples of each word. It effects on the wrong pronunciation of words by students, as in the case of [kuwaci] which is pronounced [kwaʧ], also in the word [ɛmbɛr] pronounced [əmbər]. Based on the results of the questionnaires, 90.91% of respondents agree with the addition of phonetic transcription and audio samples in KBBI Online.

The impact of invisible lexicography on the self-revision of academic English collocations

Tomasz Michta[1], Ana Frankenberg-Garcia[2]

[1]University of Bialystok, [2]University of Surrey
E-mail: t.michta@gmail.com, a.frankenberg-garcia@surrey.ac.uk

Keywords: collocations; academic writing; user study

Collocations have been found to hold back L2 writers even after many years of language instruction (Nesselhauf, 2005). L1 users have also been reported to struggle with collocations, especially when asked to produce academic or otherwise specialised collocations (Frankenberg-Garcia, 2018; Michta & Mroczyńska, 2022). Despite the extensive coverage of collocations in general and a number of specialized language dictionaries, user studies have found that learners may still struggle to find correct collocations in a dictionary even when explicitly asked to do so (Laufer 2011). Indeed, the usefulness of dictionaries hinges on the user's ability to correctly identify errors or other limitations related to the use of collocations, find the information needed to address the problem, and apply it effectively. The whole process can be cognitively disruptive, as dictionary consultation competes for time and attention with other aspects of writing (Frankenberg-Garcia, 2020).

Designed to minimize any disruption and at the same time raise awareness of limitations in the use of collocations that writers may not otherwise notice, ColloCaid - a collocational database that has been experimentally integrated into a text editor (Frankenberg-Garcia et al., 2019, 2021) – is an example of "invisible lexicography", where lexical data is brought to writers without them having to leave their writing environments. In an initial study of how it was rated by a group of university students in Spain during a controlled gap-filling exercise, Rees (2021) found that ColloCaid was perceived as less demanding in terms of the NASA Task Load Index when compared with other collocation tools and dictionaries. This study takes user analysis a step further, by evaluating the actual lexical changes motivated by ColloCaid.

After a short explanation of how the tool works, a group of 27 L2 English students at a European university were asked to use ColloCaid to revise an approximately 600-word excerpt of their BA or MA dissertation (excluding quotations) before it was seen by their supervisor. We looked at the revision data first to assess coverage, by examining the number of lemmas in the excerpts for which collocation suggestions were available. We looked at uptake, by examining the percentage of lemmas for which collocation suggestions were taken on board, and collected data on the reasons why suggestions were or were not followed. Next, the actual changes undertaken were classified according to their effect on the text in terms of the revision taxonomy developed in (Frankenberg-Garcia, 1990). Finally, we conducted guided interviews with a small sample of the participants to discuss their use of the tool in further detail. First-hand results of the study will be presented at eLex. Our findings enable us not only to come to a better understanding of the use of ColloCaid in particular, but also to gain more general insights into user reactions to invisible lexicography.

References

- Frankenberg-Garcia, A. (1990) Second Language Writing Instruction: a Study of the Effects of a Discourse-Oriented Programme upon the Ability of Skilled Writers to Improve their Written Production. PhD thesis, Edinburgh University.
- Frankenberg-Garcia, A. (2018) Investigating the collocations available to EAP writers. Journal of English for Academic Purposes 35: 93-104.
- Frankenberg-Garcia, A. (2020) Combining user needs, lexicographic data and digital writing environments. Language Teaching, 53-1:29-43.
- Frankenberg-Garcia, A., Lew, R., Roberts, J. C., Rees, G. P., Sharma, N. (2019). Developing a writing assistant to help EAP writers with collocations in real time. ReCALL, 31(1), pp. 23–39.
- Frankenberg-Garcia, A., Rees, G. P., Lew, R. (2021). Slipping Through the Cracks in e-Lexicography. International Journal of Lexicography, 34(2), pp. 206–234.
- Laufer, B. (2011). The Contribution of Dictionary Use to the Production and Retention of Collocations in a Second Language. International Journal of Lexicography, 24(1), pp. 29–49.
- Michta T., Mroczyńska, K. (2022). Towards a dictionary of legal English collocations. Siedlce: Wydawnictwo Uniwersytetu Przyrodniczo-Humanistycznego w Siedlcach.
- Nesselhauf, N. (2005). Collocations in a Learner Corpus. Amsterdam/Philadelphia: John Benjamins Publishing Company.
- Rees, G. P. (2021). Measuring User Workload in e-Lexicography with the NASA Task Load Index. In I. Kosem, M. Cukr (eds.) Electronic lexicography in the 21st century (eLex 2019): Post-editing lexicography. Book of abstracts. p. 54–55.

Woordcombinaties (Word Combinations)

Lut Colman, Carole Tiberius
Instituut voor de Nederlandse Taal
E-mail: carole.tiberius@ivdnt.org, carole.tiberius@ivdnt.org

Keywords: collocations, idioms, patterns, CALL

In this demo, we present Woordcombinaties (Word Combinations), a relatively new online lexicographic resource for advanced learners of Dutch as a second or foreign language. It answers various types of phraseological queries by combining access to collocations, idioms, conversational routines and constructions in one tool.

For collocations, we follow the example of Sketch Engine for Language Learning (SkeLL)[1], whereas the constructions are encoded according to Patrick Hanks' Corpus Pattern Analysis (CPA) technique (Hanks 2004, 2013). In Woordcombinaties this technique has been tailored to the needs of the target audience. Sentences are sorted before annotation, using a specially developed GDEX[2] configuration to enable the output of short, comprehensible and yet informative sentences. Furthermore, arguments are represented by indefinite pronouns iemand 'someone', iets 'something', ergens 'somewhere', zo 'such' (or combinations thereof) in the pattern slots where this is possible for increased readability and the corresponding semantic types and syntactic functions are accessible through a tool tip. The slots are also enriched with collocations offering a kind of advanced word sketch in the patterns. Another interesting feature is that collocations for nouns can be filtered and displayed for each sense separately. Verbs do not have this feature as following Hanks' approach, their 'sense' is encoded in the patterns' implicature.

Idioms and conversational routines are also included in Woordcombinaties. Currently idioms and conversational routines are encoded as special instances among the collocations and the patterns. However, separate access with specific search options for idioms and conversational routines is planned and currently being designed. For instance, it will be possible to search for idioms based on image categories, such as 'body parts' and 'food' for een vinger in de pap hebben 'have a finger in the pie' and less specific sense categories, such as 'have a property'. Conversational routines will be linked to speech acts, such as 'greeting' or 'apologizing'.

This way, Woordcombinaties, forms a unique point of access for anyone who wants to learn more about Dutch phraseology.

References

- Hanks, P. (2004). Corpus pattern analysis. In Proceedings of the 11th EURALEX International Congress, pp. 87-98.
- Hanks, P. (2013). Lexical analysis: Norms and exploitations. Cambridge, MA: The MIT Press.

[1] https://skell.sketchengine.co.uk/
[2] https://www.sketchengine.eu/guide/gdex/

Definition Extraction for Slovene: Patterns, Transformer Classifiers and ChatGPT

Thi Hong Hanh Tran[1,2], Vid Podpečan[2], Mateja Jemec Tomazin[3], Senja Pollak[1]
[1]Jožef Stefan Institute, [2]La Rochelle University, [3]Fran Ramovš Institute of the Slovenian Language ZRC SAZU
E-mail: hanh.usth@gmail.com, vid.podpecan@ijs.si, mateja.jemec-tomazin@zrc-sazu.si, senja.pollak@ijs.si

Keywords: Definition Extraction; Rule-based; Transformers; Language models

Definition Extraction is a Natural Language Processing task that automatically identifies the terms and their corresponding definition from the unstructured text sequences. In our research, we frame this problem as a binary classification task, aiming to detect whether a given sentence is a definition or not, using text sequences in Slovene.

The main contributions of our work are two-fold. First, we introduce a novel Slovene corpus for the evaluation of Definition Extraction named RSDO-def. The corpus was collected as a part of the project Development of Slovene in a Digital Environment – Language Resources and Technologies. The sentences were extracted from the Slovene domain-specific corpora in two ways. On the one hand, we extracted 962 sentences by random sampling (RSDO-def-random); on the other hand, in order to improve the number of definitions which represent only a small number in the randomly sampled file, we extended this initial dataset by using pattern-based extraction methods, resulting in the RSDO-def-larger dataset. Both sets contain manual annotations by linguists with three labels: Definition, Weak definition, and Non-definition. Second, we propose the benchmarks for Slovene Definition Extraction systems that use (1) rule-based techniques; (2) Transformers-based models as binary classifiers using Wikipedia-based corpus as a training data; (3) ChatGPT prompting.

We evaluate the approaches on the RSDO-def corpus. The results demonstrate that if there are only a few well-structured instances of definitions that have clear linguistic characteristics (e.g., in the strict evaluation scenario, where Weak definitions are considered as non-definitions), a rule-based technique performed better in terms of F1-score (on the Definition class) than language models or prompting. However, for less structured examples (relaxed evaluation scenarios with Weak definitions considered as definitions), ChatGPT prompting and language models were more effective than classical rule-based approaches. When comparing prompting and language model classifiers, for the Definition class, classifiers lead to higher Precision, while in terms of Recall, ChatGPT has better results.

(Invisible) pictures in online dictionaries: Shall we see them?

Anna Dziemianko
Adam Mickiewicz University in Poznan
E-mail: danna@amu.edu.pl

Keywords: online dictionary; picture; reception; retention; access; hyperlink; dictionary use

Empirical research shows graphic illustrations in dictionaries to be useful in reception and retention (Nesi, 1989; Gumkowska, 2008; Dziemianko, 2022), but their harmful effect on vocabulary learning is attested, too (Van den Broek et al., 2021). Lexicographers need to decide how to display pictures, as presentation space is constrained in hand-held portables and regular computers, on which online dictionaries are mostly accessed (Kosem et al., 2019). It is worthwhile to see whether making pictures instantly visible or hyperlinking them is more recommendable.

The aim of the paper is to determine if the presence of pictures in online dictionaries and their access path (instant/default visibility vs. hyperlinking) affect meaning reception and retention. Four research questions are posed:

1. Does the reception of meaning depend on the presence of pictures in online dictionaries?

2. Is meaning reception affected by how pictures are accessed (immediately visible vs. available by clicking/tapping)?

3. Is the retention of meaning conditioned by the presence of pictures in entries?

4. Are pictures visible in entries by default or hyperlinked ones more useful for learning meaning?

An online experiment involved 15 English nouns and consisted of a pre-test, a main test and a post-test. In the pre- and post-tests, meaning had to be explained without access to dictionaries. In the main test, the same task had to be done following the consultation of purpose-built, monolingual online entries. Three test versions were created: with pictures visible by default, with pictures available by clicking/tapping hyperlinks, without any pictures. 238 learners of English (B2 in CEFR) took part in the experiment.

To analyze the data, one-way ANOVAs were conducted for each dependent variable (meaning reception and retention). Access to pictures was a between-groups independent variable. Significant ANOVA results were analyzed with the help of the Tukey HSD test.

The results show that meaning reception was dependent on pictures ($F=21.23$, $p=0.00$, partial $\eta 2 =0.503$). It was the most successful when entries offered pictures either visible by default (82.15%) or hyperlinked (81.98%), with no difference between these two conditions ($p=1.00$). In the absence of pictures, reception was significantly (about one fourth) worse (61.06%, $p=0.00$).

Learning meaning was also affected by pictures ($F=7.99$, $p=0.00$, partial $\eta 2 =0.276$). It was largely facilitated by pictures visible by default in the entry (62.07%). Entries with hyperlinked pictures (48.21%) were no more useful than those without any pictorial support (41.04%, $p=0.38$). In these conditions, meaning retention was, respectively, about one

fourth and one third worse than when pictures were instantly visible, and these differences were statistically significant ($p>0.05$).

The study shows that understanding meaning is affected by the presence of pictures in entries (RQ1). The way of accessing pictures proves to be inconsequential in this respect (RQ2). However, remembering meaning is dependent on whether pictures are instantly visible or hyperlinked (RQ3). The former significantly enhance retention, while the latter do not; they prove to be only as good as entries without pictures (RQ4). Thus, pictures immediately visible in online entries emerge as more recommendable for learning meaning.

References

- Dziemianko, A. (2022). The usefulness of graphic illustrations in online dictionaries. ReCALL, 34(2): 218–234.
- Gumkowska, A. (2008). The role of dictionary illustrations in the acquisition of concrete nouns by primary school learners and college students of English. Collegium Balticum, unpublished BA.
- Kosem, I., Lew, R., Wolfer, S., Müller-Spitzer, C., Silveira, M. (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): 92–114.
- Nesi, H. (1989). How many words is a picture worth? A review of illustrations in dictionaries. In Tickoo, M. L. (ed.), Learners' dictionaries: State of the art. Singapore: SEAMEO, 124– 134.
- Van den Broek, G. S. E., van Gog, T., Jansen, E., Pleijsant, M., Kester, L. (2021). Multimedia effects during retrieval practice: Images that reveal the answer reduce vocabulary learning. Journal of Educational Psychology, 113(8): 1587–1608.

Ensuring vocabulary consistency for an under-resourced language with limited data

Sven-Erik Soosaar[1], Valts Ernštreits[2]
[1]Institute of the Estonian Language, [2]University of Latvia Livonian Institute
E-mail: svenerik@eki.ee, valts.ernstreits@lu.lv

Keywords: under-resourced languages; dictionary building methods; bilingual dictionaries; Finno-Ugric languages

Livonian is an indigenous language of Latvia and one of the most endangered languages in the world. Being spoken by a limited number of speakers over past centuries (from 2500 in the mid of 19th century till ca 20 today; Laakso, 2022) it has been used in limited language domains and documented scarcely and unevenly.

Building dictionaries for such a language is challenging. Although currently largest dictionary (Livonian-Estonian-Latvian; Viitso & Ernštreits, 2012) containing ca 12 000 lemmas was published a decade ago, it lacks consistency in terms of even basic, everyday use vocabulary, as it has been created manually using fieldwork data. In order to ensure consistency and develop dictionary evenly, frequency data is needed for Livonian.

Approaching consistency from the perspective of frequency is beneficial from two key aspects – firstly is ensured that consistent basic vocabulary is available for the language acquisition purposes. Secondly – it enables making lexicographic products in creation being instantly available and consistent throughout the evolvement, which is especially important for languages or combinations struggling with lack of resources.

Such an example was the Estonian-Latvian dictionary (Ernštreits et al., 2015). For the lemma selection and building statistical metadata was used from 3 separate sources (5000 list; frequency dictionary 10000 and balances corpora divided by 5000 from 5000–40000). This was done to ensure, that dictionary instantly covers the core of the vocabulary and grows evenly, not sequentially (e.g. by starting letters).

However even having corpora of Livonian texts does not allow to acquire sufficient frequency data of lemmas due to unbalanced collections, e.g. one third of corpus developed by the ULLI represents religious literature (translation of the New Testament), and another third – fairy tales and legends.

To tackle this, we explore opportunities of using frequency data from other languages related to the Livonian – Estonian as its closest linguistic relative and Latvian as a main contact source for Livonian and a language impacted by Livonian itself. Our goal is to identify missing vocabulary, so that we can specifically look for that vocabulary from sources. We also use other methods that have been proven to be useful for under-resourced languages (Mittelholcz et al., 2017).

Currently we have applied a list of 5000 Estonian basic vocabulary (ELD) to the Estonian translations in (Viitso & Ernštreits, 2012), with 3100 already included. . In the near future we are planning to apply Latvian frequency data from balanced corpus of Latvian texts (`https://korpuss.lv/id/LVK2018`), ELD data (both Estonian and Latvian), and frequency data Livonian corpus frequency data to identify missing lemmas and category of their importance (5000–40 000). Our aim is also to compare all three – Estonian, Latvian and Livonian frequency sets to determine patterns and peculiarities that should be noticed when ensuring consistency of the Livonian data. Among others we will analyze the use of different methods for coining new terms (derivation, compounding, borrowing etc).

References

- Ernštreits, V., Muzikante, M., Grīnberga, M., Ernštreits, M. (2015). Igauņu-latviešu vārdnīca = Eesti-läti sõnaraamat. Latviešu valodas aģentūra. Riia, Tallinn: Eesti Keele Sihtasutus.
- Kallas, J., Tiits, M., Tuulik, M. (2014). Eesti keele põhisõnavara sõnastik. Tallinn, Eesti Keele Sihtasutus.
- Laakso, J., (2022). The Oxford Guide to the Uralic Languages. Edited by: Bakró-Nagy et al., Oxford University Press.
- Viitso, T.-R., Ernštreits, V. (2012). Līvõkīel-ēstikīel-leļkīel sõnārõntõz. Tartu Riga.
- Mittelholcz, S. (2017). Evaluation of Dictionary Creating Methods for Under-Resourced Languages. In: Ekštein, K., Matoušek, V. (eds) Text, Speech, and Dialogue. TSD 2017. Lecture Notes in Computer Science(), vol 10415. Springer, Cham.

Ceci n'est pas un dictionnaire. Adding and Extending Lexicographical Data of Medieval Romance Languages to and through a Multilingual Lexico-Ontological Project

Sabine Tittel
Heidelberg Academy of Sciences and Humanities
E-mail: sabine.tittel@hadw-bw.de

Keywords: historical lexicography; Romance languages; Linked Open Data

$$
\begin{gathered}
([\text{lexeme} \in \text{text}] + [\text{lexeme} \in \text{dictionary resource}] + [\text{text} \in \text{corpus}]) \\
\times \text{four medieval Romance languages} \\
= \\
[\text{lexeme} \in \text{multilingual analysis} \in \text{medieval knowledge network}] \\
+ [\text{lexeme concept} \in \text{ontology}] \\
+ [\text{lexeme / text / dictionary resource} \in \text{Linked Open Data cloud}]
\end{gathered}
$$

This is a ragged attempt to put the approach of the newly launched long-term project ALMA[1] into a nutshell: Lexicographical data is a crucial element of both its empirical basis and its outcome. The latter is *not a real dictionary* but can be interpreted as a particular—*real*—elaboration of a dictionary, much like Magritte's pipe[2] yet on another abstraction level. ALMA is a representation of concept-driven lexical-semantic analyses that are deeply rooted in long-approved approaches to lexicography.

The aim of the ALMA project is to investigate the interaction between language, knowledge, and scholarship. The field of observation is the Romance cultural sphere that sees the emergence of new knowledge networks between 1100 and 1500 AC expressed in vernacular languages (ALMA focuses on medieval Italian, French, and Occitan) and exemplified by two knowledge domains ('medicine', 'law'). The arising 'scientific' languages are a particularly important part of the intellectual heritage of Europe.

Based on two domain-specific, multilingual text corpora, ALMA will elaborate lexical-semantic studies on the linguistic manifestations of the knowledge networks. The project's relations to lexicography are manifold:

(1) Re-use: ALMA confronts its data with the data of the state-of-the-art dictionaries that examine the language in all functional areas, beyond the technical vocabulary in question. ALMA finds itself in the lucky position of being able to re-use—through database access—the published as well as raw data of the pertinent dictionaries for the research field: LEI, DEAF, and DOM.[3] The funding of these dictionaries, apart from the LEI, recently ended, hence, re-use through ALMA is an excellent means to keep the valuable data alive as part of an innovative workflow.

(2) Extension: ALMA extends dictionary knowledge through lexical-semantic studies on lexical units of its specific domains based on new corpus material. The studies show

[1] *Wissensnetze in der mittelalterlichen Romania / Knowledge Networks in Medieval Romance Speaking Europe*, Academies of Sciences and Humanities Heidelberg / Bavaria / Mainz; directed by Elton Prifti / Wolfgang Schweickard (Mainz), Maria Selig (München), Sabine Tittel (Heidelberg), duration 22 years; `https://www.hadw-bw.de/alma` [2023-01-05].

[2] `https://collections.lacma.org/node/239578` [2023-01-05].

[3] *Lessico Etimologico Italiano* (LEI), *Dictionnaire étymologique de l'ancien français* (DEAF), *Dictionnaire de l'occitan médiéval* (DOM).

many dictionary-like features including senses and sub-senses in a hierarchical tree, genus-differentia definitions, contexts (corpus material) for encyclopedic illustration, graphical apparatus separated from semantics, and etymological and linguistic discussions.

(3) Further Processing: ALMA models the pertinent articles of DEAF, LEI, and DOM as Linked Open Data (following Tittel & Chiarcos, 2018). The lexical units will be mapped to historicised domain ontologies for medieval medicine and law developed by ALMA. This will create a frame-like architecture of historical Semantic Web resources for ALMA-LOD resources, strengthening the hitherto under-represented historical resources of the LOD cloud.

(4) Dissemination: The dictionary articles are made accessible as RDF resources (Cyganiak et al. 2004-2014). Via mapping to the extra-linguistic ontologies, their linguistic and historico-cultural knowledge will be introduced into a knowledge circulation beyond the scopes of lexicography and historical linguistics.

References

- DEAF. Dictionnaire étymologique de l'ancien français, founded by Kurt Baldinger, continued by Frankwalt Möhren and Thomas Städtler. Québec (Presses de L'Université Laval)– Tübingen (Niemeyer)–Berlin (De Gruyter), 1971-2021.
- LEI. Lessico Etimologico Italiano, founded by Max Pfister, directed by Elton Prifti and Wolfgang Schweickard. Wiesbaden (Reichert), 1979.
- DOM. Dictionnaire de l'occitan medieval, founded by Wolf-Dieter Stempel, continued by Maria Selig. Berlin [i. a.] (De Gruyter), 1996-2013.
- Cyganiak, R., Wood, D., Lanthaler M. (2004–2014). RDF 1.1. concepts and abstract syntax: W3C recommendation 25 February 2014. URL: `https://www.w3.org/TR/rdf11-concepts/`.
- Tittel, S., Chiarcos, Ch. (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), 7-12 May 2018, Miyazaki, Japan. Paris (ELRA), pp. 58-66.

Towards a lexical database of Dutch taboo language

Gerhard Van Huyssteen[1], Carole Tiberius[2]
[1]Centre for Text Technology (CtexT), North-West University, [2]Instituut voor de Nederlandse Taal
E-mail: gerhard.vanhuyssteen@nwu.ac.za, carole.tiberius@ivdnt.org

Keywords: Dutch, lexical database, swearword, taboo language

Over the past 45 years, at least eighteen Dutch paper-based dictionaries of taboo-language (or taboo-related language) have been published (i.e., as visible works of lexicography). However, none of these are available as (linked) lexical data that could be integrated in natural language processing (NLP) tools and applications (i.e., as invisible works of lexicography). In this paper, we describe the development of a comprehensive lexical database of taboo language (LDTL) for Dutch (TaboeLex) that can be integrated in NLP tools and applications. TaboeLex will be made available as open data, i.e., as a freely available, structured, annotated lexicon that can be linked to other data in the future. The paper focusses on the first phase of the project, namely, to define and design TaboeLex.
Warning: This paper contains content that may be offensive or upsetting.

Establishing criteria and procedures to identify conventionalized similes in Croatian

Jelena Parizoska[2], Ivana Filipović Petrović[1], Kristina Kocijan[2]
[1]Croatian Academy of Sciences and Arts, [2]University of Zagreb
E-mail: jparizoska@gmail.com, ifilipovic@hazu.hr, krkocijan@ffzg.hr

Keywords: similes; Croatian; hrWaC; Corpus Query Language; NooJ

This study deals with a subset of Croatian idiomatic expressions – similes – which follow the pattern adjective + *kao/ko* ('as') + noun (e.g. *tvrd kao kamen* lit. hard as stone 'very hard'). The aim is to establish the criteria and procedures which can be used to identify conventionalized similes in a large corpus. A set of similes thus obtained may be used in dictionary-making and/or to create a lexical database. Furthermore, corpus findings were used to create a rule-based grammar of similes in NooJ.

We conducted a study in the Croatian web corpus hrWaC (Ljubešić and Klubička, 2016). Four types of queries were designed using Corpus Query Language (CQL) in the Sketch Engine: 1) $adjective+kao+noun$, 2) $adjective+kao+any\ word+noun$, 3) $adjective+kao+1-2\ words\ in\ between+noun$, 4) $adjective+1-2\ words\ in\ between+kao+noun$. Five groups of results were obtained:

- The majority of constructions containing *ko* are similes (e.g. *gladan ko vuk* lit. hungry as a wolf 'very hungry'), whereas *kao* more commonly occurs in non-idiomatic constructions (e.g. *poznat kao grad bicikla* 'known as a bicycle town').

- Nouns and adjectives occur in the singular and plural, as well as in masculine and feminine forms.

- Some constructions contain the same adjective and different nouns, e.g. *pijan kao letva* (lit. drunk as a lath) and *pijan kao deva* (lit. drunk as a camel) 'very drunk'; *hladan kao led* (lit. cold as ice) 'very cold' or 'very unfriendly' and *hladan kao špricer* (lit. cold as a spritzer) 'calm and composed'.

- Some nouns, e.g. *pas* 'dog' occur with a number of adjectives: *umoran* 'tired', *ljut* 'angry', *gladan* 'hungry', *ružan* 'ugly', *ljubomoran* 'jealous'.

- In some similes the adjective preceding the noun may be omitted, e.g. *sretan kao malo dijete* (lit. happy as a little child) and *sretan kao dijete* (lit. happy as a child) 'very happy'.

A dictionary and a rule-based grammar of similes were created in NooJ on the basis of the results of CQL queries in hrWaC. The grammar may be used for the automatic detection of similes in a large corpus, e.g. Croatian MaCoCu (Bañón et al., 2022). It may also serve to identify other structural ypes of similes, e.g. those which follow the pattern verb $+kao/ko('as')+$ noun (e.g. *raditi kao konj* lit. work like a horse 'work very hard'). Furthermore, the rule-based grammar created for Croatian similes may be extended to other Slavic languages and adapted depending on the structural type (e.g. Slovene and Serbian vs. Polish and Russian).

References

- Bañón, M. et al. (2022). Croatian web corpus MaCoCu-hr 1.0, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, URL: `http://hdl.handle.net/11356/1516`.
- Ljubešić, N., Klubička, F. (2016). Croatian web corpus hrWaC 2.1, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, URL: `http://hdl.handle.net/11356/1064`.

Invisible lexicography enhances neural machine translation

Ilan Kernerman

Lexicala by K Dictionaries
E-mail: ilan@lexicala.com

Keywords: invisible lexicography; machine translation; parallel corpora; examples of usage; bilingual segments; learning models; training data

Neural machine translation (NMT) relies on learning models trained on masses of language data. However, in most (commercial) cases, the training data stem heavily from web-crawled parallel corpora that are often inflicted by diverse flaws. Besides having to first clean "noise" for smoother processing, a major concern is data that might originate from other machine translation systems and whose quality is not clear. In addition, details of the source and the usage license are often lacking, posing a serious setback for NMT enterprises regarding copyright issues. Another drawback is the relative scarcity of such data for languages other than English, particularly under-resourced ones.

Acknowledging these and other drawbacks in automatically harvested data (e.g. suiting specialized languages), last year two Big Tech leaders released multilingual parallel corpora they created manually from scratch, with English as source language: Amazon announced the MASSIVE dataset, containing one million 'segments' that consist of 20,000 utterances translated to 50 languages, including typical instructions and questions for training Alexa virtual assistant (Fitzgerald, 2022); and, Microsoft announced NTREX-128, a dataset based on nearly 2,000 news text segments translated by human professionals (without post-editing) to 128 languages (Federmann et al., 2022).

Against this background, legacy lexicography has a lot to offer, as it thrives on deep and meticulous research by experts into individual languages and across languages, accessible in systemic, well-structured data representation. Lexicographic resources, especially those developed for learner's dictionaries (but also others), incorporate typical linguistic patterns that are particularly valuable for training language and translation models, thereby improving the quality of their by-product services – from search engines to chatbots, etc. – including those for machine translation.

In this talk I will describe these advantages and present a recent lexicography-derived project aimed at enhancing the performance of NMT training models. We used 260,000 examples of usage from learner's dictionaries and their translations (i.e. bilingual segments) for four Korean language pairs. Besides the English Korean set, where 93% of the segments were translated directly, all the others were joined automatically via a third pivot language and were reviewed thoroughly, and revised if necessary, by professional translators. Some of the bilingual segments reoccur in a few of or in all four pairs, making them multilingual, and some also feature extra lexicographic components that enrich the results, such as domain labels from the original dictionary entries. I will discuss the superior qualities of dictionary examples of usage that illustrate typical language patterns, the software methods used to develop the data, the editorial guidelines for the translators' work, and the convergence of human created and curated lexicographic content with auto-generated data.

References

- Federmann, C., Kocmi, T., and Xin, Y. (2022). NTREX-128 – News Test References for MT Evaluation of 128 languages. Proceedings of the First Workshop on Scaling Up Multilingual Evaluation (SUMEval 2022). ACL: pp 21-24.
- Fitzgerald, J. (2022). Amazon releases 51-language dataset for language understanding. Amazon Science. URL: `https://amazon.science/blog/amazon-releases-51-language-dataset-for-language-understanding`

Virtual lexicographic laboratory in linguistic researches based on the dictionary content

Yevhen Kupriianov[1], Volodymyr Shyrokov[2], Mykyta Yablochkov[2], Iryna Ostapova[2]

[1]National Technical University Kharkiv Polytechnic Institute Kharkiv Ukraine, [2]Ukrainian Lingua-Information Fund NAS of Ukraine Kyiv

E-mail: eugeniokuprianov@gmail.com, vshirokov48@gmail.com, gezartos@gmail.com, irinaostapova@gmail.com

Keywords: virtual lexicographic laboratory; explanatoty dictionary; lexicographic text; lexicographic system; dictionary-based researches

The virtual lexicographic laboratories (shortly VLL) are the resources to perform linguistic studies on the basis of a dictionary content in digital medium. In this context we refer comprehensive explanatory dictionaries giving a thorough and in-depth descriptions for any language unit in question. However, such detailed description may complicate the search for a needed information about single unit or a group of units. Therefore, the issue is how to supply these dictionaries with the proper tools to extract any linguistic information from the dictionary content.

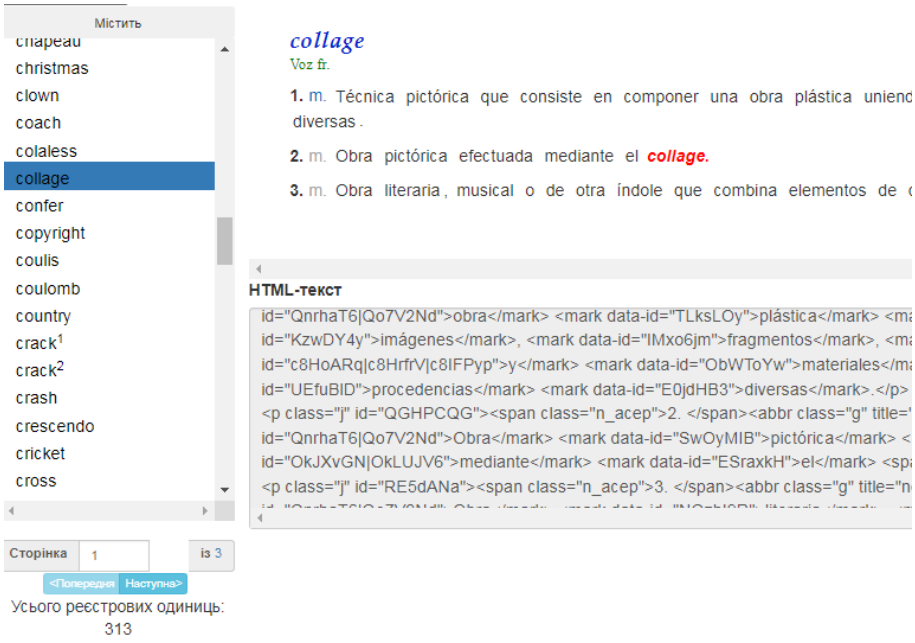


Figure 1: Sample of the foreign words used in modern Spanish.

This paper offers the authors' experience in this regard by demonstrating their own development – the VLL DLE 23 created for working with the Spanish language dictionary "Diccionario de la lengua española, 23a ed." (DLE 23). At present VLL DLE 23 allows Spanish vocabulary classification, as well as deriving a sample of the head words having common morphological, semantic and word combination properties. The current version of the VLL DLE 23 can be accessed at https://svc2.ulif.org.ua/Dics/ResIntSpanish. Examples of using the VLL interface to perform linguistic experiments are given below.

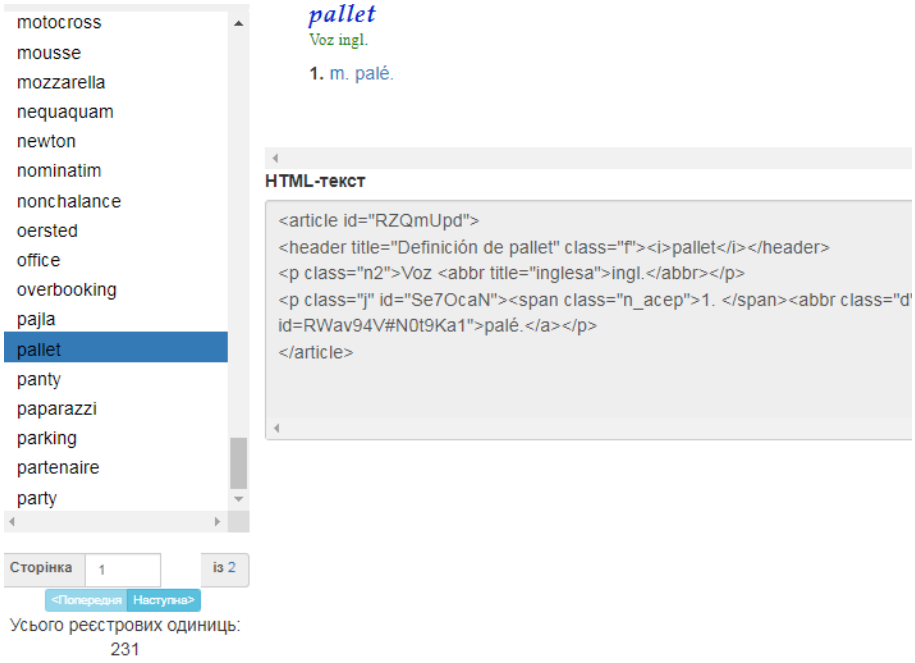


Figure 2: Monosemantic foreign words in the Spanish language.

Example 1. Let's make a sample of the words borrowed from other languages. As a result, the laboratory selected 313 units from DLE 23 (Fig. 1).

Example 2. We may want to know the availability of monosemantic words among those selected in the example 1 (Fig. 2). The total of the words in the sample is 231.

Example 3. By the user's request the VLL DLE 23 can make a sample of the words denoting some specific objects. For example, we can form a sample of the headwords denoting "instrumento de hierro" (iron tools) in Spanish (Fig. 3).

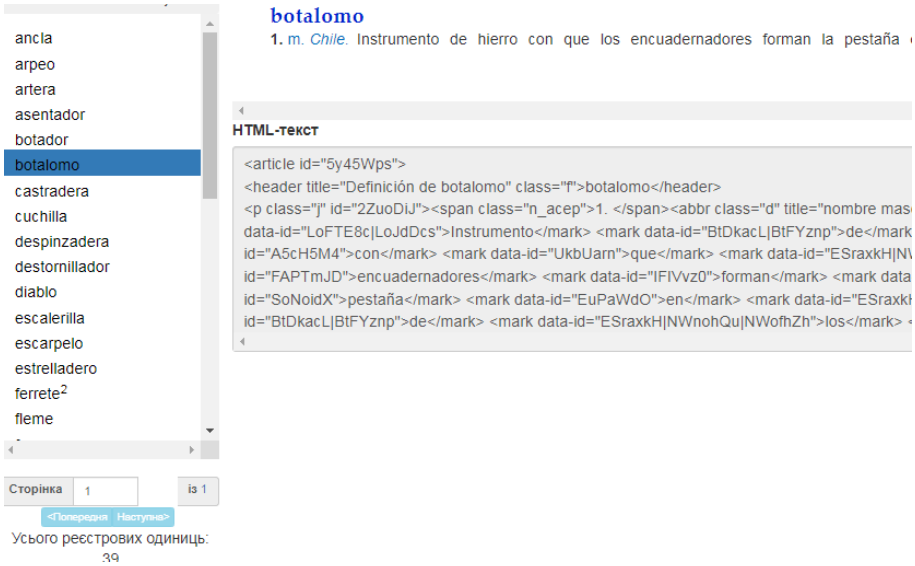


Figure 3: Sample of the words denoting different kinds of iron tools in Spanish.

Example 4. Another example is related to studying word combination features of Spanish by using different suffixes. In particular we can study the peculiarities of forming derived words. In the entries such words are described by using different definition patterns like “De manera...” (In ... way), “Acción de ...” (Action from...), “Que siente...” (who feels...) etc. The figure 4 shows the sample of the derived words having the meaning “Que siente” (Who feels smth).

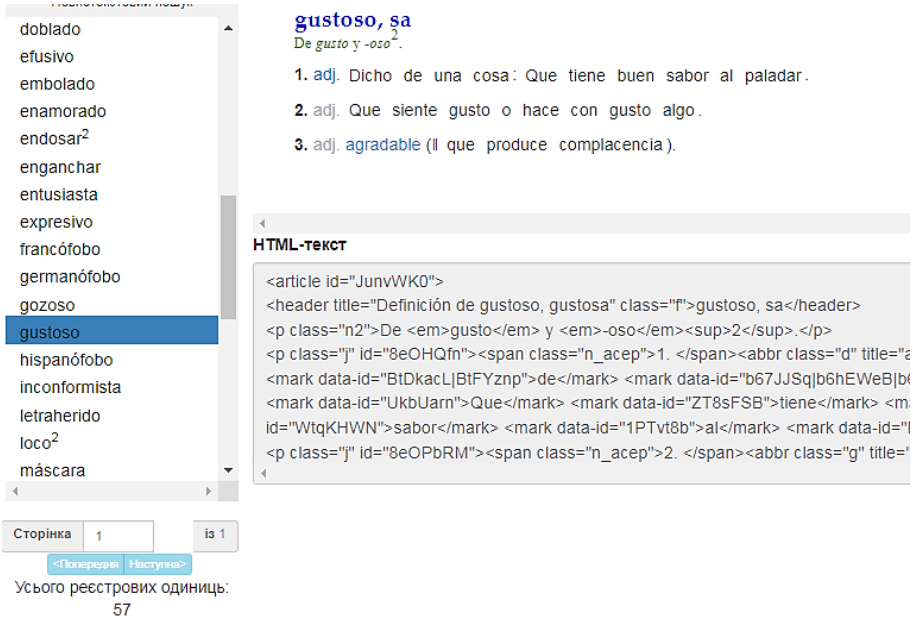


Figure 4: The sample of the derived words in the meaning of “Que siente” (Who feels smth).

The work with VLL DLE 23 project is underway. The next version will offer extended access to all the entry elements of DLE 23 and their combinations.

The Czechoslovak Word of the Week. Rejoining Czech and Slovak together in a piece of an invisible lexicography work

Michal Škrabal[1], Vladimír Benko[2], Peter Malčovský[2], Jan Kocek[1]

[1]Institute of the Czech National Corpus, Charles University, [2]Slovak Academy of Sciences, Ľ. Štúr Institute of Linguistics
E-mail: michal.skrabal@ff.cuni.cz, vladimir.benko@juls.savba.sk, peter.malcovsky@juls.savba.sk, jan.kocek@ff.cuni.cz

Keywords: Czech; Slovak; JHipster application generator; Vue front-end framework; Word at Glance interface; PostgreSQL database

The Czechoslovak Word of the Week is a joint popularization project of the Institute of the Czech National Corpus and the Ľ. Štúr Institute of Linguistics of the Slovak Academy of Sciences, that was inaugurated on the occasion of the 30th anniversary of the dissolution of Czechoslovakia (January 1, 1993). Throughout the year, each and every week, we will be publishing a new entry on the project website (`https://slovo.juls.savba.sk`), written parallelly in Czech and Slovak. We intend to draw the attention of both the Czech and the Slovak publics (especially the younger generation, for whom the former mutual intelligibility between the two languages no longer holds) to the interesting parallels, but chiefly the differences, between our two languages. We try to do so in a user-friendly and entertaining way, the central part of each entry being a language feuilleton (a very popular genre in the Czech Republic and Slovakia), supplemented with data drawn from language corpora (SYN2015, SYN2020, and ORAL v1 for Czech; prim-10.0-public-all and s-hovor-7.0 for Slovak; the Czech-Slovak section of the parallel corpus InterCorp) and the respective entries from some older monolingual and bilingual dictionaries (Bernolák, 1825; Jungmann, 1835-1839; SSJČ, 1960-1971; SSJ, 1959-1968; KSSJ, 2003; ČSS, 1981; SČS, 1967). In a way, we see the website as being a dictionary with a fixed macrostructure (52 entries including some multi-word units) and a microstructure determined by the order of the individual components (described in Škrabal & Benko, 2019: 475-476). Thus, our project could be considered a good example of "invisible lexicography" in practice. The target audience is presented with various kinds of lexicographic information unobtrusively, covertly, and "invisibly," usually without them having the feeling that they are "leafing through" a dictionary.

At this year's eLex, we would like to present not only the website itself but also the database behind it within the software presentation/demo section. Our solution uses modern web technologies: the JHipster application generator (`https://www.jhipster.tech/`) in combination with the Vue front-end framework (`https://vuejs.org/`), and the PostgreSQL database (`https://www.postgresql.org/`). The application allows the administrator to easily enter content, including importing and formatting texts from various sources (dictionary portals, Word documents, etc.), and to use audio samples from spoken corpora as well. The website itself is graphically based on the Word at Glance interface (Machálek, 2019, 2020), as the original layout was adapted to the needs of our project.

References

- Machálek, Tomáš (2019). Slovo v kostce – agregátor slovních profilů. FF UK, Praha. URL: `http://korpus.cz/slovo-v-kostce/`.
- Machálek, Tomáš (2020). Word at a Glance: Modular Word Profile Aggregator. In: Proceedings of LREC 2020, p. 7011-7016.
- Škrabal, Michal & Benko, Vladimír (2019). Make my (Czechoslovak word of the) day. In: Kosem, I. et al. (eds.). Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference. 1-3 October 2019, Sintra, Portugal. Brno: Lexical Computing CZ, s.r.o., p. 467-477

Dictionaries

- Bernolák, A. (1825). Slowár Slowenskí Češko-Laťinsko-Ňemecko-Uherskí seu Lexicon Slavicum Bohemico-Latino-Germanico-Ungaricum. Budae: Typis et Sumtibus Typogr. Reg. Univers. Hungaricae.
- [ČSS] Česko-slovenský slovník (1981). Bratislava: Veda.
- Jungmann, Josef (1835-1839): Slownjk česko-německý. Praha: Knížecí arcibiskupská knihtiskárna.
- [KSSJ] Krátky slovník slovenského jazyka. Bratislava: Veda.
- [SČS] Gašparíková, Želmíra & Kamiš, Adolf (1967): Slovensko-český slovník. Praha: SPN 1967.
- [SSJ] Slovník slovenského jazyka (1959-1968). Bratislava: Vydavateľstvo SAV.
- [SSJČ] Slovník spisovného jazyka českého (1960-1971). Praha: ČSAV.

The Central Word Register of the Danish language

Thomas Widmann
Dansk Sprognævn
E-mail: tw@dsn.dk

Keywords: lexical database; orthography; Danish language; historical lexicography

Det Centrale Ordregister (”The Central Word Register”) is a unique and innovative lexical database for the Danish language. Developed by the Danish Language Council, the Danish Society for Language and Literature and the Centre for Language Technology at the University of Copenhagen, with funding from the Agency for Digital Government, the COR assigns unique identification numbers to every lemma and form of the Danish language.

At the heart of the COR lies ”Retskrivningsordbogen”, the official orthographical dictionary of Danish, which provides the foundation for the unique identification numbers. The Danish Language Council will update this basis whenever the orthography changes, publishing the changes compared to the previous version, ensuring that the COR will always reflect the orthography of the day while ensuring that existing resources will continue to function even when the orthography changes.

The COR is divided into three levels, with Level 1 corresponding to the orthographical dictionary, Level 2 encompassing additional resources from professional language bodies and Level 3 comprising all other resources, with no restrictions on who can contribute. Version 1.0 of Level 1 was released by the Danish Language Council in September 2022. The Danish Society for Language and Literature and the Centre for Language Technology are currently working on adding a semantic component on Level 2.

The primary goal of the COR is to create a common key that enables more efficient reuse of language resources, similar to the way Denmark’s *Central Person Register* (CPR) allows different databases containing information about the inhabitants of Denmark to communicate with one another.

The COR database can be easily accessed through a downloadable CSV file or an API, allowing developers to retrieve ID numbers, lemmas, and forms in either CSV or JSON format, providing a great example of *invisible lexicography*.

The project also opens up new possibilities for historical lexicography, as the Danish Language Council intends to make its previous orthographical dictionaries available in COR format, enabling users to track the evolution of the language over time, to study historical texts in a more accurate way and to modify NLP software to work on historic texts.

We will also discuss the development of COR linkers (programs that will assign the correct COR number to every word in a text) and how these are effectively solving the problems of part-of-speech tagging and homograph resolution at once. An example of a COR linker is the Danish Language Council’s CLINK project.

Another aspect of the COR is the ability to use crowdsourcing in lexicography. Users can contribute their own data and insights, simply by publishing their with data with added COR ID numbers. This fosters greater collaboration and enables the creation of a plethora of rich, dynamic resources for the Danish language.

We will explore the benefits and potential applications of the COR and discuss the exciting possibilities this creates for the future of the Danish NLP and language research.

The impact of multiple corpus examples in English monolingual learners' dictionaries on language production

Bartosz Ptasznik
University of Warmia and Mazury in Olsztyn, Poland
E-mail: bartosz.ptasznik@uwm.edu.pl

Keywords: corpus example; online dictionary; English monolingual learner's dictionary; production; empirical study

The present contribution focuses on the topic of example sentences in English monolingual learners' dictionaries. Example sentences play a major role in pedagogical dictionaries which are written for advanced learners of English, as they illustrate how words are used by native speakers of the language and, consequently, how they should be used by second language learners. They are a necessary element of the microstructure of a dictionary entry and facilitate the process of learning a foreign language from the point of view of reception and production. In the digital era of lexicography, it has become common practice for lexicographers to supply dictionary entries with multiple corpus examples exhibiting various types of syntax and collocation patterns (for example, in the Longman Dictionary of Contemporary English). This suggests that nowadays online dictionary users are exposed to a dictionary-using environment encompassing a considerable amount of lexicographic data. By and large, dictionary users have made it clear that they would prefer to be given more example sentences in dictionary entries, given the need to improve their English language production skills. But what does "more examples" actually mean? What is the optimal number of example sentences in a dictionary entry that benefits dictionary users in practice?

References

- Frankenberg-Garcia, A. (2012). Learners' use of corpus examples. International Journal of Lexicography. 25(3), 273—296.
- Frankenberg-Garcia, A. (2014). The use of corpus examples for language comprehension and production. ReCALL. 26(2), 128—146.
- Frankenberg-Garcia, A. (2015). Dictionaries and encoding examples to support language production. International Journal of Lexicography. 28(4), 490—512.
- Ptasznik, B. (2023). More examples may benefit dictionary users. International Journal of Lexicography. 36(1), 29—55.

Invisible meaning relations for representing near equivalents

Arvi Tavast, Kristina Koppel, Margit Langemets, Silver Vapper, Madis Jürviste

Institute of the Estonian Language
E-mail: arvi.tavast@eki.ee, kristina.koppel@eki.ee, margit.langemets@eki.ee,
silver.vapper@eki.ee, madis.jyrviste@eki.ee

Keywords: data model; bilingual dictionary; approximate equivalents

The Institute of the Estonian Language (EKI) has been developing an in-house dictionary writing system Ekilex (Tavast et al., 2018, 2020) since 2017. One of its central design principles is the symmetry of its data model: the many-to-many relationship between word and meaning simultaneously accommodates semasiological and onomasiological resources. It is now being used for compiling the general dictionary of Estonian – EKI Combined Dictionary (CombiDic - Langemets et al., 2021) – as well as over 120 termbases, with lexicographers and terminologists working in oppositely oriented views on the same data. Readers can access the completed resources in the language portal Sõnaveeb ['Word Web']1 (Koppel et al., 2019).

One of our purposes has been to add more languages to the CombiDic. Russian, French and Ukrainian (in different coverages) were either incorporated during the process of creating the initial database or have been compiled later manually. As for English, we started a new project in 2021 to automatically generate a dataset of English candidate equivalents, with the aim of designing and testing the whole process for adding other languages in the future.

The project has involved different (partly parallel) phases: a) collecting available dictionary data and parallel texts and importing the best candidates, b) designing the representation of near (narrower, wider and approximate) equivalents both in the database and for the reader, c) developing a specialized view for a lexicographer working with adding a language.

a) Equivalents were gatherd by processing sentence pairs and doing word alignments using ArgMax matching method (Sabet et al., 2020), which were then gathered in frequency lists.

b) While exact equivalents are simply related to the same meaning in our symmetrical data model, near equivalents are represented using relations between meanings, with the counter-intuitive consequence that not all meanings have designations in all languages. For the reader, these meaning relations are traversed in order to render a habitual presentation of near equivalents.

c) In order to semi-automatically compile bilingual dictionaries, a specialized view was created where the lexicographer is presented with the sense distribution of the headword with approximately 50 automatically detected candidate equivalents. Lexicographer's task is to work through the list of possible equivalents and to drag and drop them to corresponding senses. If the relevant equivalents are missing, they can be added manually. The lexigographer can also assign, whether it is the case of a narrower, wider or approximate equivalent.

In this paper we will report on the achievements and the lessons learned during the project.

References

- Langemets, M. et al. (2021). Eesti keele ühendsõnastik 2021 [EKI Combined Dictionary 2021]. **CombiDic**.
- Koppel, K., Tavast, A., Langemets, M., Kallas, J. (2019). Aggregating dictionaries into the language portal Sõnaveeb: issues with and without a solution. In: Kosem, I., Zingano Kuhn, T., Correia, M., Ferreria, J. P., Jansen, M., Pereira, I., Kallas, J., Jakubíček, M., Krek, S. & Tiberius, C. (Ed.). Proceedings of the eLex 2019 conference. 1–3 October 2019, Sintra, Portugal. (434–452). Brno: Lexical Computing CZ, s.r.o.
- Sabet, M.J., Dufter, P., Yvon, F., Schütze, H. (2020). SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. arXiv preprint arXiv:2004.08728.
- Tavast, A., Langemets, M., Kallas, J., Koppel, K. (2018). Unified Data Modelling for Presenting Lexical Data: The Case of EKILEX. Proceedings of the XVIII EURALEX International Congress: EURALEX: Lexicography in Global Contexts, Ljubljana, 17-21 July 2018. Ed. Jaka Čibej, Vojko Gorjanc, Iztok Kosem, Simon Krek. Ljubljana University Press, Faculty of Arts, pp. 749–761.
- Tavast, A., Koppel, K., Langemets, M., Kallas, J. (2020). Towards the superdictionary: layers, tools and unidirectional meaning relations. Proceedings of XIX EURALEX Congress: Lexicography for Inclusion, Vol. I. Ed. Gavriilidou, Z, Mitsiaki, M, Fliatouras, A. Alexandroupolis, Greece: Democritus University of Thrace, pp. 215–223.

Theoretical Bases for Dutch-Persian Learner's E-dictionary and its Realisation

Said M.H. Abafar

Individual researcher/lexicographer
E-mail: abafar@casema.nl

Keywords: Learner lexicography; E-lexicography; Learner's e-dictionary; Communicative-productive dictionary; Cognitive-receptive dictionary; Lexicographic systems.

Considerable need for modern Dutch-Persian dictionary was the motivation to create the first Dutch- Persian learner's dictionary. The upcoming paper discusses some of the problems I encountered in my research to find and implement the suitable solution for making this e-dictionary based on the four fundaments (bases) of Learner lexicography (the linguistic, methodical, lexicographic, and technological).

To proceed from the linguistic basis, I defined, for example, which headwords should be included in the dictionary, and what type of information should be provided for each headword. The methodical basis is a specific item in the Learner lexicography. Information in the dictionary must be organized in a way that should help the users learn the words in context. Special attention must be paid to the specifications of the users' native languages (the target language) and the speakers' known difficulties with learning source language (in this case Dutch). Another valuable item is the inclusion of (interactive) exercises to the dictionary. The lexicographical basis contains two aspects: 1) the lexicographical approach, and 2) the macro- and microstructure of the dictionary and the design and layout of the dictionary articles. Finally, the technological basis (which deals the most with scientific discussions in e-lexicography) is about the ways we could develop e-dictionaries, find technological solutions, use existing resources and also make dictionaries suitable for integration into modern lexicographic platforms and (visible or invisible) into other applications.

The learners' dictionary consists of three parts: the communicative-productive, the cognitive-receptive Dutch-Persian, and the cognitive-receptive Persian-Dutch.

In the communicative-productive dictionary, a set of frequently used Dutch words has been collected and given the status of headwords. Each headword has been provided with a dictionary article (the main entry) that contains linguistic information about the headword and its use in texts. That set of information includes semantic, phonetical, orthographical, morphological, syntactic, derivation, collocations, examples, etc.

In the cognitive-receptive part, all words used in the communicative-productive part, included and provided (only) with translation equivalents. The headwords are also marked (with a hyperlink) and refer to the related article in the communicative-productive part.

All three parts are components of one dictionary and use a common database. All three components can be opened via one web-based interface. Dictionary articles are xml-based. In this way I hope to make the dictionary suitable for publication online or integration into one if the available dictionary platforms (e.g., Elexis) in the future.

While the work on all three parts of the dictionary is presently underway, the exercise component is in its design phase and is oriented towards the use of features provided by opensource tools or modern chat boxes.

These technological aspects will also be discussed in this paper.

References

- Electronic lexicography in the 21st century: post-editing lexicography. In I. Kosem et al. (eds.) Proceedings of the eLex 2021 conference. Brno, July 2021. Lexical Computing CZ s.r.o., Brno, Czech Republic. 664 p. ISSN 2533-5626.
- Electronic lexicography in the 21st century: Smart lexicography. In I. Kosem et al. (eds.) Proceedings of eLex 2019 conference. Sintra, October 2019. Lexical Computing CZ s.r.o., Brno, Czech Republic. 984 p. ISSN 2533-5626.
- 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.
- E-Lexicography: The Internet, Digital Initiatives and Lexicography/Redactors: Pedro A. Fuertes- Olivera and Henning Bergenholtz, Continuum, 2011. 341 p. ISBN: 978-1-4411-2806-5.
- Leroyer P. (2011). Change of Paradigm: From Linguistics to Information Science and from Dictionaries to Lexicographic Information Tools E-Lexicography: The Internet, Digital Initiatives and Lexicography/Redactors. Continuum. P.P. 121 – 140. ISBN: 978-1-4411-2806-5.
- Kilgarriff A. (2003). What computers can and cannot do for lexicography or Us precision, them recall. University of Brighton and Lexicography Masterclass Ltd. UK.
- Morkovkin V.V. (1990). Основы Теории Учебной Лексикографии (Foundations of the Theory of Learner Lexicography), Doctoral dissertation. Pushkin Institute of Russian language. Moscow.
- Bergenholtz, H. & Tarp, S. (eds.) (1995). Manual of Specialised Lexicography. Amsterdam/Philadelphia: John Benjamins.
- `https://elex.link/`
- `https://elex.is/`
- `https://euralex.org`

Military Feminine Personal Nouns: Corpus-based Update to the Web Dictionary of Ukrainian Feminine Personal Nouns

Olena Synchak
Ukrainian Catholic University (Lviv. Ukraine)
E-mail: o_synchak@ucu.edu.ua

Keywords: Military feminine personal noun; Web Dictionary of Ukrainian Feminine Personal Nouns (WDUF); r2u.org.ua; General Regionally Annotated Corpus of Ukrainian; GRAC; dictionary entry; subject labels

The paper investigates hundreds of newly coined feminine personal nouns from the military sphere and how corpus data can be used for their publication in online dictionaries. Particular attention is paid to the *Web Dictionary of Ukrainian Feminine Personal Nouns* (WDUF) (2022, published on r2u.org.ua) and the *Alphabet of Feminine Personal Nouns*, as well as their coverage of these lexical items in comparison with other dictionaries. The use of the *General Regionally Annotated Corpus of Ukrainian* (GRAC) in the selection of words, compilation of the dictionary entries and the frequency list of said words are presented. Due to semantic analysis, five lexico-semantic groups of military feminine terms are determined. For updating the WDUF, the author argues for the necessity of adding military subject labels to three of them. Using quantitative data from the corpus GRAC, a decision about the arrangement and quality of derivational alternatives among military feminine terms is drawn. These findings have affirmed the necessity to combine the approaches of traditional lexicography with the corpus-based ones, as well as to balance description with prescription.

References

- Vebslovnyk zhinochykh nazv ukrainskoi movy [A Web Dictionary of Ukrainian Feminine Personal Nouns]. (2022). URL: `https://r2u.org.ua/html/femin_details.html`.
- Abetka feminityviv [The alphabet of feminine personal nouns]. (2022). URL: `https://behindthenews.ua/spetsproiekti/po-toy-bik-genderu/abetka-feminitiviv-358/`
- Shvedova, M. et al. (2017–2022). GRAC: General Regionally Annotated Corpus of Ukrainian. Electronic resource: Kyiv, Lviv, Jena. URL: `http://uacorpus.org`.
- Synchak, O., Starko, V. (2022). Ukrainian Feminine Personal Nouns in Online Dictionaries and Corpora. COLINS-2022: 6th International Conference on Computational Linguistics and Intelligent Systems, Gliwice, Poland
- Slovnyk ukrainskoi movy [A Dictionary of the Ukrainian Language]. In 20 vols. Naukova Dumka, Kyiv, (2015–2022). Vols. 1–12. URL: `https://services.ulif.org.ua/expl/`.
- Slovnyk ukrainskoi movy [A Dictionary of the Ukrainian Language]. (2018–2022). URL: `http://sum.in.ua`.

Improving second language reading through visual attention cues to corpus-based patterns

Kate Challis[1], Tom Drusa

[1]Iowa State University of Science and Technology
E-mail: kchallis@iastate.edu, t.drusa@gmail.com

Keywords: second language acquisition; computer-assisted language learning; corpus-informed software; vocabulary; data driven learning

The patterns inherent to written text often remain opaque to second language learners due to the considerable cognitive demands that reading places on working memory. Learners must attend to the meaning of unknown words, the grammatical structure of sentences, and the meaning of the text as a whole – and this all simultaneously. One solution for helping learners to better attend to existing form, function, and frequency patterns within texts is through systematic visual attention cues, which may offload some of the burden on working memory. Lex-See is a Chrome browser extension that highlights words within a user-supplied text in a variety of shades and colors based on underlying corpus-based data about frequency and word class, and also provides further information about forms, definitions, and phonetic similarity, on mouse-over. Currently Lex-See is optimized for Czech, a less-commonly taught, morphologically rich language with a clear need for easily accessible corpus-informed language learning tools, but it is designed to work with any language for which lemma frequency, form, dictionary, and phonetic data can be supplied.

References

- Brezina, V., Gablasova, D. (2015). Is There a Core General Vocabulary? Introducing the New General Service List. Applied Linguistics, 36(1), pp. 1–22.
- Challis, K. (2022). Is there a core vocabulary for Czech? Introducing the Czech General Service List. doi:10.13140/RG.2.2.17678.02889.
- Frey, A., Bosse, M. L. (2018). Perceptual span, visual span, and visual attention span: Three potential ways to quantify limits on visual processing during reading. Visual Cognition, 26(6), pp. 412–429.
- Lee, H., Warschauer, M., Lee, J. H. (2020). Toward the establishment of a data-driven learning model: Role of learner factors in corpus-based second language vocabulary learning. The Modern Language Journal, 104(2), pp. 345–362.
- Plecháč, P. (2017). Euphonometer 2.0. Prague: Institute of Czech Literature, CAS. URL: `http://versologie.cz`.

Utilizing Natural Language Processing Technologies for Controlled Lexicon Building: A Pilot Study Focusing on English and Japanese Verbs

Daichi Yamaguchi[1], Hodai Sugino[1], Rei Miyata[2], Satoshi Sato[1]
[1]Nagoya University, [2]The University of Tokyo, Japan
E-mail: yamaguchi.daichi.e4@s.mail.nagoya-u.ac.jp, 89sugino1230@gmail.com,
ray.miyata@gmail.com, sato.satoshi.g9@f.mail.nagoya-u.ac.jp

Keywords: controlled lexicon building; word variation management; interchangeability of words; natural language processing application; automotive domain

This paper presents our ongoing research project to automate the process of controlled lexicon building.

A controlled lexicon is a set of approved words defined for a specific purpose, such as controlled authoring and translation (ASD, 2021; Møller & Christoffersen, 2006; Warburton, 2014). The proper use of a controlled lexicon can prevent textual variation, leading to improved text consistency and clarity. Although many controlled lexicons have been built for various purposes (Kuhn, 2014), there have been few examinations of the possibility of automating lexicon creation. Fundamentally, the process of building a controlled lexicon has not been well formalized. Miyata & Sugino (2020) presented corpus-based lexicon-building procedures and proposed the *interchangeability* of words in actual sentences as a key criterion to identify word variations. Nevertheless, their lexicon-building process mostly depends on human expertise, and the detailed steps for judging interchangeability have yet to be clarified. Because natural language processing (NLP) technologies based on deep learning have advanced rapidly, we envisage the effective use of such technologies in this process.

Hence, towards the automation of controlled lexicon building, we have formalized the lexicon-building process and examined the applicability of various NLP technologies. Following the corpus-based procedures in (Miyata & Sugino, 2020), the process of controlled lexicon building can be broadly divided into the following two steps:

Step 1. Connect words that are interchangeable to form word clusters.
Step 2. For each cluster, define one word as approved and the rest as unapproved.

In Step 1, to capture interchangeability, we quantify the different levels of word similarity using various NLP technologies:

(a) General similarity: Word embeddings, such as word2vec (Mikolov et al., 2013), trained on general domain corpora, such as web text, can be used. Conventional thesauri, such as WordNet (Princeton University, 2010), can also be used.
(b) Domain-specific similarity: Word embeddings trained on target domain corpora can be used.
(c) Domain-specific context-aware similarity: Contextualized embedding methods, such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), can be used.

For level (c), we examined the interchangeability of words in example sentences in the target domain corpus. For example, the verb “delete” can be replaced with “erase” in an example sentence “Delete the data”. If this consistently applies to other example sentences in

the target corpus, we can assume that a controlled lexicon should include either word but not both to avoid variation. To simulate human judgments regarding interchangeability, we used contextualized embedding methods to produce vector representations that encode not only target words but also their context.

In Step 2, we tested several algorithms to define the approved words based on the word frequency and the linguistic symmetricity of their antonyms. Although the frequency of words in the target corpus can be regarded as a major factor in deciding the approved words, the symmetricity of certain word pairs in a lexicon can sometimes precede frequency evidence. For example, if the verb "engage" is already defined as approved, the symmetric verb "disengage" is likely to be selected as approved instead of a synonymous verb "detach", even if the latter is more frequently observed in the corpus than the former. To capture the symmetricity of words, we devised language-specific heuristic rules that use linguistic or textual clues, such as verb constructions (e.g., *sa*-hen noun + *suru* construction) and n-gram overlap at the character level (e.g., "**engage**" and "dis**engage**").

First, we explain our overall framework for automating the process of controlled lexicon building. We then present the results of our pilot experiments applying various NLP technologies to each lexicon-building step, focusing on English and Japanese verbs observed in automotive domain corpora. The obtained lists of approved and unapproved words are next compared with a controlled lexicon manually constructed from the same corpora. These results suggest to what extent current technologies can help with specialized lexicographic work.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Numbers 19H05660 and 23H03689. The automobile manuals used in this study were provided by Toyota Motor Corporation.

References

- ASD (2021). ASD Simplified Technical English. Specification ASD-STE100, Issue 8. URL: `http://www.asd-ste100.org`.

- 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. Minneapolis, Minnesota, USA, pp. 4171–4186.

- Kuhn, T. (2014). A Survey and Classification of Controlled Natural Languages. Computational Linguistics, 40(1), pp. 121–170.

- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. CoRR, abs/1907.11692.

- Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani & K. Weinberger (eds.) Advances in Neural Information Processing Systems 26. Curran Associates, Inc., pp. 3111–3119.
- Miyata, R., Sugino, H. (2020). Building a Controlled Lexicon for Authoring Automotive Technical Documents. In Z. Gavriilidou, M. Mitsiaki & A. Fliatouras (eds.) Lexicography for Inclusion: Proceedings of the 19th EURALEX International Congress, 7-9 September 2021, Alexandroupolis, Vol. 1. Democritus University of Thrace, pp. 171–180.
- Møller, M.H., Christoffersen, E. (2006). Buiding a Controlled Language Lexicon for Danish. LSP & Professional Communication, 6(1), pp. 26–37.
- Princeton University (2010). About WordNet. WordNet. URL: `https://wordnet.princeton.edu/`.
- Warburton, K. (2014). Developing Lexical Resources for Controlled Authoring Purposes. In Proceedings of LREC 2014 Workshop: Controlled Natural Language Simplifying Language Use. Reykjavik, Iceland, pp. 90–103.

Neo^{2020+} – A New Online Resource of German Neologisms

Petra Storjohann, Merle Benter
Leibniz-Institut für Deutsche Sprache
E-mail: storjohann@ids-mannheim.de, benter@ids-mannheim.de

Keywords: neologisms; customised e-dictionary; interactive resource; user study on languara forums

The detection, analysis and documentation of neologisms in everyday communication has a long-standing tradition in German lexicography. The dictionary landscape also includes rich experience in compiling corpus-based and electronic neologism dictionaries for many years. However, resources exploiting various new technological options, explicitly integrating lexical with extra-linguistic information and based on the investigation of users' needs have so far not been on the market for reference guides for new words.

Last year, a new online resource has been drawn up and designed combining established categories with an innovative design and multiple functions. Its objective is to describe newly established lexical items of German emerging after 2020 (e.g. *Doomscrolling*, *E-Football*, *Neobroker*, *Greenflation*, *Klimakleber*, *Energiepreisbremse*). At the same time, it specifically accounts for user needs as well as for principles of cognitive lexicography in terms of contents and design. For this purpose, it was necessary to develop a good idea about its potential users initially. Hence, we conducted user studies examining numerous language questions and answers concerning new terms as addressed in internet forums. These investigations uncovered valuable insights into needs, interest and various preferred foci. Users' needs turned out to be very heterogeneous depending on the type of neologism including questions regarding the pronunciation of Anglo-neologisms, origin of calques, meaning of new items, gender of new nouns, discursive potential of politically (in)correct and ambivalent terms, correct spelling of larger compounds etc. Still above all, we identified a compelling need for a strong link between any new word and its underlying signified concept, calling for interconnecting linguistic with extra-linguistic knowledge more effectively.

With a large pool of information at hand, we were able to prioritise specific details and rethink presentation options. Our insights had a decisive input during the conceptual phase, particularly on the presentation of lexicographic data. It encompasses combined dictionary details, linking lexical-semantic data with encyclopaedic information and an interactive dashboard including data of different formats such as audios, charts, podcasts, recent linguistic papers, latest entries, current discourse domains and more. This poster presents the fundamental ideas and the design behind a new German dictionary of new words which will be implemented into an online resource. The new reference guide will be part of an existing online portal and also freely accessible.

An Unsupervised Approach to Characterise the Adjectival Microstructure in a Hungarian Monolingual Explanatory Dictionary

Enikő Héja[1], Noémi Ligeti-Nagy[1], László Simon[2], Veronika Lipp[2]

[1]Hungarian Research Centre for Linguistics, Language Technology Research Group, Budapest, Hungary, [2]Hungarian Research Centre for Linguistics, Lexical Knowledge Representation Research Group, Budapest, Hungary

E-mail: eniko.heja@gmail.com, ligeti-nagy.noemi@nytud.hu, simon.laszlo@nytud.hu, lipp.veronika@nytud.hu

Keywords: microstructure of adjectival entries; Hungarian monolingual explanatory dictionary; adjectival polysemy; data-driven word sense induction; graph-based methods; word2vec representation

Finding how to properly partition the meaning-space of a lexeme poses a well-established difficulty both in bilingual and in monolingual lexicography (Adamska-Sałaciak, 2006; Atkins & Rundell, 2008; Hanks, 2012; Véronis, 2003). This problem is even more pronounced in the case of polysemy: as far as we know, there is no widely used distributional definition of polysemy which could allow for more data-based and thus, for more objective meaning distinctions (Geeraerts (2009)).

Therefore, our basic objective is to investigate to what extent a certain quantitative technique is applicable to compile a Hungarian monolingual dictionary in the case of adjectival polysemy.

Extending the work described in Héja&Ligeti-Nagy (2022), the planned presentation is centered around three topics: first, we introduce four distributional criteria to distinguish between polysemic meanings based on the notion of near-synonymy (cf. Ploux&Vittori 1998), which is closely related to the distributional conception of synonymy (cf. Frege 1892). These criteria make it not only possible to anchor various sub-meanings to observable contexts, but also yield interpretable sense distinctions modeling human intuition. Secondly, we describe a simple graph-based unsupervised method to automatically retrieve the adjectival polysemic meanings from corpora along with their relevant nominal contexts. Most importantly, we also present to what extent our results can be deployed in the creation of a Hungarian monolingual explanatory dictionary.

Although graph-based word sense induction (WSI) is a rather elaborated branch of NLP (Biemanni, 2006; Dorow et al., 2004; Pelevina et al., 2016) we are not aware of any lexicographic work that relies on the findings of this field. This fact is even more striking in the light of the emergence of static dense vectors (Mikolov, 2013, Pennington et al., 2014) that provided us with an easy-to-use representation of words. Still, these representations are underrepresented in the graph-based WSI literature.

After extending the method in Héja&Ligeti-Nagy (2022) to yield higher coverage, we found that it may also be sufficient for lexicographic purposes. Therefore, in our planned presentation the results are put into practice and a closer look is given to what extent the technique is able to facilitate the work of expert lexicographers to craft the adjectival microstructure in the new version of *The Explanatory Dictionary of the Hungarian Language* (EDHL). The aim is to compile an up-to-date online dictionary of contemporary Hungarian (2001–2020) via corpus-driven methods (Lipp&Simon 2021). Our experiment is motivated by the fact that adjectives are difficult to divide into senses (Moon, 1987): it is hard to analyse

them in isolation because they are essentially an aspect of the modifying noun (Stammers, 2008).

According to our expectations, the automatically extracted adjectival subsenses provide the lexicographers with a ready-to-use adjectival microstructure, hugely facilitating their work. A sample of automatically extracted polysemies along with their salient nominal contexts (clustered into semantic classes) will be compared to the relevant microstructures of an existing traditional explanatory dictionary from multiple perspectives, such as coverage and most importantly, the motivatedness of meaning distinctions.

References

- Adamska-Sałaciak, A. (2006). Meaning and the bilingual dictionary: The case of English and Polish. Frankfurt: Peter Lang.
- Atkins, B. T. S., and Rundell M. (2008). The Oxford guide to practical lexicography. Oxford: Oxford University Press.
- Biemann, C. (2006). Chinese Whispers - an Efficient Graph Clustering Algorithm and its Application to Natural Language Processing Problems. In: Proceedings of TextGraphs: the First Workshop on Graph Based Methods for Natural Language Processing. 73–80. Association for Computational Linguistics, New York City (Jun 2006), https://aclanthology.org/W06-3812
- Dorow, B., Widdows, D., Ling, K., Eckmann, J.P., Sergi, D., Moses, E. (2004). Using Curvature and Markov Clustering in Graphs for Lexical Acquisition and Word Sense Discrimination. URL: https://arxiv.org/abs/cond-mat/0403693
- EDHL = Bárczi, G. & Országh, L. (eds.) (1959–1962). A magyar nyelv értelmező szótára I–VII. [The Explanatory Dictionary of the Hungarian Language], Akadémiai Kiadó, Budapest.
- Frege, G. (1892). Über Sinn und Bedeutung. In: Textor, M. (ed.) Funktion - Begriff - Bedeutung, Sammlung Philosophie, vol. 4. Vandenhoeck & Ruprecht, Göttingen.
- Geeraerts, D. (2009). Theories of Lexical Semantics, Oxford University Press.
- Hanks, P. (2012). The Corpus Revolution in Lexicography. In: International Journal of Lexicography 25.4. 398–436.
- Héja, E. and Ligeti-Nagy, N. (2022). Clique-based Graphical Approach to Detect Interpretable Adjectival Senses in Hungarian. Proceedings of TextGraphs-16: Graph-based Methods for Natural Language Processing, 35–43 October 16.
- Lipp, V. and Simon, L. (2021). Towards a new monolingual Hungarian explanatory dictionary. Studia Lexicographica 15:29, 83–96.
- Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. URL: https://arxiv.org/abs/1301.3781

- Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. Advances in Neural Information Processing Systems 26.
- Moon, R. (1987). Monosemous Words and the Dictionary. In Cowie, A. P. (ed.) The Dictionary and the Language Learner. Lexicographica Series Maior, Tübingen, Max Niemeyer Verlag, 173–182.
- Pelevina, M., Arefiev, N., Biemann, C., Panchenko, A. (2016). Making Sense of Word Embeddings. In: Proceedings of the 1st Workshop on Representation Learning for NLP. 174–183. Association for Computational Linguistics, Berlin, Germany, https://aclanthology.org/W16-1620
- Pennington J., Socher R., Manning Ch. (2014). Glove: Global vectors for word representation. In Proceedings of the Empirical Methods in Natural Language Processing (EMNLP), 1532–1543, Doha, Qatar.
- Ploux, S., Victorri, B. (1998). Construction d'espaces sémantiques a l'aide de dictionnaires de synonymes. Traitement automatique des langues 1(39), 146–162.
- Stammers, J. (2008). Unbalanced, Idle, Canonical and Particular: Polysemous Adjectives in English Dictionaries. Lexis, 1, 85–111. URL: http://journals.openedition.org/lexis/771
- Véronis, J. (2003). Sense tagging: does it make sense? In: Wilson, A., Rayson, P., McEnery, T. (eds.) Corpus Linguistics by the Lune: a festschrift for Geoffre Leech. Peter Lang, Frankfurt.
- Véronis, J. (2004). HyperLex: lexical cartography for information retrieval. Computer Speech & Language 18(3), 223–252, http://dblp.uni-trier.de/db/ journals/csl/csl18.html#Veronis04

How Lexicographers Evaluate User Contributions in The Thesaurus of Modern Slovene in Comparison to Dictionary Users

Magdalena Gapsa[1], Špela Arhar Holdt[2]
[1]Faculty of Arts, University of Ljubljana, [2]Faculty of Computer and Information Science, University of Ljubljana
E-mail: magdalena.gapsa@ff.uni-lj.si, arharhs@ff.uni-lj.si

Keywords: user involvement, responsive dictionary, synonyms, user evaluation, lexicographers

In digital lexicography, user involvement provides an opportunity to enhance the relevance, quality, and prompt accessibility of language resources. A state-of-the-art example of this is Thesaurus of Modern Slovene, which incorporates user participation to improve its automatically created content. The Thesaurus allows users to suggest new synonym candidates, as well as evaluate existing ones. User-suggested synonyms are displayed in the dictionary interface, but only those that have received lexicographical approval are included in the openly accessible dictionary database for wider usage. To shed light on the otherwise invisible lexicographic decision-making processes and develop editorial protocols that align with the opinions and needs of dictionary users, we studied the difference in how lexicographers evaluate user-suggested synonyms versus how dictionary users evaluate them. We conducted an evaluation campaign with nearly 1,000 user-suggested synonyms from the Thesaurus of Modern Slovene. The evaluation set was assessed by a total of 42 evaluators, divided into 7 user groups based on their profession or interests: lexicographers, translators, teachers, students etc. This paper focuses on the lexicographers' data presented against the background of the other groups' responses and comments. We tested 4 hypotheses about lexicographers as an evaluator group: (1) their evaluation would be more consistent and the Inter-Annotator Agreement (IAA) would be higher than in other user groups, (2) they would argue their decisions in more detail, (3) they would identify more potential problems than other groups, while being (4) more rigorous in their decisions and more reserved to include user suggestion. After the evaluation, IAA was calculated in all user groups using Krippendorff's alpha and entropy, and the evaluators' comments were classified into bottom-up categories. The data was statistically analysed and compared within each group and between groups. Results show that some initial assumptions were correct, e.g., the lexicographers gave the most arguments explaining their judgement and most often identified shortcomings in the suggested synonyms, while others proved to be wrong, e.g., they scored the second lowest IAA among all the participating groups, as well as the highest number of pairs with tied answers (where half of them chose one option and the other another). Interestingly, they also had the lowest number of user-suggestions they found entirely inappropriate for the inclusion in the database. We discuss the possible reasons for the results presented, explain the limitations of the evaluation, and emphasise the value of the observed differences for the further development of language resources, especially responsive dictionaries, of which the Thesaurus of Modern Slovene is an example.

Structuring the Dictionary Entries of Unrecorded Korean Lexical Items Based on their Type and Applicability

Jun Choi[1], Jinsan An[2], Minkyu Sung[2], Kilim Nam[2]
[1]Chonnam National University, [2]Kyungpook National University
E-mail: c-juni@daum.net, siveking@naver.com, dse4062@naver.com, nki@knu.ac.kr

Keywords: unrecorded lexical items; database type; Instant Messenger Corpus; Korean Neologism Investigation Project resources

Lexicography has now become an independent academic field, having undergone many changes from the so-called 'corpus revolution (Rundell & Stock, 1992; Hanks 2012)' to the development of 'user-generated content' (Rundell et al. 2015)', but still facing many challenges. The advent of the web may have resolved the spatial limitations of print dictionaries and built a bridge between lexicographers and dictionary users, but dictionaries now need to reflect the linguistic dynamics all the more quickly as language ever changes in the era of new media.

Unrecorded lexical items in particular, which extensively appear in instant messaging and on the web, include, amongst other things, non-standard expressions, unethical expressions, and variants of language expressions such as emoticons, with particular meanings and functions. While these are frequently used semantic units, they are excluded from headword selection. Cook (2010, 2012), Breen (2017), and Breen et al. (2018) have pointed out the necessity of including such lexical items within the scope of dictionary headwords from the perspective of natural language processing. This study aims to address such linguistic and lexicographic changes, by extracting all unrecorded Korean lexical items from a 2.4 million ecel (Korean word unit) Instant Messenger Corpus and classifying them according to their practical applicability in natural language processing. The authors define 'unrecorded lexical items' as any item that requires analysis in natural language processing in the age of artificial intelligence but has not been included in lexical resources. These items have been organized in a database, classified into three types based on their characteristics in terms of theoretical and applied linguistics.

First, the Lexicographic Data Type corresponds to the unrecorded lexical items that need to be described after existing dictionary entries. The second type, namely the Annotation Data Type, includes the items that can be used for analysis and processing based on a relatively concise description. Finally, the Equivalent Data Type regards the unrecorded items that have a standard equivalent in the dictionary.

The data under study consists of a list of potential unrecorded items compiled from the Instant Messenger Corpus (2019-2021) and the Korean Neologism Investigation Project resources (2012-2022). The list has been compared to the macrostructure of the Korean language dictionary Urimalsaem which comprises 1.2 million headwords, to extract automatically a selective list of unrecorded lexical items, which have been manually evaluated to compile the final list of unrecorded items and divided into the three aforementioned types. The next step will be to build a three-type entry dictionary based on the needs of natural language processing.

This research is hoped to contribute to the discussion on lexicographic compilation and criticism by expanding the scope of the macrostructure and the microstructure of the dictionary, and participate in the shift in perspective from the stabilized framework of print dictionaries to the new framework of the digital era.

References

- Breen, J. (2017). Extraction of Neologisms from Japanese Corpora, Australia: University of Melbourne Melbourne.
- Breen, J., Baldwin, T., & Bond, F. (2018). The Company They Keep: Extracting Japanese Neologisms Using Language Patterns. In Proceedings of the 9th Global Wordnet Conference (pp. 163-171).
- Cook, C. P. (2010). Exploiting linguistic knowledge to infer properties of neologisms. Toronto, Canada: University of Toronto.
- Cook, P. (2012). Using social media to find English lexical blends. In dalam Proceedings of the 15th EURALEX International Congress (EURALEX 2012) (pp. 846-854).
- Hanks, P. (2012), The corpus revolution in lexicography, International Journal of Lexicography, 25(4). 398-436.
- Rundell, M., & Stock, P.(1992), The corpus revolution, English Today, 8(4), 45-51.

Word sense induction on a corpus of Buddhist Sanskrit literature

Matej Martinc[1], Andraž Pelicon[1], Senja Pollak[1], Ligeia Lugli[2]
[1]Jožef Stefan Institute, [2]Mangalam Research Center for Buddhist Languages
E-mail: matejmart@gmail.com, andraz.pelicon@ijs.si, senja.pollak@ijs.si, ligeia.lugli@kcl.ac.uk

Keywords: Word sense induction; Buddhist Sanskrit; Transformer language models

This paper reports on a series of experimentations on word sense induction (WSI) we conducted on a corpus of Buddhist Sanskrit literature. Our corpus is small, comprising about 7 million words, and extremely few resources are available for the gamut of Sanskrit varieties it includes. The objective of our experiments has been to introduce a degree of automation in the labour-intensive lexicographic task of matching citations for a lemma to the corresponding sense of the lemma.

For this purpose, we construct a Buddhist Sanskrit WSI dataset consisting of 3110 sentences with manually labeled sense annotations for 39 distinct lemmas. The WSI dataset is used for fine-tuning and evaluation of three distinct transformer-based language models (Devlin et al., 2019), pretrained on a corpus of Buddhist Sanskrit literature. More specifically, each model is trained on the binary classification task of predicting whether the target lemma in two concatenated sentences containing the lemma has the same sense or not.

The binary predictions produced by the models are used for clustering of lemma sentence examples into distinct lemma senses. We propose a novel clustering solution, which relies on building a (0,1)-adjacency matrix for each lemma, in which ones indicate whether pairs of vertices (in our case sentences) are adjacent (i.e., contain lemmas with the same sense) in the graph. The rows in the matrix are used for construction of initial clusters, i.e. we create a cluster containing the target vertice and its adjacent sentences for each example. To obtain the final clusters, the initial clusters are merged by recursively combining the clusters with the largest intersection up to a predefined threshold of minimum intersection or maximum number of clusters.

The obtained clusters, representing sense distributions for each lemma, are evaluated in two 5-fold cross-validation scenarios. In the first scenario, we test how well do the obtained clusters represent the true sense distribution of new unseen (polysemous and monosemous) lemmas not used for model training. Using our novel methodology, we report the Adjusted Rand Index (ARI) score (Hubert & Arabie, 1985) of 0.208 and an F1-score (Manandhar & Klapaftis, 2009) of 80.36%. In the second scenario, we test how well do the clusters represent the true lemma sense distribution when classifier is tested on new unseen sentence examples for polysemous lemmas used for model training. Here, we report the best ARI score of 0.300 and an F1 score around 76%.

Overall, the proposed approach outperforms several state-of-the-art WSI baselines, is the first WSI solution employed for Buddhist Sanskrit and will be used for creating novel lexicographical resources for Buddhist Sanskrit. The dataset, models and code will be made freely available after the publication of the paper.

References

- Devlin J., Chang M., 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:, (pp. 4171–4186).

- Hubert, L., Arabie, P. (1985). Comparing partitions. Journal of classification, 2, 193-218.
- Manandhar, S., Klapaftis, I. (2009). Semeval-2010 task 14: Evaluation setting for word sense induction & disambiguation systems. In Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009) (pp. 117-122).

Word Sense Induction for the Automatic Construction of a Valency Dictionary of French Verbs

François Lareau, Naïma Hassert
Université de Montréal
E-mail: francois.lareau@umontreal.ca, naima.hassert@umontreal.ca

Keywords: valency dictionary; clustering; word embeddings; automatic extraction; word sense induction

Valency dictionaries such as VerbNet (Kipper et al. 2006), Verbnet (Danlos et al. 2016) or Lefff (Sagot 2010) are useful in many natural language processing applications, in particular for rule-based natural language generation. This type of dictionary indicates precisely how a predicate expresses its arguments in syntax, including information on selected part-of-speech, preposition or case. However, the way a word expresses its arguments can change significantly depending on its sense.

For example, the verb *change* requires a direct object when it means 'alter or modify', as in *The discussion has changed my thinking about the issue*, but with the sense 'undergo a change, become different', as in *She changed completely as she grew older*, then there is no object at all (examples taken from WordNet (Miller, 1995)). Therefore, a valency dictionary must distinguish at least the main senses of a lemma. Constructing this kind of resource manually, however, is very costly in both time and money, and requires highly trained staff. Our goal is thus to automate the construction of a valency dictionary, focusing on French verbs. This paper presents how we tackled an important subtask: automatically identifying the polysemy of verbs.

Since our goal is to produce a resource entirely automatically, we want to use raw data as material and rely on as little external resources as possible. This comes down to a word sense induction (WSI) task, but with an ulterior goal. Several WSI techniques have been introduced as early as the 1990s (e.g., context clustering (Schütze, 1998), word clustering (Lin, 1998) or cooccurence graphs (Véronis, 2004)). However, the field has been revolutionized with the arrival of Transformers (Vaswani et al., 2017), which can produce high quality contextualized word embeddings in several languages. In this paper, we will explore the use of transformers in WSI.

We tackled this task in two main steps: first, we extracted contextualized vectors of the sentences in the FrenchSemEval evaluation dataset (Segonne, Candito, and Crabbé, 2019) with one language-specific model, CamemBERT (Martin et al., 2019), and two multilingual models, XLM-RoBERTa (Conneau et al., 2019) and t5 (Raffel et al., 2020). This dataset is comprised of around 50 sense-annotated examples of 66 different French verbs in context. Then, we tested three unsupervised clustering algorithms that don't require to know the number of clusters beforehand: Affinity Propagation (Dueck, 2009), Agglomerative Clustering (Szekely et al., 2005) and HDBSCAN (McInnes and Healy, 2017). The best results were achieved with CamemBERT vectors clustered with Agglomerative Clustering, attaining a BCubed F_1 score (Amigó et al., 2009) of 65.2 %. As a comparison, the FlauBERT team (Le et al., 2019), also using CamemBERT vectors, attained an F_1 score of 50.02 % on the same dataset, although they used a supervised method and measured their results with the traditional F_1 score. Our experiments confirm the potential of unsupervised methods to identify verb senses, and indicate that monolingual language models are better than multilingual language models for WSI tasks involving a single language.

References

- Amigó, E., Gonzalo, J., Artiles, J., Verdejo, F. (2009). "A Comparison of Extrinsic Clustering Evaluation Metrics Based on Formal Constraints." Information Retrieval 12: 461–86.

- Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, V., Stoyanov, F. (2019). "Unsupervised Cross-Lingual Representation Learning at Scale." URL: `https://arxiv.org/abs/1911.02116`.

- Danlos, L., Pradet, Q., Barque, L., Nakamura, T., Constant, M. (2016). "Un Verbnet Du Français." TAL 57 (1): 33–58.

- Dueck, D. (2009). "Affinity Propagation: Clustering Data by Passing Messages." PhD thesis, University of Toronto.

- Kipper, K., Korhonen, A., Ryant, N., Palmer, M. (2006). "Extending VerbNet with Novel Verb Classes." Proceedings of LREC, 1027–32.

- Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, D., Schwab, B. (2019). "Flaubert: Unsupervised Language Model Pre-Training for French." URL: `https://arxiv.org/abs/1912.05372`.

- Lin, D. (1998). "Automatic Retrieval and Clustering of Similar Words." Proceedings of ACL/COLING 2: 768–74.

- Martin, L., Muller, B., Suárez, P. J. O., Dupont, Y., de La Clergerie, É., Romary, L. (2019). "CamemBERT: A Tasty French Language Model." URL: `https://arxiv.org/abs/1911.03894`.

- McInnes, L., Healy, S., Astels, J. (2017). "Hdbscan: Hierarchical Density Based Clustering." Journal of Open Source Software 2 (11): 205.

- Miller, G. A. (1995). "WordNet: A Lexical Database for English. Communications of the ACM 38 (11): 390–41.

- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Liu Matena, M. (2020). "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer." The Journal of Machine Learning Research 21 (1): 5485–5551.

- Sagot, B. (2010). "The Lefff, a Freely Available and Large-Coverage Morphological and Syntactic Lexicon for French." Proceedings of LREC, 2744–51.

- Schütze, H. (1998). "Automatic Word Sense Discrimination." Computational Linguistics 24 (1): 97–123.

- Segonne, V., Candito, M., Crabbé, B. (2019). "Using Wiktionary as a Resource for WSD: The Case of French Verbs." Proceedings of IWCS, 259–70.

- Szekely, G. J., Rizzo, M. L. (2005). "Hierarchical Clustering via Joint Between-Within Distances: Extending Ward's Minimum Variance Method." Journal of Classification 22 (2): 151–84.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Polosukhin, I. (2017). "Attention Is All You Need." NIPS 30.
- Véronis, J. (2004). "Hyperlex: Lexical Cartography for Information Retrieval." Computer Speech Language 18 (3): 223–52.

Invisible lexicography in Sepedi writing systems

Theo Bothma, Daniel Prinsloo
University of Pretoria
E-mail: theo.bothma@up.ac.za, danie.prinsloo@up.ac.za

Keywords: Invisible lexicography; Dictionary database; Writing assistant ; Corpus verification

Much research and development is being done by numerous companies on writing assistants, for example, Grammarly (`https://www.grammarly.com/`), Linguix (`https://linguix.com/`) and Ginger (`https://www.gingersoftware.com/`). Most of these writings assistants work on text production in one language only, but some provide assistance in multiple languages, such as LanguageTool, which provides assistance in more than twenty languages (`https://languagetool.org/`).

Most assistants work in resource-rich languages. In this paper, we describe some features of two pioneering writing assistants developed for Sepedi, a lesser resourced language of Southern Africa. Current and future compilation of writing assistants for Sepedi should be inspired by these two successful efforts, the Sepedi Helper (Prinsloo & Taljard, 2019; Prinsloo, 2020; Sepedi Helper, nd) and the Copulative Decision Tree (Prinsloo & Bothma, 2020), in which invisible lexicographic strategies played a major underlying role. User studies revealed that the Sepedi Helper is very useful tool for students learning Sepedi (Prinsloo & Taljard, 2019; Prinsloo, 2020). The Sepedi Helper provides access to constructing Sepedi sentences via Sepedi or English words, and makes use of a fairly basic translation dictionary. It therefore could not utilise any detailed grammatical and semantic descriptions. We foresee that a much more complex database structure with access to much more detailed grammatical and semantic features is required. In addition to standard grammatical features, we foresee that the database should also have access to and be integrated with other functionalities, which include:

1. Underlying grammatical rules;
2. Underlying key grammatical features, e.g. nouns, pronouns and concords;
3. Linking to processed and raw corpus data;
4. A text production verification system, based upon corpus queries to find exact matches of texts produced or near matches or part-of-speech matches.

In constructing a sentence, the system leads the user step by step through the complex grammatical rules and lexical items to arrive at a grammatically correct sentence. In this sense, the system also has a didactic function, i.e., computer-assisted language learning (CALL). The user involuntarily learns grammatical rules based on the choices they make. However, the user can also directly consult brief pop-up sections in which the specific language rule is explained in more detail, by clicking on a "read more" button that provides a summarised version of grammatical rules, or even drill down to more detailed descriptions or outer texts of dictionaries or grammars for further explanations. All the extra information is provided on demand (as demonstrated in the Copulative Decision Tree and the Sepedi Helper), and the user is not confronted with an information overload.

The usability of such a next generation writing assistant depends to a very large extent on the nature of an enriched lexicographic database.

References

- Prinsloo, D.J. (2020). User studies on the Sepedi Copulative Decision Tree. SALALS 38(4): 323-335.
- Prinsloo, D.J., Bothma, T.J.D. (2020). A copulative decision tree as a writing tool for Sepedi. South African Journal of African languages 40:1, 85-97.
- Prinsloo, D.J., Prinsloo, Daniel. (2021). A writing assistant en route to a full computational grammar for Sepedi. South African Journal of African languages 41(1): 12-21.
- Prinsloo, D.J., Taljard E. (2019). User studies on the Sepedi Helper writing assistant. Language Matters 50(2): 73-99.
- Sepedi Helper. nd. URL: `http://www.sepedihelper.co.za/`. Accessed January 31, 2023.

Development of a methodology and enhancements of lexicographical resources for an online Platform of Academic Collocations Dictionaries in Portuguese and English

Adriane Orenha-Ottaiano[1], Tanara Zingano Kuhn[2], Arnaldo Candido Junior[1], João Pedro Quadrado[1], Carlos Roberto Valêncio[1], Stella Esther Ortweiler Tagnin[3]

[1]São Paulo State University, [2]Research Centre for General and Applied Linguistics (CELGA-ILTEC), University of Coimbra, [3]University of São Paulo
E-mail: adriane.ottaiano@unesp.br, tanarazingano@outlook.com, -, jp.quadrado@unesp.br, carlos.valencio@unesp.br, seotagni@usp.br

Keywords: collocations; collocation dictionary; dictionary writing system

In view of the growing demand for publications of academic texts and for the dissemination of studies in national and international congresses, the development of specific lexicographic tools that can assist with academic writing is fundamental. Some resources are already available, such as dictionaries of academic language (e.g., Oxford Learner's Dictionary of Academic English; The Louvain EAP Dictionary (https://leaddico.uclouvain.be/) and writing assistant tools that have a lexicographic component (e.g., ColloCaid, Frankenberg-Garcia et al., 2019; HARTAes-vas, Alonso-Ramos & Zabala, 2022; Gracia-Salido et al., 2018). As can be seen, some languages are better equipped with this type of resources than others. However, dictionaries whose unique focus is on academic collocations are, to the best of our knowledge, still unheard of in any language. The objective of our research project is to contribute to filling this gap by creating Online Dictionaries of Academic Collocations. Initially, English and Brazilian Portuguese will be the languages covered, but we intend to include more languages in the next phases of the project. During this three-year, publicly funded research project CNPq, Process nr. 409178/2021-7, our main objectives are: a) to develop a methodology for the creation of corpus-driven academic collocation dictionaries and b) to improve an existing dictionary writing system and an end-user interface (PLATCOL, Orenha Ottaiano, 2020; Orenha Ottaiano et al., 2021a; Orenha Ottaiano & Silva 2021b) for the purposes of this project. To achieve these goals, we will first establish criteria to define academic collocations and develop a method for the automatic identification and extraction of these collocations (based on Kuhn, 2017). In addition, we will adapt PLATCOL's existing Dictionary Writing System and End-User Interface to meet the macro and microstructural characteristics of the Academic Collocations Dictionaries. By the end of the project, we will publish prototypes of these two dictionaries in an online platform. With a tested methodology of dictionary-making and fully functional adaptations to the dictionary writing system and end-user interface, we will be able to move to a new phase in which not only will we work on turning the prototypes into fully fledged dictionaries, but we can also include additional languages.

Acknowledgements

The authors gratefully acknowledge the financial support provided by CNPq, Grant nr. 409178/2021-7, Brazil, and the Portuguese National Funding Agency, FCT - Foundation for Science and Technology, I.P. (grant number UIDP/04887/2020).

References

- Alonso-Ramos, M., Zabala, I. (2022). HARTAes-vas: Lexical Combinations for an Academic Writing Aid Tool in Spanish and Basque. SEPLN (Projects and Demonstrations), pp. 22-25.
- Frankenberg-Garcia, A. Lew, R., Roberts, J., Rees, G., Sharma, N. (2019). Developing a writing assistant to help EAP writers with collocations in real time. ReCALL, 32/2, pp. 23-39.
- García-Salido, M., M. Garcia, M. Villayandre, M. Alonso-Ramos. 2018. A Lexical Tool for Academic Writing in Spanish based on Expert and Novice Corpora. In N. Calzolari et al. (eds.), Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), pp. 260–265.
- KUHN, T. Z. 2017. A design proposal of an on-line corpus driven dictionary of Portuguese for university students. Ph.D. Thesis. University of Lisbon.
- Orenha-Ottaiano, A.; Garcia, M.; Olímpio De Oliveira, M. Eugênia; L'Homme, M-C; Alonso Ramos, M.; Valêncio, C. R., Tenório, W. (2021a). Corpus-based methodology for an Online Multilingual Collocations Dictionary: First Steps. In: Kosem, I.; Michal C.; Miloš J.; Jelena K.; S. Krek & C. Tiberius (eds.). Proceedings of eLex 2021, pp. 1-28.
- Orenha-Ottaiano, A.; Silva, M. E. O. O. (2021b). A Corpus-based Platform of Multilingual Collocations Dictionaries (PLATCOL): some lexicographical aspects aiming at pre- and in-service teachers. Proceedings of the International Conference Corpus Linguistics 2021. St. Petersburg, Russia, pp. 122-132.
- Orenha-Ottaiano, A. (2020). A phraseographical methodology and model for an online corpus-based multilingual collocations dictionary platform. Research Grant (2020-2022) awarded by Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP Process ner. nº 2020/01783-2).
- Oxford Learner's Dictionary of Academic English. (2014). Oxford: Oxford University Press.
- Granger, S., Paquot, M. The Louvain EAP Dictionary. URL: `https://leaddico.uclouvain.be/`. (last access: 27-01-2023).

What gooseberries, grapes and (bad) wine have in common? Linking Dictionaries of Historical Varieties of Polish

Krzysztof Nowak, Ewa Rodek, Dorota Mika
Institute of Polish Language, Polish Academy of Sciences
E-mail: krzysztof.nowak@ijp.pan.pl, ewa.rodek@ijp.pan.pl, dorota.mika@ijp.pan.pl

Keywords: historical lexicography; plant names; LLOD; dictionary interlinking

Although general dictionaries are not expected to provide encyclopaedic knowledge, lexicographers seem also to agree that the quantity of real-world information depends on the user's familiarity with specific culture (Atkins and Rundell, 2008: 424). This is particularly true of historical dictionaries since their users will usually require more guidance in contextualising linguistic description of terms related to social life (e.g. kinship, administration, economy), culture (e.g. deonyms, "emotion words", virtues), science or crafts.

The DARIAH.Lab Project[1] aims, among others, at modelling a number of lexical resources for Polish as linked open data. Converted to digital form, this large dataset represents diachronic, diatopic, and diastratic variation and is currently being interlinked and mapped to external resources. In this paper, we discuss methodological issues which arose during the SKOS modelling (2009) of three diachronic dictionaries, namely, the Onomasiological Dictionary of Old Polish (DOPol)[2], the Dictionary of Polish Medieval Latin (DMLat)[3], and the Electronic Dictionary of XVII-XVIII century Polish (D17Pol)[4]. This case study focuses on phytonyms, since they cross the boundary between real-world and linguistic knowledge and have been traditionally defined by explicit reference to scientific taxonomies, a feature which makes them natural candidates for any LOD project.

We start with discussing the strategies adopted in defining plant names as they not only tend to differ significantly between different dictionaries, but also are often inconsistent within a single, usually paper-born work. To name just a few:

- domain labels, such as bot. for 'botanical', are applied both to terminological units and general language lexemes;
- regular polysemy (Apresjan, 1974), a phenomenon systematically occurring in phytonyms, may be encoded in the form of separate senses or embedded in the definitional string.

The formal features and inconsistencies have a major impact on the interpretation and extraction of conceptual knowledge. The lack of clear distinction between general and expert use of a term is misleading as they both entail significant differences on the categorization level. Another common feature in defining phytonyms, namely providing their Latin names, is not without its problems either. After all, the dictionaries describe language reality predating Linnaean and modern classification systems which may not fit with the premodern conceptualisation of the natural world. Consequently, the relation between a historical phytonym and its scientific designation is by no means one of identity.

[1] https://lab.dariah.pl/en/
[2] https://spjs.ijp.pan.pl/spjs/strona/kartaTytulowa
[3] http://elexicon.scriptores.pl/
[4] http://sxvii.pl/

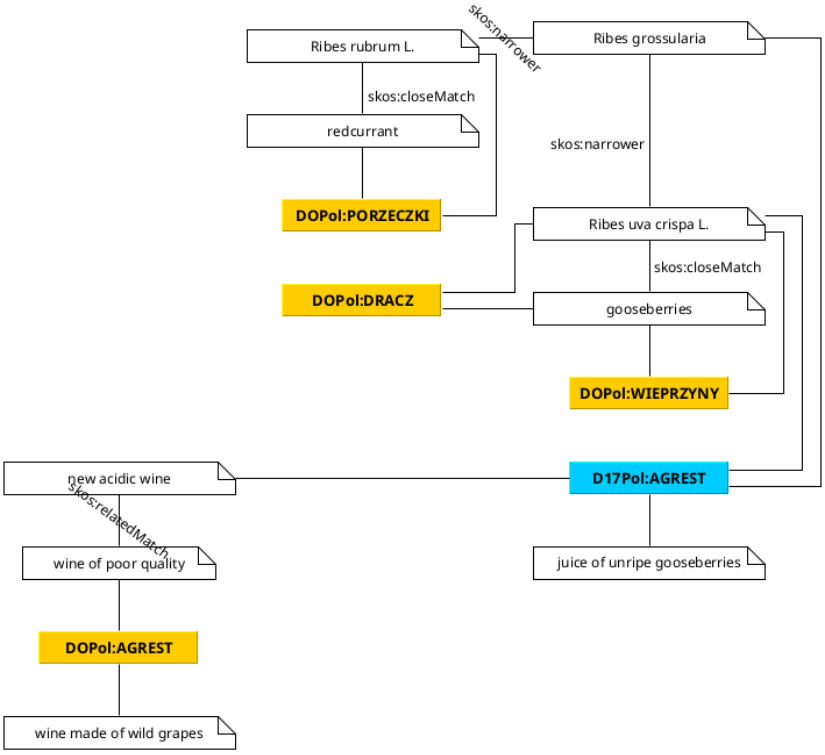


Figure 1:

These problems only accumulate when interlinking dictionaries. To demonstrate them, we will discuss the example of 14th-18th century terms referring to 'gooseberries' which shows the extent to which manual analysis and a careful interpretation of the definition scope is sometimes required (Figure 1).

References

- Apresjan J.D. (1974). Regular Polysemy, „Linguistics" 14, 5-32.
- Atkins, B. T. S., Rundell, M.. (2008). The Oxford guide to practical lexicography. Oxford; New York: Oxford University Press. SKOS - Simple Knowledge Organization System Reference. URL: `https://www.w3.org/TR/skos-reference/`. Accessed 31 Jan. 2023

Representing ideology in terminological resources

Pilar León-Araúz, Arianne Reimerink, Melania Cabezas-García, Pamela Faber

University of Granada
E-mail: pleon@ugr.es, arianne@ugr.es, melaniacabezas@ugr.es, pfaber@ugr.es

Keywords: terminological resources; ideology; multimodal knowledge representation

Cultural aspects of specialized discourse are underrepresented in terminological resources, which may respond to the complexity of reflecting the cultural component in the description of terms and concepts. Culture is generally understood as the ways of life, customs, knowledge and degree of artistic, scientific, and industrial development of a group of people. A discourse community (Swales, 1990, 2016) is related to culture in the sense that it is a group of people that communicate for a purpose and share the same goals. Politicians can be considered a discourse community that is subdivided into subcommunities according to where they stand on the political and ideological spectrum. The interrelations between the discourse community of politicians, the media and the general public make ideology an especially complicated cultural aspect to convey in terminological resources. However, the way politicians convey scientific knowledge for their specific political goals must be taken into account. A terminological resource should provide the necessary information to understand the political perspective taken in discourse on the environment or to choose the most adequate term variant to write a text on environmental issues with a specific political goal in mind.

Term variants often respond to the cognitive intention of the speaker and may influence the way a concept is perceived by the recipient (Cabré 2008). They can be used deliberately to reflect multidimensionality, imprecision or ideological attachment. Politicians are well aware of the power of term choice and use it accordingly. Despite the need for clarity and accuracy in environmental communication to act on climate change (Federici and O'Brien, 2019), as in all specialized domains, environmental concepts and terms are subject to dynamism and variation (León-Araúz, 2017). For instance, *climate change*, *climate crisis*, *climate emergency* or even the newly-coined *climate breakdown* can all be regarded as term variants of the same concept, each seeking a different reaction.

To study term variance from an ideological perspective, we used two corpora currently available in Sketch Engine (Kilgariff et al., 2004): Spanish parliamentary debates, and English parliamentary debates (ParlaMint 2.1) to identify the term variants related to climate change. We then selected a sample of excerpts balanced in terms of political ideology and annotated it according to the frames present as defined by Bolsen and Shapiro (2018). The results showed how the political spectrum changes across national parliaments and how different ideologies frame climate change through different conceptual features and discursive strategies (e.g. relating the concept to human health, economic or biodiversity issues).

Finally, we explain how to represent climate change from an ideological perspective in terminological resources in conceptual, linguistic and graphic modules. We consider ideology to be one of the cultural dimensions of terms and our intention is to explicitly represent the "ideological distance" between scientific knowledge on climate change and where certain political stances are situated on the political spectrum.

References

- Bolsen, T. and Shapiro, M.A. (2018). The US News Media, Polarization on Climate Change, and Pathways to Effective Communication. Environmental Communication, 12(2):149-163.
- Cabré, M.T. (2008) El principio de poliedricidad: La articulación de lo discursivo, lo cognitivo y lo lingüístico en Terminología (I). Ibérica 16: 9-36.
- Federici, F. and O'Brien, S. (2019). Translation in Cascading Crises. London: Routledge.
- Kilgarriff, A, Rychly, P., Smrz, P., Tugwell, D. (2004) The Sketch Engine. In Proceedings of the 11th EURALEX International Congress. Lorient: EURALEX, pp. 105-116.
- León-Araúz, P. (2017). Term and concept variation in specialized knowledge dynamics. In Multiple Perspectives on Terminological Variation, edited by Drouin, P., Francœur, A., Humbley, J., Picton, A. Terminology and Lexicography Research and Practice, 18:213-258. Amsterdam/Philadelphia: John Benjamins.
- Swales, J. (1990). Genre analysis: English in academic and research settings. Cambridge/Nueva York: Cambridge University Press.
- Swales, J. (2016). Reflections on the concepts of discourse community. Asp La revu dy GERAS 69: 7-19.

Corpus-based extraction of good example sentences with a high range of variation

Alexander Geyken[1], Ulf Hamster[1], Iryna Gurevych[2], Lothar Lemnitzer[1], Ji-Ung Lee[2]
[1]Berlin-Brandenburgischen Akademie der Wissenschaften, [2]UKP Lab, TU Darmstadt
E-mail: geyken@bbaw.de, hamster@bbaw.de, gurevych@ukp.uniformatik.tu-darmstadt.de, lemnitzer@bbaw.de, ji-ung.lee@tu-darmstadt.de

Keywords: example extraction, example variation, corpus linguistics, quadratic optimisation

The development of computer-aided systems for extracting good sentence evidence from large corpora for use as dictionary evidence was first presented by Kilgarriff (2008) and applied to German (Didakowski, 2012). The focus of these rule-based systems, as well as their follow-on systems, has been on extracting example sentences with the high quality according to a predefined set of parameters. These systems are mainly used in large reference dictionaries as initial corpus filters, from which candidate lists of good example sentences are extracted, which are then curated either on the publishing platforms without further post-processing or subsequently by lexicographic expertise.

Less consideration was given to another, no less important aspect in these systems: how to compile evidence for a word from large corpora in such a way that it reflects the range of usage of the word as much as possible, ideally representing the entire spectrum of meanings of the word. In the above-mentioned systems, this desideratum is not included. In particular, the exclusive use of a system based on the example sentence quality of a sentence may lead to the selection of many example sentences that are similar to each other. For example, semantically approximating duplicates or syntactic text templates may receive the highest scores, but together be of poor variation.

This question of how to achieve a balanced range of variation in corpus documents was the starting point for EVIDENCE, a DFG-funded project that aims to develop a solution to precisely this problem based on machine learning methods. In contrast to the rule-based systems mentioned above, the system developed in EVIDENCE is not parameterized by the developers, but the system learns interactively during annotation processes by users of the system (Boullosa 2017; Simpson & Gurevych 2020); a more recent implementation of the system is based on neural networks (Hamster/Lee 1, forthcoming).

In our approach, the tradeoff between good sentence evidence on the one hand and the greatest possible variety, on the other hand, is formulated as a dual optimisation problem. As a result of the analysis, lexicographers receive a clustered result set that can be ordered according to different parameters (semantic, syntactic, and metadata, such as text types or time periods). The preference parameters can be dependent on the existing sentence records of a lemma, but can also reflect the individual preference of the users. The latter can be helpful to identify possibly unconscious biases in the curation of sentence evidence. On the one hand, the preference parameters collected on training data can be used to apply the example sentence selection to new corpora, but on the other hand, they can also be used to obtain example sentences for words outside the headwords of the training data (generalisation).

We present an application for comparative evaluation of example sentences using the best-worst scaling ranking method, which are used to train a machine evaluation model. In addition to presenting the system developed as a web app (Hamster/Lee 2, forthcoming), we demonstrate its functionality using some selected examples.

Repository for the argument/adjunct distinction SARGADA: syntactic resource with a lexicographical background

Ivana Brač, Siniša Runjaić, Matea Birtić

Institute of Croatian Language and Linguistics
E-mail: ibrac@ihjj.hr, srunjaic@ihjj.hr, mbirtic@ihjj.hr

Keywords: Croatian language; syntax; argument/adjunct distinction; diagnostic tests; lexicographical background

Extensive preparations for and the formalization of a dictionary entry design precede the initiation of any lexicographic project and the creation of dictionary material, but certain unexpected theoretical and practical issues can often arise while creating a dictionary. For the authors of the *e-Glava* online valency dictionary (Birtić, Brač and Runjaić, 2017), this was the correct categorization of dependents into arguments or adjuncts for certain verbs. The authors' discussions on this theoretical issue during the lexicographic work eventually turned into an incentive for a full-scale research work, so the 4-year project *Syntactic and semantic analysis of arguments and adjuncts in Croatian* (with the acronym SARGADA) was launched in 2020.

The main objective of the project was to determine the criteria and tests (Forker, 2014; Toivonen, 2021) for the distinction of arguments and adjuncts in the Croatian language, as well as to apply those criteria to the building of the syntactic repository SARGADA. This repository directly arises as a by-product of the research of ambiguous syntactic parts, where it's difficult to determine an arguments/adjunct status of the syntactic phrase. Therefore, the syntactic repository SARGADA won't be similar to prototypical digital resources, like dependency treebanks (Hajič et al., 2018), valency dictionaries (Jezek et al., 2014), or lexical databases with elaborated systems for marking semantic frames (Fillmore and Baker, 2009).

This paper aims to present the process of developing the database, starting from the theoretical idea of organizing and tagging ambiguous syntactic parts. In the first phase of preparation, a list of 130 Croatian verbs was compiled. After additional analysis, it was clear that some of these verbs have different meanings that involve various valency patterns so we are operating with 130 lemmas. Those groups of syntactically ambiguous parts that occur with certain verbs were examined, and verbs in the repository are classified into so-called 13 "macrogroups". This part of the process was, to the greatest extent, methodologically compatible with standard lexicographical procedures. In the next phase, these verb lemmas were searched for in the corpora. Based on the research in various Croatian corpora, sentences were selected that will serve as examples for testing ambiguous parts in the repository, which was technically developing along with it. Finally, we have chosen a tag set of 11 syntactic tags for unquestionable parts of sentences ("manual parsing"), and also a set of 7 diagnostic tests, based on which it will be expressed numerically and in percentages whether the tested part of the sentence is an argument or an adjunct.

In the final part of the paper, examples of testing the distinction between arguments and adjuncts in the repository SARGADA will be shown. According to them, a possibility of connecting the presented records of verb lemmas with already finished dictionaries will be considered, as well as the possibility of using the methodology and the results of diagnostic tests for the next phase of the valency dictionary.

References

- Birtić, Matea; Brač, Ivana; Runjaić, Siniša. (2017). The Main Features of the e-Glava Online Valency Dictionary. Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference. Leiden, the Netherlands, 19–21 September 2017. / Eds. I. Kosem et al. Leiden: Lexical Computing CZ s.r.o., Brno, Czech Republic, pp. 43–62.

- Fillmore, Charles J.; Baker, Collin F. (2009). A Frames Approach to Semantic Analysis. The Oxford Handbook of Linguistic Analysis. Eds. B. Heine and H. Narrog. Oxford University Press, Oxford, UK/New York, New York: pp. 313–340.

- Forker, D. (2014). A Canonical Approach to the Argument/Adjunct Distinction. Linguistic Discovery, 12: 27–40.

- Hajič, Jan et al. (2018). Prague Dependency Treebank 3.5. Institute of Formal and Applied Linguistics: LINDAT/CLARIN; Charles University: LINDAT/CLARIN. URL: `http://hdl.handle.net/11234/1-2621` (accessed 2022-04-11).

- Jezek, Elisabetta et al. (2014). T-PAS: A Resource of Typed Predicate Argument Structures for Linguistic Analysis and Semantic Processing. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). European Language Resources Association (ELRA): Reykjavik, Iceland: pp. 890–895.

- Toivonen, Ida. (2021). Arguments and adjuncts across levels. Proceedings of the LFG'21 Conference. Eds. M. Butt, J. Y. Findlay, I. Toivonen. CSLI Publications: Stanford, CA: pp. 306–331.

Dicionário da Língua Portuguesa: a new lexicographic resource of Academia das Ciências de Lisboa

Ana Salgado[1,2], Alberto Simões[3], Álvaro Iriarte Sanromán[4], Rita Vieira[1], Manuela Ferreira[1], Rita Carmo[1], Conceição Pinheiro[1]

[1]ILLLP – Instituto de Lexicologia e Lexicografia da Língua Portuguesa, Academia das Ciências de Lisboa, [2]NOVA CLUNL – Centro de Linguística da Universidade NOVA de Lisboa, Portugal, [3] 2Ai – School of Technology, Instituto Politécnico do Cávado e do Ave, Barcelos, Portugal, [4]Centro de Estudos Humanísticos da Universidade do Minho, Portugal

E-mail: anacastrosalgado@gmail.com, asimoes@ipca.pt, alvaro@elach.uminho.pt, arita.v77@gmail.com, manelaferreira7@gmail.com, ritamendescarmo@gmail.com, mconceicaocpinheiro@gmail.com

Keywords: lexicography; lexicographic resource; dictionary-making process; dictionary editing system; Portuguese language

This article aims to present the *Dicionário da Língua Portuguesa* (DLP), a Portuguese lexicographic resource of the Academia das Ciências de Lisboa (ACL), which will be made available (free access) to the public in the coming months.

In an introductory section, we frame DLP in the European lexicographic scenario, in what is called the 'academy tradition' (Considine, 2014), i.e., the dictionaries produced by academies, making, then, a brief retrospective of the various editions of academy dictionaries, from the beginning to the present day. Despite the commitment and effort of several academicians, the ACL has only three dictionaries published to date, two of which are incomplete, i.e., they remained in a single volume (letter A) published in 1793 and 1976. The third, the *Dicionário da Língua Portuguesa* Contemporânea (DLPC), was published in two volumes (A-F, G-Z), which served as the starting point for the lexicographic resource that will now be made available through the Instituto de Lexicologia e Lexicografia da Língua Portuguesa (ILLLP).

The DLP project started with the conversion of a PDF file corresponding to the paper edition of the DLPC to an XML document (Simões et al., 2016) to which a customised P5 scheme of the Text Encoding Initiative (TEI) (TEI Consortium) was applied after defining the microstructural model of the lexicographic articles and the main markers/labels. This scheme conforms to the TEI Lex-0 guidelines (Tasovac et al., 2018) DLP's initial editing support was the *Oxygen XML Editor*, which was in use but has gradually been replaced by a new editing environment, LeXmart (Simões & Salgado, 2022).

The new project corresponds to a partially revised version of the DLPC with the introduction of thousands of lexicographic articles—compared to the previous edition (69,426 entries)—and a goal to reach 100,000 entries. For the construction and consultation of an ad-hoc corpus, a *Sketch Engine* license was recently purchased (Kilgarriff et al., 2014).

Our presentation will focus essentially on the following areas:

1. Presentation of the workflow of the new project, from its conception to the final product, and of LeXmart as a support for editing (revision and addition of new articles) and management of the lexicographic content (status of the articles, reports, and above all, the need to obtain reports to send to academicians).

2. Presentation of the DLP model and general characteristics (structure of dictionary articles, typographical conventions, data, metadata, etc.).

3. The recent revision work was carried out by a team of five people hired to correct the main conversion errors that persisted. The most problematic (for example, the distinction between collocations, usage examples, and citations; structural division of categorical homonyms; examples that are wrongly split by the presence of punctuation marks; decisions taken to present the spellings before and the new spelling in effect, etc.) will be duly systematised and explored here.

4. The main changes introduced compared to the previous edition, mainly as a response and adaptation of the retro-digitised version of the DLPC to the digital environment, namely the representation of articles online.

5. Measures implemented to guarantee the quality and preservation of the new resource in the long term.

References

- Considine, J. (2014). Academy dictionaries 1600–1800 [quoted from p. 2]. Cambridge: Cambridge University Press.

- Simões, A., Almeida, J. J., Salgado, A. (2016). Building a dictionary using XML technology. In Mernik, M., Leal, J. P., Oliveira, H. G. (Eds.), 5th Symposium on Languages, Applications and Technologies (SLATE'16) (14:1–14:8). Germany: Dagstuhl. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

- TEI Consortium, (Eds). TEI P5: Guidelines for Electronic Text Encoding and Interchange. [Version 4.5.0]. [Last updated on 2022-10-25]. TEI Consortium. URL: `http://www.tei-c.org/Guidelines/P5/`.

- Tasovac, T., Romary, L., Bański, P., Bowers, J., Does, J. de, Depuydt, K., Erjavec, T., Geyken, A., Herold, A., Hildenbrandt, V., Khemakhem, M., Petrović, S., Salgado, A., e Witt, A. (2018). TEI Lex-0: A baseline encoding for lexicographic data. Version 0.8.5. DARIAH Working Group on Lexical Resources

- Simões, A., Salgado, A. (2022). Smart Dictionary Editing with LeXmart. In Klosa-Kückelhaus, Annette, Engelberg, Stefan, Möhrs, Christine & Storjohann, Petra (eds.). Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag, pp. 423-434.

- Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., Suchomel, V. (2014). The Sketch Engine: ten years on Lexicography, 1: 7-36.

Other references

- ACL. (1793). Planta para se formar o Diccionario da lingoa portugueza. In Diccionario da lingoa portugueza, t. 1, A (pp. i-xx). Academia Real das Ciências de Lisboa. Lisboa: Na Officina da mesma Academia.

- ACL. (1976). Dicionário da Língua Portuguesa, 1 vol., Coelho, J. P. (Coord.) Lisboa: Academia das Ciências de Lisboa.
- ACL. (1987). Instituto de Lexicologia e Lexicografia da Língua Portuguesa. Lisboa: Academia das Ciências de Lisboa.
- ACL (2001). Dicionário da Língua Portuguesa Contemporânea, 2 vols. Lisboa: Academia das Ciências de Lisboa and Editorial Verbo.
- ACL (2023). Dicionário da Língua Portuguesa. Lisboa: Academia das Ciências de Lisboa. [New digital edition under revision.]
- Agudo, F. R. D. (1980). A Academia das Ciências de Lisboa e as relações internacionais no domínio da ciência e da cultura. Lisboa: Academia das Ciências de Lisboa.
- Agudo, F. R. D. (1986). Contribuição da Academia das Ciências de Lisboa para o desenvolvimento da ciência. Lisboa: Academia das Ciências de Lisboa.
- Amaral, I. (2012). Notas históricas sobre os primeiros tempos da Academia das Ciências de Lisboa. Lisboa: Colibri.
- Atkins, B. T. S., & Rundell, M. (2008). The Oxford Guide to Practical Lexicography. New York: Oxford University Press.
- Casteleiro, J. M. (1981). Estudo linguístico do 1.o dicionário da Academia. Memórias da Academia das Ciências de Lisboa, 22, 47–67.
- Casteleiro, J. M. (2008). Actividades lexicográficas da Academia das Ciências de Lisboa. In González Seoane, E., Santamarina, A., & Varela Barreiro, X. (Ed.), A lexicografía galega moderna. Recursos e perspectivas (pp. 315–322). Santiago de Compostela: Consello da Cultura Galega; Instituto da Lingua Galega.
- Dias, J. A. (2018). A Academia Real das Ciências de Lisboa (1779–1834) – Ciências e hibridismo numa periferia europeia. Lisboa: Colibri.
- Landau, S. I. (2001). Dictionaries. The art and craft of lexicography. Cambridge: Cambridge University Press.
- Peixoto, J. P. (1997). A ciência em Portugal e a Academia das Ciências de Lisboa. Colóquio/Ciências, 19, 71–84.
- Salgado, A. (2021). Terminological Methods in Lexicography: Conceptualising, Organising and Encoding Terms in General Language Dictionaries. Doctoral dissertation. Universidade NOVA de Lisboa. URL: `https://run.unl.pt/handle/10362/137023`.
- Simões, A., Salgado, A., Costa, R., & Almeida, J. J. (2019). LeXmart: A smart tool for lexicographers. In Kosem, I., Zingano Kuhn, T., Correia, M. Ferreira. J. P., Janson, M., Pereira, I., Kallas, J., Jakubicek, M., Krek, S. & Tiberius, C. (Eds.), Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference (pp. 453–466). Sintra, Portugal, Bron: Lexical Computing CZ, s.r.o. ISSN 2533-5626.

- Svensén, B. (2009). A Handbook of Lexicography: The Theory and Practice of Dictionary Making. Cambridge: Cambridge University Press.
- Verdelho, T. (2007). Dicionários portugueses: Breve história. Verdelho, T., & Silvestre, J. P. (Orgs.), Dicionarística portuguesa, inventariação e estudo do património lexicográfico (pp. 11–60). Aveiro, Universidade de Aveiro.
- Zgusta, L. (1971). Manual of lexicography. Prague and The Hague: Academia and Mouton.

Towards a Comprehensive Dictionary of Middle Persian

Francisco Mondaca[1], Kianoosh Rezania[2], Slavomír Čéplö[2], Claes Neuefeind[1]
[1]University of Cologne, [2]Ruhr-Universität Bochum
E-mail: f.mondaca@uni-koeln.de, kianoosh.rezania@rub.de,
slavomir.ceploe@ruhr-uni-bochum.de, c.neuefeind@uni-koeln.de

Keywords: corpus-based dictionary; middle persian; api

Middle Persian was spoken in the province Persis (Fārs) in the first millennium CE. It served as the official language of the Sasanian Empire (224 - 651 CE), a dominant power beside the Roman Empire during late antiquity. Middle Persian derives from Old Persian, the language spoken in the same area until the third century BCE. In the last centuries of the first millennium CE, Middle Persian has developed into New Persian, the major language of people in today's Iran, Afghanistan, and Tajikistan.

The project 'Zoroastrian Middle Persian: Corpus and Dictionary (MPCD)[1] will develop an exhaustive Middle Persian-English dictionary along with a digital corpus of Zoroastrian Middle Persian texts. These texts form the largest sub-corpus of the Middle Persian corpus. Based on Middle Persian codices from the 13th to 17th centuries CE, the corpus comprises approximately 54 texts with nearly 700,000 tokens. It will be annotated meticulously with layers of orthographical, grammatical, semantic, and intertextual data. The goal of this project is to offer a rich resource for the exploration of the Middle Persian language and literature through detailed and multifaceted annotations.

In our paper, we discuss the technical aspects of the project, with a specific focus on the modeling of the dictionary and the corpus. We discuss their connection and emphasize why their integration is crucial to enhancing the search functionality, which is key to the project's success. The dictionary's digital construction involves a backend infrastructure developed in Django, paired with a user interface frontend developed using React.js. To enhance the search functionality of our Middle Persian-English dictionary, we are integrating semantic search capabilities. Our ultimate objective is to render the MPCD accessible to the scholarly community. We aspire to catalyze further research in the field of Middle Persian language studies and digital lexicography by offering a user-friendly, web-based platform.

[1]https://www.mpcorpus.org

The Kosh Suite: A Framework for Searching and Retrieving Lexical Data Using APIs

Francisco Mondaca, Philip Schildkamp, Felix Rau, Luke Günther
University of Cologne
E-mail: f.mondaca@uni-koeln.de, philip.schildkamp@uni-koeln.de, f.rau@uni-koeln.de, lguent12@uni-koeln.de

Keywords: api; graphql; rest; xml; react

The Kosh Suite[1] is a comprehensive software framework for managing and accessing lexical data in XML (eXtensible Markup Language) format. With a few configuration settings, Kosh provides REST and GraphQL APIs for lexical XML data. XML, a popular format for lexical data, enables a structured way to store and share information. The use of tags in XML allows for the identification of various elements of the data, including but not limited to words, definitions, and examples. These tags enhance data understanding and navigation, providing clear demarcation of different elements and their interdependencies. XML enables parsing and processing by computer systems. This facilitates automated indexing and searching of data, which is especially useful for analyzing large datasets.

The framework's backend is based on Elasticsearch, a search engine, to index the lexical data and make it searchable. The indexed data is then exposed through two APIs per dataset: a REST API and a GraphQL API. The REST (Representational State Transfer) API, a standard for creating web services, enables communication between different systems over the internet. It follows a client-server architecture and allows read operations on the indexed data, which is compatible with a wide range of programming languages and frameworks. The GraphQL API, conversely, allows clients to precisely request the data they need, making it more flexible than a REST API. It also supports nested queries and retrieving multiple resources in a single request. Having both a REST and GraphQL API available for each dataset makes the Kosh Suite a versatile tool for managing and searching lexical data.

The Kosh Suite's frontend, developed using React.js and customized with Tailwind CSS, provides a user-friendly interface for searching indexed lexical data. A unique characteristic of the Kosh Suite is the provision to tailor search fields via JSON, enhancing the searchability and accessibility of indexed lexical data. This feature enables the users to shape the search experience to better suit their requirements, thereby improving the efficiency of data exploration. The Kosh Suite can be deployed on both frontend and backend using Docker, offering flexibility in operation and ensuring a consistent integration with various system architectures. The Kosh Suite's proven reliability in various research projects and the recent integration of a frontend component affirm its value in lexical data management and search.

[1] https://kosh.uni-koeln.de

Humanitarian reports on ReliefWeb as a domain-specific corpus

Loryn Isaacs
Universidad de Granada
E-mail: lisaacs@ugr.es

Keywords: corpus linguistics; humanitarian domain; specialized corpus; API methods

This paper presents an assessment of the content available on ReliefWeb's API for its suitability as a domain-specific corpus. ReliefWeb (`https://reliefweb.int/`) is a service managed by the United Nations Office for the Coordination of Humanitarian Affairs that aggregates publicly available documents related to current humanitarian issues. It contains nearly a million texts that span half a century and represent thousands of diverse actors. While this data could be of significant value for corpus-based research into humanitarian discourse, an analysis of its composition and features is needed to guide its utilization.

The corpus created for assessment includes the majority of reports available on ReliefWeb, focusing on those written in English from the year 2000 onward. These constitute nearly 660,000 texts and over 430 million tokens. The metadata fields offered by the API are enumerated and classified by their variability and level of representation throughout the corpus. Attention is given to fields likely to be the most useful for tracking the usage of humanitarian concepts and measuring their variability, e.g., publication date, affected countries, disaster type, and organization type. High-level trends based on these fields are described to establish specific avenues for further work.

The software and procedures used for compiling the corpus are detailed. A Python package called Corpusama is introduced as a means to orchestrate corpus creation and maintenance. It integrates the control of API data sources with a workflow that relies on open-source tools to generate vertical-formatted corpus files. Tokenization, lemmatization, and part-of-speech tagging are assigned to a neural network pipeline available in the Stanza natural language processing package (Qi et al., 2020). Prepared corpus content is queried with a local instance of NoSketch Engine, the open version of the popular corpus management system (Kilgarriff et al., 2014; Rychlý, 2007).

This assessment of ReliefWeb's API content is part of ongoing efforts to expand and refine the corpus-based methods that are used to generate concept entries for the Humanitarian Encyclopedia platform (`https://humanitarianencyclopedia.org/`). The Encyclopedia, a project by the Geneva Centre of Humanitarian Studies, offers analyses on key humanitarian concepts with a combination of corpus-based linguistic reports and input from domain experts. Applications for the ReliefWeb corpus are discussed with respect to the Encyclopedia's objectives for concept analysis and knowledge production.

References

- Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: Ten years on. Lexicography, 1, Article 1.
- Qi, P., Zhang, Y., Zhang, Y., Bolton, J., Manning, C. D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In A. Celikyilmaz

& T.-H. Wen (Eds.), Proceedings of the 58th annual meeting of the association for computational linguistics: System demonstrations. Association for Computational Linguistics.

- Rychlý, P. (2007). Manatee/Bonito-A modular corpus manager. In P. Sojka & A. Horák (Eds.), First Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2007 (pp. 65–70). Masaryk University.

Actional properties of verbs in learner's dictionaries' entries

Sarah Piepkorn[1], Laura Giacomini[2]
[1]University of Hildesheim, [2]University of Innsbruck
E-mail: piepkorns@uni-hildesheim.de, laura.giacomini@uibk.ac.at

Keywords: learner's lexicography; verbal aspect; actionality; German; English; Italian

Verbal aspect, which includes actionality as one of its dimensions, can be understood as the way in which languages present the phase structure and boundaries of situations, expressed by linguistic devices operating on different levels of language (Sasse, 2002, p. 201f.). Aspectual phenomena are closely related to verbs and their properties and are therefore a potential area of description in learner's dictionaries when providing learners with information on verbs and their behaviour. This is the subject of a doctoral project which aims at extracting information on the aspectual behaviour of verbs in German, English and Italian from corpora and integrating this information into a phraseology-centred, monolingual electronic dictionary model for language learners (named 'Phrase-based Active Dictionary' or PAD) within the larger PhraseBase project (DiMuccio-Failla & Giacomini, 2017a, 2017b; Giacomini et al., 2020; DiMuccio-Failla & Giacomini, 2022).

The present contribution focuses on the area of actionality (also 'Aktionsart') as a lexical semantic property of verbs and verb phrases, excluding other phenomena in the area of verbal aspect for this purpose. It presents an exploratory analysis of existing lexicographic practice concerning actionality: How and to what extent do verb entries in (electronic) learner's and general dictionaries of German, English and Italian already reflect actional properties of verbs and verb senses? Using Johanson's (1971, 1996, 2000) and Vendler's (1957) actional content classes (modified and refined by several authors like Comrie (1976), Rothstein (2004), Kratzer (1995), Croft (2012)) we analysed a total of seven verbs of movement among four dictionaries each as to the presence of actional information in the division into meanings, the wording of their definitions and the allocation of examples. The analysis showed that the dictionary entries contain certain types of relations to actional properties in all of the three areas, but that the information in the individual entries is non-systematic.

Based on this finding, we then present some considerations on how actional properties can be actively used in the production of verb entries within the framework of the electronic dictionary model developed in PhraseBase.

References

- Comrie, B. (1976). Aspect: an introduction to the study of verbal aspect and related problems. Cambridge [i.a.]: Cambridge University Press.
- Croft, W. (2012). Verbs: Aspect and Causal Structure. Oxford [i.a.]: Oxford University Press.
- DiMuccio-Failla, P. V., Giacomini, L. (2017a). Designing a Learner's Dictionary Based on Sinclair's Lexical Units by Means of Corpus Pattern Analysis and the Sketch Engine. 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.

Leiden, the Netherlands, 19–21 September 2017. Brno: Lexical Computing, pp. 437–457. URL: https://elex.link/elex2017/proceedings-download/.

- DiMuccio-Failla, P. V., Giacomini, L. (2017b). Designing a Learner's Dictionary with Phraseological Disambiguators. In R. Mitkov (ed.) Computational and Corpus-Based Phraseology. Second International Conference, Europhras 2017. London, UK, 13–14 November 2017, pp. 290–305. Cham: Springer. URL: https://doi.org/10.1007/978-3-319-69805-2_21.
- DiMuccio-Failla, P. V., Giacomini, L. (2022). A proposed microstructure for a new kind of active learner's dictionary. Lexicographica, 38(1), pp. 475–499.
- Giacomini, L., DiMuccio-Failla, P. V., Lanzi, E. (2020). The interaction of argument structures and complex collocations: role and challenges in learner's lexicography. In Z. Gavriilidou, M. Mitsiaki & A. Filatouras (eds.) Lexicography for Inclusion: Proceedings of the 19th EURALEX International Congress, 7–9 September 2021, Alexandroupolis, Vol. 1, pp. 285–293. Alexandroupolis: Democritus University of Thrace. URL: https://euralex.org/category/publications/euralex-2020-2021/.
- Johanson, L. (1971). Aspekt im Türkischen. Vorstudien zu einer Beschreibung des türkeitürkischen Aspektsystems. Stockholm: Almqvist & Wiksell.
- Johanson, L. (1996). Terminality operators and their hierarchical status. In B. Devriendt, L. Goossens & J. Auwera (eds.) Complex Structures: A Functionalist Perspective. Berlin: Mouton De Gruyter, pp. 229–258. URL: https://doi.org/10.1515/9783110815894.229.
- Johanson, L. (2000). Viewpoint operators in European languages. In Ö. Dahl (ed.) 6 Tense and Aspect in the Languages of Europe. Berlin/New York: De Gruyter Mouton, pp. 27–188. URL: https://doi.org/10.1515/9783110197099.
- Kratzer, A. (1995). Stage-level and individual-level predicates. In G. N. Carlson & F. J. Pelletier (eds.) The generic book. Chicago [i.a.]: Chicago University Press.
- Rothstein, S. (2004). Structuring Events: A Study in the Semantics of Lexical Aspect. Malden: Blackwell. URL: https://doi.org/10.1002/9780470759127.
- Sasse, H.-J. (2002). Recent activity in the theory of aspect: Accomplishments, achievements, or just non-progressive state?. Language Typology, 6(2), pp. 199–271. URL: https://doi.org/10.1515/lity.2002.007.
- Vendler, Z. (1957). Verbs and times. The Philosophical Review, 66(2), pp. 143–160. URL: https://doi.org/10.2307/2182371.

A Federated Search and Retrieval Platform for Lexical Resources in Text+ and CLARIN

Thomas Eckart[1], Axel Herold[2], Erik Körner[1], Frank Wiegand[2]

[1]Saxon Academy of Sciences and Humanities in Leipzig, [2]Berlin-Brandenburg Academy of Sciences and Humanities

E-mail: eckart@saw-leipzig.de, herold@bbaw.de, koerner@saw-leipzig.de, wiegand@bbaw.de

Keywords: Lexical API; Federated research infrastructure; Lexical search plattform; Dictionary integration

The Text+ consortium[1] is part of Germany's National Research Data Infrastructure (NFDI)[2] that focuses on FAIR-compliant utilization of text- and language-based research data in a distributed environment. Its data domain "lexical resources" deals with all kinds of lexical resources, including dictionaries, encyclopedias, normative data, terminological databases, ontologies etc. Many of the largest German providers of such resources are members of the consortium.

One salient goal is the integration of lexical data in a decentralized dictionary platform. Due to the heterogeneous nature of available resources, formats, levels of annotation, and technical architectures in use, the implementation will follow a federated approach based on common protocols and data formats. Query and retrieval of lexical data is based on the protocol for the Federated Content Search (FCS[3]). It builds upon and significantly extends preliminary work done in the European CLARIN[4] project (see also Figure 1).

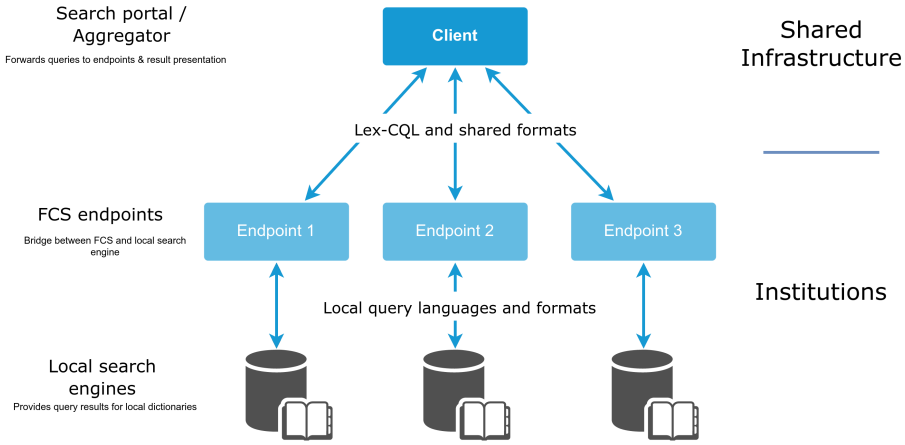


Figure 1: The general FCS architecture

The CLARIN FCS is an established federated search engine that allows querying distributed corpora by using a standardized RESTful protocol and data formats (Stehouwer et al., 2012). The extensible specification and architecture is currently focused on text corpora,

[1]https://www.text-plus.org/en
[2]https://www.nfdi.de/?lang=en
[3]https://www.clarin.eu/content/content-search
[4]https://www.clarin.eu/

but support for request and retrieval of lexical entries has long been discussed and is currently implemented in an iterative work process coordinated between Text+ and CLARIN's FCS taskforce.

The FCS specification (Schonefeld et al., 2014) will be extended with regard to announcing, querying and retrieving lexical resources. Specifically, this entails:

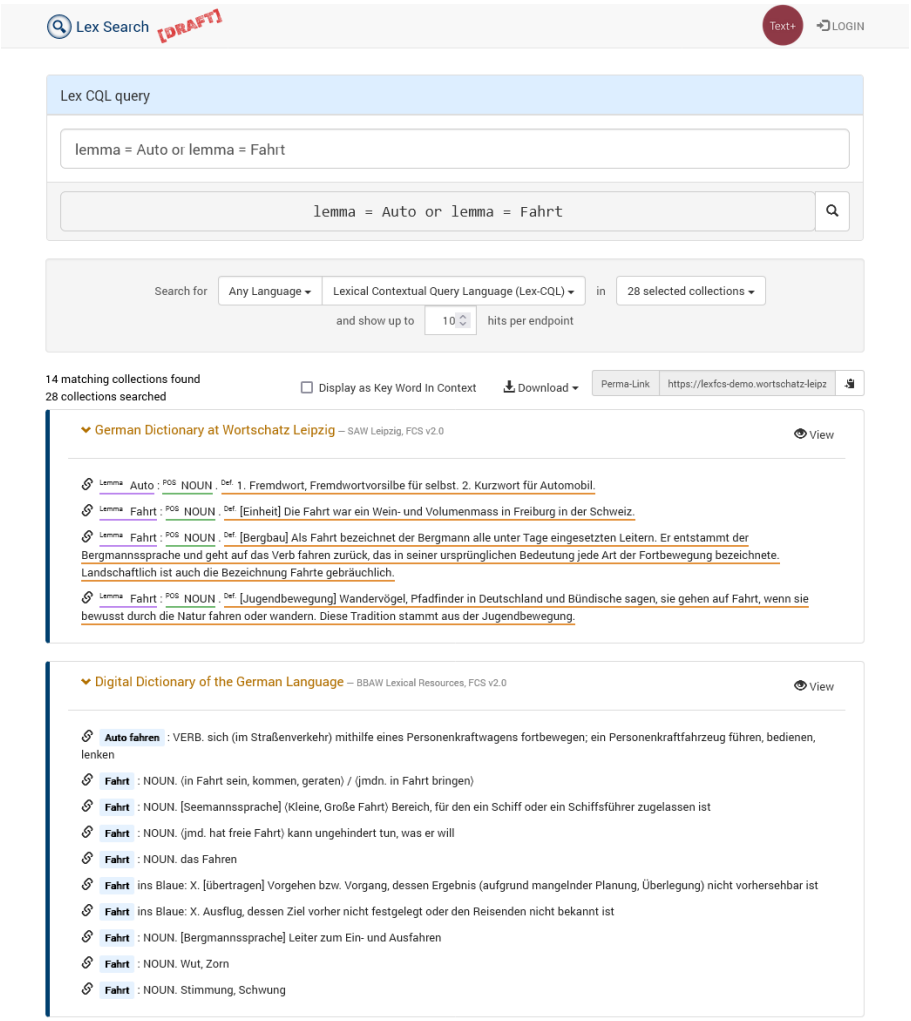


Figure 2: Screenshot of the frontend demonstrator

- Specifying the query language which is a "CQL Context Set[5]" of the Contextual Query Language[6] (standardized by the US Library of Congress) dedicated to query lexical entries. Its specification includes agreements on accessible fields of information (like part-of-speech, definitions, (semantically) related entries etc.) for a lexeme and how to combine them to complex queries. This is especially challenging due to the inherently hierarchical structure of lexical data.

[5] https://www.loc.gov/standards/sru/cql/contextSets/theCqlContextSet.html

[6] https://www.loc.gov/standards/sru/cql/

- Specifying common data formats for a unified result presentation. On the basic level, this is achieved by a mandatory KWIC representation that allows annotating information types inline and by an advanced tabular-representation of all fields in a key-value-style. It is clearly understood that in most cases these representations can only provide a simplified view of the data. It is therefore endorsed to provide records in their complex native representation as well; with examples being different TEI dialects including TEI Lex-0[7], OntoLex/Lemon[8], and other formats.
- Extending the core FCS specification while remaining compatible with the overall architecture to enable the reuse of features such as access control for restricted resources or automatic registering of endpoints within the FCS system.

All mentioned constituents of the architecture are actively worked on and are incrementally developed. Throughout specification and implementation, feedback is provided by interested parties, particularly from but not limited to the Text+ and CLARIN consortia. With a first public release in the coming months – based on the current demonstrator[9] (see Figure 2) –, we will improve the availability and visibility of various lexical resources, including some that were not easily accessible or even unknown to the general public until now.

References

- Schonefeld, O., Eckart, T., Kisler, T., Draxler, Ch., Zimmer, K., Ďurčo, M., Panchenko, Y., Hedeland, H., Blessing, A., Shkaravska, O. (2014). CLARIN Federated Content Search (CLARIN-FCS) – Core Specification. URL: `https://www.clarin.eu/content/federated-content-search-core-specification`. Accessed at 16.01.2023.
- Stehouwer, H., Durco, M., Auer, E., Broeder, D. (2012). Federated Search: Towards a Common Search Infrastructure. Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12).

[7] `https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html`

[8] `https://www.w3.org/2019/09/lexicog/`

[9] `https://hdl.handle.net/11022/0000-0007-FBF2-D?urlappend=%3FqueryType%3Dlex`

The lexicographic process revisited

Annette Klosa-Kückelhaus[1], Carole Tiberius[2]

[1]Leibniz-Institut für Deutsche Sprache, [2]Instituut voor de Nederlandse Taal
E-mail: klosa@ids-mannheim.de, carole.tiberius@ivdnt.org

Keywords: lexicographic process; lexicographic database; lexicographic data models

By the lexicographic process, we understand all the steps that are necessary for a dictionary or a lexicographic database to be published in print or electronically. This process has initially been described exclusively for print dictionaries (e.g., in Dubois, 1990; Landau, 1984; Riedel & Wille, 1979; Schaeder, 1987; and Zgusta, 1971), later Wiegand (1999) and Müller-Spitzer (2003) expanded it to electronic dictionaries, and finally the process of online dictionaries was analysed (cf. Klosa, 2013; Klosa & Tiberius, 2015; Svensén, 2009). Although the lexicographic process does not seem to be a particularly fashionable topic[1], it is important to continually readjust our ideas on it, as the process is manyfold (see the results of a study comparing 14 different projects in Europe in terms of their lexicographic process within the COST ENeL action, cf. Tiberius & Krek, 2014) and is constantly evolving.

As lexicographic institutions seem to be moving to creating a central database which contains at least a shared core of lexicographic data, we describe in this paper how the lexicographic process changes when you move away from the publication of one single dictionary to a central database that feeds various lexicographic projects and potentially also other (not necessarily) lexicographic tools and applications. In a constellation like that, there does not seem to be only one lexicographic process but there are two interacting processes (one for each individual dictionary and one for the central database). We will discuss the possible reasons for this development (emergence of linguistic linked data, possibly also shorter project running times, etc.) and some of its consequences (e.g., more stable and established formats for encoding lexicographic resources, cf. Kernerman, 2011; Depuydt et al., 2019; Parvizi et al., 2016; Tavast et al., 2018; see also initiatives such as TEI Lex-0 (Tasovac et al., 2018), Ontolex-Lemon (Cimiano et al., 2016), LMF[2] and the ongoing work on standardisation in the OASIS - Lexicographic Infrastructure Data Model and API (LEXIDMA Technical Committee[3]).

We will also take into consideration two important outcomes of the ELEXIS surveys on lexicographic practices in Europe (Kallas et al., 2019; Tiberius et al., 2022): the results show that online publication is now really the most popular publication medium for dictionaries in Europe and that crowdsourcing and gamification are not usually part of the lexicographic workflow yet. Nevertheless, we will present a new model of the lexicographic process in which these are taken into consideration.

[1]Searching for "lexicographic(a) process" in the title of all 6.482 articles listed in ELEXIFINDER (see `https://elex.is/tools-` and-services/elexifinder/) only gives two hits. There are also only two hits for "compilation process", three hits for "dictionary making process", and one hit for "computational lexicon making process" in article titles as of 02 February 2023).

[2]`https://www.iso.org/standard/68516.html` (16 January 2023).

[3]`https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=lexidma` (16 January 2023).

References

- Cimiano, P., McCrae, J. P., Buitelaar, P. (2016). Lexicon Model for Ontologies: Community Report, 10 May 2016 Specification. URL: https://www.w3.org/2016/05/ontolex/. (16 January 2023).

- Depuydt, K., Schoonheim, T., de Does, J. (2019). Towards a More Efficient Workflow for the Lexical Description of the Dutch Language. URL: http://videolectures.net/elexisconference2019_depuydt_dutch_language/. (16 January 2023).

- Dubois, C. (1990). Considérations générales sur l'organisation du travail lexicographique. In: F.-J. Hausmann, O. Reichmann, H. E. Wiegand & L. Zgusta (eds.), Dictionaries. An international handbook on lexicography. Berlin/New York: de Gruyter, pp. 1645–1672.

- 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. Proceedings of the eLex 2011 Conference, 11-12 November 2011, Bled Slovenia. URL: https://elex2011.trojina.si/Vsebine/proceedings/eLex2011-14.pdf. (16 January 2023).

- Kallas, J., Koeva, S., Kosem, I., Langemets, M., Tiberius, C. (2019). ELEXIS deliverable 1.1 Lexicographic Practices in Europe: A Survey of User Needs. URL: https://elex.is/wp-content/uploads/2020/06/Revised-ELEXIS_D1.1_Lexicographic_Practices_in_Europe_A_Survey_of_User_Needs.pdf. (16 January 2023).

- Klosa, A. (2013). The lexicographical process (with special focus on online dictionaries). In: R. H. 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.

- Klosa, A., Tiberius, C. (2015). Der lexikografische Prozess. In: A. Klosa & C. Müller-Spitzer (eds.), Kompendium Internetlexikografie. Berlin, Boston: de Gruyter, pp. 65–110.

- Landau, S. (1984). Dictionaries. The Art and Craft of Lexicography. New York: Charles Scribner's Sons. Müller-Spitzer, C. (2003) Ordnende Betrachtungen zu elektronischen Wörterbüchern und lexikographischen Prozessen. In: Lexicographica 19, pp. 140–168.

- Parvizi, A., Kohl, M., Gonzàlez, M., Saurí, 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'16), Portorož, Slovenia. URL: https://www.aclweb.org/anthology/L16-1071/. (16 January 2023).

- Riedel, H., Wille, M. (1979). Über die Erarbeitung von Lexika. Leipzig: Bibliographisches Institut. Schaeder, B. (1987) Germanistische Lexikographie. Berlin, Boston: de Gruyter Mouton. (Lexicographica. Series Maior 34).
- Svensén, B. (2009). A Handbook of Lexicography. The Theory and Practice of Dictionary-Making. Cambridge: Cambridge University Press.
- Tasovac, T, Romary, L., Banski, P., Bowers, J., de Does, J., Depuydt, K., Erjavec, T., Geyken, A., Herold, A., Hildenbrandt, V., Khemakhem, M., Petrović, S., Salgado, A., Witt, A. (2018). TEI Lex-0: A baseline encoding for lexicographic data. Version 0.8.6. DARIAH Working Group on Lexical Resources. URL: `https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html`. (16 January 2023).
- 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.
- Tiberius, C., Krek, S. (2014). Workflow of Corpus-Based Lexicography. Deliverable COST-ENeL-WG3 meeting. URL: `https://www.elexicography.eu/wp-content/uploads/2015/04/LexicographicalWorkflow_DeliverableWG3BolzanoMeeting2014.pdf`. (16 January 2023).
- Tiberius, C., Kallas, J., Koeva, S., Langemets, M., Kosem, I. (2022). An insight into lexicographic practices in Europe. Results of the extended ELEXIS Survey on User Needs. In: A. Klosa Kückelhaus, S. Engelberg, Ch. Möhrs & P. Storjohann (eds.): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag, pp. 509–521.
- Wiegand, H. E. (1998). Wörterbuchforschung. Untersuchungen zur Wörterbuchbenutzung, zur Theorie, Geschichte, Kritik und Automatisierung der Lexikographie. 1. Partial volume. Berlin/New York: de Gruyter.
- Zgusta, L. (1971). Manual of Lexicography. The Hague/Paris: Mouton.

From Structured Textual Data to Semantic Linked-data for Georgian Verbal Knowledge

Archil Elizbarashvili[1], Mireille Ducasse[2], Manana Khachidze[1], Magda Tsintsadze[1]

[1]Tbilisi State University, Georgia, [2]Univ Rennes, INSA Rennes, CNRS, IRISA, France
E-mail: archil.elizbarashvili@tsu.ge, mireille.ducasse@irisa.fr, manana.khachidze@tsu.ge, magda.tsintsadze@tsu.ge

Keywords: Data transformation; Data validation; Machine learning; Decision tree; Georgian language

The Georgian language has a difficult grammar. The verbal system, in particular, is challenging. To help foreigners learn Georgian, a linked-database of inflected forms of Georgian verbs is being built: KartuVerbs. It is accessible by a logical information system, Sparklis (Ferré, 2017), that enables powerful access and navigation as demonstrated in (Ducassé, 2020; Ducassé & Elizbarashvili, 2022). To build KartuVerbs, we started from a structured textual form of the knowledge developed by Meurer (2007) for the INESS project. INESS is an infrastructure for the exploration of syntax and semantics. It is multilingual and it has a much broader scope than KartuVerbs. However, accessing its lexicographic data is challenging for our target users. Furthermore, the work on its base for Georgian has stopped. Integrating its data into KartuVerbs both revives them and allow them to evolve. The verbs are indexed by roots, a given root in general corresponds to several verbs, each verb has inflected forms in 11 tenses. There are more than 60 possible properties. Some of them are obsolete, kept for historical reasons. There are missing pieces of information. All properties are not systematically present for every verb. Some properties, important for us, do not exist, for example the ending of a form. After filtering and reconstruction of some properties, KartuVerbs currently contains more than 5 million inflected forms related to more than 16 000 verbs for 11 tenses; each form can have 14 properties; there are more than 80 million links in the base. Response times are acceptable when running on a private machine, thus validating the feasibility of the linked-data approach. There is still a need to validate, correct and expand data. Considering the mass of data, this requires tools.

The full paper analyses the Clarino database with respect to our needs and introduces a typology of fields. It describes the transformation process to go from the structured text to the linked data. The process is in 3 blocks. The first block scraps the web pages into a CSV file. The second block aims at incrementally improving the data. The third block produces RDF data and integrates them into Sparklis. We describe how the decision tree algorithm can help improve a field that has occasional missing values. The field is the verbal noun, the lemma to represent a Georgian verb. Verbal noun is crucial for our knowledge base. The main contribution of the described work is that all the scripts of the process are freely available on the web[1]. They can be adapted to other applications. Those of the first block could be the base to scrap other textual sources for other languages or applications. Those of the third block could be used to integrate into KartuVerbs (or another linked-data application) CSV data from other sources than INESS. The scripts to implement the decision tree algorithm dedicated to missing values for verbal nouns could be customized to predict occasional missing values of other fields. Furthermore, the typology of fields can be used as

[1]https://github.com/aelizbarashvili/KartuVerbs

an analysis grid to help transform a set of data into another set of data suited for different objectives.

We are indebted to Paul Meurer who granted us a private access to a web version of the base behind the Georgian functionalities of https://clarino.uib.no/iness. We thank Mikheil Sulikashvili for his help to scrap Clarino web pages. This research PHDF-22-1840 is supported by Shota Rustaveli National Science Foundation of Georgia (SRNSFG) and by ANR Project SmartFCA, ANR-21-CE23-0023.

References

- Ducassé, M. (2020). Kartu-Verbs: A Semantic Web Base of Inflected Georgian Verb Forms to Bypass Georgian Verb Lemmatization Issues. In Proceedings of XIX EURALEX International Congress, volume 1, pp.81–89.
- Ducassé, M., Elizbarashvili, A. (2022). Finding Lemmas in Agglutinative and Inflectional Language Dictionaries with Logical Information Systems: The Case of Georgian verbs. In Proceedings of XX EURALEX International Congress.
- Ferré, S. (2017). Sparklis: An Expressive Query Builder for SPARQL Endpoints with Guidance in Natural Language. Semantic Web: Interoperability, Usability, Applicability, 8(3), pp. 405–418.
- Meurer, P. (2007). A computational grammar for Georgian. In International Tbilisi Symposium on Logic, Language, and Computation. Springer, pp. 1–15.

A Search Engine for the Large Electronic Dictionary of the Ukrainian Language (VESUM)

Tamila Krashtan
Lviv Polytechnic National University
E-mail: tamila.krashtan@gmail.com

Keywords: search engine; online dictionary; Ukrainian; VESUM

The Large Electronic Dictionary of the Ukrainian Language (also known as VESUM) is a project started in 2005 aiming at generating a morphological dictionary for the Ukrainian language, which is also used in a Ukrainian POS-tagger (Rysin and Starko, 2005-2022). It is constantly being updated with new lexical data as well as new fine-grained tags describing the words of the Ukrainian language from various grammatical and semantic perspectives.

This paper presents a new search engine developed for the Electronic Dictionary. The current Dictionary's webpage enables the search through the database using only full-word forms: either among lemmas or among all word forms. Search results show a list of matched lemmas and word forms along with some of the internal tags associated with each of the word forms.

The aim of the project is to set up a more user-friendly interface with broader search options, which at the same time provides more information contained in the Dictionary database.

To achieve this, the search engines of the two languages closely related to Ukrainian – Polish and Belarusian – were analyzed (Kieraś and Woliński, 2017; Koshchanka and Bułojčyk, 2021). The decision was made to follow the path of the Belarusian tool since it has more flexibility in search settings as well as a better UI.

The newly developed search functionality for the Ukrainian Dictionary is built upon the search engine created for the Belarusian grammar database and utilizes grammar tags defined in the VESUM database. It enables the usage of wildcards in the search queries and allows a user to set up search grammars.

The developed system provides more extensive search options and a way of displaying lemma information that is more structured and transparent both for professionals and non-linguists. It is well-suited for the addition of new tags and search parameters (including, but not limited to, conjugation classes and variations in the orthography of certain words) which will be featured in future versions of the software.

References

- Rysin, A., Starko, V. (2005-2022). Large Electronic Dictionary of Ukrainian (VESUM). Web version 6.0.1. URL: `https://r2u.org.ua/vesum/`
- Kieraś, W., Woliński, M. (2017). "Grammatical Dictionary of Polish" – an online version. Język Polski, 97(1), 84–93.
- Koshchanka, U., Bułojčyk, A. (2021). Граматычная база беларускай мовы. Unpublished.

The use of lexicographic resources in Croatian primary and secondary education

Ana Ostroški Anić, Martina Pavić, Daria Lazić, Maja Matijević
Institute of Croatian Language and Linguistics
E-mail: aostrosk@ihjj.hr, mpavic@ihjj.hr, dlazic@ihjj.hr, mmatijevic@ihjj.hr

Keywords: dictionary skills; lexicography in teaching; usage practices; usage needs

The research on the use of dictionary in the context of education has largely focused on their role as reference tools and teaching aids in foreign language learning (e.g. Boulton & De Cock, 2017; Tono, 2001), with the exception of enhancing literacy and reading skills (e.g. Beech, 2010). The development of dictionary skills is still an active part of many European curricula (Vicente, 2022). The national curriculum for the Croatian language as L1 (MZO, 2018) lists the student's active use of a children's dictionary as one of the learning outcomes as early as in the first grade of primary education.

Recently, there have been several important large-scale surveys on user needs (Kallas et el., 2019; Kosem et al., 2019), but not enough studies have looked into how teachers use dictionaries and other lexicographic resources, both in class and when preparing teaching material. This paper presents research on the use of dictionaries and other lexicographic and specialized resources, such as encyclopedias, specialized dictionaries and databases, glossaries, etc., by all Croatian primary and secondary school teachers.

An online survey has been created to investigate the extent to which teachers of different subjects use dictionaries and other resources in preparing their classes and as teaching aids in the classroom. The survey anonymously asks teachers to answer questions regarding the frequency of their use of lexicographic resources in the classroom and while preparing material, as well as their satisfaction with the dictionaries' content and structure, and their use of dictionaries in class.

The survey aims to answer three main research questions:

1. To what extent do teachers use lexicographic resources in their teaching practices, both in preparation and in the classroom?

2. How do teachers perceive the relevance and accuracy of information in lexicographic resources in relation to the curriculum they are teaching?

3. How familiar are teachers with specialized dictionaries, databases, and other lexicographic resources?

A pilot survey was first conducted with a sample of 20 participants. Based on their feedback, certain questions were amended and others were introduced. The survey will be conducted from February 1 to February 15 at a national level, and will be distributed through various social networks, teacher associations, and educational networks. The preliminary results provide interesting results, e.g. that 34,4% participants often use online dictionaries in class, while as many as 73% participants use lexicographic resources to verify the meaning of a specialized term.

The results of the survey will provide insights into the extent of dictionary usage in the classroom, as well as teacher satisfaction with their content and structure. This information will be valuable for future development of dictionaries and other educational resources, to better meet the needs of teachers and students in the classroom.

References

- Boulton, A., De Cock, S. (2017). Dictionaries as aids for language learning. International Handbook of Modern Lexis and Lexicography. Eds. Hanks, Patrick; de Schryver, Gilles-Maurice. Springer. Berlin, Heidelberg.
- Egido Vicente, M. (2022). Dictionaries in German and Spanish Primary Education Curricula: A Comparative Study. International Journal of Lexicography 35(2): 176–203.
- Kallas, J. et al. (2019). Lexicographic Practices in Europe: Results of the ELEXIS Survey on User Needs. Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference. Eds. Kosem, Iztok et al. Brno: Lexical Computing CZ, s.r.o.
- Kosem, I. 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): 92–114.
- MZO = Ministarstvo znanosti i obrazovanja. 2018. Odluka o donošenju kurikuluma za nastavni predmet Hrvatski jezik za osnovne škole i gimnazije u Republici Hrvatskoj. URL: `narodne-novine.nn.hr/clanci/sluzbeni/2019_01_10_215.html`
- Beech, J. R. (2010). Using a dictionary: Its influence on a children's reading, spelling, and phonology. Reading Psychology 25(1): 19–36.
- Tono, Y. (2001). Research on Dictionary Use in the Context of Foreign Language Learning: Focus on Reading Comprehension. Berlin, Boston: Max Niemeyer Verlag.

Exploring the Capabilities of ChatGPT for Lexicographical Purposes: A Comparison with Oxford Advanced Learner's Dictionary within the Microstructural Framework

Chayanon Phoodai[1], Richárd Rikk

[1]Károli Gáspár University of the Reformed Church in Hungary, Budapest
E-mail: chayanon507@gmail.com, rikk.richard@gmail.com

Keywords: Artificial Intelligence; Generative Models; ChatGPT; E-lexicography; Microstructure; Oxford Advanced Learner's Dictionary

Artificial Intelligence (AI) has seen success in many areas of science in the past few years. From computer science to linguistics, deep neural networks have the ability to perform better than the previous state-of-the art solutions. Indeed, generative text-based models like ChatGPT are able to imitate human writing, however its capabilities in lexicography have not been studied thoroughly. This paper compares the lexicographical data provided by ChatGPT and the Oxford Advanced Learner's Dictionary in the scope of microstructure. Two main datasets are created for manual analysis and similarity score tests. The aim is to demonstrate the effectiveness of ChatGPT in providing lexicographical data to English language learners as compared to the Oxford Advanced Learner's Dictionary.

We accomplish this by comparing the provided data related to lexicographical items, using Wiegand's item classes to identify the co-occurring items within the microstructure of both platforms. The framework of item classes provides us with a list of lexicographical items that serve as our criteria. We then examine each lexical entry individually to determine whether each lexicographical item is present in both tools. The results are presented in a comparative table as percentages. Also, using Bilingual Evaluation Understudy (BLEU) and Recall Oriented Understudy for Gisting Evaluation (ROUGE) methods we calculate the similarity between the lexicographical data provided by ChatGPT and the Oxford Advanced Learner's Dictionary. Since ChatGPT has been trained on human data, we investigate how similar its generated answers are to the ground truth.

This study provides valuable insights into the potential of AI-generated dictionary content and its applicability in pedagogical lexicography. Additionally, it highlights the challenges and limitations that need to be addressed in order to inform the development of AI models for lexicography.

Thesaurus of Modern Slovene 2.0

Špela Arhar Holdt[1,2], Polona Gantar[1], Iztok Kosem[1,3], Simon Krek[1,3], Pori Eva[1], Marko Robnik-Šikonja[2]

[1]Faculty of Arts, University of Ljubljana, [2]Faculty of Computer and Information Science, University of Ljubljana, [3]Jozef Stefan Institute

E-mail: arharhs@ff.uni-lj.si, apolonija.gantar@ff.uni-lj.si, iztok.kosem@cjvt.si, simon.krek@guest.arnes.si, eva.pori@ff.uni-lj.si, marko.robniksikonja@fri.uni-lj.si

Keywords: Thesaurus of Modern Slovene; responsive dictionary; automated lexicography; user involvement; post-editing lexicography

The paper presents the improvement of the Thesaurus of Modern Slovene from version 1.0 to 2.0. The Thesaurus, first published in 2018, introduced the concept of a responsive dictionary: a digitally-born, automatically created resource (Krek et al., 2017) that provides fast access to open data on modern language use and is gradually improved through editing, which involves both lexicographic work and user participation. The Thesaurus allows users to propose new synonym candidates and assess existing ones (Arhar Holdt et al., 2018). The initial, lexicographically unreviewed version 1.0 contained entries and synonym candidates in a form of lemmata without part-of-speech or other metadata, with semantic description temporarily replaced by automatically obtained semantic clusters and the data lacking dictionary labels, apart from terminological ones. Despite these limitations, potential users found the new resource and the concept of a responsive dictionary useful (Arhar Holdt, 2020), and data shows the consistent widespread use of the Thesaurus ever since it was published. However, the aforementioned user study also identified priorities for the first upgrade, which was funded by the Slovenian Ministry of Culture in 2021–2022. The project aimed to upgrade the dictionary interface design; establish editorial protocols for including user-suggested synonyms in the dictionary database; pilot the automatic extraction and selection of antonyms and facilitate crowdsourcing of antonyms through the dictionary interface; add dictionary labels for extremely offensive (hateful) and vulgar vocabulary and allow users to also provide dictionary labels when contributing synonyms and antonyms; and finally, supplement the dictionary database with the description of sense distribution including short definitions of senses known as ”semantic indicators” for 2,000 entries. In the paper, we present the upgraded Thesaurus of Modern Slovene, together with the methodology for each enhancement. We illustrate the challenges and solutions of lexicographic work that involved utilizing and improving automatically extracted data and assess the effectiveness of machine-supported workflows. The Thesaurus, which was automatically generated from various lexical resources, serves as a state-of-the-art example of lexical data reuse, interconnectivity, and user involvement. The methodology and insights gained from our work can be useful for other language communities pursuing similar initiatives.

References

- Arhar Holdt, Š. (2020). How users responded to a responsive dictionary: the case of the Thesaurus of Modern Slovene. Rasprave Instituta za hrvatski jezik i jezikoslovlje. Vol. 46, no. 2, pp. 465-482.

- Arhar Holdt, Š. et al. (2018). Thesaurus of Modern Slovene: By the Community for the Community. In: Čibej, Jaka, Vojko Gorjanc, Iztok Kosem, Simon Krek (eds.). Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts.). Ljubljana: Znanstvena založba Filozofske fakultete. Pp. 401-410.
- 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: KOSEM, Iztok et al. (eds.), Proceedings of eLex 2017: Lexicography from Scratch, 19-21 September 2017, Leiden, Netherlands.

Digital Cartography for Dialectal Loanwords

Gerd Hentschel[1], Peter Meyer[2]
[1]Institut für Slavistik, Universität Oldenburg, [2]Leibniz-Institut für Deutsche Sprache, Mannheim
E-mail: gerd.hentschel@uni-oldenburg.de, meyer@ids-mannheim.de

Keywords: language geography; cartography; loanword lexicography; Polish language; dialect lexicography

This paper reports on an ongoing international lexicographical project on German lexical borrowings in Polish dialects (Meyer & Hentschel, 2021). The resulting dictionary will be published as part of the Lehnwortportal Deutsch (LWPD), a freely accessible online publication platform for a growing number of dictionaries on German lexical borrowings in other languages.

A central concern of the project is to document and visualize the often remarkable diversity of both the formal and the semantic reflexes of German words of origin in their dialectal and geographical distribution. The paper presents (a) the data model for "localizing" the word senses as well as the expression variants of a given loanword, and outlines (b) the cartographic representation and (c) the search options based on this model.

(a) The data model is based on the main data source of the project, the large Dictionary of Polish Dialects (SGP), of which only about a quarter has been published in (currently) 10 volumes since 1982. Therefore, most of the empirical data have to be taken from the card index of the SGP. In a complex process of interpreting the raw data of the SGP, often involving consultation of other resources, each expression variant of a loanword and each of its attested senses is separately "localized" by assigning it a subset of an inventory of more than 500 labels designating either counties/districts or (mostly dialectal) regions. Each region label is systematically mapped to the set of all "micro-area" county/district localizations within it and comes in two flavors (somewhere in X vs. everywhere in X). Where appropriate, the localizations contain information about the extensive east-west-migration of speakers after World War II. A separate data matrix links each expression (variant) to the senses in which it is attested and is used for complex consistency checks.

(b) In each final entry, all variants and senses are accompanied by a thumbnail map, allowing the user to see the distributional variation for a given loanword at a glance. For the editors, these maps help to resolve multiple homonymy constellations based on folk etymological blends in dialectal speech and to identify borrowing paths from different German dialects into Polish ones. Users can explore the dialectal distribution of a loanword on interactive, zoomable detail maps at various levels of data aggregation, from the overall attestation down to the geographical distribution of e.g. individual form-sense pairs, with documentation (legend, citations) for each "micro-area" localization shown.

(c) Search options: The comprehensive graph-based search options of the LWPD (cf. Meyer 2019) will include the possibility to use fine-grained localization data as a search criterion, e.g. for entries with an expression variant located in a certain area

With an estimated number of more than 6,500 loanwords and approximately 20,000 word senses and 21,000 expression variants, the resulting digital dictionary will, to the best

of our knowledge, be the first loanword dictionary to include an interactive cartographic tool that illustrates the traces of extensive and intensive historical language contact on a detailed regional level.

Acknowledgements

The project described in this paper is funded by the German Research Foundation (HE 1566/16-1; ME 4968/2-1).

References

- LWPD: Lehnwortportal Deutsch, ed. Leibniz Institute for the German Language. URL: `http://lwp.ids-mannheim.de`. [21 April 2023].
- Meyer, P. (2019). Leistungsfähige und einfache Suchen in lexikografischen Datennetzen. Ein Query Bilder für lexikografische Property-Graphen. In P. Sahle (ed.) Digital Humanities: multimedial & multimodal. 6. Tagung des Verbands Digital Humanities im deutschsprachigen Raum e.V. (DHd 2019), Frankfurt am Main, Mainz, 25.3.2019 – 29.3.2019. Konferenzabstracts. Frankfurt a.M., pp. 312-314.
- Meyer, P., Hentschel, G. (2021). Charting A Landscape of Loans. An e-Lexicographical Project on German Lexical Borrowings in Polish Dialects. In Z. Gavriilidou, L. Mitits & S. Kiosses (eds.): Proceedings of XIX EURALEX Congress: Lexicography for Inclusion, Vol. II. Democritus University of Thrace, pp. 615-621.
- SGP: Słownik gwar polskich (A – iżeby), Karaś, M. et alii (eds.). Wrocław etc.; since vol. IV – Kraków, 1977-2023.

Teaching Digital Lexicography from Scratch: an Open-Source Tool for XML and HTML

Peter Meyer

Leibniz-Institut für Deutsche Sprache, Mannheim
E-mail: meyer@ids-mannheim.de

Keywords: learning tool; XML technology stack; HTML

Understanding what lexicographical data modeling is all about and how to use a data model to construct, for example, an online presentation from structured data is an important prerequisite for anyone trying to get started in modern digital lexicography. Teaching such concepts in a hands-on approach to students with minimal prior IT knowledge can be challenging and requires a wise choice of software tools. Ideally, the software should be freely available on all major platforms, be immediately usable by novices and allow students to focus on the formal content rather than on the idiosyncrasies and interdependencies of particular editors and tools.

This software demonstration discusses the didactic concept of a new open-source application, x4ml, to be released in mid-2023, that meets these requirements and uses XML / HTML as an easily accessible paradigmatic example of a technology stack for Internet lexicography. In the demo, the use of x4ml as a teaching tool will be compared with available alternatives, e.g. using a dictionary writing system (such as Lexonomy; cf. Měchura, 2017), a professional XML editor (oXygen), or a general-purpose programming editor with customizations and add-ons (Atom).

x4ml is delivered as a single-file Java binary that can be used locally or deployed on a server. The application systematically eliminates a large number of typical beginner problems, such as accidentally not saving changes, not properly connecting XML data to a schema or stylesheet, etc. This dramatically reduces the amount of IT support required in the classroom.

The user interface concept is based on the idea of a flat workspace of files. It always displays an XML document and the contents of an "XML-processing file" side by side in two editor panes. The latter can be an XML schema (DTD or RelaxNG), a query using XPath or XQuery, or a stylesheet (XSLT). Whenever a document is modified, the user is immediately presented with updated information on well-formedness, validity, and/or query or transformation results, displayed in separate output panes below the two editor panes. A single click sequentially displays processing results for all individual XML files in the workspace, rather than just the one currently displayed. XPath/XQuery statements can also be executed on the entire collection of XML documents in the workspace, rather than on individual documents, making it easy to create headword lists and implement advanced queries on the would-be dictionary. Since XSLT is very demanding for beginners, x4ml also provides a simple, yet flexible and powerful XPath/XQuery-based HTML templating system that demonstrates the basic idea of transforming XML data into arbitrary HTML. Transformation results for all XML documents in the workspace can be viewed in a dictionary preview.

A preliminary version of x4ml was used in a week-long course on modeling and representing data in digital lexicography held in 2022 for students of the EMLex international Master's program in lexicography (see Schierholz, 2010). In the demonstration we report on the experiences made and identify areas for future improvement.

References

- Měchura, M. B. (2017). Introducing Lexonomy: an open-source dictionary writing and publishing system. In Kosem, I. 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, pp. 662-679.
- Schierholz, S. J. (2010). EMLex: Europäischer Master für Lexikographie – European Master in Lexicography. Lexicographica, 26, pp. 343-350.

EDictViz: making dictionary content accessible for people with visual impairments

Geraint Paul Rees
Universitat Rovira i Virgili
E-mail: geraintpaul.rees@urv.cat

Keywords: accessibility ; online dictionaries; visual impairment; WCAG

Research on e-lexicography has shown great concern for making the content of lexicographic resources visually accessible. This is evident from research on digital typography (e.g., Hao et al., 2022), signposts (e.g., Dziemianko, 2016), the effect of advertisements (e.g., Dziemianko, 2020) as well as many other topics. However, for the 285 million people worldwide with visual impairments (WHO, 2014), many popular online dictionaries remain largely inaccessible. A great deal of the information on dictionary websites is imperceptible for this group (Arias-Badia & Torner, forthcoming). For example, sense selection menus sometimes rely on users being able to perceive arrows, other symbols, or colours. Moreover, dictionary websites are frequently incompatible with assistive technologies such as screen reader software, which reads screen content using synthesised speech, and magnifiers (Rees, 2023). For example, the webpages are often coded in such a way that they are not read in a logical order by screen readers or contain content which frequently reloads the page meaning the screen reader must start from the beginning. Unfortunately, commercial imperatives limit the extent to which the publishers of major dictionary websites can make their pages accessible to people with visual impairments.

To solve this problem, EDictViz takes lexicographic data from monolingual English learners' dictionaries and presents it in a pared-down visual format which is accessible to people with many types of visual impairments and compatible with screen-readers, magnifiers, and other assistive technologies. This presentation will first summarise some of the research which motivated the creation of EDictViz. It will then provide a demonstration of the tool and discuss some of the problems encountered during its development. Finally, plans for further empirical testing and development will be outlined.

It is hoped that the presentation will prompt debate about how to make lexicographic resources accessible for disadvantaged users, and, more generally, about the challenges of visualizing lexical data. Such a debate would not only benefit the EDictViz project but also the e-lex community as a whole.

Acknowledgements

The author is a member of the Research Group in Language and Technologies (AGUR 2021 SGR 00151) and Serra Húnter Fellow.

References

- Arias-Badia, B., Torner, S. (forthcoming). Bridging the gap between website accessibility and lexicography: Information access in online dictionaries. Universal Access in the Information Society.

- Dziemianko, A. (2016). An insight into the visual presentation of signposts in English learners' dictionaries online. International Journal of Lexicography, 29(4), pp. 490–524.
- Dziemianko, A. (2020). Smart advertising and online dictionary usefulness. International Journal of Lexicography, 33(4), pp. 377–403.
- Hao, J., Xu, H., Hu, H. (2022). A multimodal communicative approach to the analysis of typography in online English learner's dictionaries. International Journal of Lexicography, 35(2), pp. 234–260.
- Rees, G. P. (2023). Online dictionaries and accessibility for people with visual impairments. International Journal of Lexicography, 36(2), pp. 107–132.
- WHO. (2014). WHO | 10 facts about blindness and visual impairment. URL: `https://web.archive.org/web/20211125163741/https://www.who.int/features/factfiles/blindness/blindness_facts/en/` (6 February 2023)

Sketch Engine pre-processing pipelines: towards on-the-fly tokenization of user queries

Matúš Kostka, Marek Medveď
Lexical Computing
E-mail: 515157@mail.muni.cz, marek.medved@sketchengine.eu

Keywords: Sketch Engine; automatic text pre-preprocessing; pre-processing pipelines; pipeline performance

Sketch Engine is a leading corpus management system with corpora for over 100 languages. When building a corpus, the source texts are always to be processed by a group of linguistic tools making altogether a processing "pipeline" that is applied to the corpus. This pipeline usually consists of several consecutive scripts which transform the corpus content in a specific way. As an example, we introduce the main set of tools used in the majority of Sketch Engine pipelines. First, the *Uninorm* tool is applied to normalize different variants of characters (quotes, dashes, etc.) into one form, ensuring a consistent annotation of these characters throughout the whole corpus. Next, the *Unitok* tool identifies individual tokens inside the corpus plain text and the *Tag-Sentences* tool recognizes sentence boundaries and annotates them with a special <*s*> token. Next, a POS tagger is applied. It provides necessary annotation in the form of part-of-speech and lemma (base form of the word). As the last step, post-processing is applied to the final result to fix some corner cases and mistakes made by previous tools.

For each language, the Sketch Engine has a dedicated pipeline that varies in terms of its computational performance and tagging accuracy. In this paper we focus on the former aspect with the target goal being that all user queries are also automatically tokenized on-the-fly. This makes it easier for the users to formulate corpus queries without knowing the tokenization of the corpus. Even the English corpora in Sketch Engine feature some unintuitive tokenization: e.g. *don't* becomes *do* and *n't* (the latter lemmatized as *not*) so that searching for *do* retrieves both positive and negative usages and searching for *not* retrieves both the contracted and expanded form. The disadvantage of this approach is that searching for *don't* will not retrieve any results and it is up to the user to investigate further how to fix the query. For some languages (e.g. Japanese), the tokenization is principally not well defined and improving the query processing so that users do not need to match the corpus tokenization is a big advantage for the users.

For this purpose, the final pipeline needs to have a real-time performance. Our tests primarily focused on time complexity, memory usage, and CPU usage on the 49 most used pipelines inside the Sketch Engine. These three parameters significantly influence user experience and system usability. The early results show that mainstream languages usually come with state-of-the-art tools tuned for fast processing and tested on large inputs. At the same time, several different tools are available for a specific task. On the other hand, for less-resourced languages, there is usually only a single tool available that may have difficulties handling borderline cases, resulting in frequent errors or overall system slowdown. The final tests highlight these problematic pipelines and will be adjusted by our team to improve the user experience and deploying on-the-fly tokenization of user queries.

Meanma – an end-to-end, corpus-to-entry solution for historical lexicography

Mark McConville, Stephen Barrett
Glasgow University
E-mail: mark.mcconville@glasgow.ac.uk, stephen.barrett@glasgow.ac.uk

Keywords: dictionary writing software; corpus excerpting; historical lexicography; Scottish Gaelic

Meanma is a proof-of-concept system for historical lexicography, developed since 2019 for use by Faclair na Gàidhlig, the inter-university project to create a comprehensive dictionary of Scottish Gaelic on historical principles. Meanma has been explicitly designed as an end-to-end system for creating dictionary entries, from corpus excerpting through slip/citation management and lexical sense analysis, to editing and publishing dictionary entries in multiple formats. The decision to create Meanma was preceded by an investigation of existing corpus excerpting and dictionary writing softwares (CQPWeb, SketchEngine, iLEX, IDM-DPS), none of which were found to be particularly well-suited to the needs of historical lexicographers.

The principle data source for Meanma is Corpas na Gàidhlig, a 30-million-word, full-text corpus of Scottish Gaelic printed works, manuscripts and vernacular audio transcriptions – tokenised, lemmatised and part-of-speech tagged using light-touch TEI. These texts are linked to a complex metadatabase incorporating bibliographic, biographic and sociolinguistic information. In addition, texts are linked to high-resolution page scans so that lexicographers can check for transcription and annotation errors more easily.

Meanma itself consists of the following modules:

1. The excerpting system provides a search interface for Corpas na Gàidhlig allowing for both simple headword and wordform searches, as well as a range of more advanced 'restricted' searches involving time periods, dialect areas, genres/registers, part-of-speech tags, including the ability to use XPath to query the document XML structure directly. Excerpting results can be displayed in the standard concordance format (KWIC), ordered by date or randomly, as well as by wordform in a 'dictionary view' format.

2. Lexicographers can 'save' selected search results as an electronic equivalent of traditional lexicographic 'slips', and then annotate and manage these slips in different ways. Slips can be associated with a range of different citation styles (more or less concise) and translations (more or less polished), and citations can be edited in various ways using ellipsis and editorial insertions. The lexicographer can also add detailed morphosyntactic annotation. Slips can then be organised into virtual 'piles' according to any classification criteria the lexicographer thinks appropriate, as well as being exported as printable PDFs.

3. Dictionary entries are automatically created from annotated slips, and can similarly be manipulated and viewed in different ways. In particular, an entry can be associated with a complex, hierarchical structure of subsenses, each of which is exemplified by saved citations. This 'sense tree' can then be reorganised by the lexicographer to produce a finalised sense analysis of a lexeme.

Meanma is also organised into discrete, encapsulated ‘workspaces’, allowing the same underlying corpus to be used by multiple dictionary projects, and also allowing the potential for ‘sandbox’ functionality for training new lexicographers. Current priorities involve developing the proof-of-concept into a fully featured beta release by summer 2025, by refactoring the code using the CodeIgniter PHP framework, and by upscaling the database technology to better cope with a corpus of 30 million words.

Trawling the corpus for the overlooked lemmas

Nathalie Hau Sørensen, Nicolai Hartvig Sørensen, Kirsten Lundholm Appel, Sanni Nimb

Society for Danish Language and Literature
E-mail: nats@dsl.dk, nhs@dsl.dk, ka@dsl.dk, sn@dsl.dk

Keywords: Lemma selection; neologism detection; compound splitting; Named Entity recognition; word2vec

Lemma selection is a significant part of lexicographic work, also in the case of the Danish Dictionary (DDO), a corpus based online monolingual dictionary covering today more than 100,000 lemmas. The dictionary is continuously updated with new words since the first printed edition in 2003-2005, and the lexicographers base the lemma selection on well developed statistical corpus methods. However, we are aware that some lemmas are still being overlooked, namely the lemmas that are relatively low-frequent in the corpus. However, research on the use of DDO shows that low-frequency words are in fact queried by users (Trap-Jensen, Lorentzen, & Sørensen, 2014). When using the dictionary for sense annotation of 2,000 sentences in the Danish DA-ELEXIS corpus (Federico et al., 2021; Pedersen & et al., in review), we also found a substantial number of lemma candidates among the words that only occur once, and which have not yet been registered as candidates by our existing methods. Although important, it is a time-consuming process to manually sort the relevant lemma candidates among the many low-frequency words. Previous studies have examined how to solve the challenge automatically (Kerremans, Stegmayr, & Schmid, 2012; Falk, Bernhard, & Gérard, 2014), however only with a limited success for Danish (Halskov & Jarvad, 2010). In this work, we present an automatic method to detect good candidates from a corpus divided into years, in our case 2005 and onwards. From the corpus, we collect a list of potential candidates being word forms that either suddenly appear or which frequency increases during the time frame. Our aim is then to identify the most promising of the potential lemma candidates from the list. We do so by investigating the semantically similar words to the potential lemma using a semantic model: If these are already included in the dictionary, this might indicate that the potential lemma is a good candidate. For instance, the most similar words to the potential lemma *operakoncert* 'opera concert' is *nytårskoncert* 'new year concert' and *sommerkoncert* 'summer concert', both of which are already included in DDO. To address the named entities, we retrieve corpus examples for the potential lemma and use the Named Entity Recognition component of the NLP framework DaCy (Enevoldsen, Hansen, & Nielbo, 2021) to determine whether the potential lemma is eg. a PERSON or ORGANISATION in the examples, in which case the potential lemma is probably not suitable for inclusion in DDO. Lastly, as Danish is a highly compounding language, we use a compound splitter and investigate all of the resulting subtokens. If the subtokens are included in the dictionary, or are themselves suitable lemma candidates according to the steps above, then the potential lemma in question is probably also suitable for inclusion in DDO. The automatic lemma candidate identification method will be evaluated by comparing the final suggestions for lemma candidates with a similar list collected manually. This will give us an idea of whether our proposed methods are useful to increase the future coverage of DDO with the previously overlooked lemmas.

References

- Enevoldsen, K., Hansen, L., Nielbo, K. (2021). Dacy: A unified framework for danish nlp. Ceur Workshop Proceedings, 2989, 206–216.
- Falk, I., Bernhard, D., Gérard, C. (2014). From non word to new word: Automatically identifying neologisms in french newspapers. In Lrec-the 9th edition of the language resources and evaluation conference.
- Martelli, F., Navigli, R., Krek, S., Kallas, J., Gantar, P., Koeva, S., Nimb, S., Sand ford Pedersen, B., Olsen, S., Langemets, M., Koppel, K., Üksik, T., Dobrovoljc, K., Ureña-Ruiz, R., Sancho-Sánchez, J.L., Lipp, V., Váradi, T., Győrffy, A., Simon, L., Quochi, V., Monachini, M., Frontini, F., Tiberius, C., Tempelaars, R., Costa, R., Sal gado, A., Čibej, J., Munda, T., Kosem, I., Roblek, R., Kamenšek, U., Zaranšek, P., Zgaga, K., Ponikvar, P., Terčon, L., Jensen, J., Flörke, I., Lorentzen, H., Troelsgård, T., Blagoeva, D., Hristov, D., Kolkovska, S. (2023). Parallel sense-annotated corpus ELEXIS-WSD 1.1. URL: `http://hdl.handle.net/11356/1842`. Slovenian language resource repository CLARIN.SI
- Halskov, J., Jarvad, P. (2010). Manuel og maskinel excerpering af neologismer. NyS, Nydanske Sprogstudier(38), 39–68.
- Kerremans, D., Stegmayr, S., Schmid, H.-J. (2012). The neocrawler: Identifying and retrieving neologisms from the internet and monitoring ongoing change. Current methods in historical semantics, 73, 59.
- Pedersen, B. S., Nimb, S., Olsen, S., Troelsgård, T., Flörke, I., Jensen, J., Lorentzen, H. (2023, May). The DA-ELEXIS Corpus-a Sense-Annotated Corpus for Danish with Parallel Annotations for Nine European Languages. In Proceedings of the Second Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2023) (pp. 11-18).
- DDO=Den Danske Ordbog ('the danish dictionary'). (n.d.). Det Danske Sprog- og Litteraturselskab (Society for Danish Language and Literature). URL: `http://ordnet.dk/ddo`.
- Trap-Jensen, L., Lorentzen, H., Sørensen, N. H. (2014). An odd couple–corpus frequency and look-up frequency: what relationship? Slovenščina 2.0: empirical, applied and interdisciplinary research, 2(2), 94–113.

Tēzaurs.lv - the experience of building a multifunctional lexical resource

Mikus Grasmanis, Pēteris Paikens, Lauma Pretkalniņa, Laura Rituma, Laine Strankale, Artūrs Znotiņš, Normunds Grūzītis

University of Latvia IMCS

E-mail: mikus.grasmanis@lumii.lv, peterisp@gmail.com, lauma.pretkalnina@lumii.lv, laura.rituma@lumii.lv, laine.strankale@lumii.lv, arturs.znotins@leta.lv, normunds.gruzitis@leta.lv

Keywords: dictionary; lexicographic tools; infrastructure

Tēzaurs.lv is the largest Latvian electronic dictionary with more than 388,000 entries, which was started as a compilation from approximately 300 dictionaries (Spektors et. al, 2016) and other sources, and has recently been extended and developed with the addition of Latvian WordNet (Paikens et al., 2022) data (6,610 synonym sets) and 75,400 manually curated corpus examples for specific senses.

Since the last progress report on Tēzaurs.lv (Paikens et al., 2019) it has seen a significant shift in its focus and features, transforming from a traditional explanatory dictionary towards a ”3-in-1” lexical resource that augments the senses and their explanations with WordNet style (Fellbaum, 1998) links effectively making a synonym dictionary and also a translation dictionary, showing translation equivalents on a sense level. Each entry can contain multiple lexemes, including spelling variants and derivations, and also inflectional and grammatical information for them. Senses are organised in two levels - top level senses and subsenses, and each can have corpus examples attached. Both lexemes and senses can have additional data about language style, usage, domain, etc. Entries can also contain unstructured information about etymology and normative commentary.

For these new needs we have developed a lexical database system and an editor toolkit, which is also used for two other Latvian dictionaries - LLVV1 and MLVV2. The platform is based on PostgreSQL, node.js and Vue.js.

While previously the data model and tools were based on what the end user would see in a dictionary entry, the current infrastructure is designed with a focus on a maintainable structured lexical model (shown in figure 1), avoiding duplication and enabling persistent links that stay consistent even if word senses are edited or moved. For example, multi-word entities used to be listed separately in the words referring to it, duplicating the data with some accidental variation, but now both entries include the same entity. This highly structured approach simplifies exporting data for various purposes. Currently, we have TEI3 for most dictionary data and LMF4 for WordNet related data.

For dictionary end users we now provide search results based on inflectional forms and spelling variants, as well as links to phonetically similar, alphabetically adjacent or semantically linked words. For the entries of Latvian WordNet we also provide translations allocated to specific senses, which helps language learners and translators.

For lexicographers we have built a responsive online system that allows collaborative editing of the dictionary, tracking authorship of changes and version history. Where possible, editor tools allow selecting attribute values from pre-filled drop-down menus to ensure data consistency. The platform also integrates multiple external data sources - evidence from corpora (Saulite et al., 2022), searching of Princeton WordNet (Fellbaum, 1998), etc.

An advantage of this approach is its flexibility to extend it to other lexicographic data such as etymology and derivations, including them as additional data in a shared lexical resource,

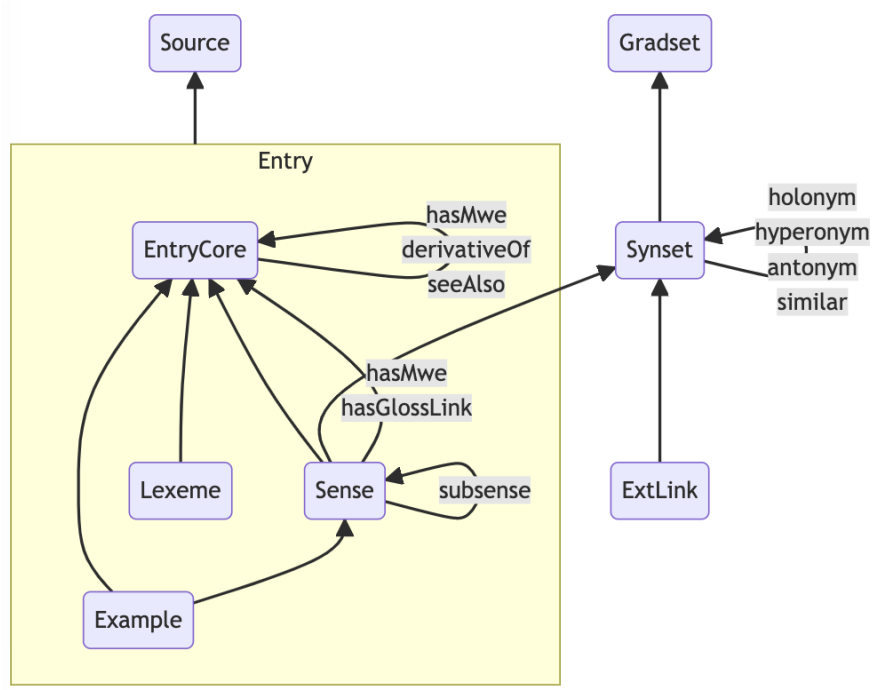


Figure 1: Conceptual data model

instead of creating a separate resource like Derinet (Vidra, 2019) which afterwards can diverge from the continuously maintained dictionary. We hope that this experience will be useful for other researchers building lexical resources and tools for maintaining them.

References

- Spektors, A., Auzina, I., Dargis, R., Gruzitis, N., Paikens, P., Pretkalnina, L., Rituma, L., Saulite, B. (2016). Tēzaurs.lv: the Largest Open Lexical Database for Latvian.

- Paikens, P., Grasmanis, M., Klints, A., Lokmane, I., Pretkalnina, L., Rituma, L., Stade, M., Strankale, L. (2022). Towards latvian wordnet. In Proceedings of the 13th Language Resources and Evaluation Conference (LREC), pages 2808–2815.

- Paikens, P., Gruzitis, N., Rituma, L., Nespore, G., Lipskis, V., Pretkalnina, L., Spektors, A. (2019). Enriching an explanatory dictionary with framenet and propbank corpus examples. In Proceedings of the 6th Biennial Conference on Electronic Lexicography (eLex), pages 922–933.

- Fellbaum, Ch. (1998). editor. WordNet: An electronic lexical database. MIT Press.

- Saulite, B. et al. (2022). Latvian national corpora collection – korpuss.lv. In Proceedings of the 13th Language Resources and Evaluation Conference (LREC), pages 5123–5129.

- Vidra, J., Žabokrtský, Z., Ševčíková, M., Kyjánek, L. (2019). DeriNet 2.0: Towards an all-in-one word-formation resource. In Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology, pages 81–89, Prague, Czechia, September 2019. Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics.

The novel Slovene COVID-19 vocabulary and its analysis from the perspective of naming possibilities and word formation

Senja Pollak[1], Ines Voršič[4], Boris Kern[2], Matej Ulčar[3]
[1]Jožef Stefan Institute, [2]ZRC SAZU, Fran Ramovš Institute of the Slovenian Language and University of Nova Gorica, [3]University of Ljubljana, Faculty of Computer science and informatics, [4]University of Maribor, Faculty of Education and Faculty of Arts
E-mail: senja.pollak@ijs.si, ines.vorsic@um.si, boris.kern@gmail.com, matej.ulcar@protonmail.com

Keywords: Covid-19, word formation, embeddings

The COVID-19 pandemic has fundamentally changed our reality and with it our linguistic reality. In our paper, we extract and analyse a sample of the novel Slovene vocabulary related to COVID-19, focusing on naming possibilities and word formation processes.

Our methodology consists of the following steps. First, we train a fastText word embeddings model (Bojanowski et al., 2017) on a Slovene corpus of 144,352 news articles about Covid-19. Next, we select the Covid-19 related input words to be used for the embeddings-based expansion. First, we use the list of Covid-19 vocabulary from The Growing Dictionary of the Slovenian Language (ed. Krvina 2014-). Next, we use the Covid-19 vocabulary from the CJVT Language Monitor (Kosem et al., 2021), and third, the list of Covid-19 vocabulary of occasional words collected by Voršič (2022). The resulting joint list contains 186 unique keywords (or key phrases) for the embeddings-based expansion. For each word (or multi-word expression) in the seed vocabulary, we extract its 200 nearest neighbours from the fastText embeddings model. Then we filter the extracted candidates by removing those that do not contain any letter of the Slovene or English alphabet, as well as all words that are already included in the Slovene lexicon Sloleks (Dobrovoljc et al. 2019), as we are interested only in the novel vocabulary. We also use Levenshtein-distance-based filtering to avoid extracting too similar words. At the end, we keep the 50 most related neighbours for each seed word and group them in a joint list by removing duplicates.

In total, 4947 lemmas were extracted. For this abstract, we analysed 843 lemmas that occur at least 5 times in our corpus and are related to Covid-19. As a result, 66 lemmas were identified as relevant results.

The analysis of naming possibilities included 149 lexemes. In addition to the 66 lexemes resulting from our embedding-based expansion process, 29 lexemes were added from The Growing Dictionary of the Slovenian Language (Krvina, 2014–) and 54 from the COVID-19 vocabulary of occasional words collected by Voršič (2022). The results show that 85.9% of the lexemes were created by word-formation processes. Among the naming possibilities derived from Slovenian, set phrases follow at 6% and neosemantisms account for 1.3%. Explicit borrowings, on the other hand, account for 6.7%.

A more detailed word formation analysis included a total of 77 relevant monosemous lexemes. 62 examples from our embedding-based extension method (out of the total list of 66 lexemes, four instances were not kept for analysis due to the fact that they were explicit borrowings from English and were not formed using word-formation processes in Slovenian) and 15 neologisms.

The analysis shows that the most frequent are systemic formations, out of which the most productive are interfixal compounds (41.56%), followed by ordinary derivatives by suffixation (41.56) and ordinary derivatives by prefixation (15.58), modificational derivatives by

suffixation (3.90%), derivatives from a prepositional phrase (2.60%), coordinate interfixal-suffixal compounds (2.60%), and subordinate interfixal-suffixal compounds (1.30%). Compared to the systemic formations, the percentage of systemically unpredictable formations is much lower, with abbreviations (2.60%), blend words 1.30%), and bicapitalizations (1.30%).

The results of our study have an impact on understanding different naming possibilities and word formation processes in Slovene on the one side, and for the expansion of the current lexicographical description of Covid-19 vocabulary on the other side.

References

- Bojanowski, P., Grave, E., Joulin, A., Mikolov, T. (2017). Enriching Word Vectors with Subword Information. Transactions of the Association for Computational Linguistics 5, 135–146.
- Dobrovoljc, K., Krek, S., Holozan, P., Erjavec, T., Romih, M., Arhar Holdt, Š., Čibej, J., Krsnik, L., Robnik Šikonja, M. (2019). Morphological lexicon Sloleks 2.0, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042.
- Kosem, I. et al. (2021). Sledilnik 1.0: Language Monitor. URL: `viri.cjvt.si/sledilnik`, accessed on 31. 10. 2021.
- Krvina, D. (ed.), (2014-). Sprotni slovar slovenskega jezika [The Growing Dictionary of the Slovenian Language]. URL: `www.https://fran.si/iskanje?FilteredDictionaryIds=132&View=1&Query=%2A`.
- Voršič, I. (2022). Priložnostne tvorjenke kot odraz dobe koronavirusa. In G. Nikolovski (ed.) Slavistična prepletanja, 3, pp. 253–266. Maribor: Univerza v Mariboru, Univerzitetna založba.

Topic and Genre Classification of a Large English Web Corpus

Jan Kraus, Vít Suchomel
Lexical Computing
E-mail: jan.kraus@sketchengine.eu, vit.suchomel@sketchengine.eu

Keywords: text types; topic; genre; annotation; web corpus; meta-data

In this paper we present a semi-manual approach to annotation of 21 topics and 5 genres in texts in the new very large English web corpus from 2021, enTenTen21.

The main goal of this work with regards to lexicography is to enable identification of topic-specific and genre-specific collocations in the corpus. Now, a lexicographer can distinguish general collocations from collocations typical for particular topics or genres. For example, 'test specificity' is a general collocation while 'test sensitivity' is especially used in the domain of health and 'allegedly stole' is typically written in the news genre.

Apart from the usual supervised learning scheme, the manual annotation part of our procedure was reduced to a minimum since the same topic label and the same genre label was assigned to all web pages coming from a website. Only 3,000 websites with the most tokens in the corpus were manually inspected to achieve high efficiency while still checking random samples of 40% corpus tokens. By this approach, 16% of the corpus was covered with a topic label and 7.5% of the corpus was covered with a genre label. It has been shown that 92% of web pages rated this way share the topic of the whole website – making our method adequate.

The manual assignment of a topic and genre to a website consists of checking the hostname of the website, its landing page and most importantly reading 3 to 10 random triples of consecutive sentences in context from the corpus. More random sentences are inspected in case the basic checks reveal suspicious content, e.g. a generic or a long hostname, machine translated text or nonsense spam-like text. The annotator should be able to decide if the content shows lexical or syntactic features typical for a recognized text type. No label is given if the person is not sure. No class is given instead of assigning multiple labels to sites with many text types. Machine generated text is removed from the corpus.

To find out if the manual annotation of topics and genres according to our scheme can be carried out by inexperienced persons with only a little training, a group of four students of applied linguistics rated 2,000 websites. The label assigned by our expert with wide experience with the task matched the label given by at least three out of four students in 89% of cases for the topic and in 86% of cases for the genre. Many of the disagreements would not in fact matter, since there is a large grey zone between some categories such as Science vs. Education and both labels would be fine in the corpus for texts from the particular website. The annotation coverage of the corpus was increased by adding a topic label to 304 websites and a genre label to 338 websites in this experiment.

Automating derivational morphology for Slovenian

Tomaž Erjavec[1], Marko Pranjić[1], Andraž Pelicon[1], Boris Kern[2], Irena Stramljič Breznik[3], Senja Pollak[1]

[1]Jožef Stefan Institute, [2]ZRC SAZU, Fran Ramovš Institute of the Slovenian Language and University of Nova Gorica, [3]University of Maribor, Faculty of Arts
E-mail: tomaz.erjavec@ijs.si, marko.pranjic@ijs.si, andraz.pelicon@ijs.si, boris.kern@gmail.com, irena.stramljic@um.si, senja.pollak@ijs.si

Keywords: derivational morphology, derivational dictionary, word formation

Word formation is a branch of linguistics which helps to analyse the lexical vitality of a given language and also shows trends of language development. Slovenian is characterised by an extremely rich morphemic structure of words, a result of multistage formation: e.g. in the first stage, the adjective mlad/young yields the noun mladost/youth, which in turn yields the adjective mladosten/youthful in the second stage, which yields the noun mladostnik/a youth (adolescent), yielding the possessive adjective mladostnikov/youth's (adolescent's) in the fourth stage. While there were some linguistic descriptions of Slovenian word formation (Vidovič Muha 1988, Toporišič 2000), there is a lack of corpus-based grounding of theoretical findings. In the field of natural language processing, several researchers (Ruokolainen et al., 2013, Cotterell et al., 2019) addressed the problem of morphological analysis, but there are no advanced approaches developed for Slovenian.

The basis for our study was the digitised trail volume (letter B) of the derivational dictionary of Slovenian (Stramljič Breznik, 2004). The dictionary gathers words in word families centred around a root, and inside those presents sequences of derivations, also split into constituent morphemes and giving the part-of-speech of the source and derived words.

From this resource we derived morphological rules, which contain the derivation with the part-of-speech and constituent morphemes (e.g. VERB:X-evati → NOUN:X-anje), paired with rules describing the derivation as a minimal transformation on, and applicable to, the level of surface forms (e.g. VERB:X-ti → NOUN:X-nje for the case of bojevati/to fight → bojevanje/figthing). The rules are also gathered in sequences as presented in the dictionary (e.g. for boj/a fight → bojevati → bojevanje). We currently concentrated on suffix rules, where we derived 1,641 rules and 1,649 sequences.

We next applied the constructed rules to part-of-speech + lemmas pairs from two sources, the reference computational lexicon of Slovene Sloleks (Čibej et al., 2022) and the lexicon derived from the very large corpus MetaFida (Erjavec, 2021), with the requirement that the generated word is also present in the lexicon. We then also connected the results into sequences. With this, the initial set of morphological chains consisting only of roots starting with the letter B is extended to all other words (e.g. izklic → izklicevati → izklicevanje) and provide the resource for linguistic analysis. From Sloleks we gathered 117,769 potential pairs and 32,823 potential sequences, while from the MetaFida lexicon we get 1,549,644 potential pairs and 496,486 potential sequences.

In order to allow for more general analysis on the level of morphemes, one of our goals is to build an automated tool for morphological segmentation. For this purpose, we first built a training set based on the derivational dictionary of Slovenian (consisting of pairs of original words and its morphologically segmented counterparts, e.g. prababičin -> prab-ič-in). Next, we trained the model based on BiLSTM-CRF and achieved F1-Score

of 83.98%, which is significantly higher than the two implemented unsupervised baselines, Morfessor (Smit et al., 2014) and MorphoChain (Narasimhan et al., 2015).

References

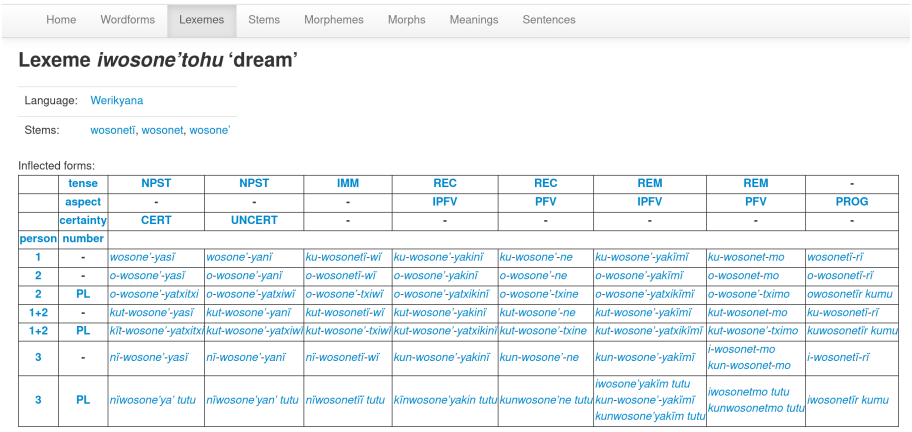
- Cotterell, R., Kumar, A., Schütze, H. (2019). Morphological segmentation inside-out. arXiv preprint arXiv:1911.04916.
- Ruokolainen, T., Kohonen, O., Virpioja, S., Kurimo, M. (2013). Supervised morphological segmentation in a low-resource learning setting using conditional random fields. In Proceedings of the Seventeenth Conference on Computational Natural Language Learning (pp. 29-37).
- Cotterell, R., Kumar, A., Schütze, H. (2019). Morphological segmentation inside-out. arXiv preprint arXiv:1911.04916.
- Čibej, Jaka; et al. (2022). Morphological lexicon Sloleks 3.0, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042.
- Erjavec, T. (2021). Corpus of combined Slovenian corpora MetaFida 0.1, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, `http://hdl.handle.net/11356/1746`.
- Narasimhan, K., Barzilay, R., Jaakkola, T. (2015). An Unsupervised Method for Uncovering Morphological Chains. Transactions of the Association for Computational Linguistics, 3, pp. 157–167.
- Ruokolainen, T., Kohonen, O., Virpioja, S., Kurimo, M. (2013). Supervised morphological segmentation in a low-resource learning setting using conditional random fields. In Proceedings of the Seventeenth Conference on Computational Natural Language Learning (pp. 29-37).
- Smit, P., Virpioja, S., Grönroos, S.A., Kurimo, M. (2014). Morfessor 2.0: Toolkit for statistical morphological segmentation. In Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics. Gothenburg, Sweden: Association for Computational Linguistics, pp. 21–24. URL `https://aclanthology.org/E14-2006`.
- Stramljič Breznik, I. (2004). Besednodružinski slovar slovenskega jezika: Poskusni zvezek za iztočnice na B. Maribor: Slavistično društvo.
- Toporišič, J. (2000). Slovenska slovnica. Maribor: Obzorja.
- Vidovič Muha, A. (1988). Slovensko skladenjsko besedotvorje ob primerih zloženk. Ljubljana: Znanstvena založba Filozofske fakultete, Partizanska knjiga.

Modeling and visualizing morphology in the CLDF/CLLD ecosystem

Florian Matter
University of Oregon
E-mail: fmatter@mailbox.org

Keywords: cldf; clld; morphology; linked data; paradigms; derivation

The CLLD (cross-linguistic linked data, Forkel et al., 2019) project has seen a massive rise in popularity; it makes it easy to create interactive data-based apps for crosslinguistic research projects. The CLDF (cross-linguistic data format, Forkel et al., 2018) is a lightweight and flexible externalization thereof, designed for sharing data. This talk will present both efforts to model morphological structure and relations in CLDF and CLLD, and feature-rich visualization methods in CLLD.



	tense	NPST	NPST	IMM	REC	REC	REM	REM	-
	aspect	-	-	-	IPFV	PFV	IPFV	PFV	PROG
	certainty	CERT	UNCERT	-	-	-	-	-	-
person	number								
1	-	*wosone'-yasī*	*wosone'-yanī*	*ku-wosonetī-wī*	*ku-wosone'-yakinī*	*ku-wosone'-ne*	*ku-wosone'-yakīmī*	*ku-wosonet-mo*	*wosonetī-rī*
2	-	*o-wosone'-yasī*	*o-wosone'-yanī*	*o-wosonetī-wī*	*o-wosone'-yakinī*	*o-wosone'-ne*	*o-wosone'-yakīmī*	*o-wosonet-mo*	*o-wosonetī-rī*
2	PL	*o-wosone'-yatxibi*	*o-wosone'-yatxiwī*	*o-wosone'-txiwī*	*o-wosone'-yatxikinī*	*o-wosone'-txine*	*o-wosone'-yatxikīmī*	*o-wosone'-tximo*	*owosonetīr kumu*
1+2	-	*kut-wosone'-yasī*	*kut-wosone'-yanī*	*kut-wosonetī-wī*	*kut-wosone'-yakinī*	*kut-wosone'-ne*	*kut-wosone'-yakīmī*	*kut-wosonet-mo*	*ku-wosonetī-rī*
1+2	PL	*kīt-wosone'-yatxitxi*	*kut-wosone'-yatxiwī*	*kut-wosone'-txiwī*	*kut-wosone'-yatxikinī*	*kut-wosone'-txine*	*kut-wosone'-yatxikīmī*	*kut-wosone'-tximo*	*kuwosonetīr kumu*
3	-	*nī-wosone'-yasī*	*nī-wosone'-yanī*	*nī-wosonetī-wī*	*kun-wosone'-yakinī*	*kun-wosone'-ne*	*kun-wosone'-yakīmī*	*i-wosonet-mo* *kun-wosonet-mo*	*i-wosonetī-rī*
3	PL	*nīwosone'ya' tutu*	*nīwosone'yan' tutu*	*nīwosonetīī tutu*	*kīnwosone'yakin tutu*	*kunwosone'ne tutu*	*iwosone'yakīm tutu* *kun-wosone'-yakīmī* *kunwosone'yakīm tutu*	*iwosonetmo tutu* *kunwosonetmo tutu*	*iwosonetīr kumu*

Figure 1:

The built-in CLDF ontology is fairly basic in terms of morphology. The presented CLDF component adds tables for storing stems, lexemes (which may have multiple stems), wordforms (which are inflected forms of stems), morphs (which may be contained in wordforms and stems), inflectional categories, inflectional values, inflections, derivational processes, and derivations. Inflections are triple links between specific morphs in wordforms, inflectional values, and stems. Derivations are links between a stem (or a root morph), another stem, and a derivational process. Optionally, a morph in the stem may be specified for representing the derivational process. Morphs are identified in wordforms and stems by an association table which includes an index, referring to the part of the morphologically parsed larger form. The same logic is used to situation stems in wordforms.

The CLDF component is accompanied by a CLLD plug-in, which largely follows the same model for morphological data. The linked data approach allows the plug-in to create highly data-rich and interactive visualizations of both inflectional and derivational morphology. It automatically generates inflectional paradigm tables for lexemes, where cells know about the morphological constituency of the wordforms they contain, and x- and y-axes have links to the inflectional categories and values they represent (Figure 1). Detail views of wordforms show their full morphological constituency, but also their stem along with their inflectional values and corresponding exponents (Figure 2). Derivational relations between

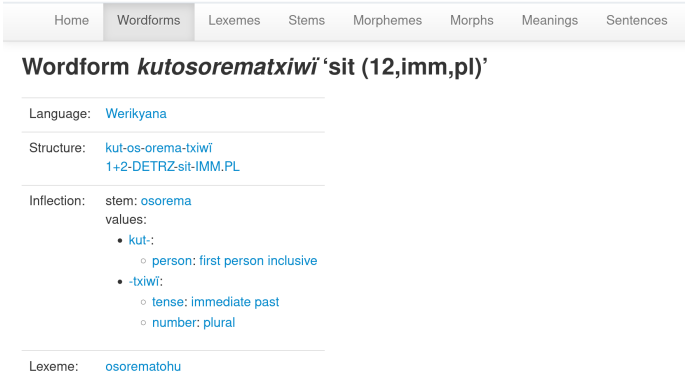


Figure 2:

stems (or roots) are visualized in tree-like fashion, where every lexicon entry shows all its "descendents" (Figure 3). This also works in reverse, where morphologically complex stems show their entire "derivational lineage", and not just the stem they are derived from (Figure 4).

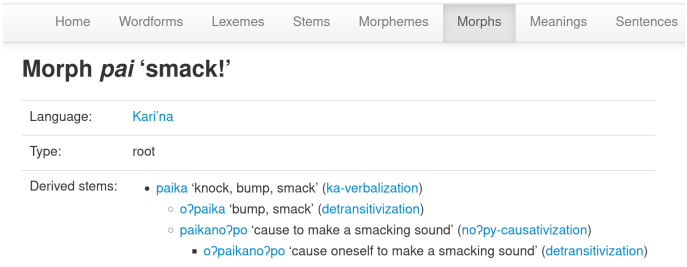


Figure 3:

This project is part of a larger effort to create a CLDF/CLLD-based workflow for digital corpus-based grammaticography, so wordforms (and morphs, and stems, and lexemes) will ultimately also have lists of corpus tokens. Features not yet implemented but planned are: other ways of stem formation (composition) and at least a basic representation of non-concatenative morphological processes.

References

- Forkel, R. et al. (2019). Clld: A Toolkit for Cross-Linguistic Databases. URL: `https://doi.org/10.5281/zenodo.3239095`.
- Forkel, R. et al. (2018). Cross-Linguistic Data Formats, Advancing Data Sharing and Re-Use in Comparative Linguistics. Scientific Data 5 (1): 180205. URL: `https://doi.org/10.1038/sdata.2018.205`.

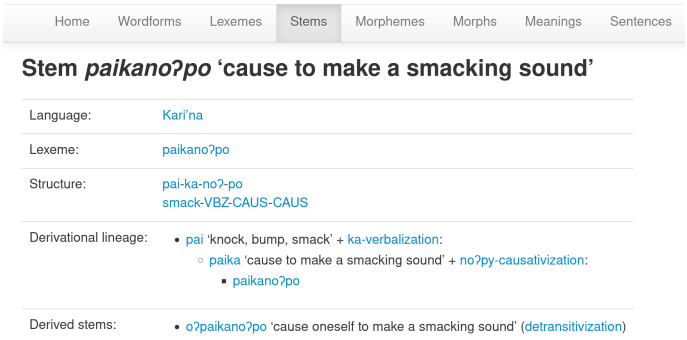


Figure 4:

DWDSmor: A toolbox for morphological analysis and generation in German, based on the DWDS lexicon and an SMOR-style grammar

Andreas Nolda

Berlin-Brandenburgische Akademie der Wissenschaften (BBAW)

E-mail: andreas.nolda@bbaw.de

Keywords: morphology; inflectional paradigms; lemmatisation; corpus annotation; online dictionaries

DWDSmor is an experimental toolbox for creating and applying a set of finite-state automata for morphological analysis and generation in German. The automata are compiled by means of Helmut Schmid's SFST compiler from an SMOR-style grammar and an SMOR-compatible lexicon (for SFST and SMOR, cf. Schmid, 2006; Schmid et al., 2004). The lexicon in turn is derived at build time from articles of the German online dictionary "Digitales Wörterbuch der deutschen Sprache" (DWDS, `https://www.dwds.de`; cf. Klein and Geyken 2010). To this aim, DWDSmor provides XSLT stylesheets mapping linguistic information from XML sources of DWDS articles to lexicon entries in SMOR format. SMOR inflection classes are deduced from grammatical information in DWDS articles (in particular, from nominative singular, genitive singular, and nominative plural *Eckforms*). For special cases (like the morphophonologically conditioned dative plural *-n*), also phonetic information in DWDS articles is taken into account. Compounding stemforms with or without linking element – which are not explicitly specified in the DWDS – are inferred from DWDS articles for compounds and their links to articles for the corresponding compound bases. Last but not least, the stylesheets make use of information on orthographic spelling variants and etymological origin in DWDS articles. The morphological grammar of DWDSmor is based on the SMORLemma implementation by Rico Sennrich (`https://github.com/rsennrich/SMORLemma`, 'lemmatiser' branch; cf. Sennrich and Kunz 2014). Like the latter, it provides a surface-level lemmatisation with morphological boundaries. Optionally, DWDS homonym and paradigm indices can be included into the analysis strings. For word-formation analysis, the grammar defines an alternative output format with word-formation bases, processes, and means in terms of the Pattern-and-Restriction Theory of word formation (cf. Nolda 2018). Once development is finished, DWDSmor shall be used for the lemmatisation and morphological annotation of DWDS corpora (cf. Geyken et al., 2017). This can either be done with standard SFST tools

or with a Python script which analyses standard input by means of a DWDSmor automaton and parses the analysis strings into lemmata and morphosyntactic categories. In addition, DWDSmor will be used for linking inflectional paradigms from DWDS articles. For this purpose, DWDSmor provides another Python script generating a set of paradigms for a given lemma (one per part-of-speech and lexical gender, optionally further partitioned by DWDS homonym and paradigm indices). Since DWDSmor is built on top of the DWDS, both corpus annotation and paradigm generation can benefit from ongoing lexicographic updates of the DWDS dictionary.

References

- Geyken, A. et al. (2017). Die Korpusplattform des „Digitalen Wörterbuchs der deutschen Sprache“ (DWDS). Zeitschrift für germanistische Linguistik 45, 327 – 344.
- Klein, W., Geyken, A. (2010). Das Digitale Wörterbuch der Deutschen Sprache (DWDS). Lexicographica 26, 79–96.
- Nolda, A. (2018). Explaining linguistic facts in a realist theory of word formation. In Essays on Linguistic Realism, ed. by Christina Behme and Martin Neef, Studies in Language Companion Series 196, Amsterdam: Benjamins, 203–233.
- Schmid, H. (2006). A programming language for finite state transducers. In Finite-State Methods and Natural Language Processing: 5th International Workshop, FSM-NLP 2005, Helsinki, Finland, September 1–2, 2005, ed. by Anssi Yli-Jyrä, Lauri Karttunen, and Juhani Karhumäki, Lecture Notes in Artificial Intelligence 4002, Berlin: Springer, 1263–1266.
- Schmid, H., Fitschen, A., Heid, U. (2004). SMOR: A German computational morphology covering derivation, composition, and inflection. In LREC 2004: Fourth International Conference on Language Resources and Evaluation, ed. by Maria T. Lino et al., European Language Resources Association, 1263–1266.
- Sennrich, R., Kunz, B. (2014). Zmorge: A German morphological lexicon extracted from Wiktionary. In LREC 2014: Ninth International Conference on Language Resources and Evaluation, ed. by Nicoletta Calzolari et al., European Language Resources Association, 1063–1067.

Using lexicography for learning mathematics

Theresa Kruse, Ulrich Heid, Boris Girnat

University of Hildesheim
E-mail: kruset@uni-hildesheim.de

Keywords: terminology; learning; mathematics

In mathematics, students struggle especially with the transition from school to university (Geisler and Rolka, 2021). We plan to investigate how lexicography and e-dictionary construction can help in this transition. In the proposed contribution, we present the concept for a seminar that uses lexicographic methods in first-year courses in mathematics.

About twenty years ago, Cubillo (2002) already used lexicography with chemistry students. They had the task to develop their own (printed) dictionaries without having received further instruction in lexicography. Since then, electronic dictionaries took root and almost replaced printed dictionaries in several fields (Fuertes-Olivera, 2016).

In introductory university courses for mathematics, students have to learn concepts, the relations between them, and typical phrases of the field. At school, however, mathematics tends to be presented as an ensemble of calculations rather than concepts. Thus, students have to learn that mathematics is basically a building constructed of definitions, theorems, and relations between them.

We plan to give the students access to and basic lexicographic training with a dictionary writing system (DWS) which is optimized for the construction of specialized dictionaries, in particular for mathematics. Therein, they can note the concepts they have learned and indicate the semantic relations between these concepts.

The relations between the concepts depend on their type. We distinguish the following concept types: Objects (e.g. *sets*, *mappings*), Properties (e.g.*complete*, *symmetric*), Theorems (e.g. *Fundamental theorem of algebra*) and Methods (e.g. *mathematical induction*, *Dijkstra's algorithm*). Between these types of concepts, different semantic relations exist, i.e. hyperonymy between different objects, theorems corresponding to different objects and properties, methods helping to determine objects with certain properties, or properties of certain (classes) of objects. In addition to these conceptual categories, there is a category of domain-specific phraseology (e.g. *if and only if*, *q.e.d.*, *corollary*).

When building their personal e-dictionaries during the course, we introduce the students to a routine for including new terms: (1) Collect the new terminology and phraseology, (2) choose a type for each term, (3) find relations between the new concepts, (4) connect the new terms to the ones already learned.

For discussing the relations between the terms, the described individual student work is accompanied by a seminar in which the students can discuss difficulties in assigning the concepts to the types and in establishing the relations. Concurrently, the lexicographic structuring of the data helps the students to gain a deeper understanding of mathematics which in turn supports the acquisition of the content as it addresses the constructivist dimension of learning (Girnat and Hascher, 2021).

Our contribution will present the concept of the lexicographic resource as well as the structure of the seminar in more detail.

References

- Cubillo, M. C. C. (2002). Dictionary Use and Dictionary Needs of ESP Students: An Experimental Approach. International Journal of Lexicography, 15(3):206–228.
- Fuertes-Olivera, P. A. (2016). A Cambrian Explosion in Lexicography: Some Reflections for Designing and Constructing Specialised Online Dictionaries. International Journal of Lexicography, 29(2):226–247.
- Geisler, S. and Rolka, K. (2021). "That Wasn't the Math I Wanted to do!"—Students' Beliefs During the Transition from School to University Mathematics. International Journal of Science and Mathematics Education, 19(3):599–618.
- Girnat, B. and Hascher, T. (2021). Beliefs von Schweizer Schülerinnen und Schülern zum konstruktivistischen und instruktivistischen Lernen im Mathematikunterricht der Sekundarstufe I – Ergebnisse eines Large-Scale-Assessments zur Überprüfung mathematischer Grundkompetenzen (ÜGK) 2016. Unterrichtswissenschaft, 49(4):525–546.

Going Beyond Standard Ukrainian: How a Corpus Informs an E-Dictionary

Maria Shvedova[1], Vasyl Starko[2], Andriy Rysin
[1]University of Jena; National University "Lviv Polytechnic", Ukraine, [2]Ukrainian Catholic University
E-mail: corpus.textiv@gmail.com, vstarko@gmail.com, arysin@gmail.com

Keywords: Ukrainian language; morphological dictionary; standard language; grammatical variants; corpus

Existing academic morphological dictionaries do not cover all the grammatical forms and variants that actually occur in the texts in Modern Ukrainian. VESUM, a morphological dictionary of the Ukrainian language designed for NLP tasks, was originally created to analyze only standard texts. However, the dictionary's functionality was significantly expanded when it was used to lemmatize a large reference corpus of the Ukrainian language (GRAC), which covers the entire period of the history of the modern Ukrainian language since early 19th century and contains texts of various styles and genres. Based on the corpus, not only many new lemmas were added to the dictionary, but also some variant grammatical forms, such as the variant forms of the accusative singular of second-declension nouns (взяти ножа, написати листа alongside with the standard ones взяти ніж, написати лист), the most frequent long forms of adjectives (гарная, гарнеє, гарнії, cf. standard гарна, гарне, гарні), short comparative forms of adverbs (гарніш, сильніш, cf. standard next гарніше, сильніше), the most frequent short forms of verbs in 3rd person singular (зна, співа cf. standard знає, співає), infinitive forms with -ть (писать, допомагать, cf. standard писати, допомагати), gerunds in -ся (стріляючися, миючися next to the standard ones стріляючись, миючись), and the most frequent imperative forms with -те (окропіте, хваліте next to the standard ones окропіть, хваліть). On the basis of the corpus, the list of words that are not recommended by the modern literary norm or not recommended as the main option, but are common in texts was expanded within the morphological dictionary. These include verbs with the prefix од- (e. g. одповісти, одчинити, одібрати; variants with the prefix від- are predominant in Modern Standard Ukrainian), active participles of the present tense with the suffixes -уч-/-юч-, -ач-/-яч- (e. g. існуючий, діючий; stigmatized in the standard language according to many normative sources). The paper discusses which periods and types of texts are typical for the use of such non-standard and variant forms, and specifies the principles of adding them to the dictionary.

From experiments to an application – the first prototype of an adjective detector for Estonian

Geda Paulsen[1,2], Ene Vainik[2], Maria Tuulik[2], Ahti Lohk[3]
[1]Uppsala university, [2]Institute of the Estonian Language, [3]Tallinn University of Technology
E-mail: geda.paulsen@eki.ee, ene.vainik@eki.ee, maria.tuulik@eki.ee, ahti.lohk@taltech.ee

Keywords: language technology; lexicography; corpus linguistics; adjective; the Estonian language

Adjectives are one of the most ambiguous word classes when it comes to defining its categorial core as PoS in Estonian lexicography (Paulsen et al. 2019: 188–189). A striking issue is for instance the problem of lexicalising participles, a phenomenon common for other languages as well (e.g. in English, certain participles tend to develop into full-blooded adjectives such as blessed or hammered). In Estonian, there are five main word classes overlapping with adjectives: nouns (via transpositional derivation forming systematic polysemy networks, see Vare 2006, Langemets 2010), verbs (mainly participles), adverbs, pronouns, and ordinals. The noun-adjective type is the largest group showing ambiguity in word class in our database over ambiforms, compiled mainly from lexicographic sources. (see Vainik et al. 2020: 122.)

In this study, we discuss the process of developing a multi-parameter application – the calculator of adjective-like corpus behaviour, available at `https://adjcalculator.pythonanywhere.com/`. The tool is based on a statistical roundup of a word (form)s corpus behaviour compared to the most typical and central aspects of the Estonian adjective. To establish the adjectival corpus profile, we use the most typical and central morphosyntactic patterns characterising adjectives and detectable in the corpus (see the experiments in Tuulik et al. 2022, Paulsen et al. 2022, Vainik et al. 2023).

The focus of the presentation is on two main optimisation issues connected to the application elaboration: 1) the scope of the overlapping PoS targeted by the calculator, and 2) the selection of the statistical method calculating the similarity of a word with the prototypical adjective. The first issue entails considerations about the scope of the application as limited by the transition zone between verbal participles and adjectives or including other PoS as well. The constituency of the set of automatically searchable test patterns will depend on this solution as well as on the criterion of optimality. The second issue involves decisions about the previously tested methods we have used for calculating the distance of a word to the adjectival profile (see Tuulik et al. 2022, Paulsen et al. 2022, Vainik et al. 2023).

The resulting product is applicable as a tool for lexicographers. It can also be adapted to other languages.

References

- Langemets, M. (2010). Nimisõna süstemaatiline polüseemia eesti keeles ja selle esitus keelevaras [Systematic polysemy of nouns in Estonian and its lexicographic treatment in Estonian language resources]. PhD thesis. Tallinn: Eesti Keele Sihtasutus.
- Paulsen, G., Vainik, E., Tuulik, M., Lohk, A. (2019). The Lexicographer's Voice: Word Classes in the Digital Era. In I. Kosem, T. Zingano Kuhn., M. Correia, J. P. Ferreria, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek & C. Tiberius (Eds.).

Proceedings of the eLex 2019 conference: Smart lexicography, 1–3 October 2019, Sintra, Portugal (pp. 319–337). Brno: Lexical Computing CZ, s.r.o. URL: https://elex.link/elex2019/wp-con-tent/uploads/2019/09/eLexi_2019_18.pdf.

- Paulsen, G.; Tuulik, M.; Lohk, A., Vainik, E. (2022). From verbal to adjectival. Evaluating the lexicalization of participles in an Estonian corpus. Slovenščina 2.0, 10/1, pp. 65–97.
- Tuulik, M., Vainik, E., Paulsen, G., Lohk, A. (2022). Kuidas ära tunda adjektiivi? Korpuskäitumise mustrite analüüs [How to recognize adjectives? An analysis of corpus patterns]. Estonian Papers in Applied Linguistics, 18, pp. 279–302. URL: http://dx.doi.org/10.5128/ERYa18.16.
- Vainik, E., Paulsen, G., Lohk, A. (2021). A typology of lexical ambiforms in Estonian. In Z. Gavriilidou, M. Mitsiaki & A. Fliatouras (Eds.), Proceedings of XIX EURALEX Congress: Lexicography for Inclusion, 7–11 September 2021, Alexandroupolis, Greece, Vol. 1 (pp. 119–130). Alexandroupolis, Greece: Democritus University of Thrace. URL: https://euralex2020.gr/wp-content/uploads/2020/11/EURALEX2020_Proceed-ingsBook-p119-130.pdf.
- Vainik, E., Paulsen, G., Tuulik, M., Lohk, A. (2023). Towards the Morphosyntactic Corpus Profile of Prototypical Adjectives in Estonian. Estonian Papers in Applied Linguistics, 19, pp. 225–244. URL: http://dx.doi.org/10.5128/ERYa19.13.
- Vare, S. (2006). Adjektiivide substantivatsioonist ühe tähendusrühma näitel. [On substantivisation of adjectives: Analysing a semantic group] E. Niit. Keele ehe. Tartu: Tartu Ülikool. Tartu Ülikooli eesti keele õppetooli toimetised; 30, pp. 205–222.

Collocations Dictionary of Modern Slovene 2.0

Iztok Kosem[1,3], Špela Arhar Holdt[1,2], Polona Gantar[1], Simon Krek[1,3]
[1]Faculty of Arts, University of Ljubljana, [2]Faculty of Computer and Information Science, University of Ljubljana, [3]Jozef Stefan Institute
E-mail: iztok.kosem@cjvt.si, arharhs@ff.uni-lj.si, apolonija.gantar@ff.uni-lj.si, simon.krek@guest.arnes.si

Keywords: collocations dictionary; responsive dictionary; crowdsourcing; examples; post-editing lexicography

In 2018, the first version of the Collocations Dictionary of Modern Slovene was published. The dictionary contained automatically extracted collocations, and their examples, using (at that point) state-of-the-art tools such as Sketch Grammar and GDEX, customised for Slovene (Gantar et al. 2016). A selection of entries was provided in a finalized form, using post-editing methodology. Over the past four years, a great deal of research related to the Collocations dictionary and the phenomenon of collocations in Slovene has been conducted, from the analysis and improvement of automatic extraction methods, lexicographic workflow, and data modelling, to user experience and participation. A project funded by the Ministry of Culture provided the opportunity to implement improved methods and new solutions into the next version of the Collocations Dictionary. In this paper, we present the methodology of compiling the second version of the Collocations Dictionary, focusing on the main novelties. One of the main differences from the first version is the method of automatic extraction, of both collocations and examples. Collocations are entirely new, i.e. they were extracted from parsed corpus data, as opposed to an extraction based on POS-tagged data which was used for the first version. A significant challenge proved to be matching the already identified collocations found in the digital dictionary database with newly extracted ones, which had to be done to prevent duplication. Among other things, this also included analysing compounds, which may have received a status of a compound in a bilingual dictionary, but were considered as legitimate collocations in a collocations dictionary. Furthermore, we decided to include significantly more syntactic structures in the second version (e.g. the first version did not include 'subject + verb' due to many bad collocation candidates); however, this meant reducing/limiting the number of collocations per structure. Importantly, the maximum number of collocations per structure varies; for example, more collocations are offered for the structures that proved more collocationally-productive in research studies (e.g. verb + noun in the accusative, adjective + noun, noun + noun in the genitive). Another more significant change, which is related to the user experience, is the enhancement of the crowdsourcing aspect of the dictionary. In the first version of the dictionary, the only crowdsourcing feature was the option to mark collocations as good or bad (using upvote and downvote) on the page of each structure. The feature was rarely used, and as shown by research, such a task is far too demanding for an average user. In the second version, we opted to introduce crowdsourcing at an example level - the users can now not only confirm the validity of the collocation in each example provided but also select the relevant sense (if sense division for a particular headword has already been made). In addition to giving a short demo of the online dictionary, we will briefly present future plans, including grouping the collocations by broader semantic groups and semantic types, improving the automatic extraction of collocations, and adding extended collocations to the entries.

Edition with Code. Towards Quantitative Analysis of Medieval Lexicography

Renaud Alexandre[1], Krzysztof Nowak[2], Iwona Krawczyk[2], Bruno Bon[1]
[1]Institut de la recherche et d'histoires des textes, [2]Institute of Polish Language, Polish Academy of Sciences
E-mail: renaud.alexandre@irht.cnrs.fr, krzysztof.nowak@ijp.pan.pl, iwona.krawczyk@ijp.pan.pl, bruno.bon@irht.cnrs.fr

Keywords: historical lexicography; digital edition; quantitative analysis

The Vocabularium Bruxellense is a little-known example of medieval Latin lexicography (Weijers 1989). It has survived in a single manuscript dated to the 12th century and currently held at the Royal Library of Belgium in Brussels. In this paper, we present the TEI-conformant digital edition of the Vocabularium and the results of a quantitative study of its structure and content. In particular, we show how our edition-with-code approach can be employed to provide insight into historical dictionary-making practices.

First, a brief discussion of a number of issues related to the TEI-annotation (TEI Consortium 2022) is offered. We focus on challenges which arise due to the discrepancy between medieval and modern lexicographic techniques. For example, a single paragraph of a manuscript may contain multiple dictionary entries which are etymologically or semantically related to the headword.

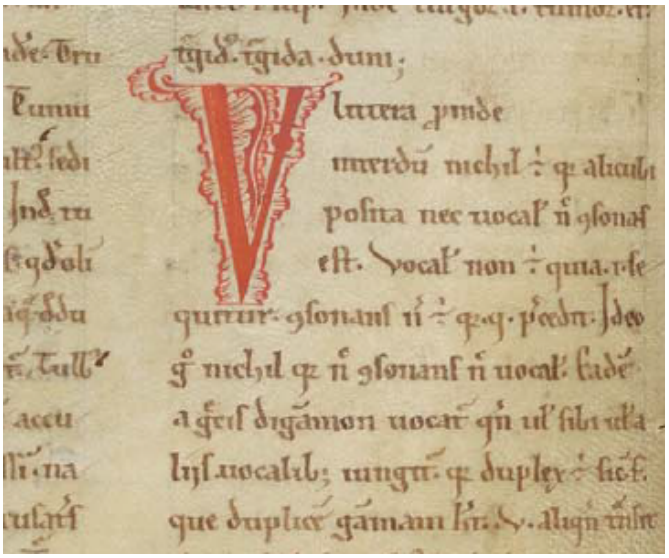


Figure 1: Ms. Bruxelles II 1049

Medieval glossaries are also less consistent in their use of descriptive devices. For instance, the dictionary definitions across the same work may greatly vary as to their form and content, and as such they systematically require fine-grained annotation.

Second, we present the TEI Publisher-based digital edition of the (Reijnders et al. 2022). At the moment, the interface provides basic browsing and Vocabularium search functionalities, making the dictionary available to the general public for the first time since the Middle Ages.

Figure 2: Digital edition of the Vocabularium Bruxellense

Thirdly, we demonstrate how a digital edition can be leveraged to enable a thorough quantitative analysis of historical dictionary-making practices. In what we call edition-with-code approach, a Jupyter notebook is used to retrieve both structural and content information from the source files of the dictionary edition. The data are next processed, summarized, and visualized in order to facilitate further research.

The results of analysis of almost 8,000 entries of the Vocabularium show, for example, that half of the entries are relatively short: a number among them contain only a one-word gloss and only 25% of entries contain 15 or more tokens.

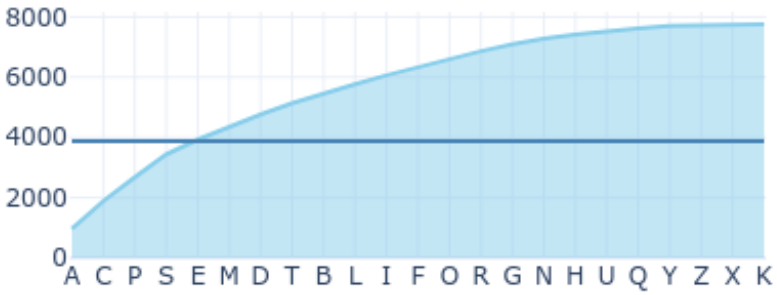


Figure 3: Cumulative sum of the number of entries

Based on the statistic analysis of nearly 1200 quotes, we were able to make a number of points concerning the function of quotations in medieval lexicographic works which is hardly limited to attesting specific language use. We observe that quotations are not equally dis-

tributed across the dictionary, as they can be found in slightly more than 10% of the entries, whereas nearly 7,000 entries have no quotations at all. The quotes are usually relatively short with only 5% containing 10 or more words. Our analysis shows that the most quoted author is by a wide margin Virgil followed by Horace, Lucan, Juvenal, Ovid, Plautus, and Terence (19). Church Fathers and medieval authors are seldom quoted, we have also discovered only 86 explicit Bible quotations so far.

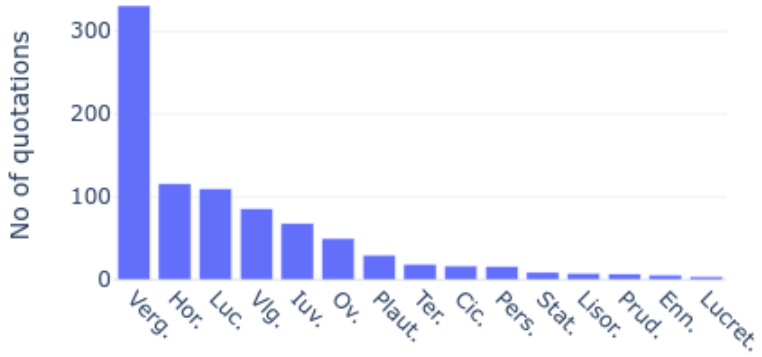


Figure 4: Cumulative sum of the number of entries

In conclusion, we argue that systematic comparative analysis of the existing editions of the medieval glossaries might provide useful insight into the development of this important part of the medieval written production.

References

- Reijnders, H. F., Krawczyk, Iwona, Alexandre, Renaud. (2022). Vocabularium Bruxellense. A Digital Edition (Version 1). Zenodo.
- TEI Consortium, eds. "9 Dictionaries." TEI P5: Guidelines for Electronic Text Encoding and Interchange. `https://tei-c.org/release/doc/tei-p5-doc/en/html/DI.html` (10 June 2022).
- Weijers, O. (1989). Lexicography in the Middle Ages. Viator 20: 139-53.

Generating English dictionary entries using ChatGPT: advances, options and limitations

Miloš Jakubíček, Michael Rundell
Lexical Computing
E-mail: milos.jakubicek@sketchengine.eu, michael.rundell@gmail.com

Keywords: ChatGPT; automatic dictionary drafting; Lexonomy; Lexonomy

In this paper we present a small English dictionary consisting of 99 sample entries generated fully automatically using the ChatGPT engine. We briefly introduce ChatGPT and the underlying machinery (an autore-gressive transformer-based neural network) but primarily focus on discussing the performance of the system, factors that influence the quality of the output and limitations that we have established. We show that while the system clearly outperforms the state-of-the-art of automatic generation for some entry components, it also has significant limitations which the lexicographic community should be aware of.

ChatGPT (Ouyang et al.,2022) is a chatbot based on the GPT3 language model (Brown et al.,2020) launched by OpenAI in November 2022. Since then, multiple new versions have been released. Even though the system had no public API at the time we were carrying out the experiments, we were able to interrogate it automatically to generate entries for 99 English single- and multi-word headwords. Because limited availability of the system made it impossible to create a bigger dictionary sample while preparing this paper, we wanted the dataset to be very diverse and therefore adapted a sample headword list used in the preparation of the DANTE lexical database for English (Convery et al.,2010).

The sample covers words of varying complexity and several parts-of-speech, as well as some multi-word expressions. We presented ChatGPT with each headword with no additional information (such as part-of-speech) and collected the response. Because the system is fine-tuned as a chatbot, we asked the following three questions for each headword H:

1. What does the word H mean?

2. Generate a dictionary entry for H.

3. Generate a dictionary entry for H including possible word forms, word senses, pronunciation, collocations, synonyms, antonyms and examples of usage.

These three questions were asked in this particular order in one conversation. As the inference of the system is generally not deterministic, we repeated this whole conversation three times independently in a new ChatGPT context, so that there would be no influence between the three runs. Altogether we thus obtained 297 entries consisting of verbatim answers to the three questions composing each conversation. They were all collected and the resulting dictionary was published in the Lexonomy platform. (Měchura et al.,2017)[1]

In the presentation and the full paper we will discuss findings that follow from our experiment including:

- quality of the results obtained for entry types and entry parts • stability and reproducibility of the results

[1] `https://lexonomy.eu/chatgpt`

- deficiencies identified in the output
- scientific and practical impact
- principal limitations and how they are manifested in the output References

References

- Ouyang, L., et al. (2022). Training language models to follow instructions with human feedback. arXiv preprint arXiv:2203.02155.
- Brown, T., et al. (2020). Language models are few-shot learners. Advances in neural information processing systems.
- Convery, C., et al. (2010). The DANTE Database (Database of ANalysed Texts of English). In: Proceedings of the XIV EURALEX International Conference. p. 293-5.
- Měchura, M., et al. (2017). Introducing Lexonomy: an opensource dictionary writing and publishing system. In: Electronic Lexicography in the 21st Century: Lexicography from Scratch. Proceedings of the eLex 2017 conference. p. 19-21.

Democratizing Digital Lexicography: a new project to facilitate the creation and dissemination of electronic dictionaries

Ligeia Lugli[1], Regiani Zacarias[2], Daniele Trevelin Donato[2]

[1]Mangalam Research Center, [2]UNESP

E-mail: ligeialugli@hotmail.com, regiani.zacarias@unesp.br, daniele.trevelin@unesp.br

Keywords: data-visualization; low-resource languages; corpus lexicography; ethnolinguistics

This poster reports on the initial results from a newly started project aimed at facilitating the creation of digital dictionaries for lexicographers working in 'low-tech' and low-resource environments. The project focusses especially on addressing the needs of lexicographers operating in Latin America, a region with many endangered languages in need of lexicographic documentation and with a thriving lexicographic community. The local lexicographic community, however, suffers from scarce technical support and inadequate training in computational methods, with the result that many face difficulties when trying to turn their dictionary-data into digital resources for online dissemination. Moreover, they all too often do not take advantage of computational methods available for complementing manually-curated lexical data with information automatically extracted from corpora. This results in lengthy manual work, which drives up the cost of their lexicographic projects and hinders their progress towards publication.

This poster reports on three aspects of our project. First, it summarises the results of a series of surveys we conducted better to understand the dictionary projects and needs of lexicographers operating in low-tech environments, especially in Latin America. The surveys gather two main pieces of information, (1) the data formats and tools used by lexicographers to record their dictionary data, and (2) the source on which their dictionary are based, with a view to differentiate between resources based on corpora and resources based on ethnolinguistics data. Second, the poster outlines the results of some preliminary experiments we carried out to devise a unified strategy to address the needs of both lexicographers working with corpora and those working with ethnolinguistics data–two groups whose methodologies differ, but also share significant aspects (Coxi, 2011). Finally, this poster sketches out some of the digital solutions we adopted to shift the communicative load of dictionaries from text to data-visualizations. This shift offers the advantage of minimising the need of costly manual editing (Lugli, 2021), but creates potential accessibility issue, as it risks to exclude visually-impaired users (Kim et al., 2021; Zong et al., 2022). The poster highlights some solutions to these issues in the context of lexicographic data-visualizations.

References

- Cox, Ch. (2011). Corpus linguistics and language documentation: challenges for collaboration. Language and Computers, 73: 239–264.

- Kim, N. W., et al. (2021). Accessible Visualization: Design Space, Opportunities, and Challenges. Computer Graphic Forum, 40 (3): 173–188.

- Lugli, L. (2021). Dictionaries as collections of data stories: an alternative post-editing model for historical corpus lexicography. In Itzok Kosem, et al. (eds.). Post-Editing Lexicography: eLex 2019, 216–231.
- Zong, J., et al. (2022). Rich Screen Reader Experiences for Accessible Data Visualization. Computer Graphic Forum, 41 (3): 15–27.

Relations, relations everywhere: an introduction to the DMLex data model

Michal Měchura[1], Simon Krek[2], Carole Tiberius[3], Miloš Jakubíček[4], Tomaž Erjavec[3]
[1]Masaryk University, [2]Jozef Stefan Institute, [3]Instituut voor de Nederlandse Taal, [4]Lexical Computing
E-mail: valselob@gmail.com, simon.krek@guest.arnes.si, carole.tiberius@ivdnt.org, milos.jakubicek@sketchengine.eu, tomaz.erjavec@ijs.si

Keywords: data modelling, standards, XML, JSON, Semantic Web, relational databases

This paper will introduce DMLex, a data model for representing machine-understandable lexicographic resources. DMLex is being developed by the LEXIDMA TC in OASIS, an organisation which oversees open standards in the IT industry. The first draft of DMLex will be entering its public review stage during the eLex conference. The purpose of this paper is to introduce DMLex to the community and to encourage participation in the public review.

General introduction to DMLEx

A data model is an inventory of data types for representing entities within a certain domain. DMLex is such an inventory for entities that exist in lexicography: dictionary entries, headwords, senses, definitions, sense relations and others. The paper will begin by laying out the two points which make DMLex different from other lexicographic data standards:

Most existing standards for representing dictionaries on computers are based on the assumption of a specific metamodel. For example, TEI (including TEI-Lex0) assumes the tree-structured metamodel of XML, and Ontolex/Lemon assumes the triple-centric, graph-structured metamodel of the Semantic Web. DMLex, on the other hand, is so abstract that it does not assume a specific metamodel. DMLex is designed to be expressible in formalisms as diverse as XML, JSON, relational database (SQL) and the Semantic Web. The DMLex standard contains recommended serialisations in these formalisms.

Many lexicographic data standards which assume a tree-structured metamodel (typically XML) are unnecessarily complex. This is because many phenomena in lexicography – subsenses, subentries, homography, entry-to-entry cross-references – are difficult to represent as tree structures (where objects are embedded inside each other) and would be more naturally represented as relational structures (where independently existing objects are connected through relations). The DMLex model only uses tree structures for representing the basic entries-and-senses hierarchy of a lexicographic resource, and relations for everything else. The result is a relatively simple, highly machine-understandable data model.

A module-by-module tour of DMLex

The DMLex specification is structured into modules. This paper will introduce the modules one by one, showing the data types that each module defines, and showing how those data types can be expressed in XML, in JSON, as a relational database, and as a Semantic Web triplestore. Each module defines data types for the following kinds of data:

- DMLex Core: the basic entries-and-senses structure of a monolingual lexicographic resource.

- DMLex Crosslingual Module: bilingual and multilingual lexicographic resources.
- DMLex Linking Module: relations between entries, senses and other objects, including semantic relations such as synonymy and antonymy, and presentational relations such as subentries and subsenses, both within a single lexicographic resource and across multiple lexicographic resources.
- DMLex Annotation Module: inline markup on various objects such as example sentences, including collocations and corpus patterns.
- DMLex Etymology Module: etymological information in lexicographic resources.
- DMLex Controlled Values Module: lists of controlled values, such as part-of-speech labels, and their mapping to external inventories.

Using DMLex

The paper will conclude with two case studies: one on representing existing born-digital dictionaries in DMLex, and one on implementing DMLex in an existing dictionary writing system.

References

- Cimiano, P., McCrae, J. P., Buitelaar, P. (2016) Lexicon Model for Ontologies: Community Report, 10 May 2016 Specification. URL: `https://www.w3.org/2016/05/ontolex/`.
- Tasovac, T, Romary, L., Banski, P., Bowers, J., de Does, J., Depuydt, K., Erjavec, T., Geyken, A., Herold, A., Hildenbrandt, V., Khemakhem, M., Petrović, S., Salgado, A., Witt, A. (2018) TEI Lex-0: A baseline encoding for lexicographic data. Version 0.8.6. DARIAH Working Group on Lexical Resources. URL: `https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html`.

From a dictionary towards the Hungarian Constructicon

Bálint Sass

Hungarian Research Centre for Linguistics, Institute for Lexicology
E-mail: sass.balint@nytud.hu

Keywords: construction; cross-reference system; constructicon; MWE

The term *constructicon* (ccn) (Fillmore, 2008) stands for the inventory of constructions (cxns) of a language – by analogy to the term *lexicon*. Accepting the position of Construction Grammar that utterances are not put together from words, but by combining cxns, it is quite straightforward that the primary unit of a lexical resource should be the cxn: a form–function pairing possibly spanning across linguistic levels. There is a considerable interest in developing (Lyngfelt et al., 2018) and investigating (Hanks and Može, 2019; Dunn, 2023) ccns nowadays.

A ccn is not a list-like structure, but rather a network of cxns, employing different kinds of inheritance and meronymy relations. Accordingly, a sophisticated cross-reference system is an important feature of ccns: from an abstract cxn to a more specific one, or from a cxn to its parts. While in traditional dictionaries phrasemes, collocations and the like are often treated only incidentally, ccns treat all kinds of meaning-bearing building blocks with equal care in a unified way as cxns, regardless of how complex they are. It is common to develop cnns by importing lexical information from existing lexical resources, especially form FrameNets.

We create a Hungarian ccn. In absence of a Hungarian FrameNet, we start from a monolingual dictionary (Pusztai, 2003) and derive the ccn to a great extent automatically: after identifying cxns in the "collocation" part of entries we lift them out and create individual entries for them on their own. Then we identify the parts of cxns, and reference the parts from the cxn and vice versa.

Interacting with a ccn, you should have the opportunity to search for cxns not just words. To avoid the user (e.g. a language learner) having to learn a formal language or a specific search tool (Sato, 2012), we introduce a new kind of search, which we call *analysed search*, suitable for ccns. The user is allowed to enter free text in a plain search box, then we apply automatic linguistic analysis (i.e. POS-tagging, dependency parsing, cxn-identification) to the text, and direct the user to the appropriate identified cxn(s). This process is performed to every types of cxns from simple or compound words to complex verbal frames.

Analysed search is supplemented by a novel referencing process called *dynamic referencing*. If the search query does not have a match, but its parts does, a "virtual" entry is created on-the-fly containing nothing but references to the parts. E.g. Hungarian counterpart of *'apple tree'* will present as an entry in the ccn, so the user will get this entry immediately, but *'papaya tree'* maybe not, so the virtual entry presented will contain a link to *'papaya'* and another to *'tree'* beyond the information that the original query is a compound construction. Instead of trying to make the ccn complete, we focus on making it easy to expand. Clearly, any expansion will influence dynamic referencing as it will decrease the need for virtual entries.

The above concepts, processes and features will be demonstrated in the presentation on a version of the Hungarian ccn created from (Pusztai, 2003). Future work includes integrating other lexical resources (e.g. Sass and Pajzs, 2010).

References

- Dunn, J. (2023). Exploring the constructicon: Linguistic analysis of a computational CxG. In: Proceedings of the Workshop on CxGs and NLP / SyntaxFest. Association for Computational Linguistics
- Fillmore, C.J. (2008). Border conflicts: FrameNet meets Construction Grammar. In: Bernal, E., DeCesaris, J. (eds.) Proceedings of the XIII EURALEX International Congress. pp. 49–68. Barcelona: Universitat Pompeu Fabra
- Hanks, P., Može, S. (2019). The way to analyse 'way': A case study in word-specific local grammar. International Journal of Lexicography 32(3), 247–269 (2019)
- Lyngfelt, B., Borin, L., Ohara, K., Torrent, T.T. (eds.). (2018). Constructicography: Constructicon development across languages. John Benjamins, Amsterdam
- Pusztai, F. (ed.). (2003). Magyar Értelmező Kéziszótár [Hungarian Monolingual Explanatory Dictionary]. Akadémiai Kiadó
- Sass, B., Pajzs, J. (2010). FDVC – creating a corpus-driven frequency dictionary of verb phrase constructions for Hungarian. In: Granger, S., Paquot, M. (eds.) Proceedings of eLex 2009. pp. 263–272. Louvain-la-Neuve: Presses universitaires de Louvain
- Sato, H. (2012). A search tool for FrameNet constructicon. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). pp. 1655–1658. European Language Resources Association (ELRA), Istanbul, Turkey

Unsupervised Sense Classification For Word Sketches

Ondřej Herman
Lexical Computing
E-mail: ondrej.herman@sketchengine.eu

Keywords: word sketch; sketch engine; word sense induction

In this article, we present a mechanism for annotating Word Sketch collocations according to word senses, which were automatically found in the source corpus text. This allows for significantly faster understanding of the behavior of the word usage by the end user.

Word Sketches (Kilgarriff et al., 2014) provide an efficient way of inspecting the collocational behavior of words in arbitrary corpora. The Word Sketch formalism is based on a grammar specification supplied by a language expert. In the Word Sketch grammar, the expert describes commonly occurring patterns present in the language. For example, the positions of modifiers relative to noun phrases, positions of objects relative to verbs or common prepositional phrase structures. These patterns are then located in the corpus and then collated according to the strength of their association. For example, for the verb fire, we can see that a common object it co-occurs with is gun, and one of the common modifiers is abruptly, among others.

Over time, Word Sketches have been extended by various features which improve the interpretability of the results, such as the Longest-Commonest Match, which shows the most common sequence of words present in the corpus along a particular collocation. For the gun as an object of fire, the sequence is fired a gun, and for the modifier abruptly of fired, it is was abruptly fired. However, the result can still be tedious to parse, as it is not obvious at a first glance which of the homonymous instances of the verb fire each of the collocations describe. In this article, we present a method of identifying sense-specific Word Sketch collocations.

To find the senses in the source corpus, we use a modified, in-house implementation of the adaptive skip-gram (Bartunov et al., 2016) model. Based on the venerable word2vec skip-gram language model, which creates an embedding vector for every word in the lexicon of the corpus, the adaptive skip-gram model generates multiple embedding vectors for each word in the lexicon, one for each sense. The senses as generated from the model are not always easily explainable, as they are expressed as numerical embedding vectors and only a list of their synonyms can be provided easily, which can be sometimes cryptic and non-descript.

To provide the sense information to the end user, we combine the respective strengths of the Word Sketch and adaptive skip-gram, combining the semantic sense information with the syntactic Word Sketch collocations. The Word Sketch gains a new view type, in which the collocations are grouped according to the distinct senses obtained from the adaptive skip-gram model.

References

- Bartunov, S., Kondrashkin, D., Osokin, A., Vetrov, D. (2016). Breaking sticks and ambiguities with adaptive skip-gram. In artificial intelligence and statistics (pp. 130-138). PMLR.

- Kilgarriff, 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.

ELEXIS Dictionary Matrix in elexiLink

Iztok Kosem[1], Tina Munda[2], Simon Krek[1]

[1]Jozef Stefan Institute, [2]Centre for Language Resources and Technologies at the University of Ljubljana (CJVT UL)
E-mail: iztok.kosem@ijs.si, tina.munda@cjvt.si, simon.krek@guest.arnes.si

Keywords: software; sense linking; dictionary matrix; elexiLink; NAISC; BabelNet

ELEXIS Dictionary Matrix has been developed in the framework of the ELEXIS project between 2018 and 2023 and received funding from the European Union's Horizon 2020 research and innovation programme. The dictionary matrix is a universal repository of linked senses and other lexicographic information found in existing lexicographic resources included in the ELEXIS infrastructure. Querying linked dictionaries is available through the user-friendly interface elexiLink[1], which is based on REST API capabilities. There are two modes of linking dictionaries: intra- and interlingual linking. The intralingual linking connects monolingual dictionaries of the same language between them. NAISC[2] (McCrae & Buitelaar 2018) is the software that computes those links by way of calculating the similarities between entries (the lemma + POS) in one and the other dictionary of the same language. The interlingual mode produces a mapping between two dictionaries in a cross-lingual scenario. The BabelNet linker[3] software (Martelli et al., 2022) links a definition from an ELEXIS dictionary with an English definition in the BabelNet semantic network (Navigli & Ponzetto, 2012). Through BabelNet acting as a pivot, the dictionaries in the ELEXIS infrastructure are linked at the sense level. The linking process is as follows: a dictionary, which has to comply with the TEI Lex-0 schema (for conversion see Elexifier[4] (Repar et al., 2020)), is uploaded to Lexonomy[5] (Měchura, 2017). There, the settings for linking (for either mode) are set and the request for linking is made. Upon the retrieval of the result, the links can be seen both in the Lexonomy UI and the elexiLink interface. The dictionaries that will be linked have to be available under an open-access license since the linking result will be publicly available in the interface. In the software demonstration, we will present the functionalities of the interface, describe the dictionaries currently included in elexiLink and present future plans for the service.

References

- Martelli, F., Navigli, R., Velardi, P., Kumar Ojha, A., McCrae, J.P. (2022). Cross-lingual Lexical Resource Linking Web Service (software). ELEXIS Deliverable 2.4. URL: `https://elex.is/wp-content/uploads/ELEXIS_D2_4_Cross-lingual_Lexical_Resource_Linking_Web_Service_software.pdf`.

[1]`https://matrix.elex.is`
[2]`https://github.com/insight-centre/naisc`
[3]`https://github.com/elexis-eu/BabelNet-linker`
[4]`https://elexifier.elex.is/`
[5]`https://lexonomy.elex.is`

- McCrae, J.P., Buitelaar, P. (2018). Linking Datasets Using Semantic Textual Similarity. Cybernetics and Information Technologies, 18(1), pp. 109–123. URL: http://www.cit.iit.bas.bg/CIT_2018/v-18-1/10_paper.pdf.
- 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. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial intelligence, 193, 217-250.
- Repar, A., Krek, S. (2020). Tools for the automatic segmentation and identification of lexicographic content. ELEXIS Deliverable 1.3. URL: https://elex.is/wpcontent/uploads/2020/02/ELEXIS_D1_3_Tools_for_the_automatic_segmentation_and_identification_of_lexicographic_content.pdf.

From Russian to Ukrainian: the r2u dictionary portal

Vasyl Starko
Ukrainian Catholic University
E-mail: vstarko@gmail.com

Keywords: dictionary portal; Ukrainian language; Ukrainian lexicography; Russian-Ukrainian dictionary

The r2u dictionary portal (r2u.org.ua[1]) was launched in 2007 with an electronic publication of a lexicographical work that best embodies the motto behind the portal's abbreviation: "Ad fontes. From Russian to Ukrainian." The work in question is the academic Russian-Ukrainian Dictionary edited by Ahatanhel Krymsky and Serhii Yefremov (1924–33). It was the last large dictionary published before the Soviet authorities launched a policy to Russianize and impoverish the Ukrainian language. This dictionary was partly censored, partly destroyed, branded in large portions as "bourgeois nationalistic," banned from public use after publication, and removed from public libraries. Even 90 years later, this dictionary remains the richest source of the Ukrainian lexis suppressed and eliminated under the Soviets, while Ukrainian lexicographic works published after 1991 have generally been slow to shed the Soviet legacy.

Over time, the r2u collection has been complemented with a number of other valuable lexicographical resources, including other Soviet-banned dictionaries and some modern works. With over 350,000 entries, including more than 265,000 unique ones, the r2u dictionary portal now offers full-text search capability for a collection of select dictionaries between Russian and Ukrainian, as well as several monolingual Ukrainian dictionaries. This collection caters to a wide and varied audience, as r2u processed over 2.4 million queries from more than 130,000 different IP addresses over the past year. A number of auxiliary dictionaries have also been made available for download as individual pdf files.

The paper discusses the rationale behind the creation and development of the r2u portal and highlights its most important dictionaries and features. One of the modern dictionaries on r2u is the Large Electronic Dictionary of Ukrainian (VESUM[2]). With over 415,000 lemmas (and counting), it is the most comprehensive morphological dictionary of Ukrainian. In parallel to providing complete paradigms of Ukrainian words to human users via the r2u web interface, it also functions as a machine-readable dictionary and serves as the backbone of the Pravopysnyk Language Tool[3], an advanced Ukrainian grammar and style checker. VESUM has also been used to tag large Ukrainian corpora, including the 1.5-billion-word GRAC corpus (uacorpus.org[4]). Other r2u dictionaries have been a valuable source of material for VESUM.

Thus, the dictionaries the Soviet authorities tried to erase were revived in electronic form on the r2u portal and then used as part of modern invisible lexicography.

[1] https://r2u.org.ua/
[2] https://r2u.org.ua/vesum/
[3] https://languagetool.org/uk
[4] http://uacorpus.org/Kyiv/ua

Probing visualizations of neural word embeddings for lexicographic use

Ágoston Tóth[1], Esra Abdelzaher[2]

[1]University of Debrecen, Faculty of Humanities, Institute of English and American Studies, Department of English Linguistics, [2]University of Debrecen, Institute of English and American Studies, Doctoral School of Linguistics
toth.agoston@arts.unideb.hu, esra.abdelzaher@gmail.com

Keywords: sense delineation,; word embedding visualization,; BERT

The distributional properties of words have been useful in unveiling word senses that are not present in dictionaries (Abdelzaher and Toth, 2020). They have also been proven to form distinct clusters corresponding to different word senses (Wiedemann et al., 2019; Schmidt and Hofmann, 2020). BERT word representations (Devlin et al., 2019) capture and predict the distributional properties of a word's use in a particular context. When word embeddings are considered vectors, they span a high-dimensional space, which is not readily useful for the human user, but the observation of 2D visualizations of these vectors may directly help the work of lexicographers.

This study argues that visualizing the distributional characteristics of headwords as seen in example sentences can help lexicographers find sense categories. It will also facilitate the detection of variations within and across sense categories and the selection of representative examples. The study uses t-distributed stochastic neighbor embedding (t-SNE) for dimension reduction. This optimizes a low-dimensional representation so that it preserves small pair-wise distances at the expense of long pairwise distances; as a result, local patterns stand out nicely. The example sentences in this study are cited from *Oxford Learners' Dictionary* (OLD) and the *British National Corpus* (BNC).

The automatically generated clusters revealed BERT's sensitivity to semantic and syntactic variations in word use. In many cases, sentences instantiating different senses appeared in different clusters. Moreover, contextual variations within the same sense category were reflected in multiple clusters. For instance, OLD sentences such as *mouth lifted in a wry smile* and BNC sentences such as *mouth twisted sardonically* formed clusters different from *stuff his mouth with pasta* and *forked food into his mouth*. Although the four sentences instantiate the same sense of *mouth* as part of the face (OLD: sense #1), the use of the mouth to eat is different from its use to make facial expressions. Second, verbal and nominal uses of the same word clustered in different parts of the space. The clusters and the distance between them clarified syntactic variations within the same sense category, for instance, *V risk of N* and *V risk that* (sense #1 of *risk*.n in OLD), and across sense categories, for example, *risk N on N* (sense #1 of *risk*.v in OLD) and *risk V-ing* (sense #2 of *risk*.v in OLD).

While this study is exploratory in nature, the method presented here can also be turned into a tool that helps lexicographers address the challenge of sense delineation, which is one of the most difficult tasks (Kilgarriff, 1998) as it involves abstracting senses from corpus citations (Kilgarriff, 2007). It could also be useful in selecting potential example sentences - as additional examples or alternatives to existing ones - that express the same semantic and syntactic patterns of word uses.

References

- Abdelzaher, E., Tóth, Á. (2020). Defining crime: A multifaceted approach based on lexicographic relevance and distributional semantics. Argumentum, 16, 44-63.
- Devlin, J., Chang, M., Lee, K., Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of NAACL, 4171-4186. Minneapolis, Minnesota: Association for Computational Linguistics.
- Kilgarriff, A. (1998). The hard parts of lexicography. International Journal of Lexicography, 11(1), 51-54.
- Kilgarriff, A. (2007). Word sense disambiguation. In P. Agirre, E., & Edmonds (Eds.), Word sense disambiguation: Algorithms and applications. Springer.
- Schmidt, F., Hofmann, T. (2020). BERT as a Teacher: Contextual Embeddings for Sequence-Level Reward.
- Wiedemann, G., Remus, S., Chawla, A., Biemann, C. (2019). Does BERT Make Any Sense? Interpretable Word Sense Disambiguation with Contextualized Embeddings. Proceedings of the 15th Conference on Natural Language Processing, KONVENS.

Research results and outcomes of the project "A Phraseographical Methodology and Model for an Online Corpus-Based Multilingual Collocations Dictionary Platform"

Adriane Orenha-Ottaiano[1], Maria Eugênia Olímpio de Oliveira Silva[2], José Manuel Pazos Bretaña[3], Carlos Roberto Valêncio[1], João Pedro Quadrado[1], Zhongmei Xiong[3]
[1]São Paulo State University, [2]University of Alcalá de Henares, [3]University of Granada
E-mail: adriane.ottaiano@unesp.br, eugenia.olimpio@uah.es, jmpazos@gmail.com, carlos.valencio@unesp.br, jp.quadrado@unesp.br, lluviadiosa@hotmail.com

Keywords: collocations; collocations dictionary; multilingual platform; Dictionary Writing System

This presentation aims to show some results and outcomes achieved within the scope of a project funded by The São Paulo Research Foundation (Process 2020/01783-2). It had the purpose of developing a phraseographical methodology and model for an online corpus-based Multilingual Collocations Dictionary Platform (MULTPLATCOL), in English, Portuguese, French, Spanish, and Chinese. The follow-up proposal will also include Italian, German, and European Portuguese. MULTPLATCOL is aimed to be customised for different target audiences according to their needs: language learners, pre- and in-service teachers, translators, material developers and researchers or lexicographers (Bothma et al., 2012; Fuertes-Olivera et al., 2014; Tarp, 2015). In this talk, we will present and discuss some data results and analysis in English, Portuguese, Spanish and Chinese, related to the project methodology, and the development of an in-house Collocations Dictionary Writing System.

The methodology developed for the MULTPLATCOL relies on the combination of automatic methods to extract candidate collocations (Garcia et al. 2019a). The automatic approaches take advantage of NLP tools to annotate large corpora with lemmas, PoS-tags and dependency relations in the five languages. Using these data, we applied statistical measures (Evert et al. 2017; Garcia et al. 2019b) and distributional semantics strategies to select the collocation candidates (Garcia et al. 2019c) and retrieve corpus-based examples (Kilgarriff et al. 2008). We also followed Garcia et al. (2019c) to carry out an automatic translation of the collocations (Orenha-Ottaiano et al. 2021). All automatically extracted data have been carefully post-edited by the lexicographers involved in this investigation.

The results regarding, for example, the number of extracted collocation candidates in the first phase is significantly high (a total of 309,838 collocation candidates in all languages) and demands a lot of effort from the whole team to carefully review them. Regarding the aforementioned methodological aspects, they have all proved to be efficient for all languages, except for Chinese. The treatment of the Chinese has created several difficulties due to the morphological and grammatical particularities of this language. For example, in the automated process of most languages, a simple question on how to delimit the extension of a word can raise unsuspected problems in Chinese, where the indispensable task of segmentation is not completely trivial (Pazos Bretaña et al., in press).

With respect to the software Collocations Dictionary Writing System (COLDWS), it was developed to specifically write and produce the Collocation Dictionaries. It comprises a database, a web interface for collaborative work, and some management tools. This way, all automatically extracted data were automatically inserted into the COLDWS and have been post-edited by the lexicographers. Afterwards, they will be exported to an end-user platform. The COLSDWS, which will also be presented in this paper, is one of the practical

and satisfactory outcomes of this project, being fully operational and particularly effective. An end-user platform model, which will be functionally integrated with the COLSDWS, has also been designed and is in process of adjustments. It still requires some improvements to be in line with our goal of building a platform that generates a more ambitious, customized and interactive lexicographic work, more suitable to the specificities and the idiosyncrasies of its users and their lexicographic needs.

Acknowledgements

We gratefully acknowledge the financial support provided by The Sao Paulo Research Foundation (FAPESP), Process ner 2020/01783-2.

References

- Bergenholtz, H., Tarp, S. (2003). Two opposing theories: On H.E. Wiegand's recent discovery of lexicographic functions. Hermes. Journal of Linguistics,31, pp. 171-196.
- Bond, F., Paik, K. (2012). A survey of wordnets and their licenses. In Proceedings of the 6th Global WordNet Conference (GWC 2012). Matsue. 64–71.
- Bothma, T., Tarp, S. (2012). Lexicography and the Relevance Criterion, Lexikos, 22, pp. 86-108.
- Evert, S., Uhrig, P., Bartsch, S., Proisl, T. (2017). E-VIEW-affilation–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, 531-549.
- Fuertes Olivera, P. A., Tarp, S. (2014). Theory and Practice of Specialised Dictionaries. Lexicography versus Terminography, Berlín/Boston: Walter de Gruyter.
- Garcia, M., García-Salido, M., Alonso-Ramos, M. (2019a). Towards the automatic construction of a multilingual dictionary of collocations using distributional semantics. In I. Kosem, T.Z. Kuhn, M. Correia, J.P. Ferreira, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek, C. Tiberius (eds.) Proceedings of eLex 2019: Smart Lexicography. Sintra, Portugal, pp. 747-762.
- Garcia, M., García-Salido, M., Alonso-Ramos, M. (2019b). A comparison of statistical association measures for identifying dependency-based collocations in various languages. In A. Savary, C. Parra Escartín, F. Bond, J. Mitrović, V. B. Mititelu (eds.) Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019). Florence, Italy, pp. 49-59.
- Garcia, M., García-Salido, M., Alonso-Ramos, M. (2019c). Weighted compositional vectors for translating collocations using monolingual corpora. In G. Corpas Pastor, R. Mitkov (eds.) Computational and Corpus-Based Phraseology. Cham, Switzerland: Springer, pp. 113-128.

- Kilgarriff, A., Husák, M., McAdam, K., Rundell, M., Rychly, P.. (2008). GDEX: Automatically Finding Good Dictionary Examples in a Corpus. In Bernal, E., J. DeCesaris (eds), Proceedings of the 13th EURALEX International Congress. Barcelona: Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra, 425-432.
- Orenha-Ottaiano, A., Garcia, M., Olímpio De Oliveira, M., L'Homme, M-C, Alonso Ramos, M.; Valêncio, C. R.; Tenório, W. (2021). Corpus-based methodology for an Online Multilingual Collocations Dictionary: First Steps. In: Kosem, I.; Michal C.; Miloš J.; Jelena K.; S. Krek, C. Tiberius (eds.). Proceedings of eLex 2021, pp. 1-28.
- Pazos Bretaña, J. M., Orenha-Ottaiano, A., Xiong, Z. (in press). PLATCOL, Plataforma Multilingüe de Diccionarios de Colocaciones: el caso del chino. Estudios de Traducción.
- Tarp, S. (2015). La teoría funcional en pocas palabras, Estudios de Lexicografía, no 4, pp. 31-42.
- Tarp, S. (2014). Theory-Based Lexicographical Methods in a Functional Perspective. An Overview [Theoriebasierte lexikographische Methoden aus funktionaler Perspektive. Ein Überblick / Méthodes en lexicographie théorique du point de vue fonctionnel. Une vue d'ensemble]. Lexicographica, 30(1), pp. 58-76.

The SERBOVERB Language Resource and Its Multifunctionality

Saša Marjanović
University of Belgrade - Faculty of Philology
E-mail: leksikograf@gmail.com

Keywords: SerboVerb; inflection; Serbian; verb; resource; dictionary; equivalents

Serbian verb inflection is quite complex. The paradigm of the average Serbian verb in the active voice gathers about a hundred inflectional forms (Šipka 2005). The relationship between these inflectional forms and their basic (lemma) form — which is conventionally used to represent the entire verb paradigm — is only predictable in a small number of inflectional classes. Hence, mastering Serbian verb inflection can be quiet challenging for average L2 Serbian speakers (cf. Krajišnik, 2011; Bošnjak Botica, 2013; Bošnjak Botica, Jelaska, 2018; Babić, 2021). The task is rendered even more difficult by the fact that some inflectional forms are hard to match to their lemma form. The existing Serbian dictionaries, both mono- and bilingual, where L2 speakers might search for an inflection information, are not well tailored to the needs of average L2 speakers: they list verbs generally only in the lemma form, while the forms relevant for establishing the entire paradigm are very often lacking (cf. Marković, 2014). Although there are different ways to process Serbian verb inflection in printed dictionaries so as to satisfy all the prototypical receptive, productive and cognitive needs of target users, we believe that the most appropriate solution, up-to-date, is to be found within the electronic lexicography.

In this paper, therefore, we present an innovative electronic lexicographically relevant language resource, intended for the L2 speakers of the Serbian language, where the Serbian verb inflection is processed in a dynamic way and with which all the aforementioned static paper-based consultation difficulties are eliminated. It is an electronic conjugator of the Serbian language — SerboVerb — with an accompanying dictionary module, which includes core equivalents in more than 20 languages, and with an additional gamification module. The entire resource, which currently offers paradigms for more than 20,000 verbs, but is planned for more than 34,000 verbs, can be accessed for free via the website (`https://serboverb.com`), but also via Google Play store and the App store in the form of a mobile app intended to be used under the Android and iOS operating systems respectively. The resource has been developed since 2017 as part of the scientific research project of the University of Toulouse – Jean Jaurès (Toulouse, France) as a result of intensive bilateral cooperation between experts and volunteers from the University of Toulouse and the Faculty of Philology of the University of Belgrade (Serbia). Moreover, the megastructure of this innovative resource, its macrostructure, as well as elements of the resource microstructure and the ways in which the mediostructure of its components is achieved is presented in this paper. The aim of the paper is to show — taking into account the main factors in the lexicographic processing of verb inflection, i.e. exhaustiveness, simplicity and availability — the multifunctionaly of the SerboVerb language resource.

References

- Babić, B. (2021). Unutarjezičke greške u nastavi srpskog jezika kao stranog. Novi Sad: Filozofski fakultet.

- Bošnjak Botica, T. (2013). Opća načela podjela na glagolske vrste u hrvatskome u perspektivi drugih bliskih jezika. Lahor 15, 63–90.
- Bošnjak Botica, T., Jelaska, Z. (2018). Načela kategorizacije glagolskih skupina i vrsta u hrvatskom, francuskom i rumunjskom jeziku. U: Poglavlja iz romanske filologije. U čast akademiku Augustu Kovačecu o njegovu 80. rođendan (N. Lanović i dr., ur.). Zagreb: FF press, 117–144.
- Krajišnik, V. (2011). Rječnik u nastavi srpskog kao stranog jezika. Anali Filološkog fakulteta 23/2, 245–258.
- Marković A. Gramatika u srpskim rečnicima. Dragićević Rajna (ur.). Savremena srpska leksikografija u teoriji i praksi. Beograd: Filološki fakultet, 2014, 69–91.
- Šipka, D. (2005). Osnovi morfologije. Beograd: Alma.

Operationalising and representing conceptual variation for a corpus-driven encyclopaedia

Santiago Chambó, Pilar León-Araúz
University of Granada
E-mail: santiagochambo@ugr.es, pleon@ugr.es

Keywords: conceptual analysis; conceptual variation; corpus-driven encyclopaedia; lexical data visualisation

The humanitarian domain is a multidisciplinary and recently professionalised field that comprises numerous specialised organisations ran by people with different professional, organisational and cultural backgrounds (Eberwein and Saurugger, 2013). This diversity plays a role in how humanitarians conceptualise their domain (Stroup, 2012; Sezgin and Dijkzeul, 2015), giving rise to highly unstable concepts such as RESILIENCE (Béné et al., 2012), EVIDENCE (Knox Clarke and Ramalingan, 2014) and LOCAL ORGANISATION (Khan and Kontinen, 2022). Although domains develop shared terminologies, conceptual variation can affect the intensions and extensions of general and specialised concepts (Hampton, 2020), resulting in fuzzy, highly diverse and multidimensional conceptualisations (León-Araúz, 2017, 215).

The Humanitarian Encyclopedia (HE; humanitarianencyclopedia.org) is a descriptive reference work of the humanitarian domain, which adopted a Frame-based Terminology (FBT) approach (Faber 2015; 2022) to conceptual analysis driven by systematic extraction and curation of lexical data from corpora. To offset biases and possible lacunae arising from background diversity, entries in the encyclopaedia are written by authors who combine their expert knowledge about a concept with conceptual reports based on lexical data provided by linguists (Humanitarian Encyclopedia 2021).

FBT describes concepts by modelling lexical data into propositions (i.e., predicate-argument structures) and elucidating semantic relations and related concepts from definitions (Faber, 2012; Vintar and Martinc, 2022), other knowledge rich contexts (KRCs; León-Araúz and Reimerink 2019) and multi-word terms (Rojas-Garcia and Cabezas-García 2019). In this study, KRCs are extracted from a corpus of humanitarian texts through semantic sketch grammars (León-Araúz et al., 2018; Martín et al., 2020; Martín and Trekker, 2021) and systematic querying with additional knowledge patterns (Marshman, 2022) in Sketch Engine. Additionally, long and diverse clauses may be subsumed inductively into manageable conceptual labels with qualitative analysis software. Conceptual variation is then operationalised by linking lexical data, corpus metadata and conceptual propositions, thereby substantiating claims with textual evidence. This method provides quantitative variables to establish a semantic core and marginal characteristics that can be disaggregated. This enables conceptual analysts to make comparisons of intensions and extensions diachronically, between actors (e.g., types of humanitarian organisation) and other corpus metadata available.

Scholars like San Martín (2022) have studied conceptual variation to design a framework for flexible definitions across several scientific disciplines. Our approach is similar, but it enhanced with data visualisations due to the different types of corpus metadata. Data visualisation enables conceptual analysts to (1) identify areas of conceptual variation and (2) communicate findings to entry authors. For example, to represent quantitative differences in intensional characteristics between subcorpora, radar charts are a good option. However,

they may have to be supplemented by other visualisations, especially when there are too many variables to display. For PROCESS concepts, Sankey diagrams and parallels sets are useful to represents frequency distribution for arguments but fail to display disaggregation effectively unless they are presented in an interactive environment. In this paper, we illustrate our method to describe and represent conceptual variation with a selection of concepts that constitute entries in the HE.

References

- Béné, Ch., Wood, R.G., Newsham, A., Davies, M.. (2012). Resilience: New Utopia or New Tyranny? Reflection about the Potentials and Limits of the Concept of Resilience in Relation to Vulnerability Reduction Programmes. IDS Working Papers 2012 (405): 1–61.
- Eberwein, W.D., Saurugger, S. (2013). The Professionalization of International Non-Governmental Organizations. In Routledge Handbook of International Organization, edited by Bob Reinalda, 257–69. Abingdon-on-Thames, England: Routledge.
- Faber, P. (2012). A Cognitive Linguistics View of Terminology and Specialized Language. Berlín, Germany: De Gruyter Mouton.
- Faber, P. (2015). Frames as a Framework for Terminology. In Handbook of Terminology: Volume 1, edited by Hendrik J. Kockaert and Frieda Steurs, 14–33. Amsterdam, The Netherlands: John Benjamins Publishing Companyi.
- Faber, P. (2022). Chapter 16. Frame-Based Terminology. In Theoretical Perspectives on Terminology: Explaining Terms, Concepts and Specialized Knowledge, edited by Pamela Faber and Marie-Claude L'Homme, 23:353–76. Terminology and Lexicography Research and Practice. Amsterdam, The Netherlands: John Benjamins Publishing Company.
- Hampton, J. A. (2020). Investigating Differences in People's: Concept Representations. In , edited by T. Marques and A. Wikforss, 67–82. Oxford, UK: Oxford University Press.
- Humanitarian Encyclopedia. 2021. Methodology. Humanitarian Encyclopedia. 2021. URL: `https://humanitarianencyclopedia.org/methodology`.
- Khan, A.K., Kontinen, T. (2022). Impediments to Localization Agenda: Humanitarian Space in the Rohingya Response in Bangladesh. Journal of International Humanitarian Action 7 (1): 14.
- Knox Clarke, P., Ramalingan, B.. (2014). Meeting the Urban Challenge: Adapting Humanitarian Efforts to an Urban World. London: ALNAP/ODI.
- León-Araúz, P. (2017). Term and Concept Variation in Specialized Knowledge Dynamics. In Multiple Perspectives on Terminological Variation, edited by Patrick Drouin, Aline Francoeur, John Humbley, and Aurélie Picton. Amsterdam, The Netherlands: John Benjamins Publishing Company.

- León-Araúz, P., Reimerink, A. (2019). High-Density Knowledge Rich Contexts. Argentinian Journal of Applied Linguistics - ISSN 2314-3576 7 (1): 109–30.
- León-Araúz, P., Martín, A.S. (2018). The EcoLexicon Semantic Sketch Grammar: From Knowledge Patterns to Word Sketches. In Proceedings of the LREC 2018 Workshop, edited by I. Kernenman and S. Krek. Miyazaki, Japan: Globalex.
- Marshman, E. (2022). Chapter 13. Knowledge Patterns in Corpora. In Theoretical Perspectives on Terminology: Explaining Terms, Concepts and Specialized Knowledge, edited by Pamela Faber and Marie-Claude L'Homme, 23:291–310. Terminology and Lexicography Research and Practice. Amsterdam, The Netherlands: John Benjamins Publishing Company.
- Rojas-Garcia, J., Cabezas-García, M. (2019). Use of Knowledge Patterns for the Evaluation of Semiautomatically-Induced Semantic Clusters. In , 121–40.
- Martín, A.S. (2022). A Flexible Approach to Terminological Definitions: Representing Thematic Variation. International Journal of Lexicography 35 (1): 53–74.
- Martín, A.S., León-Araúz, P. (2013). Flexible Terminological Definitions and Conceptual Frames. In , 1061:1–17. Montreal, Canada.
- Martín, A.S., Trekker, C. (2021). Adapting Word Sketches for Specialized Knowledge Extraction. In Proceedings of ASIALEX 2021, 64–87. Jarkarta, Indonesia: ASIALEX.
- Martín, A.S., Trekker, C., León-Araúz, P. (2020). Extraction of Hyponymic Relations in French with Knowledge-Pattern-Based Word Sketches. In Proceedings of the Twelfth Language Resources and Evaluation Conference, 5953–61. Marseille, France: European Language Resources Association.
- Sezgin, Z., Dijkzeul, D., eds. (2015). The New Humanitarians in International Practice: Emerging Actors and Contested Principles. London, UK: Routledge.
- Stroup, S. (2012). Borders among Activists: International NGOs in the United States, Britain, and France. Borders among Activists. Ithaca, NY: Cornell University Press.
- Vintar, Š., Martinci, M. (2022). Framing Karstology: From Definitions to Knowledge Structures and Automatic Frame Population. Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication 28 (1): 129–56.

Lexicographic considerations in the coding of inquisition transcripts of Medieval Latin

Dr David Zbíral, Dr Gideon Kotzé, Dr Robert L.J. Shaw

Masaryk University

E-mail: david.zbiral@mail.muni.cz, gidi8ster@gmail.com, robert.shaw@mail.muni.cz

Keywords: historical source modeling; meaning representation; semantics; syntax; annotation; NLP; corpus linguistics; medieval Latin

The utility of linguistic annotation for corpus linguistics and lexicography is well known (Faaß, 2017). Often, these annotation layers are produced by automated tools that, while having vastly superior coverage to humans, also produce errors. The ideal input is that of an expert of the language itself, with intimate knowledge of its contextual use. In the case of medieval Latin, some of the most well-versed experts are historians of the era. Working with source texts, it is in their interest to place them in the correct context and relate them to known facts. As part of the implementation of this effort, we are currently annotating transcribed medieval inquisition registers using the recently developed Computer-Assisted Semantic Text Modelling (CASTEMO) approach (Zbíral and Shaw, 2022). CASTEMO models semantics via formalized data statements based on the structure of a quadruple of subject(s), predicate(s) and two objects, inspired by well-known ideas on the concept of meaning representation, such as the RDF and OWL standards, the Semantic Web (Berners-Lee et al., 2001), and Quantitative Narrative Analysis (Franzosi, 2009). Apart from shallow semantics, the lexical and compositional semantic properties of concepts and actions are also modeled, all of which can have modifiers describing properties. With the inclusion of extratextual information and the coding of epistemic levels and modality in the document-oriented database, a knowledge graph is also produced that is useful for contextualization and deeper understanding. Linked to the corpus, this allows us to incorporate textual context, as well as the output of various natural language processing tools, further enriching and complementing the annotation layer.

A fortunate side effect of this approach is the analysis of language itself, which creates a linguistic resource that can be useful for corpus linguistics and lexicography. Since natural language sentences are annotated with a meaning representation, the resource can, on one level, also be presented as a type of semantic treebank.

To illustrate the utility for corpus linguistics and lexicography, consider the Latin verb, ”facere” (”to make, do, or accomplish; become [passive]”), which has a multitude of senses in different contexts. A part-of-speech tagger can be used to exclude non-verbal homonyms. We can lemmatize verbs in order to group all the different forms together in a single query. We can now filter occurrences of the verb based on the semantic profiles in which they occur, by considering valency information, text labels, or specific combinations of entities, actions and properties. For example, we may be interested in cases where the second actant may also involve groups or locations, or where the first actant is typically introduced by ”cum” and involves persons or groups. This would be achieved by querying the statement entry, extracting the IDs related to actants, and querying the entries for the actants with those IDs.

Through the aforementioned as well as quantitative analyses methods, we show how this modeling approach can assist the lexicographer in building suitable meaning representations, while comparing it with current standards of lexicographic solutions.

References

- Berners-Lee, T., Hendler, J., Lassila, O. (2001). The Semantic Web. Scientific American 284(5): 34-43.
- Faaß, G. (2018). Lexicography and corpus linguistics. In: Fuertes-Olivera, Pedro A. (ed.), The Routledge Handbook of Lexicography. London/New York: Routledge. DOI: 10.1093/ijl/ecab008
- Franzosi, R. (2009). Quantitative Narrative Analysis. Thousand Oaks: Sage.
- Zbíral, D. and Shaw, R.L.J. (2022). "Hearing Voices: Reapproaching Medieval Inquisition Records" Religions 13, no. 12: 1175.

Rapid Ukrainian-English Dictionary Creation Using Post-Edited Corpus Data

Vojtěch Kovář, Vlasta Ohlídalová, Marek Blahuš, Miloš Jakubíček, Michal Cukr

Lexical Computing

E-mail: vojtech.kovar@sketchengine.eu, vlasta.ohlidalova@sketchengine.eu, marek.blahus@sketchengine.eu, milos.jakubicek@sketchengine.eu, michal.cukr@sketchengine.eu

Keywords: ukrainian dictionary; rapid dictionary development; post-editing

For decades, language corpora have served as source data for building dictionaries. In the last decade, corpora were also used for automatic generation of various dictionary parts: headword lists, examples, collocations. These automatic outputs were then post-edited by professional lexicographers to ensure the data quality in the resulting dictionary. With the advancement of technology, it is now possible to create whole dictionaries using this scenario of automatic generation and postediting by native speakers (not necessarily professional lexicographers). The methodology was used before (Baisa et al., 2019); we have improved the process and used it in a new project aimed at creating a Ukrainian-English dictionary using a 3-billion-word Ukrainian corpus. The corpus was built from the web, cleaned, deduplicated and annotated automatically for part-of-speech and lemmas. In the next phases, we generated parts of the dictionary data from the corpus and arranged their post-editing by native speakers in a specialized web interface. The particular parts of the process corresponded to the entry structure in the resulting dictionary and included the following steps:

- Headword selection. The dictionary headwords were automatically extracted from the corpus using document frequency and manually classified into correct headwords, tagging/lemmatization errors, non-standard forms, foreign words and proper names. From the original 100,000 corpus words, we gained over 50,000 headwords suitable for the dictionary.
- Revision of incorrect headwords. The words marked as errors in tagging or lemmatization were manually corrected, not to miss any frequent word due to an error in the automatic tools.
- Word form selection. We have automatically extracted word forms for each headword from the corpus, and manually removed the incorrect ones.
- Audio recording. One of our native speakers recorded the pronunciation of all the correct headwords.
- Sense annotation. The native speakers were given list of frequent collocations for each headword, and were asked to classify the collocations into senses. We then used the collocations to mark theparticular senses in the corpus automatically. Later, each sense was manually given a short description (disambiguator) and one or more English translations, pre-generated automatically from 3 different translation APIs.
- Thesaurus selection. For each headword, we generated a distributional thesaurus from the corpus and our native speakers classified the items as synonyms, antonyms, similar words and the rest.

- Examples selection and translation. Candidate example sentences for each headword were generated from the corpus and automatically translated using the DeepL API. Then both the Ukrainian sentences and the English translations were post-edited by our native speakers.

In the full paper we will report on particular tools and techniques used for the automatic drafting as well as on lessons learnt during the post-editing process that are generally applicable in the context of rapid dictionary development.

References

- Baisa, V., Blahuš, M., Cukr, M., Herman, O., Jakubíček, M., Kovář, V., Medveď, M., Měchura, M., Rychlý, P. and Suchomel, V. (2019). Automating dictionary production: a Tagalog-English-Korean dictionary from scratch. In Proceedings of the 6th Biennial Conference on Electronic Lexicography. Brno, Czech Republic: Lexical Computing CZ s.r.o., p. 805-818. ISSN 2533-5626.

Adding Information to Multiword Terms in Wiktionary

Thierry Declerck[1], Lenka Bajčetić[2], Gilles Sérasset[3]

[1]DFKI GmbH, Multilingual Technologies, Germany, [2]Innovation Center of the School of Electrical Engineering in Belgrade, Serbia, [3]Université Grenoble Alpes, CNRS, France
E-mail: declerck@dfki.de, lenka.bajcetic@gmail.com, gilles.serasset@imag.fr

Keywords: Multiword terms; Wiktionary; lexical enrichment; linguistic linked data

We describe an approach aiming at enriching English multiword terms (MWTs) included in Wiktionary by generating lexical information gained by using, filtering and combining available lexical descriptions of their lexical components.

We started our work with the generation of pronunciation information, as we noticed that a vast majority of English MWTs in Wiktionary are lacking this type of information. While designing a potential evaluation dataset for the pronunciations generated by our approach, we noticed that only around 3% of MWTs are carrying pronunciation information. We also discovered that other complex lexical constructions (affix + word, or word + affix) are often lacking pronunciation information. We collected for the evaluation dataset 6,768 MWT entries with pronunciation (compared with 252,082 MWT entries that are lacking such information). Our approach for generating pronunciation information for MWTs consisted in combining the pronunciation information included in the lexical description of their components. Results of this work can be integrated in other lexical resources, like the Open English WordNet (McCrae et al., 2020), where pronunciation information has been added for now only for single word entries, as described in (Declerck et al., 2020a).

A specific issue emerged for the generation of pronunciation information for MWTs that contain (at least) one heteronym. For this type of lexical entry a specific processing is needed, disambiguating between the different senses of the heteronym for extracting the appropriate pronunciation of this one lexical component to be selected to form the overall pronunciation of the MWT. An example of such a case is given by the Wiktionary entry “acoustic bass”, for which our algorithm has to specify that the pronunciation /beIs/ (and not /bæs/) has to be selected and combined with /@”ku:.stIk/. It is important to mention that Wiktionary often lists several pronunciations for various variants of English. In this work we focus on the standard, received pronunciation for English, as encoded by the International Phonetic Alphabet (IPA).

Since there are cases for which we need to semantically disambiguate one or more lexical components of a MWT for generating its pronunciation, our work can also lead to the addition of disambiguated morphosyntactic and semantic information of those components to the lexical description of MWTs, and thus enrich the overall representation of the MWTs entries beyond the generation of pronunciation information. This is a task we have started to work on.

In this paper, we describe first briefly the way multiword terms (MWTs) are introduced in Wiktionary. We summarize then the various approaches we followed for both designing an evaluation dataset and generating pronunciation information, dealing for now with the English edition of Wiktionary. We discuss issues we encountered, and which lead to the consultation of related resources, like DBnary (Sérasset & Tchechmedjiev, 2014; Sérasset, 2015) and WikiPron (Lee et al., 2020). While the cooperation with DBnary has been already established and resulted in improvements of our approach and an adaptation of DBnary

itself, which we describe in some details, we are starting with the formulation of suggestions for adaptation for WikiPron.

References

- McCrae, J.P., Rademaker, A., Rudnicka, E., Bond, F. (2020). English WordNet 2020: Improving and Extending a WordNet for English using an Open-Source Methodology. In Proceedings of the LREC 2020 Workshop on Multimodal Wordnets (MMW2020). Marseille, France: The European Language Resources Association (ELRA), pp. 14–19. URL: https://aclanthology.org/2020.mmw-1.3.

- Declerck, T., Bajcetic, L., Siegel, M. (2020a). Adding Pronunciation Information to Wordnets. In Proceedings of the LREC 2020 Workshop on Multimodal Wordnets (MMW2020). Marseille, France: The European Language Resources Association (ELRA), pp. 39–44. URL: https://aclanthology.org/2020.mmw-1.7.

- Sérasset, G., Tchechmedjiev, A. (2014). Dbnary : Wiktionary as Linked Data for 12 Language Editions with Enhanced Translation Relations. In 3rd Workshop on Linked Data in Linguistics: Multilingual Knowledge Resources and Natural Language Processing. Reyjkjavik, Iceland, p. to appear. URL: http://hal.archives-ouvertes.fr/hal-00990876.

- Sérasset, G. (2015). DBnary: Wiktionary as a Lemon-Based Multilingual Lexical Resource in RDF. Semantic Web, 6, pp. 355–361.

- Lee, J.L., Ashby, L.F., Garza, M.E., Lee-Sikka, Y., Miller, S., Wong, A., McCarthy, A.D., Gorman, K. (2020). Massively Multilingual Pronunciation Modeling with WikiPron. In Proceedings of the 12th Language Resources and Evaluation Conference. Marseille, France: European Language Resources Association, pp. 4223–4228. URL: https://www.aclweb.org/anthology/2020.lrec-1.521.

Development of Evidence-Based Grammars for Terminology Extraction in OneClick Terms

Marek Blahuš, Ondřej Matuška

Lexical Computing
E-mail: marek.blahus@sketchengine.eu, ondrej.matuska@sketchengine.eu

Keywords: terminology extraction; corpus linguistics; IATE

This paper presents the performance and underlying principles of a new generation of terminology extraction grammars for the OneClick Terms system, which use rules inspired by patterns observed in existing terminology databases in order to improve the coverage of discovered term candidates.

The OneClick Terms system automatically extracts terminology from text using corpus-based contrastive technology in the Sketch Engine corpus management system. In order to generate term candidates, the software must be provided with a language-specific term grammar.

The term grammar contains a carefully crafted set of rules (expressed in CQL, the Corpus Query Language (Evert, 2005; Sketch Engine) describing noun phrases manifested by the presence of a head noun, but with a variable internal morphosyntactic structure.

The discovered term candidates are scored by comparing the normalized frequencies with a large reference corpus (Jakubíček et al., 2014). A recent extension to the system allows for bilingual terminology extraction from aligned documents (Kovař et al., 2016) based on co-occurrences in aligned segments being ranked using the logDice association score (Rychlý, 2008). Currently, 26 languages are supported.

Most pre-existing term grammars were prepared using a linguistic judgement and could only match term candidates of a limited variety and length. Within the present work, we adopted a strictly empirical approach to study an existing manually curated terms base, the IATE, developed by the Translation Centre for the Bodies of the European Union with terms in 24 languages (Zorrilla-Agut et al., 2019).

We used Sketch Engine to build a special ”term corpus” with terms in IATE and to tag it for parts of speech and morphology. A report detailing the frequency distribution of POS tags and their combinations in the term corpus was generated and served as the foundation for a collaborative effort of a corpus linguist and a speaker of the language, aimed at generalizing the observed patterns and turning them into matching rules expressed in CQL. The resulting term grammar is said to be ”evidence-based” due to the fact that it was created with the purpose of maximizing the recall of the terms found in an existing terminology database.

Although subject to certain limitations (such as ambiguous tagging, compromises for precision, reasonable limit for the number of rules), grammars devised in this way are capable of covering significant portions of IATE terms (e.g. 74% for English). By the time of submitting this abstract, we had generated evinced-based term grammars for 6 languages (English, Estonian, French, German, Italian and Spanish). By comparison and opinions of OneClick Terms users, the quality of term candidates has arguably improved for all of them.

References

- Jakubíček, M., et al. (2014). Finding terms in corpora for many languages with the Sketch Engine. In: Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics. p. 53-56.
- Kovař, V., Baisa, V., Jakubíček, M. (2016) Sketch engine for bilingual lexicography. International Journal of Lexicography, 29.3: 339-352.
- Rychlý, P. (2008). A Lexicographer-Friendly Association Score. In: RASLAN. p. 6-9.
- Evert, S. (2005). The CQP query language tutorial.
- Sketch Engine: CQL – Corpus Query Language: `https://www.sketchengine.eu/documentation/corpus-querying/`
- Zorrilla-Agut, P., Fontenelle, T. (2019). IATE 2: Modernising the EU's IATE terminological database to respond to the challenges of today's translation world and beyond. In: Terminology, 25(2), pp. 146–174.

Constitution of a substandard French online dictionary from and for the francophone rap corpus RapCor in order to teach dia-variation in French as a Foreign Language (FLE) at university level

Alena Polická[1], Anne-Caroline Fiévet[2], Laurent Canal[1]
[1]Masaryk University, [2]EHESS Paris
E-mail: podhorna@phil.muni.cz, acfievet@gmail.com, canallaurent66@gmail.com

Keywords: substandard French; CEFR; didactics; corpora building; lexical training data

Slang is often mentioned in the CEFR (Council of Europe, 2020) at C1 level and above (B2 for the most common slang words) but students at B1/B2 level find it very useful in practice. However, despite the use of pedagogical tools in linguistics or literature as they are offered in universities in the Czech Republic, students are unable to find the opportunities to learn present-day French and the various forms it can take. The RapCor corpus (Podhorná-Polická, 2020) would allow them to discover the lexicographical resources of slang words in order to understand, mainly orally, what their classmates are saying, for example, or their favorite rap songs. It would deepen the students' knowledge and in return, the students could broaden the corpus by adding new annotations. When integrating the text on account of these annotations and pre-processing of the software, we can offer suggestions to clarify polysemies, competing graphic forms to unify under a single lemma, etc., that would strengthen their knowledge in contextual comprehension and reinforce their sociopragmatic skills (Beeching & Woodfield, 2015).

Rap song lyrics represent a valuable source for attesting to the lexical innovations that are being spread amongst the younger generation (as exemplified in various dictionaries of contemporary slang, cf. Tengour 2013, Goudaillier 2019). Since 2006, rap song texts have been used in sociolexicology and translation seminars at the Institute of Romance Languages and Literatures at Masaryk University, in order to teach dia-variation in authentic, current and student-appealing contexts. In addition to these annual seminars taking on a qualitative approach, a linguistic corpus was put in place in 2009 from the available texts in booklets of French-language rap albums, adopting a more quantitative approach. First, the process of scanning the lyrics, from cutting up the text through the optical character recognition (OCR) and tagging of individual parts of the song, metadata on the performers, to comparing the differences between what is written and what is sung, was offered as an extracurricular activity to the students.

Next, students had to use the TreeTagger (TT) software to see segments of the song text, annotated and lemmatised with the different parts of speech using a built-in dictionary. Then, the results had to be cleaned and developed relative to an internal set of labels. The task of adding unknown lemmas to the TT or identifying substandard meanings in locutions with standard isolated elements became very difficult for students who could not discern certain subtleties, as observed in the online version of RapCor published on SketchEngine, the RapCor1288 (containing 1288 songs).

For this workshop on invisible lexicography, we aim to show the different stages of 'back and forth' between professors and students of our new course that are in line with the two approaches previously explained (quantitative and qualitative) and a concrete practical result (a new dictionary, based on the TT dictionary but prepared using our data, the FrenchTagger, which was developed by our colleagues at the Faculty of Computer Science). We will also

establish a typology of didactic approaches relating to the issues in students' lexicographic work.

References

- Council of Europe. (2020). Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume, Council of Europe Publishing, Strasbourg. URL: `www.coe.int/lang-cefr`.
- Beeching, K., Woodfield, H. (2015). Researching sociopragmatic variability. Perspectives from Variational, Cross-linguistic and Interlanguage Pragmatics. Basingstoke: Palgrave/Macmillan.
- Goudaillier, J.P. (2019). Comment tu tchatches ! Dictionnaire du français contemporain des cités. Paris, Maisonneuve & Larose, (4 éditions 1997, 1998, 2011 et 2019).
- Podhorná-Polická, A. (2020). RapCor, Francophone Rap Songs Text Corpus. In Horák, Aleš; Rychlý, Pavel; Rambousek, Adam. Proceedings of the Fourteenth Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2020. Brno: Tribun EU, s. 95-102.
- Tengour A. (2013). Tout l'argot des banlieues. Le dictionnaire de la zone en 2 600 définitions. Paris, Les éditions de l'Opportun.

Author Index