eLex 2017

Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference.

edited by

Iztok Kosem
Carole Tiberius
Miloš Jakubíček
Jelena Kallas
Simon Krek
Vít Baisa

Leiden, the Netherlands, 19–21 September 2017

elex.link/elex2017

Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference.

edited by

Iztok Kosem
Carole Tiberius
Miloš Jakubíček
Jelena Kallas
Simon Krek
Vít Baisa

published by

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

proofreading

Nikki Tagg Nama

Leiden, September 2017

ISSN 2533-5626

Acknowledgements

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

PROGRAMME SPONSOR

idm

SPONSORS

Dictionary.com

Collins

Oxford Dictionaries

SUPPORTING INSTITUTIONS

KDICTIONARIES

Amsterdam University Press

globaLex

CONFERENCE COMMITTEES

Organising Committee

Carole Tiberius
Iztok Kosem
Jelena Kallas
Miloš Jakubíček
Simon Krek
Ondřej Matuška

Scientific Committee

Andrea Abel
Valentina Apresjan
Špela Arhar Holdt
Iana Atanassova
Gerhard Budin
Nicoletta Calzolari
Lut Colman
Paul Cook
Patrick Drouin
Kseniya Egorova
Edward Finegan
Thierry Fontenelle
Polona Gantar
Yongwei Gao
Radovan Garabik
Alexander Geyken
Antton Gurrutxaga
Kris Heylen
Miloš Jakubíček
Jelena Kallas
Ilan Kernerman
Maria Khokhlova
Svetla Koeva
Iztok Kosem
Vojtěch Kovář
Simon Krek
Michal Kren
Jette Kristoffersen
Margit Langemets
Lothar Lemnitzer
Robert Lew
Pilar León-Araúz
Nikola Ljubešić
Henrik Lorentzen
Tinatin Margilitadze
Stella Markantonatou
John P. McCrae
Amalia Mendes
Michal Boleslav Měchura
Julia Miller
Victor Mojela
Monica Monachini
Orion Montoya
Sara Može
Christine Möhrs
Chris Mulhall
Carolin Müller-Spitzer
Roberto Navigli
Lionel Nicholas
Vincent Ooi
Noam Ordan
Magali Paquot
Danie Prinsloo
Adam Rambousek
Michael Rundell
Balint Sass
Roser Sauri
Jane Solomon
Egon Stemle
Kristina Štrkalj Despot
Arvi Tavast
Carole Tiberius
Yukio Tono
Lars Trap Jensen
Agnes Tutin
Tamás Váradi
Serge Verlinde
Elena Volodina
Piotr Zmigrodzki

TABLE OF CONTENTS

From Thesaurus to Framenet

Sanni Nimb[1], Anna Braasch[2], Sussi Olsen[2], Bolette Sandford Pedersen[2], Anders Søgaard[3]

[1]The Society for Danish Language and Literature, Chr. Brygge 1, 1219 Copenhagen K
[2]University of Copenhagen, Centre for Language Technology, Njalsgade 136, 2300 Copenhagen S,
[3]University of Copenhagen, Department of Computer Science, Sigurdsgade 41, 2200 Copenhagen N
E-mail: sn@dsl.dk, braasch@hum.ku.dk, saolsen@hum.ku.dk, bspedersen@hum.ku.dk, soegaard@di.ku.dk

Abstract

High-quality semantic data from a Danish thesaurus linked with valency information from a Danish dictionary allows us to compile a frame lexicon (Berkeley FrameNet style) for Danish in a very efficient way. In the paper we present the thesaurus as well as the dictionary and argue that they both represent valuable background information for assigning semantic frames to the Danish vocabulary. The resulting partial frame lexicon is tested in an annotation task where the semantic role inventory from English is directly transferred and made available for annotations of Danish. While simply aiming at reaching the highest possible frame coverage of the Danish vocabulary by reusing existing English frame and role inventories, we discuss the advantages and the drawbacks of the proposed method. The gained experiences from the work will be considered when scaling up the framenet resource to cover all verbs.

Keywords: Thesaurus; FrameNet; frame lexicon, Danish; annotation

1. Introduction

This article describes how we combine information from a monolingual Danish dictionary, Den Danske Ordbog (henceforth DDO) and a newly compiled Danish thesaurus, "Den Danske Begrebsordbog" ('The Danish Concept dictionary', Nimb et al., 2014a, henceforth the thesaurus), in order to compile standardized lexical-semantic data in the form of a partial Danish Frame lexicon compliant with the Berkeley FrameNet (BFN). The partial lexicon is tested in an annotation task carried out on already sense-annotated corpus data. The results from the pilot test are used to provide feedback to our method before we scale up the frame lexicon to cover all verbs in the thesaurus (financed 2016–2017 by the Carlsberg Foundation). We ask ourselves the following questions: How satisfying is the coverage of the generated frame lexicon based on thesaurus data, and how well can the roles described for English cover the semantics of Danish sentences?

Figure 1 illustrates the inter-linked background data.

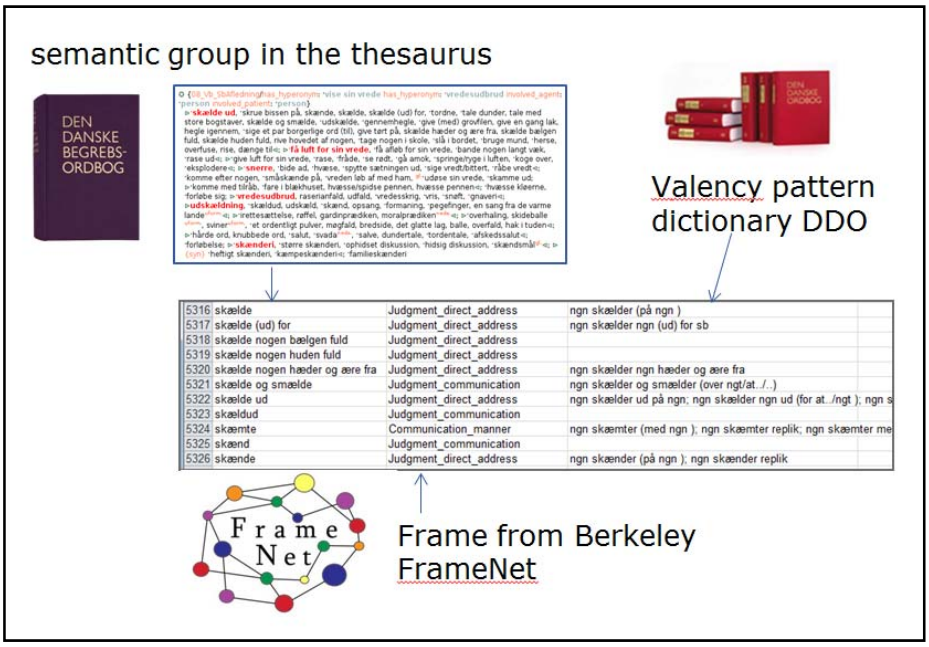


Figure 1: Linked data: The word groups in a Danish thesaurus combined with the valency information in a Danish dictionary constitute the background for the framenet.

We first introduce the research project of which the pilot frame lexicon project is a subpart, including a presentation of the sense-annotated SemDaX corpus that has been established in the project and which guides the choice of semantic coverage of the lexicon we compile. In Section 3 we discuss how role semantic information supplements the semantics of sense annotations and argue that the BFN model is well-suited for our purpose. In Section 4 we present the lexical data we use from the dictionary and the thesaurus and present our method for compiling Danish frame data. We furthermore describe how we tested the frame lexicon in an annotation task. In Section 5 we discuss the results: Finally we draw an overall conclusion and outline future plans in Section 6.

2. The Danish FrameNet in a broader context

Our method has evolved within a research project on semantic processing ("Semantic Processing across Domains", financed by the Danish Research Council 2013–2017) where several annotation tasks were carried out and used in machine learning experiments (Pedersen et al., 2014; 2016). The project focuses on Danish as a relatively low-resourced language and aims at increasing the level of semantic resources available for the Danish HLT community. A primary project goal is to provide semantically-annotated text corpora of Danish and to let these serve as training data for advanced machine learning algorithms which particularly address data scarcity and domain adaptation as central focus points. A corpus of 100,000 words has been sense-annotated with so-called supersenses (cf. Martínez Alonso et al.,

2016) and a smaller part of this has been annotated with semantic roles (frame elements) based on the frame lexicon that we describe below. The supersense annotations guided the first selection of relevant corpus data for our pilot frame semantics study on cognition and communication events.

2.1 The SemDaX corpus

The supersenses used to annotate the SemDaX corpus are based on the Princeton Wordnet lexicographical classes[1] which have become an international standard in coarse-grained sense tagging. The number of annotated sentences in SemDaX is 3,300, of which 60% have been annotated by two or more annotators, based on which a gold standard was developed. The SemDaX corpus[2] consists of various textual domains: newswire, blogs, chat, forum, magazine and written Parliament debates (Martínez et al., 2015; Olsen et al., 2015).[3]

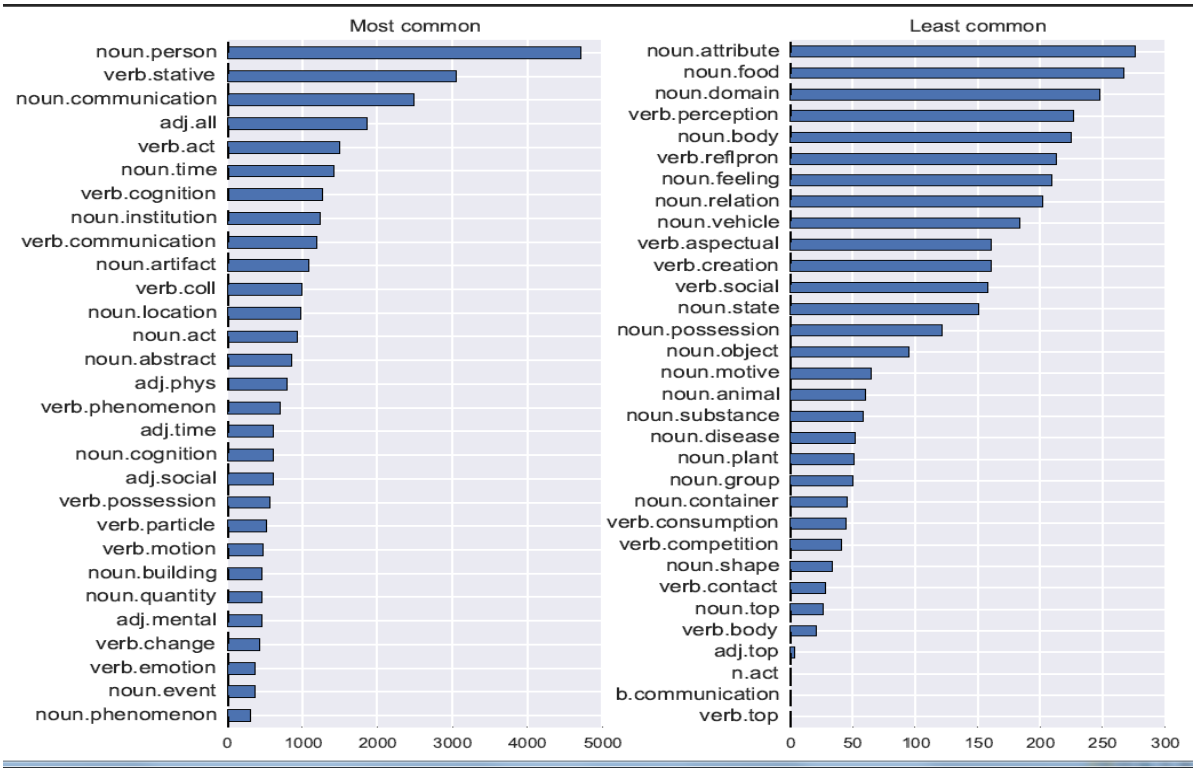


Figure 2: Most and less frequent supersenses in the complete annotated corpus (cf. Olsen et al., 2015)

[1] Cf. https://wordnet.princeton.edu/man/lexnames.5WN.html.

[2] Available for research at https://github.com/coastalcph/semdax.

[3] The texts have been extracted from the CLARIN reference Corpus, Asmussen 2012.

The most frequent supersenses in the corpus across word classes are 'noun.person', 'noun.communication' and 'verb.stative' (mainly constituted by the verb *være* (to be)), followed by supersenses for act, time, cognition and communication. It is interesting that the supersenses have a very different distribution across the various textual domains, revealing to a certain degree what the texts are mostly about. The supersense 'noun.person' is the most frequent in newswire and magazines, but much less frequent in chats, where the most frequent supersense instead is 'verb.stative' mainly constituted by the verb *være* (to be). Abstract supersenses such as 'noun.abstract' and 'noun.act' are much more frequent in Parliament debates than in the other text types. The least frequent supersenses in the corpus are either very specific ones, e.g. 'verb.body', 'verb.competition', 'noun.plant' and 'noun.disease', or abstract supersenses that the annotators, judged by the low inter-annotator agreement, found difficult to understand, such as 'noun.attribute', 'noun.relation' and 'noun.domain'.

A point of great interest to our lexicon project is the frequency of the verb supersenses. Apart from the supersenses 'stative' and 'act', 'verb.cognition' and 'verb.communication' are the most common, and put together these two categories are as frequent as the most frequent verb category, 'verb.stative'.

2.2 Selecting the frame lexicon vocabulary from the thesaurus

The supersense annotations in SemDaX enabled us to focus directly on very frequently occurring events describing communication and/or cognition. This choice was based on a comparison of the most likely supersenses of verbs in the thesaurus chapters, see Table 1, with the frequency of the different supersenses in SemDaX as illustrated in Figure 2.

The chapters which contain a rather high number of verbs and verbal nouns compared to the average of 2% are the following: '5 Relation, property', '8 Location, motion', '9 Volition, act', '10 Emotions', '11 Thinking', '12 Communication','15 Social life', and '21 Economy, finances'. A comparison with the most frequent supersenses of verbs in Figure 2 ('stative', 'communication', 'cognition', and 'act') led to the decision that in order to obtain enough sentences to annotate, the best choice would be the chapters '11 Thinking', '12 Communication' and parts of chapter '13 Science´and '15 Social life' which we, based on our detailed knowledge of the thesaurus, estimate to contain mainly the very frequent supersenses 'cognition' and 'communication'. Although 'act' verbs are typically found in chapter '9 Volition, act', they are likely to also occur in a large variety of other chapters and therefore not suitable for our task. The chapters '8 Location, motion', '10 Emotions' and '20 Economy, finances' were discarded because the corresponding supersenses 'verb.motion', 'verb.emotion' and 'verb.possession' are not among the most frequent in Figure 2. Chapter 5 was discarded even though it contains many stative verbs which are frequent in texts,

simply due to the fact that the BFN model focuses on the part of the vocabulary describing human activity.

Chapter in thesaurus	**Percentage of all verbs and verbal nouns**	**Expected to contain verbs with the following supersense:**
1 Natur og miljø (nature, environment)	1,4 %	phenomenon, act
2 Liv (life)	5 %	phenomenon, stative, body
3 Rum, form (space, form)	2,5 %	change, stative, contact
4 Størrelse, mængde, tal, grad (size, amount, number, degree)	4 %	change, quantity, relation
5 Forhold, egenskab (relation, property)	6,6 %	stative, phenomenon, relation, change, aspectual
6 Tid (time)	2,7 %	Time
7 Sanseindtryk, tilstandsformer (sense impression, material state)	4,1 %	Perception
8 Sted og bevægelse (location, motion)	9 %	Motion
9 Vilje og handling (volition, act)	11,8 %	Act
10 Følelser (emotions)	8.4 %	Emotion
11 Tænkning (thinking)	7 %	Cognition
12 Tegn, meddelelse, sprog (communication)	6 %	Communication
13 Videnskab (science)	1,4 %	Cognition
14 Kunst og kultur (arts, culture)	1,7 %	Creation
15 Socialt liv (social life)	8,6 %	social, competition, communication
16 Mad og drikke (food and drinks)	1,7 %	Consumption
17 Sport og fritid (sports and leisure)	3,6 %	body, creation, motion, competition
18 Samfund (society)	5,1 %	Social
19 Apparater, teknik (artifacts/instruments, technique)	3 %	creation, communication
20 Økonomi, finans (economy, finances)	7,1 %	possession, social
21 Ret, etik (law court, ethics)	2,1 %	Social
22 Religion (religion)	0,5 %	Cognition

Table 1: Number of verbs and verbal nouns in the 22 thesaurus chapters, and their estimated supersense types. They constitute a total of 44,607 word and expressions (=20% of whole thesaurus)

3. FrameNet as semantic model

While supersense annotations supply us with very coarse-grained semantic information at sense level, role-oriented semantic annotations are needed if we want to label in a formalized way who does what, where and when. An ongoing discussion in the Danish group has been whether to adopt a deep-syntactic approach to role-labeling as taken in PropBank (Palmer et al., 2005) and VerbNet (Schuler 2005) or a more semantically-driven, frame-based approach to roles as provided by BFN, where both the frame inventory and the frame elements describe verb semantics at quite a detailed level: what kind of act (of about 1,000 possible) is carried out, and who are the participants (e.g. speaker and addressee). Figure 3 shows the BFN interface with descriptions of frames, English lexical units and search facilities.

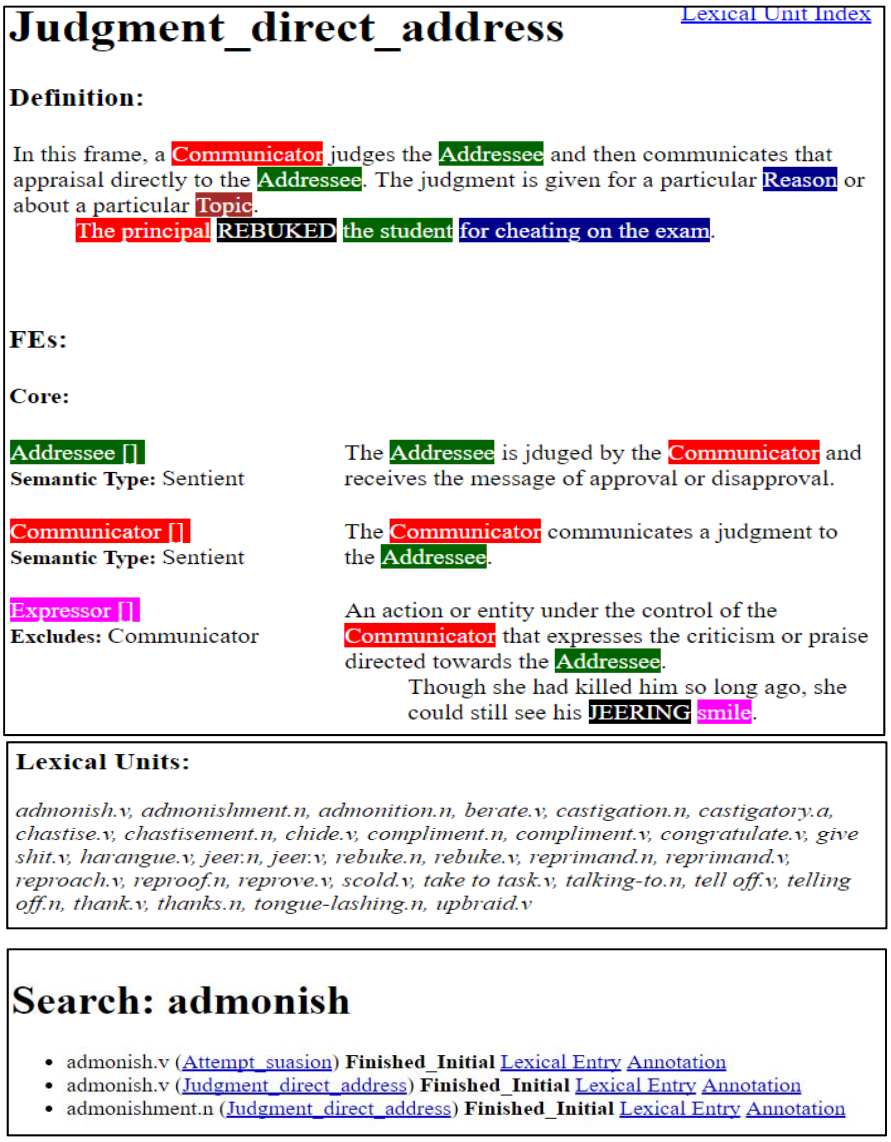


Figure 3: Frame description from BFN (the frame Judgment_direct_address), including lexical units and also the search facility where different frames of the same verb, here *admonish*, are presented, one of which is the above frame. Cf. Berkeley FrameNet

In recent years, frame-semantic *parsing* has received increased interest in the NLP community, and in spite of BFN's relatively fine-grained inventory of frames and frame elements, this approach has also proven manageable in practical tasks (cf. Section 2a). Frame-semantic parsing was introduced to the NLP community in SemEval 2007, with the introduction of a standard bench-marking corpus for English. Parsing models, such as the two-stage parsing model of Dipanjan Das et al. (2014), have been applied to various tasks, both within research and industry. Two examples of tasks that benefit greatly from frame-semantic parsing are knowledge base population (Søgaard et al., 2015) and document summarization (Schluter & Søgaard, 2015). Frame-semantic parsing is also likely to instigate break-throughs in question answering, relation extraction, and dialogue systems. Consequently, framenets are currently being built for a number of languages since it is seen as an important resource in a particular language's composite set of HLT resources. However, one major bottleneck for the application of frame-semantic parsers is still the lack of resources for many languages. Johannsen et al. (2015) therefore discusses cross-lingual adaptation of frame-semantic parsing models induced from the English corpus, to other languages such as Danish, German and Greek. While such work can potentially make the above technologies available for languages other than English, the models developed in Johannsen et al. (2015) were evaluated by using datasets that were not adjudicated, and where annotators did not have access to associations between trigger words and frames in the target languages. In comparison, our method suggests that annotators are presented with a list of the most *likely* frames to choose from.

Taking both the BFN as well as a semantic resource of the target language as starting points for the development of a new framenet, is not in itself a novel approach. Swedish FrameNet (Heppin & Gronostaj, 2012; 2014) applies BFN as the initial structural backbone of the resource but bases the sense inventory on a monolingual Swedish resource, SALDO. In contrast, other framenets like Japanese FrameNet (Ohara 2014) and French FrameNet (Candito et al., 2014) rely more solely on a lexical mapping from BFN, enriching and supporting the resource subsequently with corpus data in the target language.

4. Compilation of a Danish Frame Lexicon

The thematic divisions in the thesaurus allow us to identify and extract large groups of near synonymous verbs within our "pilot" fields, communication and cognition. The thesaurus covers approx. 200,000 words and expressions, covering 80% of the approx. 136,000 senses described in DDO (Nimb et al., 2014b). DDO was compiled as a printed dictionary in the 90s. Today the dictionary is online and continuously extended with new words and expressions.

The thesaurus is divided into 22 named chapters and 888 named sections inspired by the division in Dornseiff (2004), but adjusted to the Danish language community of today. Each section arranges the DDO vocabulary according to semantics in lists of

synonyms and near synonyms. In the source document (not in the printed book) the lists of synonyms and near synonyms are clustered in 8,300 coarse-grained semantic groups across word classes in an annotated XML structure, making it possible to identify and extract large semantic groups of words of the type 'person', 'artifact', 'event' etc. in each named section. By the use of these formal annotations we extracted all groups described with the semantic relation 'involved agent' in the chapters '11 Thinking', '12 Communication' and furthermore some sections in '13 Science' (concerning studies and science) and '15 Social life' (Sections like '15.19 Acknowledgement', '15.20 Flattery', and '15.24 Scolding' with many communication verbs). We assumed that to a large extent these sections together would cover the verb vocabulary of cognition and communication, and thereby also the verbs annotated with these supersenses in SemDaX.

The 'involved agent' groups in the thesaurus include both verbs and verbal nouns, but since verbal nouns are annotated with a broad supersense 'noun.communication' covering both the act sense and the result, as well as semiotic artifacts in SemDaX, they are not automatically identifiable in the corpus, and we chose not to include them in the annotation task. In the lexicon, the verbal nouns are assigned frames corresponding to the verbs from which they are derived.

In Figure 4 we present an 'involved agent' group from the XML document.

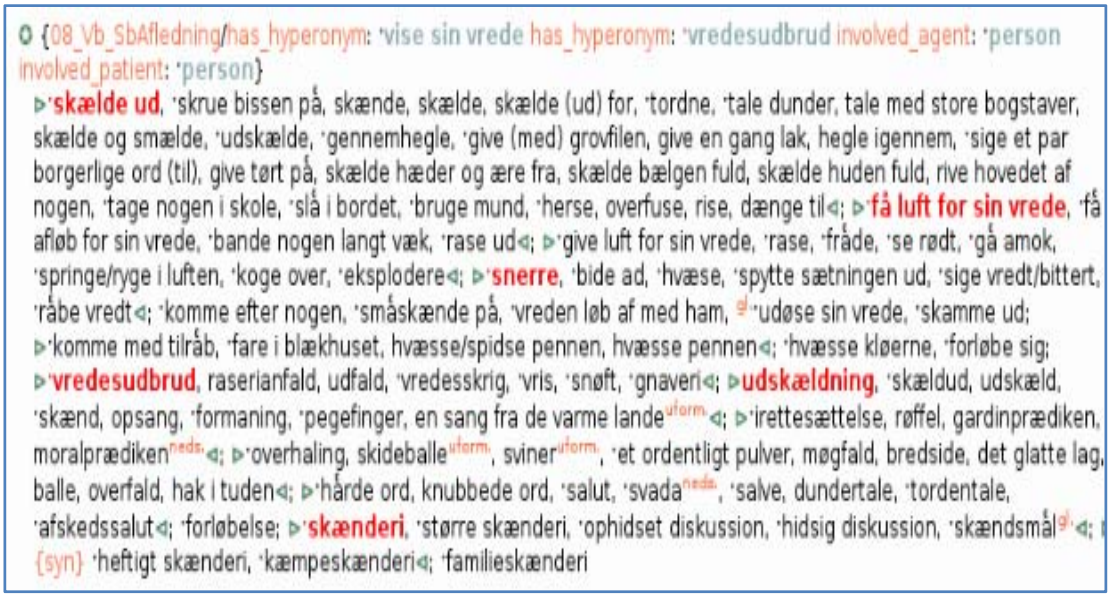

0 {08_Vb_SbAfledning/has_hyperonym: 'vise sin vrede has_hyperonym: 'vredesudbrud involved_agent: 'person involved_patient: 'person}
▷'skælde ud, 'skrue bissen på, skænde, skælde, skælde (ud) for, 'tordne, 'tale dunder, tale med store bogstaver, skælde og smælde, 'udskælde, 'gennemhegle, 'give (med) grovfilen, give en gang lak, hegle igennem, 'sige et par borgerlige ord (til), give tørt på, skælde hæder og ære fra, skælde bælgen fuld, skælde huden fuld, rive hovedet af nogen, 'tage nogen i skole, 'slå i bordet, 'bruge mund, 'herse, overfuse, rise, dænge til◁; ▷'få luft for sin vrede, 'få afløb for sin vrede, 'bande nogen langt væk, 'rase ud◁; ▷'give luft for sin vrede, 'rase, 'fråde, 'se rødt, 'gå amok, 'springe/ryge i luften, 'koge over, 'eksplodere◁; ▷'snerre, 'bide ad, 'hvæse, 'spytte sætningen ud, 'sige vredt/bittert, 'råbe vredt◁; 'komme efter nogen, 'småskænde på, 'vreden løb af med ham, gl 'udøse sin vrede, 'skamme ud; ▷'komme med tilråb, 'fare i blækhuset, hvæsse/spidse pennen, hvæsse pennen◁; 'hvæsse kløerne, 'forløbe sig; ▷'vredesudbrud, raserianfald, udfald, 'vredesskrig, 'vris, 'snøft, 'gnaveri◁; ▷udskældning, 'skældud, udskæld, 'skænd, opsang, 'formaning, 'pegefinger, en sang fra de varme lande uform◁; ▷'irettesættelse, røffel, gardinprædiken, moralprædiken neds◁; ▷'overhaling, skideballe uform, sviner uform, 'et ordentligt pulver, møgfald, bredside, det glatte lag, balle, overfald, hak i tuden◁; ▷'hårde ord, knubbede ord, 'salut, 'svada neds, 'salve, dundertale, 'tordentale, 'afskedssalut◁; 'forløbelse; ▷'skænderi, 'større skænderi, 'ophidset diskussion, 'hidsig diskussion, 'skændsmål gl◁; {syn} 'heftigt skænderi, 'kæmpeskænderi◁; 'familieskænderi

Figure 4: 'Involved agent group' from the Danish Theaurus. The header contains annotations and introduces a large list of verbs and verbal expression with the sense 'skælde ud' ('to scold', initiated by 'skælde ud') followed by a list of verbal nouns with the same sense (initiated by 'vredesudbrud')

As stated above, each word and expression in the thesaurus is linked to a DDO sense via a common identification number; this opens up a large variety of combined lexical

data across the two resources, one of which we exploit here by transferring valency patterns from DDO to the verbal groups in the thesaurus.

We extracted approx. 7,000 words and expressions, constituting about 16% of all verbs and verbal nouns in the thesaurus XML document (see Table 1 above). This indicates that we find many synonymous and near synonymous words and expressions within the semantic areas of cognition and communication. There seems to be some kind of parallel between frequency in Danish texts and frequency in the Danish lexicon, also when we compare other chapters in Table 1 with the supersense frequencies in Figure 2. When we often talk about a theme or concept it seems to influence the variety of words and expressions that we use in order to do it.

In Table 2 we see the extract of the same data, now supplied with valency patterns from DDO via the shared identification numbers, and supplied with the information on the corresponding frame in BFN.

section title, word/expression (= 'to scold') from the thesaurus		**shared ID number**	**valency pattern from DDO**	**frame from BFN**
Skælde ud	*skælde ud*	21074700	ngn skælder ud på ngn; ngn skælder ngn ud (for at/ngt); ngn skælder (ngn) ud over ngt/at	Judgment_direct_address
Skælde ud	*skrue bissen på*	21074701	NGN skruer bissen på (over for NGN)	Judgment_direct_address
Skælde ud	*Skælde*	21010806	ngn skælder (på ngn)	Judgment_direct_address
Skælde ud	*skælde (ud) for*	21033375	ngn skælder ngn (ud) for sb	Judgment_communication
Skælde ud	*skælde nogen bælgen fuld*	21034458	NONE	Judgment_direct_address
Skælde ud	*skælde nogen huden fuld*	21074699	NONE	Judgment_direct_address
Skælde ud	*skælde nogen hæder og ære fra*	21090433	ngn skælder ngn hæder og ære fra	Judgment_direct_address
Skælde ud	*skælde og smælde*	21074701	ngn skælder og smælder (over ngt/at)	Judgment_communication

Table 2: Lexical units from the thesaurus linked to valency patterns from DDO and supplied with frames from BFN

By focusing on one semantic area at a time (made possible via the chapter grouping in the thesaurus), the lexical data considered are likely to be assigned the same frame, or at least a closely related frame, from BFN. In the work process, the Danish word or expression is translated to an English equivalent (via Gyldendal's Danish English Dictionary), and the equivalent (or a more common synonym) is searched for in the lexical unit index of BFN, leading to one or more frame possibilities, see Figure 5. The frame description is studied carefully before it is assigned, see Figure 3 above. It has to be verified whether it covers the Danish lexical unit *skælde ud*, e.g. by comparing the Danish valency pattern and the role inventory of the frame.

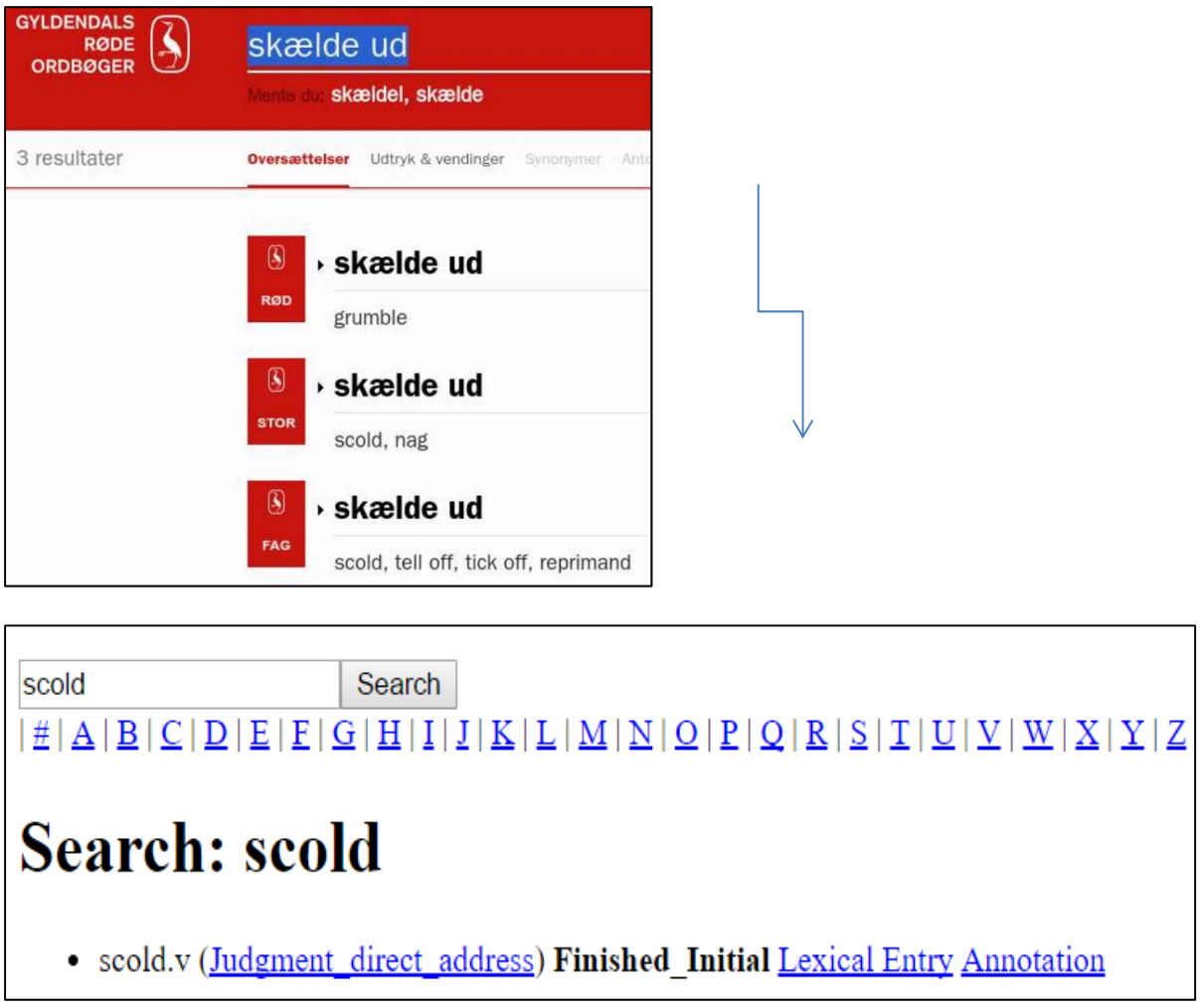


Figure 5: Translation of the Danish verb *skælde ud*. Equivalent 'to scold' used as input to manual search for a relevant frame in BFN (Judgment_direct_address, see Figure 3)

To cover approx. 3,500 words and expressions describing the semantic area of communication (Chapter 12 and part of 15), we used the following 52 BFN frames: Be_in_agreement_on_action, Be_in_agreement_on_assessment, Attempt_suasion, Attention, Become_silent, Bragging, Chatting, Commitment, Communicate, Communication_manner, Communication_noise, Communication_response, Contacting,

Deny_permission, Discussion, Education_teaching, Encoding, Gesture, Going_back_on_a_commitment, Grant_permission, Hearsay, Intentional_deception, Judgement_communication, Judgment_direct_address, Justifying, Label, Linguistic_meaning, Manipulate_into_doing, Mention, Name_conferral, Permission, Prevarication, Publishing, Quarreling, Questioning, Reading_aloud, Reassuring, Reporting, Request, Respond_to_proposal, Response, Reveal_secret, Silencing, Spelling_and_pronouncing, Statement, Suasion, Summarizing, Telling, Text_creation, Translate, Verification, Warning.

To cover approx. 2,600 words and expressions describing the semantic area of cognition (Chapter 11 and part of 13), we used the following 54: Adding_up, Adducing, Annoyance, Attention, Awareness, Becoming_aware, Categorization, Certainty, Cogitation, Coming_to_believe, Coming_up_with, Correctness, Creating, Differentiation, Education_teaching, Estimating, Evoking, Examination, Expectation, Experiencer_focus, Experiencer_obj, Experimentation, Feigning, Grant_permission, Grasp, Intentional_deception, Intentionally_act, Judgment, Just_found_out, Linguistic_meaning, Make_cognitive_connection, Manipulate_into_doing, Memorization, Memory, Mental_property, Opinion, Perception_active, Purpose, Questioning, Reading_activity, Reading_perception, Reasoning, Regard, Reliance_on_expectation, Remembering_experience, Remembering_information, Remembering_to_do, Research, Resolve_problem, Reveal_secret, Scrutiny, Sign, Topic, Trust.

In both cases the number of used frames constitute only about 5% of the 1,073 frames described in BFN. By focusing on only one semantic area at a time—first communication, then cognition—we made it possible for the lexicographer to gain confidence in the different frame descriptions, enabling her to distinguish between semantically closely related frames and to carry out a more homogenous assignment of frames. The information on valency patterns from DDO was crucial when it came to the lexicographer's clarification of the scenario in question in Danish, and her choice of exactly the one English frame which would cover the sense and the connected constituents as described in the valency pattern in the best way.

4.1 The annotation task

For the annotation task, sentences in SemDaX with verbs already annotated with the supersenses cognition and/or communication were extracted and assigned frames.

The annotation tool by Johannsen presents the annotator with the corresponding frame of the verb which has to be confirmed or rejected. In case of more than one frame for a given verb, the set of frames are listed and the annotator selects the right one after having checked the lexicon (which often presents the verb with different collocates, e.g. the verb *indsamle* ('to collect') with the noun *viden* ('knowledge') in Table 3) or/and BFN.

Danish lexical unit	**Frame from BFN**	**Danish valency pattern**
indsamle viden (lit. 'collect knowledge' ('study'))	Scrutiny	ngn indsamler ngt
indse ('understand')	Be_in_agreement_on_assessment	ngn indser at sætn/ngt
indse ('realize')	Coming_to_believe	ngn indser at sætn/ngt
indse ('realize')	Coming_to_believe	ngn indser ngt
indskole ('do introductory schooling')	Education_teaching	
indskrive ('register/inscribe')	Text_creation	ngn indskriver ngn/ngt
indskyde ('add')	Mention	ngn indskyder ngt/at sætn

Table 3: Alphabetic extract from the lexical unit index of Danish words and expressions

Once the most appropriate frame is selected, its role inventory (transferred from BFN to the annotation tool) is studied in the BFN descriptions of the frames (in case of doubts) and used for annotation, based on the assumption that the inventory covers the set of Danish roles as well due to the relative similarity between the two languages and linguistic communities.

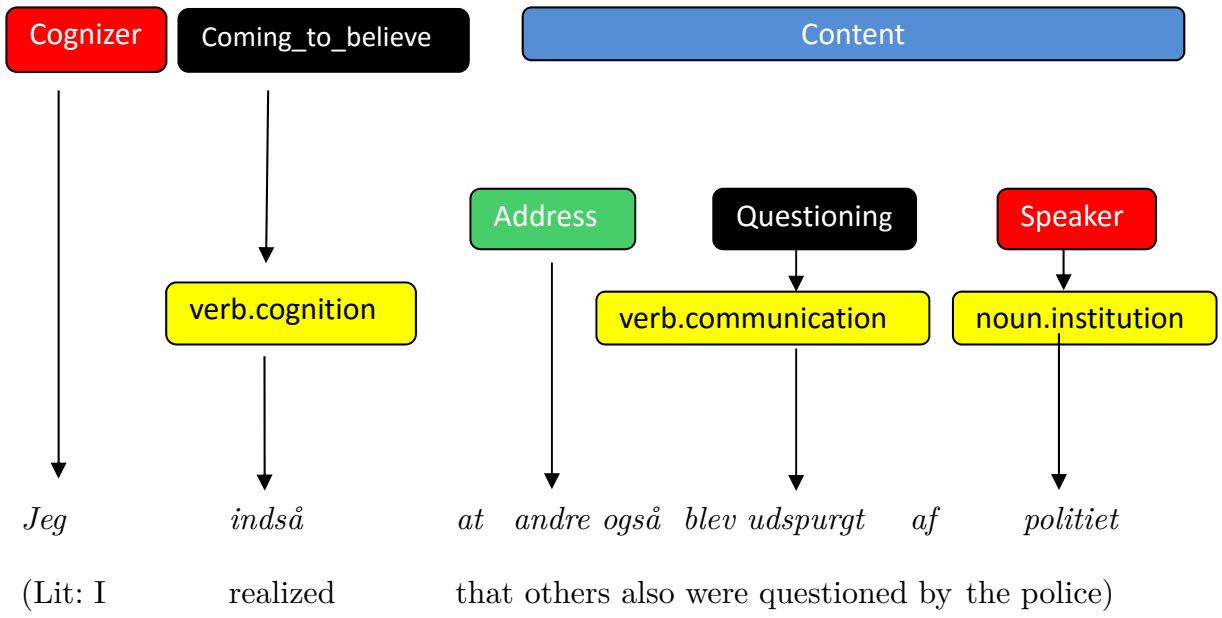


Figure 6: BFN frames and roles annotated on top of Danish supersense annotations

Regarding the sentence in Figure 6: "*Jeg indså at andre også blev udspurgt af politiet*" ('I realized that others were also questioned by the police'), the annotator is presented to two options (via the annotation tool) for *indse* ('realize'/supersense verb.cognition), namely Be_in_agreement_on_assessment and Coming_to_believe. The latter is

chosen (after having checked the lexical unit index in Table 3 in case of doubt), and the core roles of the frame are studied in BFN and annotated in the sentence as well, in this case Cognizer ("Jeg" ('I')) and Content (the complement clause "*at andre også blev udspurgt af politiet*" ('that others were also questioned by the police')). Furthermore the main verb *udspurgt* ('to question', 'to pump', verb.communication) is annotated with the frame Questioning, of which the present roles in the phrase are Speaker (*politiet* ('the police')) and Addressee (*andre* ('others')).

In total, 440 cognition and communication verbs in SemDaX were annotated and will later be used in different machine learning experiments.

5. Discussion of method

We argue that the very fact that DDO is corpus based—as is the thesaurus since it uses DDO as its lexical backbone—makes both resources well qualified as background resources for creating lexical frames. But the method also has some pitfalls, as we will demonstrate.

5.1 The advantages and disadvantages of using the DDO valency patterns

DDO is corpus based. This includes the description of the valency patterns which is established on the study of a set of randomly chosen concordance examples, typically 100–200 sentences, for high-frequent verbs with many senses, up to 1,000 examples. One could thereby claim that the valency patterns function as a sort of condensed extract of the verbs' linguistic behavior in real text, including the semantic roles they typically occur with, similar to that for which we would expect to seek and annotate in the SemDaX corpus. They contribute with very important information when the frame lexicon is compiled. But a drawback is the differences between SemDaX and the corpus used to compile DDO in the 90s. The sentences we annotate constitute newer texts (2008–2011) and cover a wider range of (new) text domains than does DDO, such as blogs and chat from the Internet.

The valency patterns in DDO describe to the dictionary user whether the verb in the same sense also might be construed as a phrasal verb with a particle (presented in brackets), whether the constituents of the verb are facultative (presented in brackets) or not, whether they are introduced by an obligatory or facultative preposition, selectional restrictions such as 'person' or 'not person', or maybe instead a phrase or an infinitive construction. Sometimes additional selectional restrictions are mentioned, e.g. 'animal'. The sense of a verb might even have several valency patterns, each of them with facultative complements or particles. The patterns aim at making the dictionary user able to construct well-formed sentences in Danish with the verb in question, but they are not described by an unambiguous, formalized pattern; they depend on human interpretation.

When used in combination with the semantic grouping from the thesaurus to compile the frame lexicon, the valency patterns function as a clear indicator of which type of frame to assign from BFN. The different patterns of a semantic group also support one another, making the picture even clearer. The constituents in the patterns are strongly connected to the (core) roles described for each frame in BFN. Altogether the exact scenario evoked by the Danish word in question becomes quite clear through the comparison of valency descriptions and the frame.

Valency pattern in DDO	**Lit.**	**English equivalent**	**Frame from BFN**
NOGEN skælder (NOGEN) ud over NOGET/at	somebody scolds (somebody) out over something / that = somebody scolds (somebody) because of something/ because he/she	somebody scolds somebody because of something/ because he/she	Judgment_direct_adress
		somebody nags about something/that somebody	Judgment_communication
NOGEN skælder ud på NOGEN	somebody scolds out at somebody	somebody scolds somebody	Judgment_direct_adress
NOGEN skælder NOGEN ud (for at/NOGET)	somebody scolds somebody out (for that /for something	somebody scolds somebody (for doing) (for something)	Judgment_direct_adress

Table 4. The valency patterns of *skælde ud* ('scold') in DDO is complex, involving several facultative complements, and it is therefore likely that the Danish verb is to be assigned more than just one BFN frame

It is important to underline that there is no one-to-one correspondence between senses and valency patterns in DDO on the one side, and frames in BFN on the other side. The same sense of a verb in DDO might be assigned more than one frame in our

lexicon. It complicates the process that the choice of frame might depend on whether or not facultative elements of the valency pattern correspond to semantic roles. The phrasal verb skælde ud (lit. 'scold out' ('scold' as in 'scold somebody for something')) is such a case, as shown in Table 4. BFN distinguishes between scenarios where somebody is criticizing a person directly in front of him or her. In this case the frame is Judgment_direct_adress. Scenarios where somebody is talking negatively about something, e.g. what a person who is not present did (= nagging about somebody) the frame is instead Judgment_communication. The Danish verb skælde ud covers both senses (as seen in Table 4, Gyldendal translation), and only the presence of specific semantic roles clarifies the sense in question. It is not clarified in the definition of the word sense in DDO that this is the case. In many cases the thesaurus presents such ambiguous senses in more than one section. E.g. skælde ud is also mentioned in chapter '10 Emotions' in the section '10.26 Unsatisfied' together with other verbs with the sense 'to complain', 'to nag', and would have been assigned the frame Judgment_communication if words from this section had been included in our frame lexicon vocabulary.

5.2 The advantages and disadvantages of using the thesaurus as input to a framenet

The thesaurus presents a large variety of lexical data in the form of extensive lists of near synonymous words and multiword units. It often displays the same sense of DDO in different, more or less fixed expressions. Thereby the thesaurus supplies us with far more multiword units than does DDO. E.g. in the case of facultative particles in the valency patterns, the thesaurus presents two lexical units where DDO only provides us with one. The DDO verb sense of *printe* ('to print'/'to print out') with the valency pattern "NGN printer NGT (ud)" ('somebody prints something (out)') thereby results in two synonymous lexical units, corresponding to print and print out in English, listed together in the thesaurus in the same semantic group with other synonymous verbs (*udprinte*, *udskrive* and *skrive ud*).

Given that DDO is corpus-based, the lexical data represents a small 'summary' of the behavior of the verb in real text, in line with the valency patterns but more focused on the lexical semantic restrictions.

Add to this that the thesaurus very often covers several aspects of a DDO sense by presenting it in more than just one section or chapter. Thereby it also sums up the different aspects of a word quite similar to that which we would probably discover by annotating large amounts of text (as it is done in the BFN project).

Furthermore, BFN is in many ways similar to a thesaurus as also stated in Ruppenhofer et al. (2016): "*Each lexical unit is linked to a semantic frame, and hence to the other words which evoke that frame. This makes the FrameNet database similar*

to a thesaurus, grouping together semantically similar words". But it is important to underline that, although we find some consistencies between section divisions across the two resources, the thesaurus and BFN are profoundly very different in their way of dividing the vocabulary into sections and chapters, and frames, respectively. BFN has 'scenarios' and the core role inventory of these as the overall division criteria. As stated in Ruppenhofer et al. (2016), "*The frames represent story fragments, which serve to connect a group of words to a bundle of meanings; for example the term avenger evokes the Revenge frame, which describes a complex series of events and a group of participants*". As an example, BFN does not distinguish between negative and positive directly expressed judgments: *to compliment* and *to scold* both evoke the frame Judgment_direct_address. The 'story', or 'scenario', as well as the participants are the same; in both cases we deal with a judgment scenario. As a consequence, it distinguishes between scenarios where the participants are not the same: when a person complains about somebody who is not present and thereby not constituting the role of the addressee, the evoked frame is Judgment_communication, but when the person complained about at the same time is the addressee in the scenario, the evoked frame is Judgment_direct_adress. Likewise, antonymous words describing the same type of cognitive event, such as the verbs 'to forget' and 'to remember', are also considered to belong to the same frame. In other words, the same frame is evoked by lexical units no matter whether these are negated or not in the phrase.

In contrast, the thesaurus divides the vocabulary according to domains (football, food, movies), but also according to traditional sense division criteria. Antonomy is an important aspect, and in some chapters most of the sections could be seen as having opposite meanings to one another, covering concepts of 'thin' as opposite to 'thick', 'angry' opposite to 'happy', 'early' to 'late', 'strong' to 'weak' etc. This is also the case for cognition and communication verbs in Chapters 11, 12, 13 and 15. E.g. the Danish lexical units of Judgment_direct_adress are found in different sections such as '15.19 Approval', '15.20 Flattering' and '15.24 Scolding'. Likewise, the thesaurus contains the two sections '11.37 Remembering' and '11.38 Forgetting' while Framenet, as stated above, has only one frame covering both, namely Remembering_information. Table 5 describes the different division criteria in the two resources. Svendsen (2017: 26) proposes that we should consider adopting the method suggested by the Swedish FrameNet project (Friberg Heppin & Gronostaj 2012) who split up such frames according to positive and negative meanings, due to the fact that the Swedish lexical resource (SALDO), just like the thesaurus distinguishes clearly between such senses.

Not surprisingly we had some cases of Danish verbs that were difficult to assign an English frame. The verbs *misforstå* ('misunderstand'), *mistolke* ('misinterpret') and near synonymous words are some of these cases. Svendsen (2017) points out other problems of the language transfer method, e.g. caused by reading too much meaning into the BFN frames when they are assigned to the Danish vocabulary. We will not study and discuss in this paper whether the problems are due to differences between the Danish and English vocabulary, or rather to the fact that BFN is still being

developed and therefore does not cover all possible scenarios yet. In general, we found that the Danish semantic areas we chose were in fact surprisingly well-covered in BFN, but we also found a certain lack of frames concerning what you could describe as 'acts one does not carry out', like *undlade* ('to leave undone') or 'acts one does not succeed with', like *overvurdere* ('to overestimate, to overrate'). Also frames for domain terms were missing, like *anonymisere* ('to anonymize'); naturally BFN does not yet cover all types of domains and terminology. When the whole thesaurus has been assigned frames, we will study the vocabulary left without frame assignment. Likewise it will be necessary to look at BFN frames which have not been applied to any Danish verbs.

Criteria to division in the thesaurus → Criteria to division in BFN ↓	15.19 Anerkendelse ('approval') only positive includes both talking about and talking directly to the person	15.24 Skælde ud ('to scold') only negative Includes both talking about and talking directly to the person
Judgement_communication Both positive and negative Not directly to judged person	*berømme* ('to praise')	skælde og smælde ('to nag'), *bande langt væk* ('curse somebody up and down')
Judgement_direct_adress Both positive and negative Directly to judged person	*komplimentere* ('to compliment')	*overfuse* ('heap/pour abuse on'), *gennemhegle* ('to dress someone down')

Table 5: BFN and DT use different criteria when dividing into frames and sections respectively

5.3 Frame and role coverage in the annotation task

Before initiating the annotation task, we studied the list of the approx. 1,600 verbs in SemDaX which are annotated as either cognition or communication. By doing so, we found that approx. 20% words at a first glance did not seem to belong to any of these semantic classes. Some had a much broader sense which was used with a communication or cognition sense in the corpus while depending on a very specific context; others were ad hoc figurative senses. Such cases are typically neither

represented in DDO nor in the thesaurus vocabulary. Dictionaries do not fully cover all senses of words as they are represented in corpora. When lexicographers describe the senses of a lemma, they focus on prototypical word use and normally discard senses with very low frequency. In the case of a set of quite similar, but rare sense instances, they try to merge them into one overall sense description whenever possible, and they normally discard ad hoc figurative use. Framenet projects like BFN and the Japanese FrameNet project which annotate texts instead of focusing on lexical units from a resource, do not encounter this problem. The cognition and communication verbs in SemDaX were by far the most cases described in DDO and the thesaurus, but some were presented in sections in the thesaurus which we did not consider to be relevant in the first place when we extracted communication and cognition groups. Chapter 19 of the thesaurus which covers artifacts and devices, and therefore also the sections on telephones and computers, is one such case. It describes an important part of the communication vocabulary which in BFN corresponds to the frames Communication_means and Contacting. These words will, in a future digital version of the thesaurus, be included in Chapter 12 on communication, and in this way, our project also gives feedback to the thesaurus project. The words and their corresponding frames were added to our lexicon before we initiated the annotation task.

If we turn to the results of the annotation task, the assignment was, in by far the most cases, easy to carry out and clearly facilitated by the reduced set of possible frames suggested by the annotation tool via the lexicon data. But approx. 20% of the cases gave us some interesting challenges. E.g. it turned out that some of the possible frames of the most frequent verbs in Danish were missing due to the fact that not all verb senses of highly frequent verbs with many senses in DDO are covered by the thesaurus. When the thesaurus was compiled, the aim was to include the highest number of different lemmas as possible and not to cover all senses of the same lemma as described in DDO. This has apparently led to a too narrow representation of some of the very frequent cognition and communication verbs in our pilot frame lexicon. When we expand it, these verbs will be assigned a bigger variety of frames according to their many senses in DDO. The thesaurus will once again benefit from the study: some highly polysemous verbs will have to be added to extra sections.

Interestingly enough, some verbs from the semantic area cognition in the SemDaX corpus turned out to have a communication sense. These verbs are not part of the communication vocabulary in the thesaurus since they depend so strongly on the linguistic context (they occur only together with direct speech/discourse), that it would be almost impossible to decode their communication sense for the user. One example is the verb *mene* ('to find', 'to think') as in "*Jo, vejret ser ud til at holde, mente han*" ('yes, the weather conditions seem to last, he found' (='he said'). We also find verbs from other semantic areas having communication senses in this context: *slutte* ('finish'), *fortsætte* ('to continue') and *begynde* ('to begin') as in "*Jeg har haft en drøm, begyndte han*" ('I had a dream, he started' (= 'started to say') and *gabe* ('to

yawn') as in "*nu må vi se at få sovet lidt, gabte moren*" ('now we ought to sleep, the mother yawned' (='said while she yawned')). In order to significantly improve our lexicon, we must fully cover such verbs which we have completely disregarded in the first place, since we focused entirely on the thesaurus vocabulary. Most of them are in fact easily identifiable by their valency pattern in DDO which describes the possibility of direct speech.

Once the correct frame was selected, the English role inventory proved to fulfill our requirements and was in fact rather easy to apply. Most phrases, however, contained rather few realised roles, for instance, the addressee was often absent in communication phrases.

While annotating the corpus sentences, another question arose: should the annotator stick to the most 'literal' frame that the verb evokes, which is normally also integrated in the frame lexicon, or should she rather try to represent the underlying meaning in the phrase? One example is the phrase: "*Nogle personer kan du lære at leve med, andre ikke*" ('Some people you are able to learn to live with, others you are not'). Should *lære* ('to learn') in this case be annotated with the frame Grasp (with the roles Cognizer and Phenomenon) or rather with the frame Tolerating (with the roles Experiencer and Content)? In the frame lexicon, only the collocation *lære at kende* ('get to know') is described, but not *lære at leve med* ('learn to live with'). We find that this case illustrates very well why the many collocations in the thesaurus are well-suited as input to a frame lexicon; in this case *lære at leve med* is candidate to occur in the thesaurus in the same group as verbs like *tolerere* (to tolerate) and its synonyms in a future version.

6. Conclusion

Overall we can conclude from our method that any possible frame of a word that the lexicographer would even think of when assigning the frames from BFN, should better be included in the lexicon right away in order to provide the annotators with a maximal set of frames for a given word. When the full thesaurus data (that is, verbs from all 888 sections) has been assigned frames, we hope to have covered a very large variety of frame possibilities of the DDO senses. Our annotation tool did not give access to the full set of frames in BFN, only to the frames assigned to each verb in our lexicon. Even though it was very clear that the predefined and manageable set made the distinctions between frames much easier to grasp and thereby facilitated the annotation process, we soon understood that it is necessary to have access to the full set of frames in BFN in order to also be able to annotate the ad hoc language use often found in corpora but not described in dictionaries.

The frame annotations are used to train a semantic parser; however, the number of annotated sentences (440) is currently rather small for this task, and we therefore plan an extension. We also plan to look deeper into the frequency of the different frames

and roles in the Danish texts, in order to compare frequency across text domains as has been done in the supersense annotation task.

7. Acknowledgements

The project "Semantic Processing across Domains" and the partial frame lexicon is financed by the Danish Research Council (2013-2017). The Carlsberg Foundation is funding the extension of the lexicon in 2016/2017.Thomas Troelsgård has carried out the extraction of XML data from the thesaurus and the dictionary. Anders Johannsen has developed the annotation tool.

8. References

Candito, M., Amsili, P., Barque L., Benamara, F., De Chalendar, G., Djemaa, M., Haas, P., Huyghe, R., Yannick Mathieu, Y., Muller, P., Sagot, B. & Vieu, L. (2014). Developing a French FrameNet: Methodology and First Results. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland.

Das, D., Chen, D., Martins, A., Schneider, N. & Smith, N. (2014). Frame-semantic parsing. *Computational linguistics*, 40(1), pp. 9–56.

Friberg Heppin, K. & Toporowska Gronostaj, M. (2012). The rocky road towards a swedish framenet - creating SweFN. In *Proceedings of the Eighth conference on International Language Resources and Evaluation (LREC'12)* Istanbul, Turkey, pp. 256–261.

Friberg Heppin, K. & Toporowska Gronostaj, M. 2014. Exploiting FrameNet for Swedish: Mismatch? *Constructions and Frames*, 6(1), pp. 52-72.

Johannsen, A., Martinez Alonso, H. & Søgaard, A. (2015). Any-language frame semantic parsing. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal: Association for Computational Linguistics, pp. 2062-2066.

Kipper Schuler, K. (2005). *Verbnet: a broad-coverage, comprehensive verb lexicon.* Ph.D. thesis, Philadelphia, PA, USA. AAI3179808.

Martínez Alonso, H., Johannsen, A., Nimb, S., Olsen, S. & Pedersen, B. (2016). An empirically grounded expansion of the supersense inventory. In *Proceedings of Global Wordnet Conference 2016.*

Martínez Alonso, H., Johannsen, A., Olsen, S., Nimb, S., Sørensen, N., Braasch, A., Søgaard, A. & Pedersen, B. S. (2015). Supersense tagging for Danish. In: *Proceedings of the 20th Nordic Conference of Computational Linguistics NODALIDA 2015. Vol. 109*, Linköping University Electronic Press, NEALT Proceedings Series, Vol. 23.

Nimb, S., Lorentzen, H. & Trap-Jensen, L. (2014b): The Danish Thesaurus: Problems and Perspectives. In: A. Abel, C. Vettori & N. Ralli (eds.) *Proceedings of the XVI EURALEX International Congress: The User in Focus.* 15-19 July

2014. Bolzano/Bozen 2014: EURAC Research, pp. 191-199.

Ohara, K. H. (2014). Relating Frames and Constructions in Japanese FrameNet. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, pp. 2474-2477.

Olsen, S., Pedersen, B. S., Martínez Alonso, H. & Johannsen, A. (2015). Coarse-grained sense annotation of Danish across textual domains. In *Proceedings of the Workshop on Semantic resources and Semantic Annotation for Natural Language Processing and the Digital Humanities at NODALIDA 2015*, Linköping University Electronic Press, Sweden.

Palmer, M., Gildea, D. & Kingsbury, P. (2005). The Proposition Bank: A Corpus Annotated with Semantic Roles, *Computational Linguistics Journal, 31:1,* Association for Computational Linguistics.

Pedersen, B. S., Martínez Alonso, H., Braasch, A., Johannsen, A., Nimb, S., Olsen, S., Søgaard, A. & Sørensen, N. (2016). The SemDaX Corpus - sense annotations with scalable sense inventories. In: *Proceedings of the 10th edition of the Language Resources and Evaluation Conference, (LREC'16)*, Portorož, Slovenia.

Pedersen, B. S., Nimb, S., Olsen, S., Søgaard, A. & Sørensen, N. (2014). Semantic Annotation of the Danish CLARIN Reference Corpus. In: *Proceedings from isa-10, 10th Joint ACL - ISO Workshop on Interoperable Semantic Annotation* pp. 25-29, Reykjavik, Iceland.

Ruppenhofer, J., Ellsworth, M., Petruck, M. R. L., Johnson, C. R., Baker, C. F. & Scheffczyk, J. (2016). *FrameNet II: Extended Theory and Practice* (Revised November 1, 2016.) https://framenet.icsi.berkeley.edu/fndrupal/the_book.

Schluter, N.-E. & Søgaard, A. (2015). Unsupervised extractive summarization via coverage maximization with syntactic and semantic concepts. In: *The 53rd Annual Meeting of the Association for Computational Linguistics (ACL). Vol. 2* Association for Computational Linguistic, pp. 840-844.

Søgaard, A., Plank, B., Martinez Alonso, H. (2015). Using frame semantics for knowledge extraction from Twitter. In: *Proceedings of he twenty-ninth AAAI Conference on Artificial Intelligence: AAAI 2015.* Association for the Advancement of Artificial Intelligence, pp. 2447-2452.

Svendsen, M.M. (2017). *Constructing a FrameNet for Danish as a tool for lexicographers.* Unpublished thesis, Aarhus University, Denmark.

Websites:

Asmussen, J. (2012).The CLARIN Reference Corpus: Accessed at http://cst.ku.dk/Workshop311012/sprogtekno2012.pdf

Berkeley FrameNet: Accessed at: http://framenet.icsi.berkeley.edu.

Johannsen, Anders. FrameNet annotation tool: Accessed at https://github.com/andersjo/framenet-annotation.

Dictionaries:

DDO. Accessed at: http://ordnet.dk/ddo, Society for Danish Language and

Literature, Copenhagen, Denmark.

The Danish thesaurus: Nimb, S., Lorentzen, H., Troelsgård T., Theilgaard, L., Trap-Jensen, L. (2014a): *Den Danske Begrebsordbog*, Society for Danish Language and Literature, Copenhagen, Denmark.

Dornseiff, F. (2004): *Der deutsche Wortschatz nach Sachgruppen*, 8. Auflage. Berlin/New York: Walter de Gruyter.

Gyldendal Danish English Dictionary. Accessed at: ordbog.gyldendal.dk.

Bilingual Dictionary Drafting: Bootstrapping WordNet and BabelNet

David Lindemann[1,2], Fritz Kliche[2]

[1] The Bilingual Mind, UPV/EHU University of the Basque Country,
Justo Vélez de Elorriaga 1, 01006 Vitoria-Gasteiz (Spain)
[2] IwiSt Institute for Information Science and Natural Language Processing,
Universität Hildesheim, Universitätsplatz 1, 31141 Hildesheim (Germany)
E-mail: david.lindemann@ehu.eus, fritz.kliche@uni-hildesheim.de

Abstract

In this paper, we present a simple method for drafting sense-disambiguated bilingual dictionary content using lexical data extracted from merged wordnets, on the one hand, and from BabelNet, a very large resource built automatically from wordnets and other sources, on the other. Our motivation for using English-Basque as a showcase is the fact that Basque is still lacking bilingual lexicographical products of significant size and quality for any combination other than with the five major European languages. At the same time, it is our aim to provide a comprehensive guide to bilingual dictionary content drafting using English as pivot language, by bootstrapping wordnet-like resources; an approach that may be of interest for lexicographers working on dictionary projects dealing with other combinations that have not been covered in lexicography but where such resources are available. We present our experiments, together with an evaluation, in two dimensions: (1) A quantitative evaluation by describing the intersections of the obtained vocabularies with a basic lemma list of Standard Basque, the language for which we intend to provide dictionary drafts, and (2) a manual qualitative evaluation by measuring the adequateness of the bootstrapped translation equivalences. We thus compare recall and precision of the applied dictionary drafting methods considering different subsets of the draft dictionary data. We also discuss advantages and shortcomings of the described approach in general, and draw conclusions about the usefulness of the selected sources in the lexicographical production process.

Keywords: Bilingual Lexicography; Bilingual Dictionary Drafting; WordNet; BabelNet

1. Computational lexicography and WSD in multilingual settings

1.1 Starting Point

According to *Ethnologue* data, more than 400 languages have one million or more first-language speakers. If we check the availability of bilingual dictionaries for these languages, we observe that many language pairs, even those involving one of the top ten languages of the world, remain uncovered. For Basque, for instance, a European language with about one million speakers, bilingual dictionaries of significant size and

quality are available today for Spanish, French, English, Russian, and German.

Lacking suitable lexicographical resources for all other language pairs, a dictionary user may follow two main strategies: they may use more than one bilingual dictionary, i.e. retrieve the desired information via hub, and thus perform double lookups or trust their knowledge in the hub language. This we may call the 'traditional' approach. Alternatively, they may also rely on automatically built bilingual dictionary-like resources for the required language pair, or place their query in machine translation backed web portals that work with automated algorithms and use English as a hub.

For the first case, there are a number of disadvantages linked to the required availability of the respective dictionaries, and to the disposition to spend the required time for multiple lookups in one process of lexical information retrieval. Its ease and its efficacy for the user is what makes the second strategy appealing.

One fundamental problem applies to both strategies. Mistakes in the retrieval of translation equivalents due to lexical semantics issues, and different distributions in the lexicalization of concepts between languages, are doubtlessly frequent, and discussions regarding *asymmetric lexicalization* are thus a real classic in metalexicographical writing (for the cited concept, see Hartmann, 1990; cf. also Wiegand, 2002; Gouws, 2002). Furthermore, if two different bilingual dictionaries are needed for looking up possible equivalents, the risk of being misled may also be doubled.

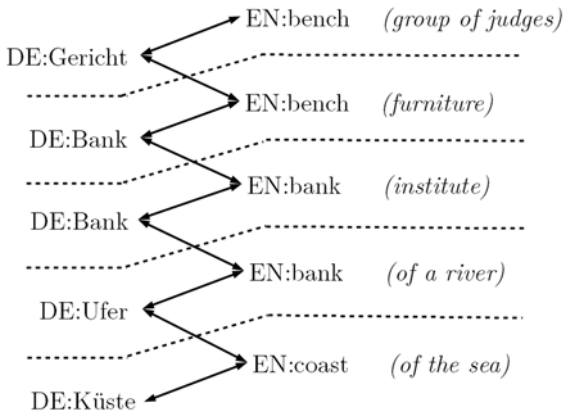


Figure 1: Asymmetric lexicalization of concepts

Asymmetric lexicalization can be illustrated by the examples given in Figure 1, where arrowed lines represent possible translation equivalences between senses that correspond to the lemma-strings preceded by the German or English language code, and dotted lines divide concepts; glosses are given in brackets to disambiguate concepts. Arrows that cross dotted lines represent mismatched translation equivalences that erroneously seem possible according to the character strings they link to each other. Without further information (dotted lines and glosses), all the equivalences between lexical items represented here are equally possible. The inventory

of senses shown here is far from complete, and the game could be continued (for example, *Gericht* may also mean an edible 'dish', while *dish* in turn also may denote a vessel used for serving food, which in German can be called 'Geschirr', which is an item that also may denote horse or ox harnesses, etc.).

The figure also does not show distinctions between (1) homonomy (German $Bank_1$ vs. $Bank_2$, English $bank_1$ vs. $bank_2$, (2) polysemy ($bench_1$ vs. $bench_2$), and (3) a splitting of senses, which is not necessary for a German monolingual but is necessary for a German-English bilingual dictionary entry (*Ufer*, 'egde' of a river vs. of the sea, a lemma that in German monolingual dictionaries is usually not marked as polysemous). Here, we see only text strings annotated as nouns of a certain language; the required condition for the mismatched equivalences to occur. A good bilingual dictionary of course provides the user with useful homonym or sense disambiguating advice in order to avoid such misleading pairings.

Problems related to a look-up process misled by asymmetric polysemy structures also may apply to the second case; in fact, this is the main shortcoming observed when employing algorithms that interlink entries of two bilingual dictionaries, using their shared language as hub (for example, as stated by Saralegi et al., 2012). Also, in parallel corpus processing, the semantic disambiguation of polysemous lexical items (WSD) has still to be regarded as a central problem; users who lack a suitable bilingual dictionary and thus stick to statistical machine translation engines, must deal with errors related to homonymy and polysemy in the results they obtain.

1.2 Bilingual Dictionary Drafting Methods: A Brief Overview

If we thus decide to develop lexicographical resources for new language pairs in order to overcome these shortcomings, we can employ 'traditional' methods: namely, the manual compilation of bilingual dictionaries starting from scratch. However, this is a very labour-intensive task, only feasible for lexicographical products that satisfy commercial criteria (which is definitely not the case for dictionaries of a language such as Basque) or grow in publicly well-funded environments. To reduce the level of manual effort required for bilingual dictionary making, a further development of (semi-)automatic dictionary drafting methods seems worthwhile.

For a rough classification of (semi-)automatic methods to obtain bilingual word pairs as candidates for an inclusion as translation equivalents into a bilingual dictionary (see Varga et al., 2009 for a survey of related work), we can distinguish between corpus-based methods on the one hand, and methods that rely on transferring data from existing lexical resources, on the other. Both approaches may be combined, e.g. as in Saralegi et al. (2012), where the equivalent pairs obtained by linking the content of two bilingual dictionaries are ranked according to distributional similarity in a bilingual text corpus.

In addition, we can group translation equivalent drafting methods according to the following qualitative feature: whether it results in bilingual word lists, i.e. lists of equivalent candidates, or whether it is capable of linking word-sense disambiguated lexical items to each other, i.e. of linking word senses, for a bilingual dictionary draft that includes WSD. Bilingual data found in the WordNet-related multilingual lexical resource MCR 3.0 (Gonzalez-Agirre et al., 2012), as shown in Table 1, demonstrates, for the instances of the noun *banku* in Basque WordNet, how equivalences may be extracted from this kind of resource and including a discrimination of word senses, i.e., a grid that avoids mismatches of the kind illustrated in Figure 1.

Basque Synset	**English Synset**	**MCR ontology classes**
banku_1; banketxe_1;	bank_9; bank_building_1	banking; artifact_1; artifact; Building+; Artifact+ Function+ Object+
aulki_3; banku_2;	bench_1	furniture; furniture_1; artifact; Artifact+; Artifact+ Furniture+ Group+ Instrument+ Object+
banku_3; banketxe_2;	depository_financial_institution_1; bank_2; banking_concern_1; banking_company_1	banking; organization_1; group; Corporation+; Function+ Group+ Human+
banku_4;	bank_3	factotum; object_1; object; LandArea+; 1stOrderEntity+ Natural+ Object+ Place+
banku_5;	bank_6	finance; assets_1; possession; CurrencyMeasure+; Function+
banku_6;	bank_5	money; income_1; possession; Keeping+; Artifact; Function+ MoneyRepresentation+ Part+
banketxe_3; banku_7;	banking_industry_1; banking_system_1	industry; industry_1; group; Corporation+; Function+ Group+ Human+

Table 1: Synsets containing *banku* in EusWN and aligned English data

On multiple occasions, lexical data have been transferred from dictionaries to build wordnets from scratch, using the Princeton WordNet concept grid as the starting point (i.e., the 'expand method'), or to enrich already existing wordnets; advantages and shortcomings of this approach have been discussed widely (Vossen, 2002; Fišer & Sagot, 2015, among others). A major problem regarding this approach is, again, a mismatched merging of word senses that belong to homonymous or polysemous dictionary headwords and WordNet concepts.

Automated drafting of bilingual dictionary content may significantly ease the manual effort required to make dictionaries from scratch. As earlier experiments have shown, even for a relatively marginal language-pair like German-Basque, one can obtain equivalent candidates for around two thirds of the initial lemma list (Lindemann et al., 2014). But, in any case, it is not only the recall on the initial word lists that automated drafting methods may offer, but it is also, of course, the precision, that is, in our case, the adequacy of the draft equivalent pairs that makes the difference: for the

production of a dictionary that deserves this name, as long as automated efforts continue to fail to achieve precision rates approaching 100%, manual editing of the draft data seems indispensable.

The English Princeton WordNet and Basque WordNet, the two resources used for the experiments described in this paper, were manually built, or at least manually validated. Thus, we should expect high precision, and experiments carried out in the past confirm this assumption even for pivoted bilingual dictionary drafting. Lindemann et al. (2014) evaluated a German-Basque dictionary drafting experiment that involved data from English and Basque WordNets, and from GermaNet (Hamp & Feldweg, 1997), version 8. They found that the rate of equivalences assessed as false did not reach 10%, and another 10% was assessed as partly correct (for the partial matching of compounds) or nearly so, i.e. fuzzily correct. These precision rates were surpassed only by the data from cross-language links attached to *Wikipedia* page titles, and by the Basque equivalents in German *Wiktionary.*[1] However, the latter two resources yielded a much lower recall on the list used as gold standard for German dictionary headwords.

WordNet	**Noun items**	**Verb items**	**Adjective items**	**Adverb items**	**Synsets**
Princeton 3.0 (PWN)	147,245	25,051	30,082	5,580	118,408
Basque 3.0 (EusWN)[2]	40,420	9,469	148	0	30,263

Table 2: Statistics from MCR 3.0

Basque WordNet (Pociello, Agirre & Aldezabal, 2011) was built by semi-automatic means following the 'extend model', i.e. by defining Basque lexicalizations for concepts present in Princeton WordNet (Miller et al., 1990). After a semi-automatic drafting by transfer from Basque dictionaries, the workflow for the construction of this resource involved a manual validation of the whole content. In Basque WordNet 3.0 (henceforth EusWN), concepts are aligned one-to-one to Princeton WordNet 3.0 (PWN) synsets. EusWN can thus be regarded as a translation of PWN. Table 2 contains the overall statistics for both resources.[3] It is clear that EusWN covers no more than about 25% of the concepts represented in PWN.

[1] Also one of the assessed parallel corpus word alignment tools led to results with a precision higher than 90%, but with a very conservative parameter setting, that allowed a recall not higher than 5%.

[2] Not all EusWN synsets contain lexical items; in the case they are not linked to any Basque lexical item, they are, however, semantically annotated. See Pociello et al. (2011) for reference.

[3] The content from both WordNets and documentation are available at http://adimen.si.ehu.eus/web/MCR/.

BabelNet (Navigli & Ponzetto, 2010) is an automatically built multilingual resource. It contains data extracted from a wide range of sources, some automatically, some manually built or manually validated. Just as in WordNet, the basic unit in the data model is the synset node, which is identified by a unique number. Just as in MCR and Open Multilingual WordNet (Bond & Foster, 2013), two of the approaches used to build a multilingual WordNet, all concepts exist in English, and as soon as lexicalizations and other item types in languages other than English that belong to these concepts are available, they become linked to one of these.

BabelNet 3.7	**Noun items**	**Verb items**	**Adjective items**	**Adverb items**	**Synsets**
English (overall)	11,303,752	58,644	112,518	19,545	6,667,885
English (English-Basque intersection)	5,010,332	15,132	1,310	190	2,469,915
Basque	2,727,673	9,558	443	54	2,469,915

Table 3: Statistics for BabelNet 3.7

One of the additional values of BabelNet is that content extracted from numerous resources[4] appears as merged to BabelNet synsets that ideally should be unique for each concept, i.e. duplicate concepts should be merged to a single synset. The extraction and merging tasks are performed by algorithms which are regularly improved and updated, along with the inclusion of more data. Table 3 contains statistics for the BabelNet content as for version 3.7, released in 2017.

2. From Bootstrapping to Evaluation

Having in mind the reasons discussed above, for the research presented in this paper we concentrate on a transfer-based method that allows for the extraction of sense-to-sense equivalences. We have been bootstrapping and evaluating bilingual lexical data from English and Basque WordNets, on one hand, and from BabelNet, on the other. The underlying approach is as simple as extracting lexicalizations in two languages for the same concept (i.e., that share a common unique synset ID), and quantitatively and qualitatively evaluating the obtained bilingual dictionary draft. The rates for the recall of the extracted data on a basic lemma list and for precision in terms of translation equivalence, give us clues related to both uses at the same time: for a possible use as draft content in dictionary making, and for what we have to expect when using web portals that present automatically gathered data as a reference dictionary.

[4] A complete list of the sources for the lexical items in BabelNet is available at: http://babelnet.org/about.

We downloaded the BabelNet 3.7 indices dump which stores the BabelNet corpus as an *Apache Lucene* index.[5] We retrieved its content using the Java API which is available for download on the BabelNet website.[6] We collected the synsets that contain at least one lexical entry for English and one for Basque, and found 2,469,915 synsets for this intersection. For each synset we collected (1) the BabelNet synset ID; (2) the English and Basque lexical items; (3) the English glosses; (4) metadata on the "type" of the synset, which is either "named entity" or "concept"; and (5) the source of the lexical item. Additionally, the synset ID includes (6) a marker for part of speech. We wrapped our scripts into a processing pipeline, both for reproducibility of the results and for an easy adaptation to other language pairs (or language sets).

For all intersection calculations, lexical items are taken into account as graphically normalized strings. All upper case letters have been converted to lower case, and all hyphens or spaces between multiword lexical units have been suppressed, in order to harmonize graphical variants found in the sources. For example, the Basque term for *death penalty* appears in the data in three graphical variants (*heriotza-zigor, heriotza zigor, Heriotza zigor*), each of which are normalized to a one-word form, *heriotzazigor*. This form is not documented in the data, but in general, noun+noun compounds in Basque also may appear as one single word (*eguzki-lore, eguzki lore* or *eguzkilore*, literally 'sunflower').[7] Some items, namely those stemming from Wikipedia and Wikidata, may contain a short sense-disambiguating gloss in brackets, in addition to the lexical item itself, as in *gotiko (hizkuntza)*, and *gotiko (estiloa)*, 'gothic language' vs. 'style'. These glosses have been suppressed for the same reason: in the respective synset, the strings *gotiko* and *Gotiko* appear, with no gloss; after normalization, all four are treated as duplicates, and therefore as one unique lexical item.

In general, we found inconsistencies regarding the initial case of lexical items. In principle, Basque orthography is more regular than English, as a range of nouns that are not considered named entities (proper names) in English have an uppercase initial letter (e.g. names of languages, days of the week, months, etc.). But, aside from this, many inconsistencies have been found in the Basque lexical items stemming from BabelNet sources other than WordNet. For instance, Basque terms in software localization (Microsoft Terminology) bear initial upper case; even verbs such as *Bidali*, 'send' or *Onartu*, 'accept', presumably because these equivalent pairs were defined to serve as localized flags for buttons on a website or software application. For items that represent Basque Wikipedia page titles, we have also found inconsistencies: around 30% have a lower case initial letter, but this feature seems not to be consistently

[5] https://lucene.apache.org/core

[6] http://babelnet.org/download

[7] Unlike the two separated variants of this compound, the merged single word is not found in the normative wordlist of Standard Basque (Euskaltzaindia, 2010), although it is frequent in corpora. In other cases, in turn, a merged compound is listed as the standard form (*aireontzi*, 'airplane'). For the experiments presented here, multiword units are merged in general.

related to the noun type.

We have not used the noun type filter built into BabelNet, that is, the tags "named entity" and "concept" present in the synsets, to evaluate the effect of that filter. Consequently, lexicalizations for named entities (proper nouns) also may appear in the counts presented in Table 4 if a common noun is homographous (e.g. Basque (and Spanish) *Lima* to *lima*, 'lime' *Gaza* to *gaza*, 'gauze'). It should also be mentioned here that Basque nouns erroneously tagged as common nouns instead of proper nouns in the corpus processing (e.g. *Praga*, 'Prague', *Polisario*) at this stage, have not been manually removed from EusLemStd, a Basque lemma inventory used for quantitative evaluation (see Section 3).

The quantitative and qualitative evaluation has been carried out using built-in features of the *TshwaneLex* software application,[8] into which we have imported all lexical data on hand. This allowed us to merge all data according to a pre-defined XML schema, and, at the same time, to keep all evaluation steps reproducible.

3. Quantitative Evaluation

In this section, we give an account of intersecting sets of (1) the extracted lexical data stemming from (a) WordNet and (b) BabelNet, and (2) the entries of EusLemStd, a frequency headword list used here as gold standard for a Basque lemma inventory. This word list is produced by computational means; it contains common nouns, verbs, adjectives and adverbs that appear as headwords in at least one of the standard reference dictionaries for Basque, as well as in at least one of the two major monolingual corpora, a hand-selected reference corpus, and a large web corpus (see Lindemann & San Vicente, 2015). The qualitative evaluation of random subsets of this intersection is presented in Section 4 below.

Headwords: intersecting sets		
EusLemStd ∩ EusWN ∩ BabelNet	18,004	(31.0%)
EusLemStd ∩ EusWN	18,122	(31.3%)
EusLemStd ∩ BabelNet	23,194	(40.0%)
EusLemStd	57,919	(100.0%)

Table 4: Intersection of EusWN, BabelNet, and EusLemStd (headword strings).

[8] http://tshwanedje.com/tshwanelex/

In Table 5, we quantify the intersection of (1) EusWN/PWN concepts, (2) BabelNet concepts and EusLemStd; that is, synsets that contain at least one item found on the Basque reference lemma list.

Concepts: intersecting sets	**Noun synsets**	**Verb synsets**	**Adjective synsets**	**Adverb synsets**	**Synsets**
EusWN ∩ EusLemStd	21,533	2,894	106	0	24,533
BabelNet ∩ EusLemStd	31,028	2,914	293	25	34,260

Table 5: Intersection of EusWN, PWN, and EusLemStd (concepts)

The Basque lexical items found in BabelNet stem from the sources listed in Table 6. The table contains the overall numbers of items, as well as the numbers of strings that also appear in EusLemStd. Lexical items homographous to each other inside or across parts of speech count here as one unique string.

Source	**EusLemStd intersection unique items**	**EusLemStd intersection total items**	**BabelNet 3.7 total Basque items**
All Sources	23,194	67,221	2,737,728
Open Multilingual WordNet	18,060	39,343	48,934
Wikidata	7,347	8,159	190,764
Wikipedia	6,646	6,849	182,967
BabelNet	2,215	3,989	2,255,355
Wikipedia Redirections	3,254	3,565	51,440
OmegaWiki	2,485	2,816	5,625
Wiktionary	1,464	1,629	2,188
Microsoft Terminology	581	735	3,887
GeoNames	75	79	1,879
WikiQuotes	29	29	218
WikiQuotes Redirections	28	28	96

Table 6: Basque lexical items in BabelNet 3.7 (concepts and named entities)

If we relate these figures to the amounts of synsets, for the intersection of the Basque BabelNet with EusLemStd, we find a distribution of Basque lexical items per synset as shown in Table 7. Note that synsets that contain a standard lemma also may contain further items not found on EusLemStd. Synsets tagged as "named entity" in BabelNet have been filtered from the subsets quantified in this table.

Source	EusLemStd intersection items/synset	EusLemStd intersection total synsets	BabelNet 3.7 total Basque synsets
All Sources	2.28	29,420	2,469,915
Open Multilingual WordNet	1.59	24,786	28,699
Wikidata	1.02	8,004	87,922
Wikipedia	0.87	7,883	81,777
BabelNet	3.44	1,161	1,755,914
Wikipedia Redirections	0.85	4,210	11,598
OmegaWiki	1.08	2,607	3,970
Wiktionary	1.09	1,496	1,656
Microsoft Terminology	1.07	689	3,108
GeoNames	0.00	0	4
WikiQuotes	0.51	57	61
WikiQuotes Redirections	1.33	21	24

Table 7: Basque concepts in BabelNet 3.7 (tagged as "concept" in BabelNet)

4. Qualitative Evaluation

4.1 WordNet

For the translation equivalences obtained from WordNet, we have carried out a qualitative evaluation for (1) a random set of noun and verb synsets that contain only monosemous Basque items; that is, items that occur only in one synset, and (2) a random set of other synsets; i.e., those that also contain polysemous Basque lexical items, as we presumed a higher degree of fuzzy or false matchings for polysemous items. For adjectives, we have not evaluated the monosemous items separately, as the number of synsets containing only these does not even reach a dozen. The adequacy of the semantic matching between Basque and English equivalents has been assessed on a scale of three values, as formerly used in similar studies (Fišer et al., 2012; Lindemann et al., 2014):

(1) OK, for a correct matching, in the sense that the Basque lexical item could be used in a dictionary entry for denoting the pertaining concept without any changes,

(2) FUZZY, for a fuzzy semantic matching, which means that the lexical item does not match the pertaining concept in a way that could be used in a dictionary entry, but that its semantic distance to the ideal equivalent is to be regarded as small; it may be a hyponym or hypernym, a meronym or a holonym of an ideal equivalent. For verbs, equivalents that are semantically very close but with incompatible valencies (e.g. regarding transitivity) are also assessed as FUZZY. A paraphrase of this value could be "the lexicographer has to intervene here,

but it is not a completely false equivalent."[9]

(3) FALSE, for a lexical item that provides nothing usable for a lexicographer when editing the entry.

For 300 synsets, the adequacy of the corresponding 546 lexical items has been assessed. The distribution of the assessment values is summarized in Table 8.

The data taken into account for assessment are the English glosses and example sentences, and the English and Basque lexical items. During the assessment process, the semantic relations or ontology classes of a synset could also be displayed. We assess the equivalents as for a translation from Basque to English, which is the direction contrary to the editing process of EusWN. Consequently, we do not assess here whether the group of English items could have been translated to Basque in a more appropriate way than via the Basque items found.[10] Critical in this context are nominal derivations from Basque verbs, often employed in EusWN as equivalent of English nouns that denote actions or results of actions, but that are not treated as lemma in Basque dictionaries, and consequently neither in EusLemStd, as for example the nominal derivations *xahutze*, *ahaitze* for the English 'wastage'.

EusWN/PWN equivalences	**Nouns**	**Verbs**	**Adjectives**	**All POS**
Total synsets intersect. EusWN/EusLemStd	21,533	2894	106	21,533
• Monosemous	6,058	201	11	6,270
• Polysemous	15,343	2,693	95	18,131
Synsets evaluated	100	100	100	300
• Monosemous	50	50	16	
• Polysemous	50	50	84	
Synsets all items OK	87%	75%	94 (94%)	85%
• Monosemous	45 (90%)	37 (74%)		
• Polysemous	42 (84%)	38 (76%)		
Synsets OK/FUZZY	98%	94%	96 (96%)	96%
• Monosemous	49 (98%)	48 (96%)		
• Polysemous	49 (98%)	46 (92%)		
Synsets 1+ FALSE	2%	7%	4 (4%)	4%
• Monosemous	1 (2%)	2 (4%)		
• Polysemous	1 (2%)	5 (10%)		

Table 8: Qualitative evaluation of equivalents extracted from EusWN

[9] The equivalence assessed in this way is not to be confused with *fuzzynymy*, which is a semantic relation encoded in EuroWordNet, that holds when the tests for synonymy, homonymy and meronymy "fail but the test *X has some strong relation to Y* still works" (Vossen, 2002: 37). Fuzziness here includes all somehow close relations apart from cross-language synonymy (in the sense of adequacy as dictionary translation equivalent), i.e. including homonymy.

[10] However, we have unsystematically annotated the assessed data with free text comments and proposals for more appropriate equivalents. These annotations may be used in the future as notes for preparing a more systematic and complete survey.

4.2 BabelNet

The qualitative evaluation of Basque-English equivalences found in BabelNet differs from the process described above in some points. As explained above, together with the Basque lexical items we have extracted the tags denoting their respective source and stored the data in the database used for evaluation. This allows the disambiguation of the evaluation results according to the source of the pertaining item, as shown in Table 10.

Since we found the rendering of the automatic sense merging carried out for building BabelNet a particularly interesting detail, we have introduced a fourth assessment value, MERGE_ERROR. This value was assigned in cases where the random synset displayed for evaluation was found to contain lexical items that denote (and glosses that describe) two different concepts. For example, one synset contains lexical items and definitions of the English noun *underground* that refer to the word sense 'tube, metro', as in "The London Underground", and to the word sense 'resistance, underground' with the definition "a secret group organized to overthrow a government…", while both senses in PWN appear in distinct synsets. As for the translation equivalence, this value has thus to be regarded a variant of FALSE.

BabelNet 3.7	**OK**	**FUZZY**	**FALSE**	**MERGE ERROR**	**(Asses-ments)**
All Sources	*1,211 (88.9%)*	*63 (4.6%)*	*44 (3.2%)*	*44 (3.2%)*	*1,362*
Open Multilingual WordNet	717 (89.2%)	49 (6.1%)	28 (3.5%)	10 (1.2%)	804
Wikidata	57 (93.4%)	0 (0.0%)	1 (1.6%)	3 (4.9%)	61
Wikipedia	194 (87.8%)	5 (2.3%)	6 (2.7%)	16 (7.2%)	221
BabelNet	3 (100.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	3
Wikipedia Redirections	13 (52.0%)	3 (12.0%)	4 (16.0%)	5 (20.0%)	25
OmegaWiki	75 (91.5%)	2 (2.4%)	0 (0.0%)	5 (6.1%)	82
Wiktionary	132 (92.3%)	4 (2.8%)	5 (3.5%)	2 (1.4%)	143
Microsoft Terminology	20 (87.0%)	0 (0.0%)	0 (0.0%)	3 (13.0%)	23
GeoNames	0	0	0	0	0
WikiQuotes	0	0	0	0	0
WikiQuotes Redirections	0	0	0	0	0

Table 10: Qualitative evaluation of BabelNet equivalences for sources

As the reader will observe, the qualitative assessments made for items stemming from different sources diverge significantly. For a dictionary draft, we may accept only items from particular sources, as the encoding of lexical data in BabelNet allows such filtered extraction. Lexical items that originally are titles of redirection pages in *Wikipedia* and *Wikiquotes*,[11] in general, should only match fuzzily or very fuzzily to the pertaining concept. This is because, in their original resource, their reason to be is that there is no other page in that resource that matches better. The redirections in *Wikipedia* that link to *turkey* in the sense of 'turkey meat', for example, include *Turkey Sandwich*, *Cooking a turkey*, *Turkey meat*, and *Turkey dinner*, i.e. two-word units, and even phrases of a different part of speech. Depending on the desired application, such fuzzy matchings may be more or less useful; as translation equivalents, most of them will not serve.

The evaluation results for BabelNet synsets, according to part of speech, are collected in Table 11. In principle, we can also relate the evaluation data disambiguated by source to the parts of speech, both for lexical items and for items grouped as synset. For space reasons, we concentrate here on giving a complete account of the outcome for synsets, as this already provides a good overview of the value a dictionary draft based on BabelNet can have in a lexicographical workflow. The assessments for the 1,184 lexical items that have been evaluated in total are distributed as follows: 1,056 OK (89.2%), 58 FUZZY, 39 FALSE, and 31 MERGE ERROR. As these items belong to 625 different synsets, the average number of Basque lexical items found per synset in this random subset of the English-Basque BabelNet is 1.89.

BabelNet 3.7	**Nouns**	**Verbs**	**Adjectives**	**Adverbs**	**Total**
Assessed synsets	200	200	200	25	625
All items OK	179 (89.5%)	163 (81.5%)	188 (94.0%)	23 (92,0%)	553 (88.5%)
1+ items OK, and 1+ items FUZZY	3 (1.5%)	14 (7.0%)	2 (1.0%)	0 (0.0%)	19 (3.0%)
1+ items OK, and 1+ items FALSE	2 (1.0%)	3 (1.5%)	0 (0.0%)	0 (0.0%)	5 (0.8%)
All items FUZZY	5 (2.5%)	9 (5.5%)	8 (2.0%)	0 (0.0%)	22 (3.5%)
1+ items FUZZY, and 1+ items	1 (0.5%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (0.5%)
All items FALSE	5 (2.5%)	8 (4.0%)	1 (0.5%)	2 (8.0%)	16 (2.6%)
MERGE_ERROR	5 (2.5%)	3 (1.5%)	1 (0.5%)	0 (0.0%)	9 (1.4%)

Table 11: Qualitative evaluation of BabelNet equivalences for synsets and part of speech

[11] As for BabelNet 3.7, there is nearly no Basque data found from Wikipedia and Wikiquote Redirections (cf. Section 3 above).

While displaying random noun synsets, in 30 cases the synset referred to a named entity, and the corresponding English lexical items were proper nouns. The reason for these to appear in our evaluation data in all cases was the fact that the Basque equivalent contained a string homographous to a EusLemStd entry, as for example the Basque common noun *datu*, 'date', homograph to "a title for chiefs, sovereign princes, and monarchs in [...] Regions of the Philippines" (*Wikipedia*), or 'materia', which also is the title of an album recorded by an Italian music band. In these cases, we skipped the evaluation of the synset and went on to the next (so that 230 noun synsets have been evaluated in total), but we also performed a second test: Whether the synset was listed as "named entity" (in opposition to "concept") in BabelNet. For all 30 cases, the result was positive, so that we may conclude that named entities are labelled properly in BabelNet. But, in principle, cross-class homograph nouns may appear merged as one in BabelNet (which was not the case in the random subset we evaluated);[12] this is the reason why we wanted to have all string homographs to EusLemStd entries evaluated.

As mentioned above, the algorithms used for concept merging, as for BabelNet 3.7, lead to some mismatched junctions. The intended lexicographic use of BabelNet data is to regard a number of translation equivalences as noisy or false. The problematic aspect for this regarding mismatches, however, is the fact that the unique ID that serves for highlighting the wrongly merged synset will not be stable: as soon as the sense merging algorithm is improved, the concept must be split again. The stability of synset IDs is a central feature for linking concepts across different resources, which we will discuss in the following section.

5. Interoperability Issues and Lexicographic Postprocessing

In this section, we want to give a brief overview of some of the issues related to data model interoperability and the representation of lexical semantic relations. We cannot discuss all issues in detail here; nevertheless, the following general comments may serve as orientation for making a transfer based dictionary drafting, with wordnet-like concept-oriented resources for bootstrapping.

Converting a concept-oriented collection of lexical data into a headword-oriented dictionary draft is a computationally trivial transformation task. As mentioned in Section 2, we are able to represent our dictionary draft datasets in XML, as illustrated in Figure 2 below. In connection with this transformation, we have to mention two issues, which are far from trivial, for lexicographers: (1) the modelling of homography,

[12] While unsystematically browsing BabelNet, we found e.g. Dexter Raymond Mills, Jr., a.k.a. *Consequence*, an American rapper from Queens, New York, merged to the common noun synset *consequence, aftermath*, which is the one the Basque equivalents found here refer to. We also have had a look at the four items extracted from GeoNames that are present in the Basque BabelNet and classified as concept (cf. Table 7); contrary to their classification, all four are place names, and thus, named entities.

i.e. on which level we distinguish between homograph headword strings that point to dictionary entries related to different parts of speech (cf. in English *sound*$_{\text{N}}$, *sound*$_{\text{V}}$, and *sound*$_{\text{ADV}}$), and (2) the modelling of a distinction between homonymy and polysemy (cf. Section 1.1).

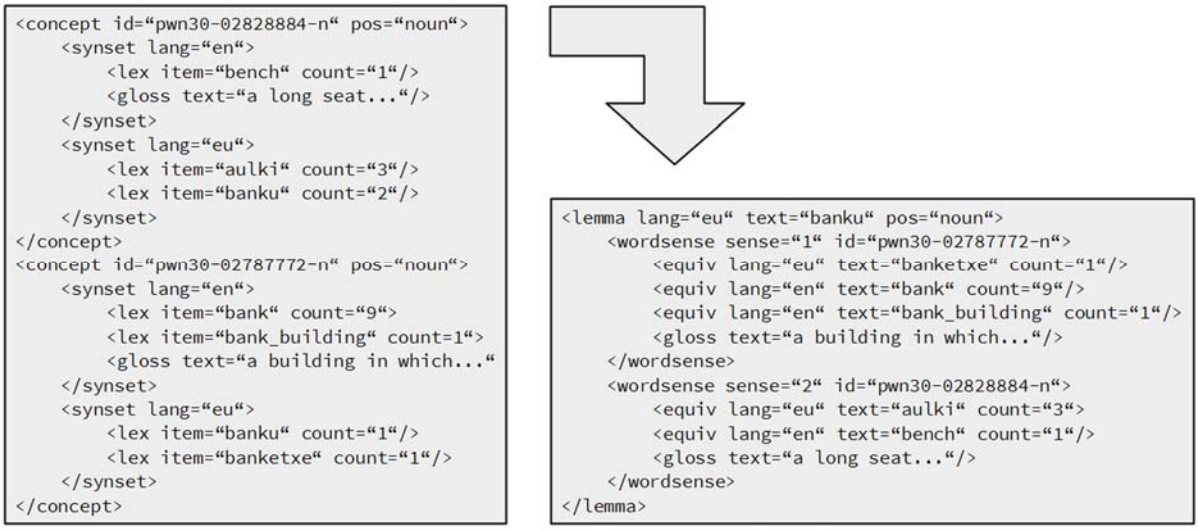


Figure 2: XML transformation

The distinction between homograph lemma-part of speech (lempos) entities is not problematic, since part of speech is encoded in the synset ID, and the transformation described here does not lead to dictionary entries with mixed-up parts of speech. On the contrary, homonymy and polysemy are treated equally in the data model of PWN and EusWN. In the case of the examples discussed in Section 1.1, as a consequence, the homonyms *bank* (institution) and *bank* (of a river) would appear in the same entry, just as do the two senses of *bench* (group of judges, furniture). If a disambiguated representation of these two different phenomena is desired, it has to be introduced in a further postprocessing step. This might work semi-automatically, e.g. by comparison to lists of items flagged as homonyms in dictionary headword lists.

Regarding the bits of XML code shown in Figure 2, we have to point out, of course, that it is a simplified presentation of what is possible. Here we just include the text attributes (alternatively representable as text values) for lexical items and abbreviated glosses. WordNet and BabelNet include more information linked to synsets, which may be used as microstructural item types in a dictionary; chiefly example sentences, domain flags, ontology classes, and semantic relations, and in BabelNet also images. For lexicographic purposes, Benjamin (2016) describes a more sophisticated cross-language mapping between lexical items, instead of (only) between synsets, in order to be able to relate every item-to-item link to more fine-grained classes of (quasi-)synonymy relations. The inclusion of more item types into a dictionary data model that is compatible with wordnet-like resources is a very attractive field to explore. Also, further item types linked to synset-IDs in a multilingual dictionary database potentially represent an extension to the source wordnet, at the same time.

A central issue which is also linked to data modelling is the internal representation of polysemy (besides its disambiguation from homonymy) that results from a transformation as illustrated in Figure 2. Two questions arise: (1) Does the draft dictionary entry contain all word senses of a lemma we want to represent? (2) Is the splitting of word senses found in the draft entry suitable for the dictionary for which we want to produce a draft, or is it (a) too fine-grained, (b) redundant, or are we (c) missing further distinctions?

Regarding question (1), we see no straightforward way to ascertain the respective answer other than via classical lexicography (i.e. manual work). However, we are preparing experiments to address that issue lemma by lemma with semi-automated quantitative comparisons to polysemy structures in existing dictionaries. Such comparisons will be helpful for question (2a,b), in case the sense splitting in the draft data significantly exceeds the number of senses found in reference dictionaries, or vice versa (2c). Before having conducted such bulk comparisons, our analysis of random subsets of the draft data suggests that the phenomena (2a,b) are frequent. One explanation lies in the 'expand' method of wordnet building and is connected to *genuine* and *false autohyponymy*, i.e. the same lexical item appearing in synsets that are hyponyms to each other (Pociello et al., 2011: 135–137). Examples of these include the translation *zahar*, 'old' for the English synset containing *moth-eaten, dusty, stale,* "lacking originality of spontaneity; no longer new", or *edan*, 'drink' for *drink, booze, fuddle*, "consume alcoholic beverages". While genuine autohyponyms should be maintained as different senses in a bilingual dictionary, for lexicographic purposes, false autohyponyms should be merged. A possible strategy for sense merging by PWN's own means is an automated classification as subsenses to one sense of homograph cohyponyms, i.e. lexical items that are graphically identical and share the same hypernym, or a common ancestor even higher in the hierarchy (cf. Miller, 1998: 42).[13]

The problems (2a-c) in computational linguistics are commonly referred to as *granularity* of word senses; different computational applications require more fine or coarse grained word senses (Prakash et al., 2007), and the same, of course, is true for dictionaries that serve different functions. In other words, requirements and strategies for a postprocessing of wordnet sense granularity will be closely related to the lexicographic project at hand. In any case, to merge senses will be technically more feasible than to introduce any splitting.

It should be clear that the problems we find for working with WordNet as a resource for lexicography are closely related to the nature of that resource, and the functions for which it was developed. Lexicography is explicitly not among these functions,

[13] In order to avoid "unmotivated cohyponyms", in other wordnet-like projects, a "crossed classification" of synsets is introduced, i.e. a classification of the same synset node in two different places in the hierarchy allowed in GermaNet, such as *banana* (a) as edible fruit and (b) as cultivated plant (Kunze, 2010, p. 507); such double classifications of the same concept could regularly be transformed into subsenses in a dictionary entry.

although WordNet has become a de-facto standard resource for monolingual and multilingual e-dictionary projects of all kinds.[14] Benjamin (2016: 28–31) mentions related problems not directly linked to data models but mostly to the original functions of WordNet: (1) The glosses linked to PWN synsets often do serve for disambiguating word senses, but not in a way that could be regarded adequate for publishing in a dictionary entry. (2) Some wordnets of languages other than English have been built automatically and contain a significant amount of errors, which is not problematic for some NLP applications, but it is, of course, for lexicography; and it becomes highly problematic if noisy data are just reproduced in a dictionary portal without being marked as possibly wrong. (3) The criterion that defines synonymy in wordnets is relatively weak in the sense that it allows too many cross-language equivalence links (between all members of a synset in language A to all members of a synset in language B). In other words, well-defined subclasses of the synonymy relation should be introduced systematically. Aside from that, the author mentions that when building a wordnet by the 'expand method', (4a) some synsets are filled with explanatory phrases instead of lexical items that serve as dictionary lemma, and (4b) a concept must exist in PWN to be expanded to the new wordnet. Finally, (5) the restrictive licensing of some wordnets makes bulk bootstrapping, and in some cases even isolated experiments, impossible.

6. Conclusions and Further Work

By bootstrapping wordnets and BabelNet, we have built a bilingual dictionary draft from scratch that includes a grid of lempos: entities and word senses, each of which furnished with one or more lexical items in two languages, and covering up to 40% of a previously defined list of Basque dictionary headwords. By the quantitative and qualitative evaluation of these draft data we have verified our initial hypothesis regarding the precision of the obtained translation equivalent pairs. Comparing the rendering of WordNet data versus BabelNet data, we come to the following two main conclusions:

(1) In terms of recall on our initial Basque lemma list, BabelNet yields significantly higher rates than EusWN alone (around 40% compared to 30%), and, at the same time, the precision we have measured by manual assessments stays on a very similar level, close to 90%. This, of course, is recall and precision regarding an English-Basque dictionary draft, and if we wanted to produce new dictionaries for uncovered language pairs with English as pivot, we would have to also take into account the data for these third languages. As an example of a lexicographically uncovered language pair, we have measured the recall for Slovene translation equivalents on Basque lemmata (EusLemStd) comparing

[14] A list of dictionary websites that use WordNet data is found at https://wordnet.princeton.edu/wordnet/related-projects/.

bootstrapped dictionary draft data from wordnets and from BabelNet, with encouraging results. By linking EusWN to SloWNet (3.0 2015 version, Fišer et al., 2012), 66% of the synsets that contain EusLemStd lemmata also contain Slovene lexical items (16,291 synsets); on the other hand, 78% of the BabelNet synsets that contain EusLemStd Basque lemmata contain also Slovene items (22,864 synsets). As we have done here for Basque-English, a qualitative evaluation of the drafted Slovene-English mappings would be necessary, in order to predict the precision of a Basque-Slovene dictionary draft.

(2) Both EusWN and BabelNet 3.7 synsets are identified by unique ID codes that may be copied into the dictionary draft, following the goals discussed in Section 5 above. There is no guarantee for the stability of BabelNet synset mergings, and consequently of the corresponding synset ID codes, at least as for the current version 3.7, as we have pointed out in Section 4.2. The same problem also applies to WordNet data, but with an announced solution. EusWN synsets are linked one-to-one to PWN synsets, and their ID numbers correspond to an Interlingual Index that has been adapted from the sense inventory of Princeton WordNet 3.0, which means that it will not necessarily be compatible with future Princeton WordNet versions nor updated versions of other wordnets. As a possible solution, we are looking forward to the implementation of a stable, version-independent Global WordNet Grid (Vossen et al., 2016), a list of unique concept identifiers that will serve as central sense index across languages and future updates of wordnets.

In any case, we have shown that if a bilingual dictionary project starts from scratch, it makes sense to include a drafting of a word sense grid and translation equivalents in the workflow, starting with wordnet-like concept-oriented resources. Apart from the more obvious and doubtlessly very important advantage of reducing the manual effort in dictionary content editing, we point out a benefit, closely linked to the data model used that underlies the resources used here for bootstrapping. As soon as the lexicographical process goes on, i.e. the lexical data obtained from the dictionary draft are being edited, enriched, and linked to other lexicographic item types, they can be used for reciprocally enriching the resources of the wordnet-family, by retro-bootstrapping and inclusion, or by the definition of cross-resource links. Necessary conditions for a continuous mutual enrichment of this kind are the stability of synset IDs on the wordnet side, and the maintenance of an interoperable data model on the dictionary side.

For the Basque language, without taking into account the licence constraints that still apply in some cases, based on wordnets today we are able to produce bilingual dictionary drafts with about 70 languages. By bootstrapping BabelNet, we can obtain drafts with many more; we would start with a quantitative analysis of the mutual coverage (intersection) of every possible language pair in this very big resource. This, as we have shown, does not significantly lower the precision of its content in comparison to its nucleus, the multilingual wordnet, in spite of growing more and

more. If we connect any two languages by the methods described here, in both cases, i.e. using wordnets and using BabelNet, the English language functions as hub. Therefore, it makes sense to first evaluate the quality of the mappings between the desired languages and English, as we have done here for Basque.

For the part of a standard Basque dictionary headword list that today can be covered by the methods described here, a manual editing would allow to discover and to fill sense gaps, to improve the description of senses, and to correct errors. For the part of the list that is not covered, links to concepts that exist in English concept-based resources will have to be set. In some cases, for a Basque word sense no matching concept is listed in the originally English-based resources; the "discovered" concept will serve as an amendment to those, and so to a human and machine readable conceptualisation of our world.

7. Acknowledgements

The research leading to these results has received funding from the Basque Government (Research Group IT665-13). Funding is gratefully acknowledged.

8. References

Benjamin, M. (2016). Problems and Procedures to Make Wordnet Data (Retro)Fit for a Multilingual Dictionary. In *Proceedings of the Eighth Global WordNet Conference*. Bucharest, Romania, pp. 27-33.

Bond, F. & Foster, R. (2013). Linking and Extending an Open Multilingual Wordnet. In *Proceedings of the The 51st Annual Meeting of the Association for Computational Linguistics*

Euskaltzaindia. (2010). *Hiztegi batua*. Donostia: Elkar

Fišer, D., Gantar, P. & Krek, S. (2012). Using explicitly and implicitly encoded semantic relations to map Slovene Wordnet and Slovene Lexical Database. In *Semantic Relations-II. Enhancing Resources and Applications. Istanbul, Turkey*

Fišer, D., Novak, J. & Erjavec, T. (2012). sloWNet 3.0: development, extension and cleaning. In Proceedings oft he 6th International Global Wordnet Conference. *Matsue, Japan*, pp. 113-117.

Fišer, D. & Sagot, B. (2015). Constructing a poor man's wordnet in a resource-rich world. *Language Resources and Evaluation*, 49(3), pp. 601–635.

Gonzalez-Agirre, A., Laparra, E. & Rigau, G. (2012). Multilingual Central Repository version 3.0. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC'12*. Istanbul, Turkey.

Gouws, R. (2002). Equivalent Relations, Context and Cotext in Bilingual Dictionaries. *Hermes*, 28(1), pp. 195–209.

Hamp, B. & Feldweg, H. (1997). GermaNet - a Lexical-Semantic Net for German. In *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications.* Madrid, Spain, pp. 9–15.

Hartmann, R. R. K. (1990). The not so harmless drudgery of finding translation equivalents. *Language & Communication*, 10(1), pp. 47–55.

Kunze, C. (2010). Lexikalisch-semantische Ressourcen. In K.-U. Carstensen, C. Ebert, C. Ebert, S. J. Jekat, R. Klabunde & H. Langer (eds.), *Computerlinguistik und Sprachtechnologie: eine Einführung*. Heidelberg: Spektrum.

Lindemann, D. & San Vicente, I. (2015). Building Corpus-based Frequency Lemma Lists. *Procedia - Social and Behavioral Sciences*, *198*, pp. 266–277.

Lindemann, D., Saralegi, X., San Vicente, I., Manterola, I. & Nazar, R. (2014). Bilingual Dictionary Drafting. The example of German-Basque, a medium-density language pair. In *Proceedings of the XVI EURALEX International Congress. EURALEX 2012*. Bolzano, Italy, pp. 563–576.

Miller, G. A. (1998). Nouns in wordnet. In C. Fellbaum (Ed.), *WordNet: An electronic lexical database*. Cambridge MA: MIT Press, pp. 24-45.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D. & Miller, K. J. (1990). Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography*, 3(4), pp. 235–244

Navigli, R. & Ponzetto, S. P. (2010). BabelNet: Building a Very Large Multilingual Semantic Network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Stroudsburg, PA

Pociello, E., Agirre, E. & Aldezabal, I. (2011). Methodology and construction of the Basque WordNet. *Language Resources and Evaluation*, 45(2), pp. 121–142.

Prakash, R. S. S., Jurafsky, D. & Ng, A. Y. (2007). Learning to merge word senses. In *Proceedings of EMNLP-CoNLL 2007*

Saralegi, X., Manterola, I. & San Vicente, I. (2012). Building a Basque-Chinese Dictionary by Using English as Pivot. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation. LREC'12*. Istanbul, Turkey

Varga, I., Yokoyama, S., & Hashimoto, C. (2009). Dictionary generation for less-frequent language pairs using WordNet. In *Literary and Linguistic Computing*, 24(4), pp. 449–466.

Vossen, P. (2002). EuroWordNet General Document. University of Amsterdam.

Vossen, P., Bond, F., & McCrae, J. (2016). Towards a truly multilingual Global

Wordnet Grid. In *Proceedings of the Eighth Global WordNet Conference. Bucharest, Romania*, pp. 419-426.

Wiegand, H. E. (2002). Equivalence in bilingual lexicography: criticism and suggestions. *Lexikos*, 12(1), pp. 239–255.

The Main Features of the *e-Glava* Online Valency Dictionary

Matea Birtić, Ivana Brač, Siniša Runjaić

Institute of Croatian Language and Linguistics, Ulica Republike Austrije 16, HR-10000 Zagreb, Croatia

E.mail: mbirtic@ihjj.hr, ibrac@ihjj.hr, srunjaic@ihjj.hr

Abstract

E-Glava is an online valency dictionary of Croatian verbs. The theoretical approach to valency follows the German tradition, particularly that of the VALBU dictionary, with some minor changes and adjustments. The main principle of our valency approach is to link valency patterns to specific verb meanings. The verb list is compiled semi-automatically on the basis of the Croatian Frequency Dictionary and Croatian language textbooks. Currently, *e-Glava* contains descriptions of 57 psychological verbs with 187 meanings and 375 valency patterns. The lexicographic articles are written in Tschwanelex. A Document Type Definition editing module has been used, and the description of verbs follows a three-level linguistic schema prepared for lexicographers. Verbs are distributed throughout 34 semantic classes, and examples are extracted manually from Croatian corpora. Fully processed data for each semantic class will be publicly available in the form of a browsable HTML dictionary. The paper also presents a comparison between *e-Glava* and other cognate resources, as well as a summary of its main advantages, disadvantages, and potential applied uses.

Keywords: Croatian language; valency dictionary; e-dictionary; syntax

1. Introduction

Sentence structure and the syntactic behaviour of verbs were perhaps the most intriguing and interesting topics for early grammatical descriptions and, later, linguistic descriptions of language. Valency properties are relevant to both theoretical and applied linguistic considerations. One way to apply valency theory to real linguistic data is by processing valency e-lexicons and e-dictionaries and corresponding lexical databases intended for use by both humans and computers.

This paper will show the main features of one such e-dictionary, which was created for the Croatian language: *e-Glava*[1]. At present, *e-Glava* is a browsable HTML valency dictionary of Croatian verbs, and it represents the public results of the first phase of the *Valency Database of Croatian Verbs* project. It is accessible at http://valencije.ihjj.hr. It currently contains 57 verbs belonging to the semantic class of psychological verbs, with 187 meanings and 375 valency patterns. *E-Glava* is intended to serve as a tool for researchers interested in valency patterns of Croatian verbs, as well as a tool for teachers and students of Croatian as a second language and as an additional resource for linguistic data linking.

[1] In Croatian, *glava* means 'head'. It is also an abbreviation composed of *gla-* (short for *glagolska* 'verbal') and *-va* (short for *valencija* 'valency').

The first part of this paper is an introduction. The second part describes *e-Glava*'s approach to valency. The third part shows how the verb list was compiled, how the verbs were distributed throughout the semantic classes, and how corpora were used. The fourth part describes the layout of lemmas. The fifth part provides a brief description of the computational basis of e-*Glava*. In the sixth part the approaches of other online valency dictionaries are compared to e-*Glava*. The seventh and final part is a conclusion outlining the main advantages and disadvantages of *e-Glava*.

2. The approach to valency

The model of verb valency used in *e-Glava* is based on the fruitful results of German valency research and their lexicographic application in valency dictionaries (Helbig & Schenkel, 1973; Engel & Schumacher, 1978; Schumacher et al., 2004). Our direct model was the German VALBU valency dictionary (Schumacher et al., 2004), and its online version E-VALBU.

There are a number of other online dictionaries or lexical databases (also for Croatian) that process the syntactic environments of lexical units and valency in different frameworks. Differences and similarities of these databases to *e-Glava* will be described in the sixth section.

We have chosen a theoretical model based on the German valency tradition for two reasons: some previous theoretical discussions and lexicographic descriptions of verb valency in Croatian have also been written following the same tradition, such as Samardžija (1986) or Filipović (1993); and the model is simple enough that lexicographers with different backgrounds can master it. The basic assumption of VALBU's approach is the identification of valency complements at the level of sublemmas or meanings, not at the level of verb or lemma. The same principle is used in the description of verbs in *e-Glava*. A verb has one or several meanings, and each meaning or sublemma is linked with one or several valency patterns. We assume the sentence analysis used in traditional valency frameworks, whereby the verb is the center of the sentence. All syntactic (nonverbal) phrases, except for conjunctions and particles, are either complements or adjuncts. The verb selects the complement of a specific morphological form, which must have a special semantic relationship to a part of the meaning of the verb. Complements can be obligatory or optional, while adjuncts are never obligatory. Valency descriptions deal with optional and obligatory complements, while adjuncts are not part of the description. However, the practice is to record some common adjuncts as additional information belonging to the sublemma. Valency descriptions begin with the extraction of a part of a sentence that has been identified as a complement. Each complement is described as a morphological, syntactic, and semantic unit. The introduction of the morphological layer of analysis departs from the VALBU model, which describes complements only syntactically and semantically.

2.1. The syntactic level

We assume that 10 complement classes are needed at the syntactic level, i.e., for valency description in the narrow sense: Nominative Complement, Genitive Complement, Dative Complement, Accusative Complement, Instrumental Complement, Prepositional Complement, Adverbial Complement, Predicative Complement, Infinitive Complement and Sentential Complement.2 The VALBU model assumes eight classes of complements: Nominative Complement, Genitive Complement, Dative Complement, Accusative Complement, Prepositional Complement, Adverbial Complement, Predicative Complement and Sentential Complement. The Croatian model has one additional case (Instrumental Complement) due to the Croatian case system.3 Like the German model, we also use the Prepositional, Adverbial and Predicative Complements. One point at which we differ considerably from the VALBU model is in our treatment of Sentential and Infinitive Complements.

In the German model, complement sentences are viewed as a realization of either Case / Prepositional or Verbal Complement (*Verbativergänzung*). If sentences co-occur with verbs that otherwise take Case or Prepositional Complements, they are considered part of a specific Case or Prepositional Complement. If a sentence appears as a complement of a verb that does not take a Case or Prepositional Complement, the sentences together with the infinitives belong to the Verbal Complement. In a way, the VALBU model views sentences only as realizations of some other complement. In our model, all sentences as complements of verbs are regarded as a unique class of Sentential Complements, while infinitive complements belong to a separate class of Infinitive Complements. In the following passages, we will describe the 10 classes of complements in detail.

2.1.1. Nominative Complement

The Nominative Complement corresponds to the traditional concept of the subject. The majority of verbs in Croatian have a Nominative Complement and only a few do not. The Nominative Complement is always obligatory. Verbs which lack any complements are avalent verbs. There are also verbs which have one or two complements, neither of which belong to the Nominative Complement. Also, not all noun phrases with the nominative case belong to the Nominative Complement.
Nominative nouns or pronouns in copular sentences (*Žena je profesorica* (nom[4]) 'The woman is a teacher') or in secondary predication (*Marko je postao bogataš* (nom)

[2] Samardžija (1986) also assumes that for description of Croatian valency patterns ten complement classes are needed.

[3] The Croatian case system has seven cases, but there are five Case complement classes. The vocative is, today, never a case on an argument, and the locative case is always found within a prepositional phrase.

[4] Abbreviation used in this paper: nom = nominative; gen = genitive; dat = dative; acc = accusative; inst = instrumental; pl = plural; sg = singular; fem = feminine gender; pres = present tense; past = past tense.

'Marko became a rich man') belong to the Predicative Complement. In Croatian, which is a null pronoun language, pronominal subjects do not need to be expressed in the first and second person ('*Došla sam.*' Came – 1sg past fem 'I came', '*Jedeš.* (Eat – 2sg pres 'You are eating.'). We assume the Nominative Complement is also present in these sentences, though not realized. In such cases, a personal pronoun is added in parentheses following the verb. The way these examples are processed is illustrated in Figure 1.

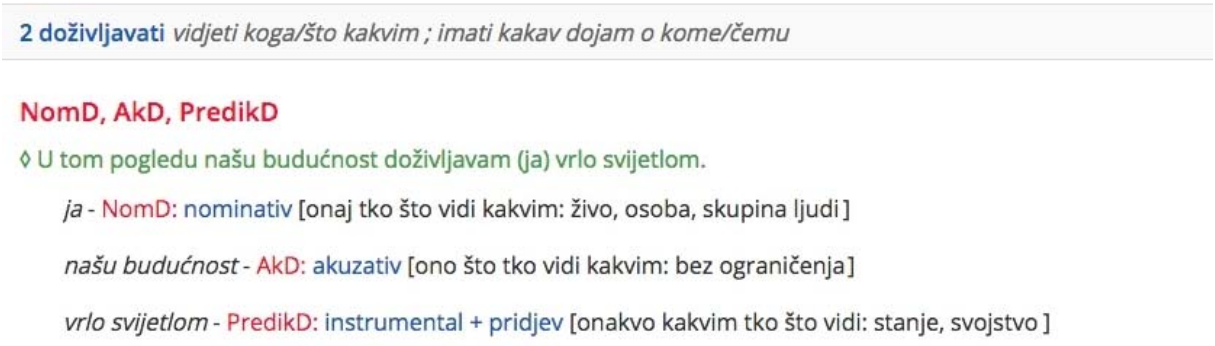


Figure 1: The layout of the verb with the unexpressed Nominative Complement

2.1.2. Genitive Complement

The Genitive Complement mostly corresponds to the genitive object (e.g., *Svijet se sjeća pape Wojtyle* 'The world remembers Pope Wojtyla'). Also, in processing psychological verbs, we decided to define the complements of some existential verbs as the Genitive Complement (*Ovdje nedostaje etike* (gen pl) *i morala* (gen pl) 'Ethics and morality are lacking here'). Genitive noun phrases with existential verbs are considered partitive genitive. Prototypical instances of partitive genitives are found in the object position where the genitive form replaces the accusative. Despite the similarities, we do not consider the partitive genitive in the object position as a separate (Genitive) complement, but rather a realization of the Accusative Complement. In the case of existential verbs, we find the partitive genitive only in some verb meanings, while other meanings use the nominative case. Thus, the nominative and genitive are not interchangeable in some verb meanings. This is why we have introduced a separate complement in the case of several existential verbs.

2.1.3. Dative Complement

The Dative Complement includes indirect objects and logical subjects marked with the dative case (*Oko se divi ljepoti* (dat) 'The eye admires beauty'; *Vrti mi* (dat) *se* 'I am dizzy'). The Dative Complement can be both obligatory and optional. Apart from being complements, nouns marked with the dative case are frequently adjuncts as well. Logical subjects in the dative case, dative experiencers, or dative stimuli with psychological verbs (*Blanki* (dat) *je dosadila duga kosa* 'Blanka is bored with long hair'); equatational datives (*Lijeva strana odgovara desnoj* (dat) 'The left side

corresponds to the right'); predicative datives (*Maslina pripada voću* (dat) 'The olive belongs to [the category of] fruit'); and some directional datives (*Prišao je djevojci* (dat) *na šanku* 'He approached the girl at the bar') are considered obligatory Dative Complements. Dative nouns with a thematic role of recipient frequently belong to the optional category of Dative Complements (*Ona mi* (dat) *se žalila na bolove* 'She complained (to me) of her pain'). The ethical dative is an adjunct (*Ona mi* (dat) *se danas dobro osjeća* 'She (to/for me) feels well today').

2.1.4. Accusative Complement

The Accusative Complement corresponds to the direct object. Not all noun phrases marked with the accusative case are part of the Accusative Complement. Some belong to the Adverbial Complement, also known as 'measure accusatives' (*Kaput je stajao hrpu novaca* (acc) 'The coat cost a pile of money'); or to adjuncts – very often to manner adjuncts (*Hodali su ruku pod ruku* 'They walked arm in arm'). However, cognate objects belong to the Accusative Complement due to their argumental properties (see Birtić & Matas Ivanković, 2009). As stated above, the partitive genitive and the genitive of negation in the object position are considered Accusative Complements.

2.1.5. Instrumental Complement

The Instrumental Complement comprises indirect objects in the instrumental case (*Ronaldo se ponosi sinom* (inst) 'Ronaldo is proud of his son') and of nominal phrases with the semantic role of instrument, which traditional grammars consider adjuncts (*Razveseljavali su nas svojim pričama* (inst) 'They cheered us up with their stories'; *Marko se oženio Ivanom* (inst) 'Marko married Ivana'). Some nouns in the instrumental case are part of a Predicative Complement (*Svi ga doživljavaju svecem* (inst) 'They all consider him a saint'). Also, many nouns in the instrumental case belong to adjuncts (*Hodao je ulicom* 'He walked down the street'). Instrumental Complements with divalent verbs are mostly obligatory, while Instrumental Complements with trivalent verbs are mostly optional.

2.1.6. Prepositional Complement

The Prepositional Complement is a complement described by traditional grammars as a prepositional object (*Zaljubila sam se u tebe* 'I fell in love with you'; *Ne ljute se svi roditelji na svoju djecu* 'Not all parents get angry at their children'). Prepositional phrases also belong to the category of Predicative Complements (*Smatrali su ga za prijatelja* 'They consider him a friend'); Adverbial Complements (*Ona živi u Londonu* 'She lives in London'); or frequently to the category of adjuncts (*Više se ne uzrujavam zbog sitnih pogrešaka* 'I do not get upset about minor errors anymore').

2.1.7. Adverbial Complements

Although most of adverbial phrases are optional adjuncts, it has been observed that some adverbials cannot be omitted, and their presence is decisive for the grammaticality of a sentence (Samardžija, 1986; Silić & Pranjković, 2005, Palić, 2011, Belaj & Tanacković Faletar, 2017). Such adverbials express location (*Ona živi u Londonu* 'She lives in London'; *Bacili su knjigu na stol* 'They threw the book on the table'); manner (*Ponašaju se nepristojno* 'They behave rudely'); cause (*Ta prava ne proistječu iz Ustava* 'These rights do not arise from the constitution'); measures of time and quantity (*Sjednica je trajala tri sata* 'The session lasted three hours'); and results (*Dijete na mlijeko reagira proljevom* 'The child reacts to milk with diarrhea'). The Adverbial Complement is obligatory or optional, but the separation between the optional Adverbial Complement and the adjunct is very complex, and depends mostly on the researcher's intuition and the chosen theory.

2.1.8. Predicative Complement

The Predicative Complement includes syntactic phrases considered part of the predicate, e.g. nouns and adjectives in copular sentences (*Profesor je šutljiv/budala* 'The professor is quiet / a fool') or part of secondary predications (*Oni svi su ga smatrali glupim / budalom / za budalu* 'They all consider him stupid / a fool / as a fool'). The Predicative Complement is realized by noun or adjective phrases in the nominative or instrumental case, by *kao*-phrase '*as*-phrase', prepositional phrase, or adverb. The Predicative Complement is always obligatory.

2.1.9. Infinitive Complement

In our approach, the Infinitive Complement represents a separate class of complements, although it is part of other complements in some models (e.g., in VALBU). Infinitives are often complements of modals and verbs that express phases of an action. Some verbs are not strictly modal, but they attain a modal component of meaning when used with an infinitive (*Bojim se ući* 'I am afraid to enter').

2.1.10. Sentential Complement

The Sentential Complement includes all sentences as complements of verbs. As mentioned above, the VALBU model considers some sentences as part of case and Prepositional Complements, while others (with verbs that do not take case or Prepositional Complements) belong to the Verbal Complement. We decided to keep all sentential complements in a separate complement class regardless of their co-occurrence with verbs which do or do not take case complements for two reasons. Firstly, sometimes it is difficult to decide whether a sentential complement actually substitutes another case complement. Hence, it is easier for a lexicographer to describe a syntactic environment of a verb. Secondly, from the viewpoint of the user,

it is easier to notice that a verb can take sentential complement instead of case complement if the information is conceptually and visually separated.

2.2. The morphological level

In addition to a syntactic description through 10 classes of complements, each complement is also described morphologically. *E-Glava* regards morphology as the realization of syntax. It is defined that syntactic (valency) complements are realized by four major morphological categories and a number of subcategories. The major morphological categories needed to morphologically describe syntactic complements in Croatian are (1) prepositions, (2) cases, (3) sentential realizations and (4) other. Prepositions include all Croatian prepositions, which amounts to 199. Cases include all Croatian morphological cases (nominative, genitive, dative, accusative, instrumental, and locative[5]) except for the vocative case, which is never realized on verbal arguments; it is always an independent phrase. Sentential realizations include the Croatian conjunctions (*da*, *što*, *kako*, *gdje*, *li*, WH-word, *neka*, *kao*+) and other elements by which a sentence can be introduced next to a verb (quotes and the zero conjunction). Quotes (marked with the word NAVOD[6]) and the zero conjunction (0) are listed alongside conjunctions. The fourth morphological category (other) includes (4.a.) adverbs and adverbial phrases, (4.b.) the infinitive, (4.c.) *kao*-phrase, (4.d.) quantificational phrases, and (4.e.) adjectives. As is apparent from the list above, morphological categories are not distributed in any meaningful way, but by functional principle. Some morphological realizations are mainly typical for some complements: *kao*-phrase and adjectives are frequently realizations of predicative complements.

2.3. Semantic level

Complements are semantically described in two layers: the verb-specific description of a participant, i.e. an individual semantic role, and the assignment of a semantic category to a specific complement.[7] For each complement, the individual semantic role is defined on the basis of the definition of the verb's meaning. Semantic categories can be chosen from a list amounting to 34 categories, most of which have been adopted from the VALBU dictionary. Categories such as animate, person, animal, plant, etc. are not organized hierarchically, so both animate and person must be chosen for each complement which can refer to a person. A more developed approach to semantic categories would be a hierarchically ordered tagset of semantic labels, which will be considered for introduction in the next phase of the project. The semantic category is not recorded if any noun can qualify as a realization of a specific

[5] Locative does not refer to a complement, but to a morphological subcategory, because for the description of locative prepositional phrases, the locative case must be chosen together with a specific preposition. The locative case never appears outside prepositional phrases in the Croatian language.

[6] Navod 'quote, quotation'.

[7] A similar kind of verb-specific description is also provided by VDE (2004), and some similar features can be found in FrameNet's descriptions of participants (Herbst, 2007: 25–26).

complement. In such cases, the complement is described as 'without restrictions'. Figure 2 below shows the semantic description of the nominative complement that appears with the verb *bojati se* 'fear' (*Marko se boji neprijatelja* 'Marko fears the enemy').

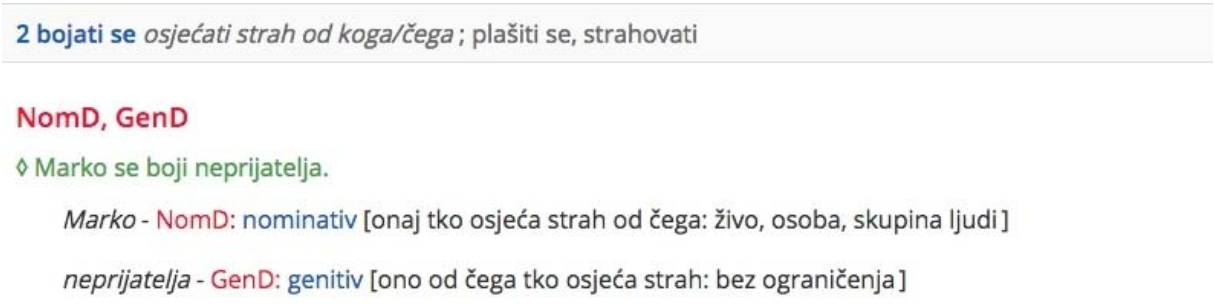


Figure 2: Two-layered semantic description of the Nominative Complement of the verb *bojati se* 'fear'

In addition, every semantic category can be preceded by the label *pren.*, which means 'figurative'. In cases where words are used metaphorically or metonymically, the figurative label is used.

3. Verb list, semantic classes and the usage of corpora

A verb list of approximately 900 of the most frequent verbs necessary for mastering Croatian at the B1 level according to the Common European Framework of Reference for Languages was extracted. The final list of verbs was compiled semi-automatically by comparing a verb list extracted manually from an older Croatian language resource, *Hrvatski čestotni rječnik* (Croatian Frequency Dictionary, Moguš, Bratanić & Tadić, 1999), and a verb list from more comprehensive textbooks of Croatian as a second language (e.g. Čilaš Mikulić et al., 2011; 2012; 2013).[8]

This list of 900 verbs intended for processing in e-*Glava* is distributed among 34[9] semantic classes and 91 subclasses. It is a well-known fact that verbs have several

[8] The lists of verbs used in texts are compiled at the end of the textbooks.

[9] Semantic classes in *e-Glava*: verbs of thinking, knowledge and learning; verbs of motion; verbs of communication; verbs of creation and transformation; verbs of positional change and placement; psych verbs; verbs involving the body (somatic verbs); verbs of social interaction; verbs of possession, taking and giving; verbs of change in possession; verbs of change in state; verbs of removing, separating and disassembling; verbs of ruling, control and influence; verbs of perception; verbs of effort and intention; verbs of emission; verbs of killing and hurting; verbs of placement in space; verbs of ingesting; aspectual verbs; verbs of carrying and sending; verbs related to money; general actions; verbs of combining and attaching; verbs of keeping and caring; verbs of inhabiting and staying; verbs of fighting; verbs of usage; verbs of happening; verbs of lingering and rushing; existential verbs; verbs of relations; verbs of judgment and success; weather verbs; and verbs of sounds made by animals.

meanings, and that the most frequent meaning does not always correspond to the prototypical one, so it is important to choose which criteria are to be used for classification. We decided to classify the verbs according to the first meaning written in two monolingual Croatian language dictionaries: *Školski rječnik hrvatskoga jezika* (Croatian School Dictionary) (Birtić et al., 2012), and the *Hrvatski jezični portal* (Croatian Language Portal) online dictionary (http://hjp.znanje.hr/). If these dictionaries did not have the same meaning written in the first place, we followed *Školski rječnik*, because it is a corpus-based dictionary (Birtić et al., 2012: xii). Our general classification is inspired by Levin (1993), but it relies more on verb semantics than syntax as compared to Levin's approach, which classifies verbs mainly on the basis of syntactic alternations. As will become clear below, each verb belongs to one prototypical semantic class, but their different meanings also allow them to belong to other semantic classes. This multiple categorisation is enabled through the ability to choose a semantic class at different levels in the description. The prototypical semantic class is written next to the lemma, and possible changes in semantic class are recorded next to the sublemma, i.e., a specific meaning of the verb.

As the verbs are processed according to their semantic classification, not according to alphabetical order, semantic class is considered a module (Klosa, 2013) or a phase in the lexicographic process. The advantage of this approach is that it enables the observation of syntactic and semantic differences between similar verbs, or of syntactic alternations in the same semantic class, such as the well-known syntactic alternations in psychological verbs (psychological verbs can express an experiencer either as subject or object, and in Croatian, a language with morphological cases, the experiencer can be realized as a noun in the nominative, accusative, or dative case). An additional advantage is that the combined processing of verbs of the same semantic class enables non-native speakers to learn how to presuppose valency patterns according to the semantic group the verb belongs to.

The processing of verbs in *e-Glava* is based on two Croatian corpora: *Hrvatska jezična riznica* (Croatian Language Repository) and *Hrvatski mrežni korpus* – hrWaC (Croatian web corpus – hrWaC), but is not directly linked to any (annotated) corpora. The Croatian Language Repository, which is also compiled at the Institute of Croatian Language and Linguistics, did not comprise annotated corpora when the project *e-Glava* begun, but its annotation has recently started. Manual corpora research is relevant at the three stages of verb processing. Firstly, corpus is a tool which enables us to check definitions of verb meanings already noted in existing dictionaries. It helps us to find the meanings of the verbs that have not yet been recorded. Secondly, after all the meanings of a verb have been identified, the corpus is searched to find valency patterns which belong to each meaning. Finally, the corpus examples are selected manually and entered into a database.

4. The three-level description of verbs in e-Glava

E-Glava describes verbs on three levels. The first level provides information regarding the verb overall, the second level introduces different meanings of the verb, and the third level is a valency description.

4.1. The first level

The first level consists of a verb lemma in the infinitive, except for inherent reflexive verbs, which are entered with the reflexive particle *se*. Each lemma or verb is connected with four sections: a grammatical block, the prototypical semantic class of the verb and its subclass, idioms and collocations, and notes. The grammatical block encompasses verb inflections (first person singular present, third person plural present, masculine perfect participle, feminine perfect participle and masculine passive participle), and an aspect label. The aspect label includes abbreviations for imperfective, perfective and biaspectual values. In *e-Glava*, the semantic class of a verb is visualized directly below the lemma and above the verb inflections. The idiom and collocation block is placed at the end of the lemma visualization. It consists of a collocation or an idiom (e.g. *mrziti iz dna duše* 'to hate from the depths of one's soul'); its definition (*jako mrziti koga ili što* 'to strongly hate someone or something'); and a usage example (*Ako idete na posao, mrzit ćete budilicu iz dna duše* 'If you go to work, you will hate your alarm clock from the depths of your soul'). The note block contains information that applies to the verb overall, not to one of its meanings or a separate valency pattern (for example, the remark that a specific verb is non-standard or is used only in a specific style).

4.2. The second level

The second description level consists of different meanings of verbs, which are introduced by numbered sublemmas (e.g., *1 mrziti, 2 mrziti* 'hate'). Each sublemma is connected with a reflexive label, a definition, a possibility of changing a verb's semantic class, and additional information. The reflexive label has two values: reflexive and zero. The reflexive value mostly serves to mark the reflexivity of reflexive verbs which are not *reflexiva tantum* or inherently reflexive, i.e., those entered with particle *se*. All reflexive verbs that are not inherently reflexive are treated as sublemmas, i.e., as one of the meanings of the verb. Definitions consist of three parts: a stylistic label, paraphrase definitions (two can be entered) and synonyms. The stylistic label (e.g., historical, poetic) precedes the definition.

An illustration of the first and second levels of the description of the verb *vrijeđati* 'offend, insult, irritate', with an introduction of the separate sublemmas for particular meanings, is provided in Figure 3.

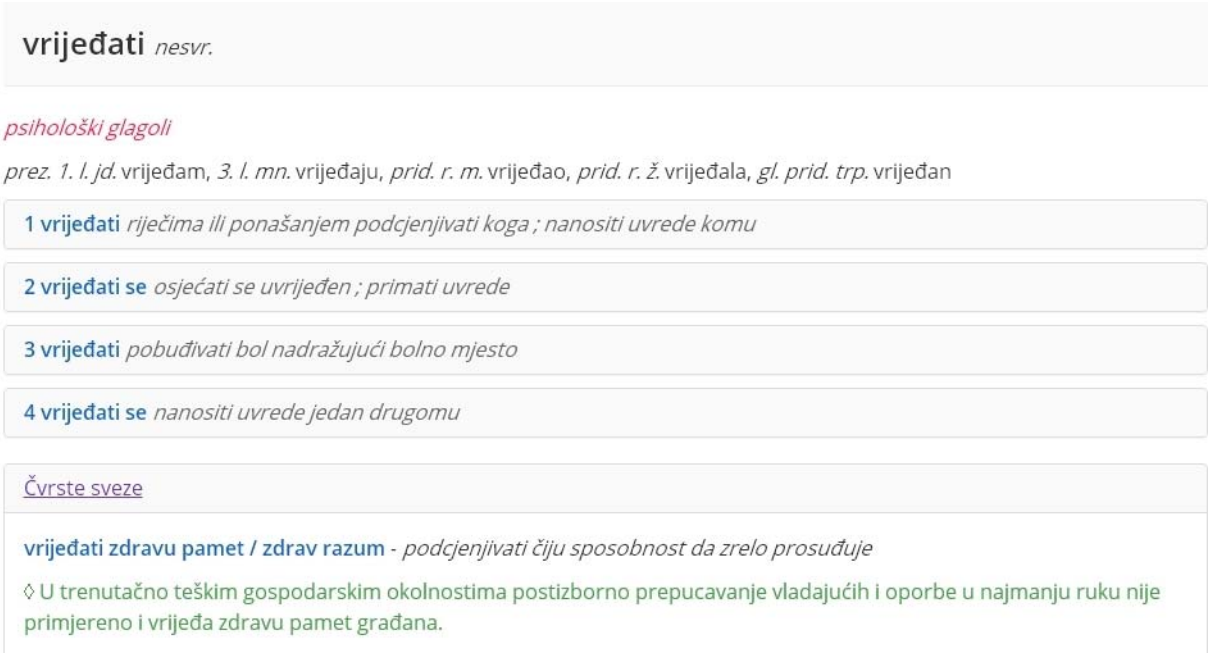


Figure 3: An illustration of the first and second level of the description of the verb *vrijeđati* 'offend, insult, irritate'

4.3. The third level

Clicking on a sublemma brings the user to the third level, which contains the valency analysis. The valency analysis consists of an example sentence and parts of sentences recognized as valency complements. Valency analyses contain a morphological, syntactic and semantic description of a complement (in square brackets). Above the detailed valency analysis, valency patterns are written as abbreviations of complements (e.g., NomD, InfD).[10] Each meaning can be associated with several valency patterns, and each valency pattern can be linked to several examples.

This is illustrated in Figure 4, which provides the complete processing of the verb *živcirati* 'to upset someone/to become irritated'. This illustration shows the sentence examples, which are introduced with a diamond. The example section, shown below, is connected to the syntactic, morphological and semantic descriptions with a hyphen.

[10] NomD is an abbreviation for *Nominativna dopuna* 'Nominative Complement', InfD is an abbreviation for *Infinitivna dopuna* 'Infinitive Complement'.

živcirati *nesvr.*

psihološki glagoli

prez. 1. l. jd. živciram, *3. l. mn.* živciraju, *prid. r. m.* živcirao, *prid. r. ž.* živcirala, *gl. prid. trp.* živciran

1 živcirati *izazivati u kome živčanost*

NomD, AkD

◊ Mene zaista živciraju takve špekulacije.

takve špekulacije - NomD: nominativ [ono što u kome izaziva živčanost: bez ograničenja]

mene - AkD: akuzativ [onaj u kome što izaziva živčanost: živo, osoba, skupina ljudi]

◊ Našu javnost najviše intrigiraju i živciraju Balkan i balkanizacija.

Balkan i balkanizacija - NomD: nominativ [ono što u kome izaziva živčanost: bez ograničenja]

našu javnost - AkD: akuzativ [onaj u kome što izaziva živčanost: živo, osoba, skupina ljudi]

RečD, AkD

◊ Douglasa posebno živcira što se u svakoj prigodi nježno drže za ruke.

što se u svakoj prigodi nježno drže za ruke - RečD: što [ono što u kome izaziva živčanost: propozicija]

Douglasa - AkD: akuzativ [onaj u kome što izaziva živčanost: živo, osoba, skupina ljudi]

NomD, AkD (InstD)

◊ Ona me je živcirala neobičnošću i ekscentričnošću.

ona - NomD: nominativ [ono što u kome izaziva živčanost: bez ograničenja]

me - AkD: akuzativ [onaj u kome što izaziva živčanost: živo, osoba, skupina ljudi]

neobičnošću i ekscentričnošću - (InstD): instrumental [ono čime što u kome izaziva živčanost: bez ograničenja]

◊ Populistička lokalna stranka živcirala je ostatak nacije svojim istrijanstvom.

populistička lokalna stranka - NomD: nominativ [ono što u kome izaziva živčanost: bez ograničenja]

ostatak nacije - AkD: akuzativ [onaj u kome što izaziva živčanost: živo, osoba, skupina ljudi]

svojim istrijanstvom - (InstD): instrumental [ono čime što u kome izaziva živčanost: bez ograničenja]

2 živcirati se *biti živčan, katkad izazvan čime ; osjećati se živčan*

NomD

◊ Ja se uvijek živciram kad kasnim.

ja - NomD: nominativ [onaj tko je živčan: živo, osoba, skupina ljudi]

◊ Hrvatska će se još živcirati u žici stvorenih obveza.

Hrvatska - NomD: nominativ [onaj tko je živčan: pren. geografsko mjesto]

NomD (PrijeD)

◊ Ministrica pravosuđa nije se živcirala oko rasprave o reformi pravosuđa.

ministrica pravosuđa - NomD: nominativ [onaj tko je živčan: živo, osoba, skupina ljudi]

oko rasprave o reformi pravosuđa - (PrijeD): oko + genitiv [ono čime je izazvano da je tko živčan: bez ograničenja]

Obavijest: U ovome značenju (živcirati se 2) uz glagol se često pojavljuje dodatak uzroka ostvaren prijedložnom skupinom s prijedlogom *zbog* i imenicom u genitivu (*zbog* + gen.), npr. *Više se ne živciram zbog sitnih pogrešaka kao prije.*

Čvrste sveze

Figure 4: The complete layout of the verb *živcirati* 'to upset someone/to become irritated'

5. The computational basis of e-Glava

In 2013, a newly formed team of researchers initiated the *Valency Database of Croatian Verbs* project at the *Institute of Croatian Language and Linguistics*, and a linguistic model had been chosen by the end of 2014. Valency had been researched at the Institute prior to this, but the outcomes of these descriptions were compiled as non-structured or linear data. As a part of preparation[11] we had to re-evaluate the entire concept, and the team had to decide whether to develop its own customized Content Management System (CMS) or to use an existing lexicographic package.[12] Considering the fact that there was no funding for the project, and that the team members had previous experience in compiling dictionaries using TshwaneLex, we began to develop a three-level linguistic schema for a valency dictionary in TshwaneLex (see Section 4), which we considered a computerisation phase of our lexicographic process. Accordingly, we began writing new lexicographic entries in the prepared TschwaneLex schema for 57 psychological verbs. The I.T. department attempted to make the dictionary entry writing process as precise and user-friendly as possible for researchers and lexicographers, mostly through the implementation of drop-down menus and controlled multiple choice options for all linguistic features.

After this small dictionary of psychological verbs was compiled, it was made publicly available in order to receive initial feedback from fellow researchers and other interested parties. Although the dictionary grammar was developed using a Document Type Definition (DTD) editing module of TshwaneLex and an ODBC connection, and the DTD was automatically transcribed into a PostgreSQL database environment, the project team still had to make some adjustments before the data could be presented on an internet platform. We decided to export the native XML file for all verbs within the semantic class that were marked "completed" to an easily-accessible SQL database. This process made the part of the dictionary that we consider completed, automatically browsable through a web-based search engine using PHP and HTML5. This gave researchers the ability to make verbs currently being described (the semantic class of verbs of moving and putting) available by

[11] Klosa (2013) has defined six phases in computer lexicographic process for online dictionaries under construction: the phase of preparation, the phase of data acquisition, the phase of computerization, the phase of data processing, the phase of data analysis, and the phase of preparation for online release. The phase of preparation is partly described in this section and in Section 3 (criteria for choosing verbs for a verb list). In the phase of data acquisition we decided to use the *Croatian Language Repository*, the *Croatian web corpus – hrWaC*, and the *Croatian Frequency Dictionary* as primary sources. Our secondary sources were textbooks of Croatian as a second language, *Školski rječnik hrvatskog jezika* and *Hrvatski jezični portal* (Section 3). The corpus designed especially for the purpose of *e-Glava* was omitted from this project. The phase of computerization and data processing is described in this section (5): the choice of dictionary writing system and the specification of database system. The phase of data analysis is presented to a lesser extent in last part of Section 3 (the usage of corpora) and mostly in Section 4. Finally, the phase of preparation for online release is described at the end of this section (5). As Klosa (2013) states, following Klein (2004): "all phases of the computer-lexicographical process merge giving yet unknown flexibility to the lexicographer."

[12] For more details, see Birtić & Nahod (2016: 103–105).

exporting an updated XML file, which then goes "live" on the website. In addition to this first version, which is browsable by lemma, an advanced search function is being developed which will enable users to search by specific categories, such as valency complements, morphological forms, or semantic features.

6. A comparison of e- Glava and other online dictionaries and lexicons

This section compares the main features of *e-Glava* to those of some other well-known online dictionaries (FrameNet, FrameBank, VALLEX, Crovallex, VALBU).

One of the most well-known online dictionaries is UC Berkeley's FrameNet, which is based on the theory of frame semantics (Fillmore & Baker, 2010). The most notable difference between descriptions of verbs in *e-Glava* and descriptions of nouns, adjectives, and verbs in FrameNet is their ordering and the hierarchy of their syntactic, morphological and syntactic descriptions. While *e-Glava* begins its valency description with the syntactic level, followed by morphological and semantic layers, FrameNet begins from the semantic layer in accordance with the theory of frame semantics (Fillmore & Baker, 2010). FrameNet derives grammatical function (external argument, object and dependent) and phrase type algorithmically (Ruppenhofer et al., 2016: 41) based on frame element label (semantic role), position in the sentence, and part of speech. Deriving grammatical functions from the position of phrases in sentences is not possible for Slavic languages with free word order. We believe that detailed descriptions of both morphology and syntax are essential for languages with rich morphological systems. For example, the Russian FrameBank also employs morphological descriptions. As can be concluded, *e-Glava* differs considerably from FrameNet in several respects: it deals only with verbs; its starting point is syntax; examples are extracted manually (FrameNet automatically extracts examples from the British National Corpus); and word order is not taken in account.

Semantic and syntactic verb descriptions are a part of the Russian FrameBank (Lyashevskaya & Kashkin, 2011; Lyashevskaya, 2012) and the Czech VALLEX (Kettnerová, Lopatková & Bejček, 2012; Lopatková et al., 2006). The differences between e-*Glava* and FrameBank or VALLEX are less significant than the differences between e-*Glava* and FrameNet. Unlike FrameNet, FrameBank and e-*Glava* take morphology into account. FrameBank and *e-Glava* share some units of description: e.g., the morphosyntactic features of elements in FrameBank and morphological descriptions in e-*Glava*; the lexical-semantic class of elements in FrameBank and semantic categories in e-*Glava* (e.g., human, animate); and the division of complements into optional and obligatory. FrameBank also includes the syntactic rank of elements / grammatical functions (e.g., subject, object, predicate, peripheral and clause) and the semantic roles of arguments (Agent, Patient and Instrument). FrameBank consists of examples taken randomly from the annotated Russian National Corpus (Lyashevskaya & Kashkin, 2011), while *e-Glava* is not linked to any

annotated corpora. In *e-Glava,* examples are chosen intentionally as the best fit for meanings and valency descriptions. FrameBank is based on Construction Grammar (Goldberg, 1995) and the Moscow Semantic School (e.g., Apresjan, 1995).

E-Glava also shares similarities with the Valency Lexicon of Czech Verbs VALLEX which also focuses on the most frequent verbs and their meanings. VALLEX and *e-Glava* share the same general approach to valency: valency patterns are identified at the level of particular verb meanings, not at the level of the verb. VALLEX also provides information on the number of complements, functors, or semantic roles, their morphological realizations, and the obligatoriness of complements (Kettnerová, Lopatková & Bejček, 2012). The same information is provided by *e-Glava*, except that semantic descriptions in *e-Glava* use an individual semantic role and semantic category, not general semantic roles (functors). Both e-dictionaries provide some additional information about idioms, reflexivity, reciprocity and aspect. Reflexivity and aspect values are approached differently in VALLEX and *e-Glava*. Imperfective and perfective verbs are considered the same entry in VALLEX, whereas the perfective and imperfective variants of verbs are considered two separate entries in *e-Glava*. Each imperfective verb in e-*Glava* does not need to have its perfective pair entered by default: each verb lemma is entered independently depending on its frequency of appearance. *E-Glava* enters *reflexiva tantum* or inherent reflexive verbs as separate lemmas, whereas all other reflexive verbs are considered sublemmas of lemma. VALLEX also records *reflexiva tantum* as separate lemmas, but in addition to this, it treats derived reflexives[13] as separate lemmas as well (for more on this, see Oraić Rabušić & Bošnjak Botica, 2016; Kettnerová & Lopatková, 2014). VALLEX entries are also manually taken from the Czech National Corpus. VALLEX divides verbs into 22 semantic classes according to their prototypical meaning, which is based on intuition (Lopatková et al., 2006: xxiii). As we have already stated, verb classification in *e-Glava* is performed in a more systematic and precise manner than in VALLEX. Verbs belong to a prototypical semantic class and can be linked to one or more additional semantic class. In contrast, VALLEX associates each verb only with one semantic class. For some other Slavic languages electronic valency dictionaries or dictionaries including verb descriptions are available, e.g., the Polish Valency Dictionary (Walenty) (Przepiórkowski et al., 2014), Slovene Lexical Database (Gantar & Krek, 2011).

Needless to say, it is very important to mention another online Croatian valency dictionary, CROVALLEX, developed by Mikelić Preradović (2008; 2010). It describes 1,739 verbs with 5,118 valence frames classified into 72 semantic classes and subclasses (173 in total). The number of verb lemmas exceeds the designated number of verbs in *e-Glava*. Just like VALLEX, CROVALLEX also enters only *reflexiva*

[13] Derived reflexives are verbs derived from a corresponding non-reflexive verb, but their meaning is so distant from their non-reflexive counterpart that they must be viewed as a separate verb.

tantum and derived reflexives as separate lemmas. As in *e-Glava*, valency is related to meaning, and a valency frame example and class is defined for each meaning. Slots in valence frames are filled with functors, which can be inner participants and free modifications. Functors roughly correspond to deep cases (Agent, Patient, Recipient, Result and Origin) and can appear in a sentence only once. There are about 30 free modifications, and they can appear in a valence frame more than once. According to descriptions in CROVALLEX, inner participants and free modifications can be optional or obligatory (despite the term free). The valence frame is notated with abbreviations of functors. Obligatory vs. optional status is marked in superscript, while morphological form is marked in subscript with the abbreviation of a functor. If the approaches to valency used in both Croatian dictionaries are compared, it can be said that *e-Glava* is more syntax-oriented than CROVALLEX, in which semantic description prevails despite the presence of both syntactic and morphological descriptions. Both share the principle of defining complements on the level of meaning. Verb meanings are finer-grained in *e-Glava*, as they are defined and divided on the basis of Croatian corpora, and do not rely only on available dictionaries. At the level of sentence periphery, CROVALLEX provides more phrases which are considered adjuncts in *e-Glava*. In CROVALLEX, idioms and collocations are listed as a part of verb meaning, while in *e-Glava* they form a separate unit. In terms of semantic classes, CROVALLEX defines a new semantic class for each (different) meaning, but, as opposed to *e-Glava*, it does not specify the prototypical semantic class of a verb.

Finally, although we have followed VALBU quite consistently, there are some points in our treatment of valency in which we depart from our model, as has been mentioned in several parts of this paper. Specifically, we treat reflexive verbs differently: VALBU enters each reflexive verb as a separate lemma entry; we have added morphological descriptions, which are justified for languages with rich morphology; we have introduced semantic verb classes; and we treat Sentential and Infinitive Complements quite differently from VALBU.

7. Conclusion

In conclusion, we would like to outline what has been done so far and set out the main advantages and disadvantages of *e-Glava*. The first version of *e-Glava* is available online and is accessible for free. It offers a detailed description of the syntactic and semantic interface of one semantic class of verbs. Additionally, many verb meanings that are not found in dictionaries of the contemporary Croatian language are described in *e-Glava* thanks to its corpus-based analysis. Consequently, semantic switches and new uses are described. Since it is sometimes an intricate task to properly assign a semantic role to a specific participant, we decided to use semantic (conceptual) categories, e.g., person, animal, place, etc. We also believe these categories to be more intuitively recognizable for dictionary users without formal linguistic expertise. The main disadvantage of *e-Glava* is its manual extraction

of examples and descriptions, which is time-consuming, resulting in slow project progress. On the other hand, this kind of lexicographic work guarantees better and more reliable descriptions.

When we think about other possible usage advantages, it occurs that mastering verb valency is a very important part of language learning, in particular when it comes to learning Slavic languages. *E-Glava* allows non-native speakers to check verb meanings, syntactic patterns, and their morphological realizations. Consequently, *e-Glava* might become a useful tool for learning Croatian as a second language. However, learners should possess a basic understanding of Croatian, as all definitions with simple metalanguage are written in the Croatian language. To master a second language at a higher level, an understanding of idiomatic phrases is also important. Idioms are included and visually represented in a special field separate from the syntactic patterns, and so their meanings can be easily explained to learners.

E-Glava's data could also become an additional resource for linguistic data linking in comprehensive research on the Croatian language. Its detailed descriptions can be used as the starting point for various lexical resources, as the syntactic, semantic and morphological levels are represented as structured data. Related ongoing projects at the Institute of Croatian Language and Linguistics, such as the *Croatian e-Dictionary* (*MREŽNIK*), the *Croatian Collocation Database*, and the *Croatian Metaphor Repository*, could certainly benefit from it. Moreover, *e-Glava*'s research team is open to providing all project data in open format to the greater NLP community in Croatia if they consider it usable for the morphosyntactic and semantic tagging and parsing of corpora or for other processes.

8. References

Apresjan, J.D. (1995). *Izbrannye trudy, tom I. Leksicheskaja semantika.* Moskva: Jazyki Russkoj Kul'tury, Vostochnaja Literatura.

Belaj, B. & Tanacković Faletar, G. (2017). *Kognitivna gramatika hrvatskoga jezika: sintaksa jednostavne rečenice.* Zagreb: Disput.

Birtić, M. & Matas Ivanković, I. (2009). Akuzativne dopune uz neprijelazne glagole: što su unutrašnji objekti?. In *Rasprave Instituta za hrvatski jezik i jezikoslovlje*, 35(1), pp. 1–19.

Birtić, M. & Nahod, B. (2016). An outline of the online valency dictionary of Croatian verbs. In Karolína Skwarska & Elżbieta Kaczmarska (eds.) *Research of Verbal Valency in Slavic Languages in the Past and Present = Výzkum slovesné valence ve slovanských zemích včera a dnes.* Praha: Slovanský ústav AV ČR, v.v.i., pp. 103–116.

Birtić, M. et al. (2012). *Školski rječnik hrvatskoga jezika.* Zagreb: Institut za hrvatski jezik i jezikoslovlje & Školska knjiga.

CROVALLEX 2.0008: *The Croatian Valency Lexicon of Verbs.* (2008.) at: http://theta.ffzg.hr/crovallex/ (1-10 May 2017).

Čilaš Mikulić, M. et al. (2011). *Hrvatski za početnike 1*: *vježbenica i gramatički pregled hrvatskoga kao drugog i stranog jezika*. Zagreb: Hrvatska sveučilišna naklada.

Čilaš Mikulić, M. et al. (2012). *Razgovarajte s nama*! *B1-B2*: *vježbenica, gramatika i fonetika hrvatskog jezika za niži srednji stupanj*. Zagreb: FF press.

Čilaš Mikulić, M. et al. (2013). *Razgovarajte s nama*! *A2-B1*: *vježbenica, gramatika i fonetika hrvatskog jezika za niži srednji stupanj*. Zagreb: FF press.

E-VALBU: *das elektronische Valenzwörterbuch deutscher Verben*. Accessed at: http://hypermedia.ids-mannheim.de/evalbu/index.html. (10 May 2017)

Engel, U. & Schumacher, H. (1978). Kleines Valenzlexikon deutscher Verben. Manheim: Institut für deutsche Sprache.

Filipović, R. (ed.) (1993) *Teorija valentnosti i rječnik valentnosti hrvatskih glagola, Kontrastivna analiza engleskog i hrvatskog jezika IV*. Zagreb: Zavod za lingvistiku Filozofskoga fakulteta u Zagrebu.

Fillmore, C.J. & Baker, C.F. (2010). A frames approach to semantic analysis. In Heine, B. & Narrog, H. (eds.) *Oxford Handbook of Linguistic Analysis*. New York: Oxford University Press, pp. 313–341.

Gantar, P. & Krek, S. (2011). Slovene lexical database. In D. Majchraková & R. Garabík (eds.) *Natural language processing*, *multilinguality*: *sixth international conference*. Modra, Slovakia, pp. 72–80.

Goldberg, A. E. (1995). *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago & London: The University of Chicago Press.

Herbst, T. (2007). Valency complements or valency patterns? In T. Herbst & K. Götz-Votteler (eds.) *Valency: Theoretical, Descriptive and Cognitive Issues*. Berlin: Walter de Gruyter, pp. 15–35.

Helbig, G. & Schenkel, W. (1973). Wörterbuch zur Valenz und Distribution deutsher Verben, Leipzig VEB Bibliographisches Institut.

Hrvatska jezična riznica = Croatian Language Repository. Accessed at: http://riznica.ihjj.hr/philologic/. (10 May 2017)

Hrvatski mrežni korpus: hrWac = Croatian web corpus. Accessed at: http://nl.ijs.si/noske/all.cgi/first_form?corpname=hrwac. (10 May 2017)

Kettnerová, V., Lopatková, M. & Bejček, E. (2012). The Syntax-Semantics Interface of Czech Verbs in the Valency Lexicon. In R. Vatvedt Fjeld & J.-M. Torjusen (eds.) *Proceedings of the 15th EURALEX International Congress*. Oslo: Department of Linguistics and Scandinavian Studies, University of Oslo, pp. 434–443.

Kettnerová, V. & Lopatková, M. (2014). Reflexive Verbs in a Valency Lexicon: The Case of Czech Reflexive Morphemes. In A. Abel, Ch. Vettori, & N. Ralli (eds.) *Proceedings of the XVI EURALEX International Congress: The User in Focus*, *EURALEX 2014*. Bolzano: Institute for Specialised Communication and Multilingualism, pp. 1007–1022.

Kipper Schuler, K. (2005). *Verbnet*: *A Broad-Coverage*, *Comprehensive Verb Lexicon*. Doctoral dissertation, University of Pennsylvania.

Levin, B. (1993). *English Verb Classes and Alternations*. The University of Chicago Press, Chicago – London.

Lopatková, M. et al. (2006). *Valency Lexicon of Czech Verbs VALLEX 2.0.* ÚFAL Technical Report TR-2006-34. Available at: https://ufal.mff.cuni.cz/~lopatkova/literatura/06-TR-vallex-2.0.pdf.

Lyashevskaya, O. & Kashkin, E. (2011). FrameBank: a database of Russian lexical constructions. In *Communications in Computer and Information Science Springer*, Vol. 542. Springer Verlag: Berlin & Heidelberg, pp. 337–348.

Lyashevskaya, O. (2012). Dictionary of Valencies Meets Corpus Annotation: A Case of Russian FrameBank. In R. Vatvedt Fjeld & J.-M. Torjusen (eds.) *Proceedings of the 15th EURALEX International Congress*. Oslo: Department of Linguistics and Scandinavian Studies, University of Oslo, pp. 1023–1030.

Klosa, A. (2013). The lexicographic process (with special focus on online dictionaries). In R. F. Gouws et al. (eds.) *Dictionaries. An International Encyclopedia of Lexicography*. Supplementary Volume: *Recent Developments with Focus on Electronic and Computational Lexicography.* De Gruyter.

Mikelić Preradović, N. (2008). *Pristupi izradi strojnog tezaurusa za hrvatski jezik*, doctoral thesis, Faculty of Humanities and Social Sciences, University of Zagreb.

Mikelić Preradović, N. (2010). Semantic classification of verbs in CROVALLEX. In S. Lagakos at al. (eds.) *Recent Advances in Computer Engineering and Applications* 1. Harvard University, Cambridge, USA, pp. 53–59.

Moguš, M., Bratanić, M. & Tadić, M. (1999). *Hrvatski čestotni rječnik*. Zagreb: Zavod za lingvistiku Filozofskog fakulteta & Školska knjiga.

Oraić Rabušić, I. & Bošnjak Botica, T. (2016). Chorvatský vs český model valenčního popisu. In K. Skwarska & E. Kaczmarska (eds.) *Research of Verbal Valency in Slavic Languages in the Past and Present = Výzkum slovesné valence ve slovanských zemích včera a dnes.* Praha: Slovanský ústav AV ČR, v.v.i., pp. 305–318.

Palić, I. (2011). O glagolima koji vežu obvezatne adverbijalne dopune u bosanskome jeziku. *Suvremena lingvistika,* 37(72), pp. 201–217.

Przepiórkowski, A. et al. (2014). Extended phraseological information in a valence dictionary for NLP applications. In *Proceedings of the Workshop on Lexical and Grammatical Resources for Language Processing* (*LG-LP 2014*). Dublin, Ireland, pp. 83–91.

Ruppenhofer, J. et al. (2016). FrameNet II: Extended Theory and Practice. https://framenet2.icsi.berkeley.edu/docs/r1.7/book.pdf (30 April 2017)

Samardžija, M. (1986). *Valentnost glagola u suvremenom hrvatskom književnom jeziku.* Doctoral thesis, Faculty of Humanities and Social Sciences, University of Zagreb.

Schumacher, H. et al. (2004). *VALBU – Valenzwörterbuch deutscher Verben*. Tübingen: Gunter Narr Verlag.

Silić, J. & Pranjković, I. (2005). *Gramatika hrvatskoga jezika za gimnazije i visoka učilišta*. Zagreb: Školska knjiga.

VDE (2004): *A Valency Dictionary of English: A Corpus-Based Analysis of the Complementation Patterns of English Verbs, Nouns and Adjectives.* Berlin–New York: Mouton de Gruyter.

VALLEX 3.0: *Valency Lexicon of Czech Verbs.* Accessed at: http://ufal.mff.cuni.cz/vallex. (30 April 2017).

Specifying Hyponymy Subtypes and Knowledge Patterns: A Corpus-based Study

Juan Carlos Gil-Berrozpe, Pilar León-Araúz, Pamela Faber

University of Granada
Department of Translation and Interpreting, Buensuceso 11, 18071 Granada, Spain
E-mail: juancarlosgb@correo.ugr.es, pleon@ugr.es, pfaber@ugr.es

Abstract

The organization of a terminological knowledge base (TKB) relies on the identification of relations between concepts. This involves making an inventory of semantic relations and extracting these relations from a corpus by means of knowledge patterns (KPs). In EcoLexicon, a multilingual and multimodal TKB on the environment, 17 semantic relations are currently being used to link environmental concepts. These relations include six subtypes of meronymy, but only one subtype of hyponymy (*type_of*). However, a recent pilot study (Gil-Berrozpe et al., in press) showed that the generic-specific relation could also be subdivided. Interestingly, these preliminary results indicated that hyponymy subtypes were constrained by the ontological nature of concepts, depending on whether they were entities or processes. The new proposal presented in this paper expands the scope of our preliminary research on hyponymy subtypes to include concepts belonging to a wider range of semantic categories, and examines the behavior of knowledge patterns used to extract hyponymic relations. In this research, corpus analysis was used to explore the correlation of concepts in many different categories with KPs as well as with hyponymy subtypes. Thanks to these constraints, it was possible to formulate a more comprehensive inventory of generic-specific relations in the environmental domain.

Keywords: hyponymy subtypes; knowledge patterns; corpus analysis; concept nature

1. Introduction

In recent years, the study of terminology and specialized language has been undergoing a 'cognitive shift' (Faber, 2009: 111), which places a greater focus on conceptual representation and knowledge organization. In this line, descriptive theories of terminology (Cabré, 1999; Temmerman, 2000; Faber, 2009) now reflect dynamic phenomena (such as variation or multidimensionality) and emphasize the importance of hierarchical and non-hierarchical relations.

A crucial factor in the organization of a terminology knowledge base (TKB) lies in the relations between its terms (Barrière, 2004a). These semantic relations can be discovered through corpus analysis and the use of knowledge-rich contexts (KRC). Such contexts are highly informative since they provide conceptual information and domain knowledge (Meyer, 2001), and usually codify semantic relations in the form of knowledge patterns (KPs) (Meyer, 2001; Condamines, 2002; Barrière, 2004b; Agbago & Barrière, 2005; León-Araúz, 2014).

In recent years, much research has targeted the development of semi-automatized procedures for extracting KRCs (Jacquemin & Bourigault, 2005; Bielinskiene et al., 2012; Schumann, 2012), especially for hyponymic term pairs. Although recent work has focused on other conceptual relations, such as meronymy, function, and causality (Marshman, 2002; Girju et al., 2003; León-Araúz et al., 2016), hyponymy is a complex relation that requires a more in-depth study. As the backbone of hierarchical organization, it entails both categorization and property inheritance (Barrière, 2004a). Moreover, it is characterized by a variety of nuances and dimensions that should be further exploited (Gil-Berrozpe & Faber, 2016).

To explore the viability of our proposal, a pilot study (Gil-Berrozpe et al., in press) was conducted to ascertain whether the generic-specific relation could be subdivided in EcoLexicon[1] (Faber et al., 2014, 2016), a multilingual and multimodal TKB on environmental science. For this purpose, the EcoLexicon English Corpus[2] was processed with Sketch Engine (Kilgarriff et al., 2004), where the Word Sketch (WS) module was used. WSs are automatic corpus-derived summaries of a word's grammatical and collocational behavior (Kilgarriff et al., 2004). In this pilot study, we reconstructed the taxonomies of ROCK (an entity) and EROSION (a process). The resulting hierarchies were based on the analysis of (i) the default *modifier* WS, from which hyponymy can be extracted by analyzing the composition of multiword terms; (ii) a customized WS based on hyponymic KPs, where hyponymy was explicitly conveyed in the texts. The results showed that hyponymy subtypes were based on the semantic category of the concept, and were constrained by the nature of the concept, namely, whether it was an entity or a process.

This paper presents the results of a new study on hyponymy subtypes that includes concepts belonging to a wider range of semantic categories (e.g. activities, chemical elements, landforms, etc.), and analyzes the behavior of the knowledge patterns used to extract hyponymic relations. Accordingly, corpus analysis was used to explore the correlation of concepts in a variety of different categories with KPs as well as with hyponymy subtypes. These constraints led to a more comprehensive inventory of generic-specific relations in the environmental domain, as well as to a more accurate way of extracting them.

The rest of this article is organized as follows. Section 2 briefly presents the EcoLexicon TKB and explains how hyponymy refinement can enhance its conceptual networks. Section 3 explains the materials used and the methods followed to analyze semantic categories in relation to hyponymic KPs and hyponymy subtypes. In Section 4, the results of our research are presented and discussed. Section 5 highlights the conclusions that can be derived from this study and outlines plans for future research.

1 http://ecolexicon.ugr.es/

2 Part of this corpus (23 million words) is now available in Sketch Engine's Open Corpora (https://the.sketchengine.co.uk/open/).

The bibliography cited is followed by three appendices in which semantic categories, hyponymic knowledge patterns, and hyponymy subtypes are defined and exemplified.

2. Hyponymy refinement in EcoLexicon

EcoLexicon is a TKB on environmental science that is based on the theoretical premises of Frame-Based Terminology (Faber, 2012, 2015). Its objective is to facilitate user knowledge acquisition through different types of multimodal and contextualized information, in order to respond to cognitive, communicative, and linguistic needs. This resource is available in English and Spanish, although five more languages (German, Modern Greek, Russian, French and Dutch) are currently being added. To date, EcoLexicon has a total of 3,601 concepts and 20,212 terms.

EcoLexicon has a visual interface with different modules for conceptual, linguistic, and graphical information (Figure 1). Once a concept has been selected, it is represented in the center of an interactive map. Also displayed are the multilingual terms for that concept, as well as different conceptual relations between all the concepts belonging to the same network.

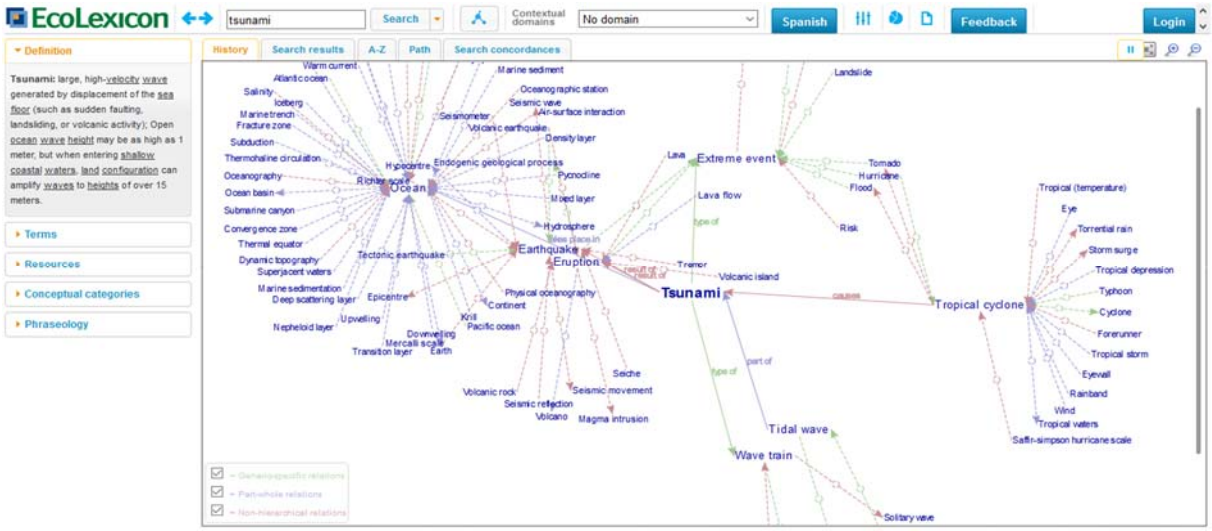


Figure 1: Visual interface of EcoLexicon (conceptual network of TSUNAMI).

The conceptual relations in EcoLexicon are classified as follows: (i) generic-specific relation (1 type); (ii) part-whole relations (6 types); (iii) non-hierarchical relations (10 types). Evidently, the generic-specific or hyponymic relation, which only has one subtype, would benefit from a more fine-grained representation since this would enhance its informativity and help to eliminate noise, information overload, and redundancy in the conceptual network (Gil-Berrozpe & Faber, 2016). Hyponymy is a semantic relation of inclusion whose converse is hyperonymy (Murphy, 2006: 446), and it can be refined by specifying subtypes (Murphy, 2003) or by establishing 'facets' and/or 'microsenses' (Cruse, 2002: 4-5).

Our pilot study (Gil-Berrozpe et al., in press) based hyponymy refinement on the following criteria: (i) the correction of property inheritance according to concept definitions; (ii) the creation of umbrella concepts; (iii) the decomposition of hyponymy into subtypes. As previously mentioned, our results indicated that hyponymy subtypes were based on whether the concept was an entity (ROCK) or a process (EROSION). For example, natural entities, such as ROCK, were found to have different sets of hyponyms based on formation (e.g. SEDIMENTARY ROCK, IGNEOUS ROCK), composition (SILTSTONE, SANDSTONE), and location (PLUTONIC ROCK, VOLCANIC ROCK).

3. Materials and methods

Our study analyzed hyponymic KPs as well as hyponymy subtypes. In both cases, the main information source was the EcoLexicon English corpus (67,903,384 words), which was uploaded to Sketch Engine. Apart from the default options, the system also permitted the creation of customized word sketches by storing CQL queries in new sketch grammars.

The corpus was thus compiled by implementing hyponymic sketch grammars developed by León-Araúz et al. (2016). These grammars are based on the KPs that generally reflect hyponymy in real texts. Simple examples of such KPs are *HYPERNYM such as HYPONYM*, *HYPONYM is a kind of HYPERNYM*, *HYPONYM and other HYPERNYM*, etc. These patterns were formalized as regular expressions combined with POS-tags, which resulted in 18 hyponymic sketch grammars. Table 1 shows a summarized version of the KPs.

1. HYPONYM ,|(|:|is|belongs (to) (a|the|...) type|category|... of HYPERNYM // **2.** types|kinds|... of HYPERNYM include|are HYPONYM // **3.** types|kinds|... of HYPERNYM range from (...) (to) HYPONYM // **4.** HYPERNYM (type|category|...) (,|() ranging (...) (to) HYPONYM // **5.** HYPERNYM types|categories|... include HYPONYM // **6.** HYPERNYM such as HYPONYM // **7.** HYPERNYM including HYPONYM // **8.** HYPERNYM ,|(especially|primarily|... HYPONYM // **9.** HYPONYM and|or other (types|kinds|...) of HYPERNYM // **10.** HYPONYM is defined|classified|... as (a|the|...) (type|kind|...) (of) HYPERNYM // **11.** classify|categorize|... (this type|kind|... of) HYPONYM as HYPERNYM // **12.** HYPERNYM is classified|categorized in|into (a|the|...) (type|kind|...) (of) HYPONYM // **13.** HYPERNYM (,|() (is) divided in|into (...) types|kinds|... :|of HYPONYM // **14.** type|kind|... of HYPERNYM (is|,|() known|referred|... (to) (as) HYPONYM // **15.** HYPONYM is a HYPERNYM that|which|... // **16.** define HYPONYM as (a|the|...) (type|category|...) (of) HYPERNYM // **17.** HYPONYM refers to (a|the|...) (type|category|...) (of) HYPERNYM // **18.** (a|the|one|two...) (type|category|...) (of) HYPERNYM: HYPONYM

Table 1: Hyponymic knowledge patterns (León-Araúz et al., 2016)

3.1 Hyponymic KPs and semantic categories

When the customized hyponymic sketch grammars were applied to the English EcoLexicon corpus, this created a filtered subcorpus, which was only composed of hyponymic concordances. This was accomplished by applying the CQL query *[ws(".*-n","\"%w\" is the generic of...",".*-n")]*. The resulting subcorpus contained a total of 938,386 potential hyponymic concordances (Figure 2).

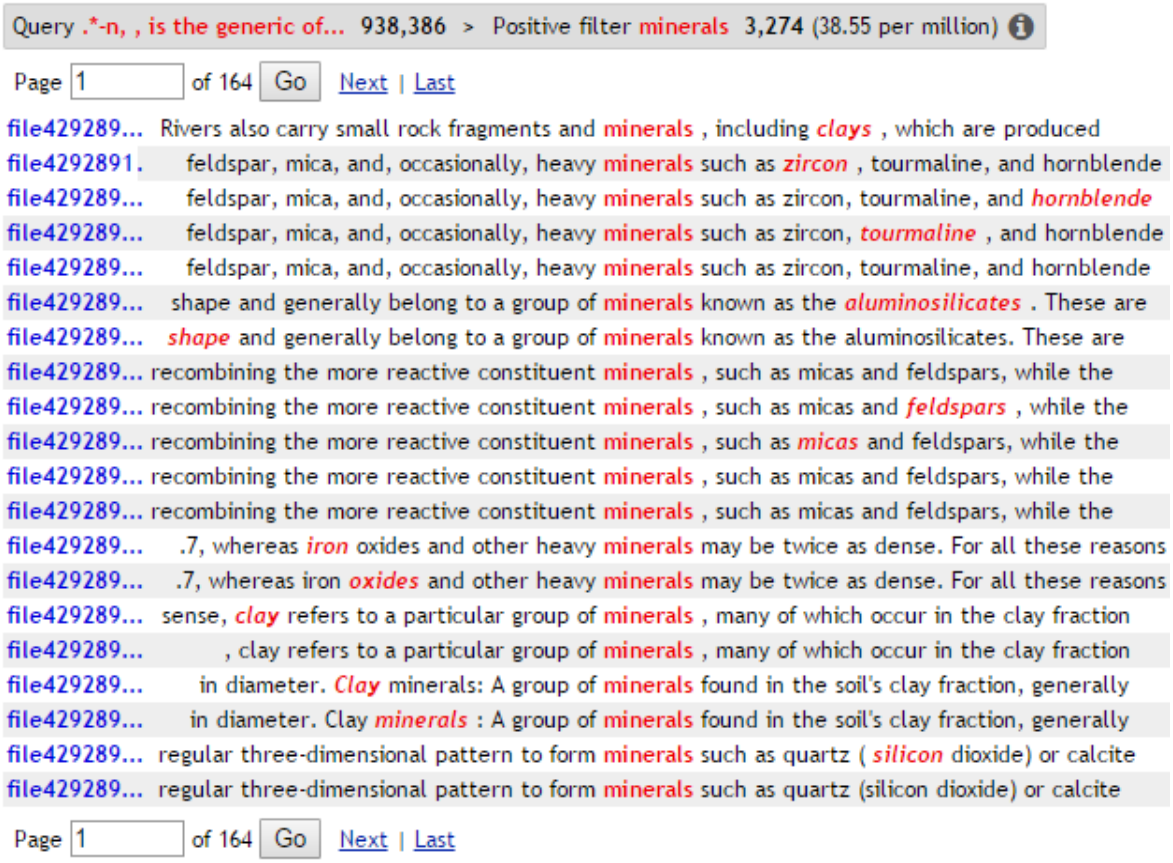


Figure 2: Concordances retrieved from the hyponymic subcorpus

However, after filtering the hyponymic concordances in the EcoLexicon corpus with the customized word sketch, a manual process of data extraction was required. Since the customized word sketch was composed of 18 grammars describing a wide range of permutations and paraphrases of the hyponymic KPs, it was necessary to manually collect and analyze a representative sample of this information. Furthermore, the hyponymic subcorpus contained various identical sentences (since multiple hypernym-hyponym pairs in the same concordance were shown several times). There were also false positives that had to be eliminated from the results.

A randomized portion of the hyponymic subcorpus was examined, from which a set of 3,133 positive hyponymic concordances were selected to be the basis of the KP analysis. The extracted information was subsequently classified for analysis (Figure 3).

No.	Hypernym(s) [HYPER]	Hyponym(s) [HYPO]	Activated semantic category	Hyponymic pattern	Hyponymic pattern type
2635.	Acacia	Acacia tortilis, Capparis decidua	lifeform	types of HYPER, mainly HYPO	selection
1.	academic field	geography, architecture, psychology	domain	HYPO and other HYPER such as HYPO	itemization + exemplification
1585.	acid	H2SO4	element	HYPER such as HYPO	exemplification
1584.	acidic species	H2SO4, HCl, HF	element	HYPER such as HYPO	exemplification
2714.	acidic surface oxide	strong carboxylic, weak carboxylic	element	# types of HYPER, namely HYPO	enumeration + selection
692.	acidification	episodic acidification	process	HYPER *be* classified into # types: HYPO	enumeration + classification
2495.	acidification	episodic acidification	process	HYPER, especially HYPO	selection
1722.	acrylamide	N-alkylacrylamide	element	HYPER such as HYPO	exemplification
1064.	acrylic acid	alkyl acrylate, methacrylate	element	HYPER such as HYPO	exemplification
2904.	active region	swash zone	location	HYPER, such as HYPO	exemplification
1378.	active substance	clay, charcoal, diatomaceous earth	substance	HYPER such as HYPO	exemplification
405.	active volcano	Mount Spur	landform	HYPER, such as HYPO	exemplification
414.	active volcano	Mount Erebus	landform	HYPER, such as HYPO	exemplification

Figure 3: Extract of the hyponymic KP table

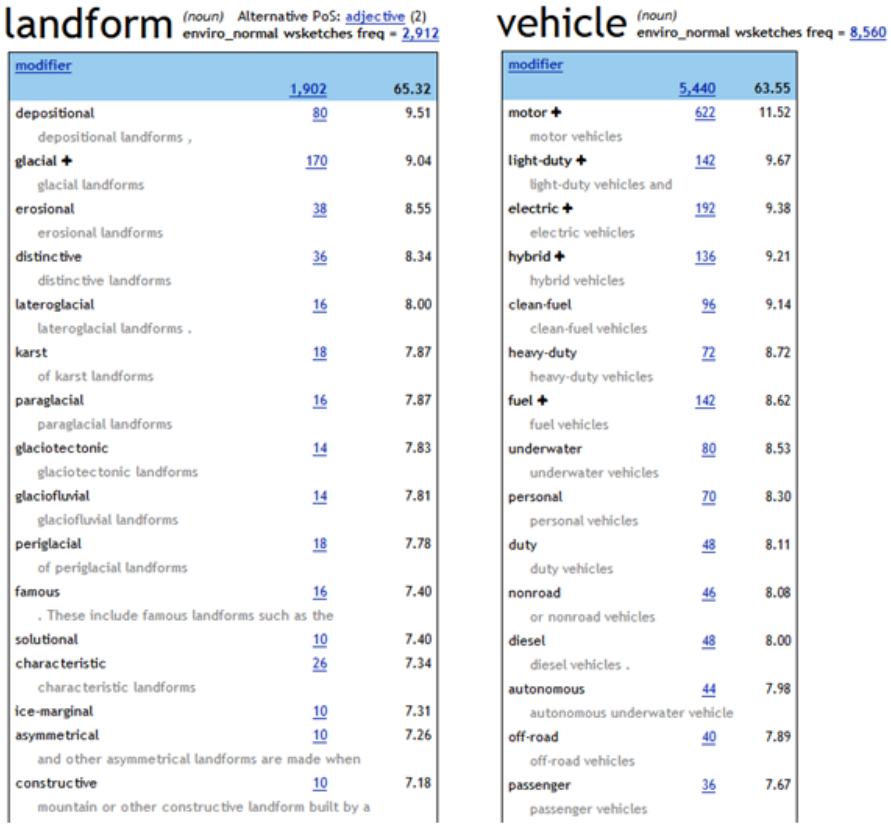
As shown in Figure 3, the hyponymic KP table contained the following categories: (i) ID number of the concordance; (ii) hypernym in the concordance; (iii) hyponym(s) in the concordance; (iv) semantic category of the hypernyms/hyponyms; (v) hyponymic KP expressing the generic-specific relation; (vi) type of hyponymic KP. A list of semantic categories and a list of pattern types were also formulated in order to classify and filter the information. As previously mentioned, our research objective was to examine the correlation between hyponymic KPs and the semantic category of concepts. It was thus necessary to create an inventory of semantic categories (Section 4.1).

3.2 Hyponymy subtypes and semantic categories

In the KP study (Section 3.1), the compilation of hypernym-hyponym pairs was performed by filtering KPs, rather than by focusing on semantic categories. However, in the case of hyponymy subtypes, emphasis was placed on selecting different concept types so as to generate a list of hyponymy subtypes that was as comprehensive as possible. Since our previous results seemed to indicate that hyponymy subtypes depended on the nature of the concept (Gil-Berrozpe & Faber, 2016), we wished to confirm this hypothesis by using more fine-grained semantic categories (e.g. *activity*, *landform, chemical element,* etc.).

It was thus necessary to perform a second compilation of hypernym-hyponym pairs, though this time with a greater focus on semantic categories. For this reason, we extracted 109 hypernyms of concepts belonging to a wide range of semantic categories: 32 natural entities, 32 artificial entities, 21 natural processes, 17 artificial processes, and seven hybrid processes (which could be considered natural or artificial depending

on their respective agents or methods). These 109 hypernyms were then analyzed using the default *modifier* word sketch in Sketch Engine. This gave us a set of hyponyms characterized by their modifier (Figure 4).



landform *(noun)* Alternative PoS: adjective (2)
enviro_normal wsketches freq = 2,912

modifier	1,902	65.32
depositional	80	9.51
depositional landforms ,		
glacial ✚	170	9.04
glacial landforms		
erosional	38	8.55
erosional landforms		
distinctive	36	8.34
distinctive landforms		
lateroglacial	16	8.00
lateroglacial landforms .		
karst	18	7.87
of karst landforms		
paraglacial	16	7.87
paraglacial landforms		
glaciotectonic	14	7.83
glaciotectonic landforms		
glaciofluvial	14	7.81
glaciofluvial landforms		
periglacial	18	7.78
of periglacial landforms		
famous	16	7.40
. These include famous landforms such as the		
solutional	10	7.40
characteristic	26	7.34
characteristic landforms		
ice-marginal	10	7.31
asymmetrical	10	7.26
and other asymmetrical landforms are made when		
constructive	10	7.18
mountain or other constructive landform built by a		

vehicle *(noun)*
enviro_normal wsketches freq = 8,560

modifier	5,440	63.55
motor ✚	622	11.52
motor vehicles		
light-duty ✚	142	9.67
light-duty vehicles and		
electric ✚	192	9.38
electric vehicles		
hybrid ✚	136	9.21
hybrid vehicles		
clean-fuel	96	9.14
clean-fuel vehicles		
heavy-duty	72	8.72
heavy-duty vehicles		
fuel ✚	142	8.62
fuel vehicles		
underwater	80	8.53
underwater vehicles		
personal	70	8.30
personal vehicles		
duty	48	8.11
duty vehicles		
nonroad	46	8.08
or nonroad vehicles		
diesel	48	8.00
diesel vehicles .		
autonomous	44	7.98
autonomous underwater vehicle		
off-road	40	7.89
off-road vehicles		
passenger	36	7.67
passenger vehicles		

Figure 4: *Modifier* word sketches of LANDFORM and VEHICLE

Furthermore, it was necessary to manually select the relevant information in order to avoid matches that were not necessarily terms (e.g. FAMOUS LANDFORM, seen in the *modifier* word sketch of LANDFORM in Figure 4). A total of 1,912 hypernym-hyponym pairs were extracted and inserted in a classification table (Figure 5).

ID	Hypernym [HYPER]	General semantic category	Hyponym [HYPO]	Specific semantic category	Hyponymy subtype
NE10	acid	natural entity	abscisic acid	element	effect-based hyponymy
NE02	element	natural entity	abundant element	element	amount-based hyponymy
HP02	contamination	hybrid process	accidental contamination	phenomenon	method-based hypoynymy
NE10	acid	natural entity	acetic acid	substance	composition-based hyponymy
NP11	precipitation	natural process	acid precipitation	phenomenon	patient-based hyponymy
NE16	soil	natural entity	acid soil	substance	composition-based hyponymy
HP04	reaction	hybrid process	acid-base reaction	process	agent-based hyponymy
NE03	compound	natural entity	acidic compound	element	composition-based hyponymy
NP19	absorption	natural process	active absorption	process	method-based hypoynymy
NE23	dune	natural entity	active dune	mass of matter	activity-based hyponymy
AP09	management	artificial process	adaptive management	activity	method-based hypoynymy
NP20	radiation	natural process	adaptive radiation	process	method-based hypoynymy
HP04	reaction	hybrid process	addition reaction	process	method-based hypoynymy
NP08	melting	natural process	adiabatic melting	change of state	method-based hypoynymy
NE21	continent	natural entity	adjacent continent	mass of matter	location-based hyponymy
NE22	land	natural entity	adjacent land	mass of matter	location-based hyponymy

Figure 5: Extract of the hyponymy subtype table

The hyponymy subtype table in Figure 5 has the following categories: (i) ID number of the hypernym; (ii) hypernym; (iii) general semantic category of the hypernym; (iv) hyponym; (v) semantic category of the hyponym; (vi) hyponymy subtype derived from the hypernym-hyponym pair. As in the corpus study, our objective was to explore the correlation between hyponymy subtype and concept type, expressed in the form of semantic categories. For this reason, it was necessary to create an inventory of semantic classes (Section 4.2).

4. Results and discussion

As part of this research, two sets of hypernym-hyponym pairs were analyzed: (i) 3,133 pairs extracted from the corpus with customized hyponymic grammars; (ii) 1,912 pairs extracted from word sketch data using the default *modifier* word sketch. In both cases, concepts were classified in semantic categories. Although most of the semantic categories coincided in both data sets, there were certain categories exclusive to each set.

4.1 Hyponymic KP analysis: general results

Figure 6 shows the distribution of the 3,133 concepts extracted for hyponymic KP analysis. As can be observed, 21 semantic categories were found. (See Appendix A for the description and typical examples of each category.)

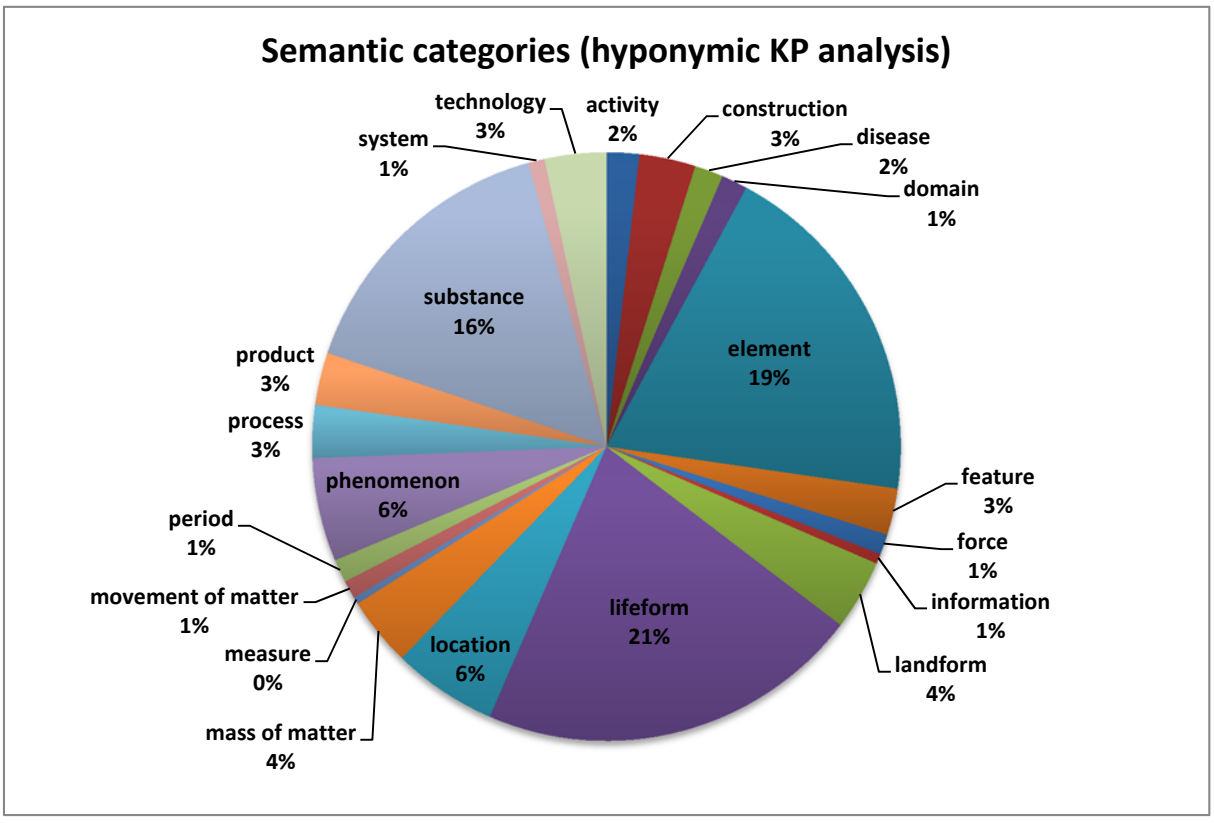


Figure 6: Semantic categories of the concepts of the hyponymic KP analysis

The results of our study showed that the semantic categories of the main concept types were lifeform, chemical element and substance, whose percentages were significantly higher than those of the other categories.

In regard to hyponymic KPs, 125 patterns were identified. KPs that expressed hyponymy in a similar way were placed in the same category. Figure 7 shows the distribution of these 125 patterns in 10 categories. (See Appendix B for a description of each knowledge pattern with examples.)

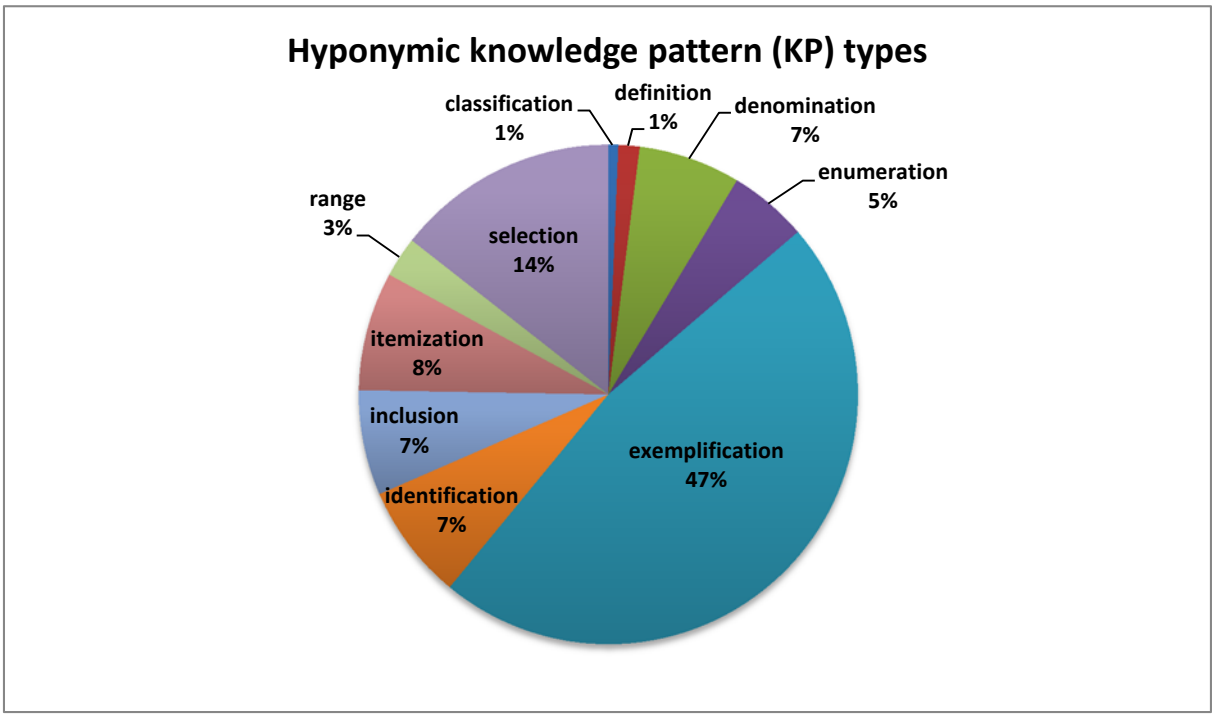


Figure 7: Hyponymic knowledge patterns

As reflected in our results, the most frequent hyponymic pattern types were exemplification KPs, selection KPs, and itemization KPs, though patterns expressing any sort of exemplification were clearly the most predominant.

4.1.1 Correlations between hyponymic KPs and semantic categories

Exemplification KPs (Figure 8), by far the most frequent pattern, comprised almost half of the sample analyzed. Because of the quantity of information in these patterns, they were typical of the most common semantic categories, namely: chemical element, lifeform, and substance. The second most significant group of categories included location, phenomenon, landform, and construction. The other semantic categories were found in significantly fewer concordances.

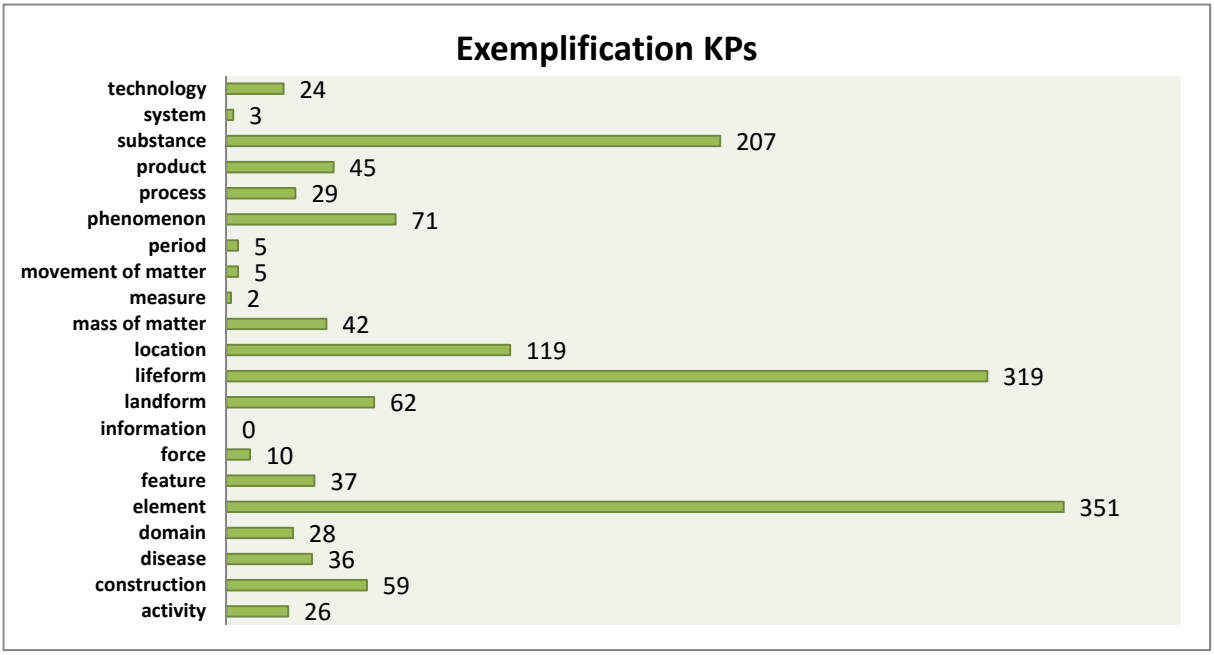


Figure 8: Exemplification KPs per semantic category

Since exemplification KPs were the most common, the only conclusion that can be derived is that the occurrences of exemplification KPs per semantic category are proportional to the ratios of semantic categories shown in Figure 6.

As for selection KPs (Figure 9), itemization KPs (Figure 10), and inclusion KPs (Figure 11), lifeform, chemical element, and substance were also the most prominent semantic categories.

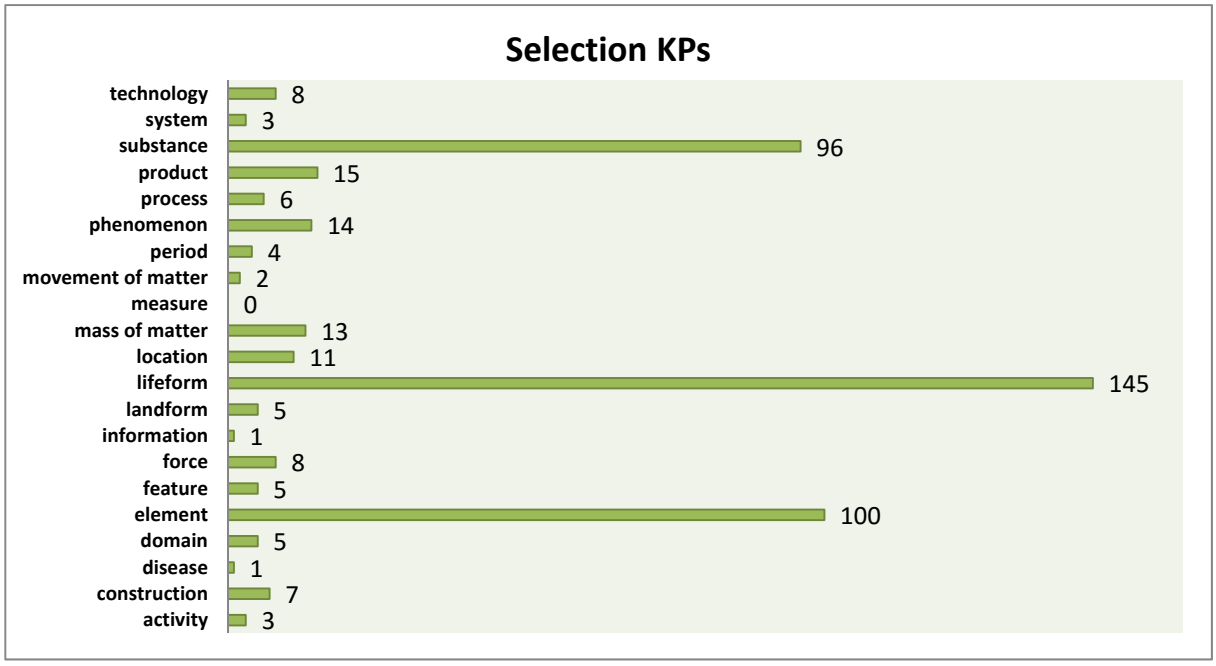


Figure 9: Selection KPs per semantic category

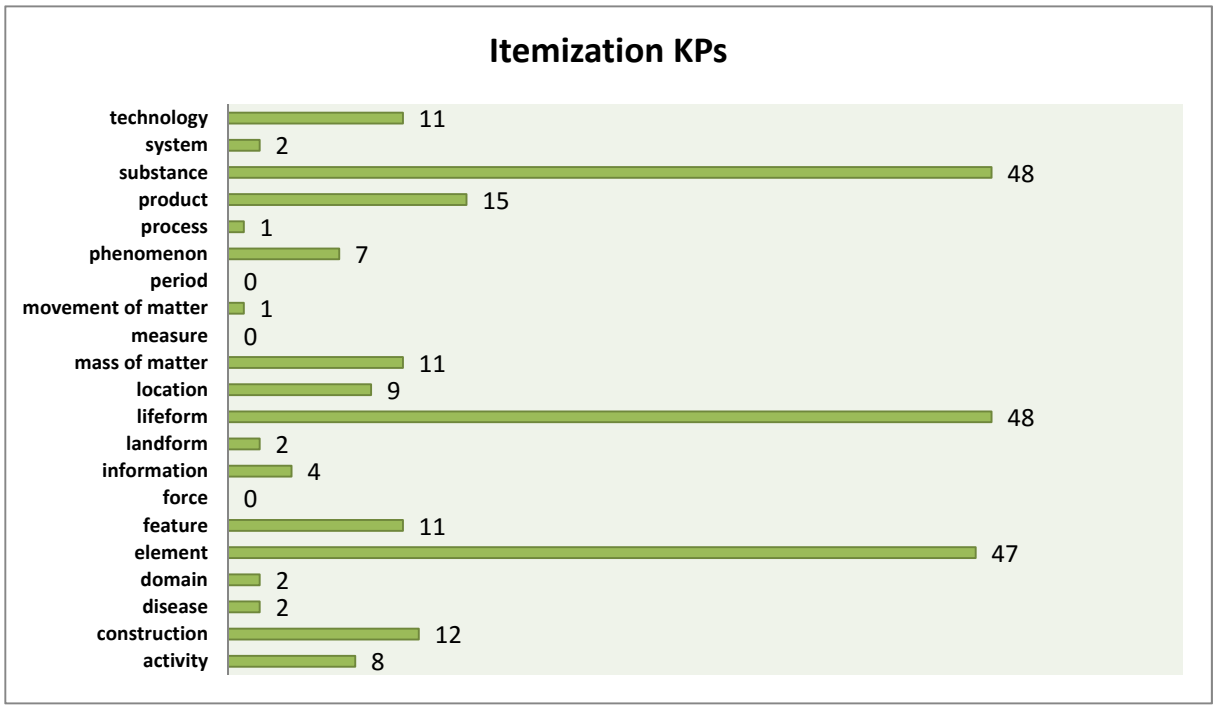


Figure 10: Itemization KPs per semantic category

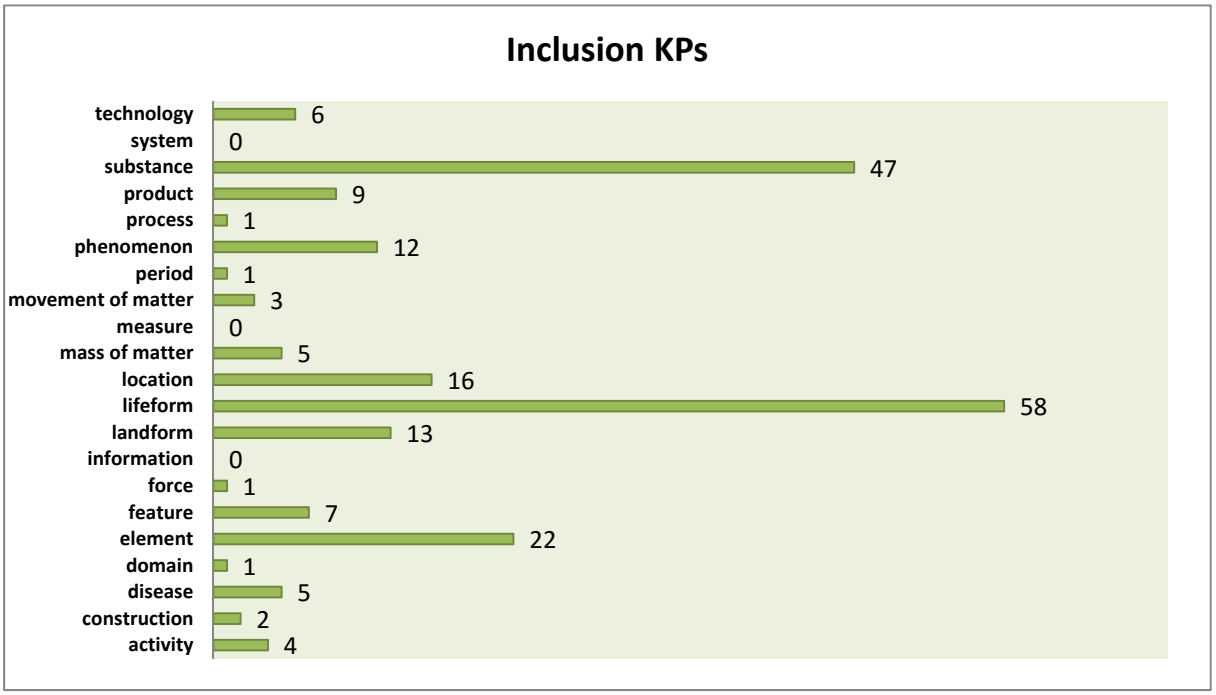


Figure 11: Inclusion KPs per semantic category

The predominance of these patterns could be a matter of statistics, since these concepts are the most frequent in the English EcoLexicon corpus. However, another possibility is that this phenomenon is related in some way to discourse type and function since most of the texts in the corpus are research articles, textbooks, and encyclopedias, whose functions are to facilitate the acquisition of specialized environmental knowledge.

With regard to identification KPs (Figure 12) and denomination KPs (Figure 13), the category of phenomenon held the second position, only surpassed by chemical element, and followed by lifeform and substance. In addition, the categories of process and technology also had a significant presence. As in the previous cases, this showed that identification KPs and denomination KPs are also activated by semantic categories in relation to the ratios shown in Figure 6. However, the significantly greater frequency of phenomenon, process and technology also indicates that these hyponymic KPs could be related to complex concepts that need an identifying or denominating structure (HYPO is a HYPER, a type of HYPER is a HYPO, types of HYPER are called HYPO) in order to better explain them.

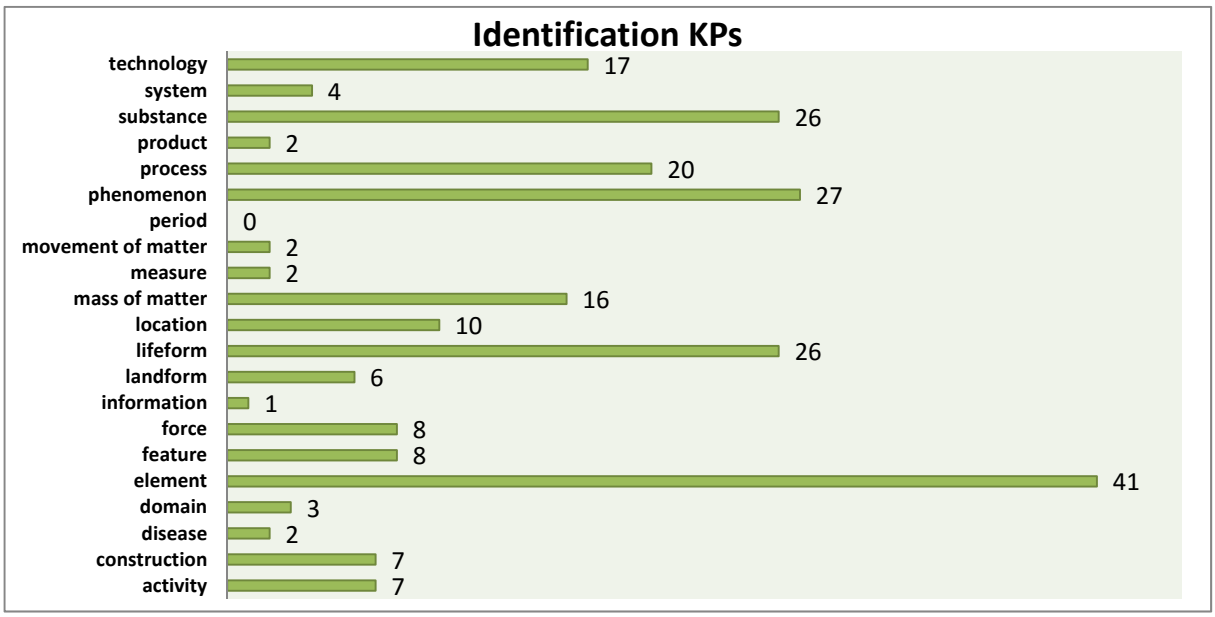


Figure 12: Identification KPs per semantic category

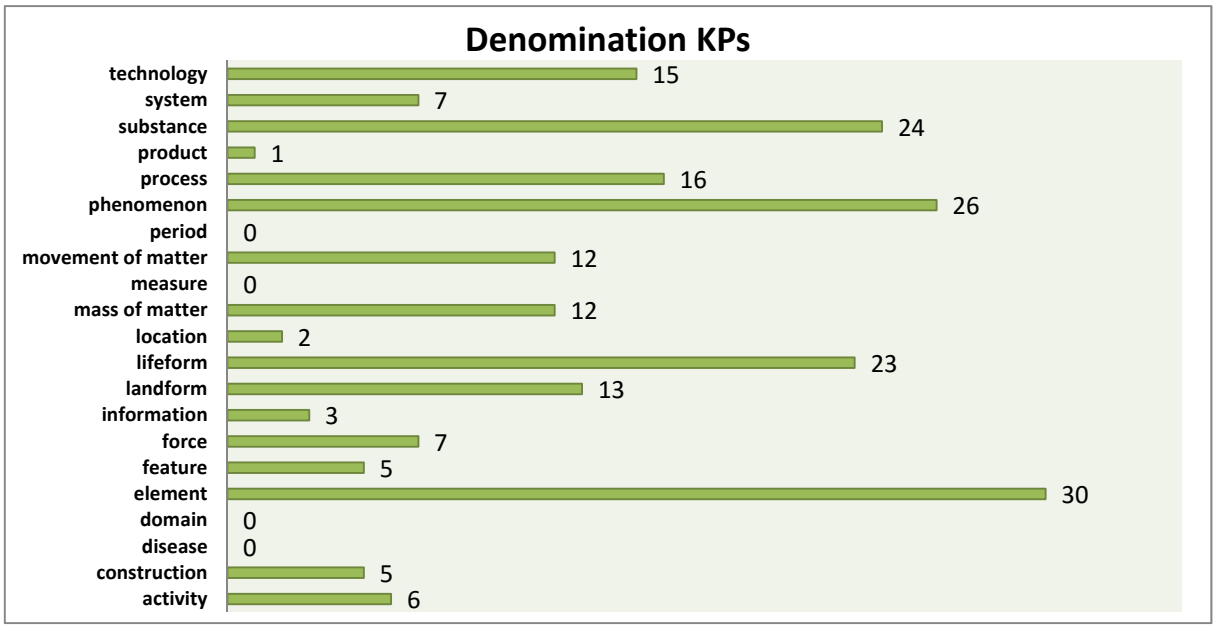


Figure 13: Denomination KPs per semantic category

This could also be true of definition KPs (Figure 14), where the categories of technology and phenomenon share second position, after substance. Once again, the KP expressions in this category specifically define a concept (HYPO: a HYPER, HYPO: a type of HYPER) in terms of its superordinate.

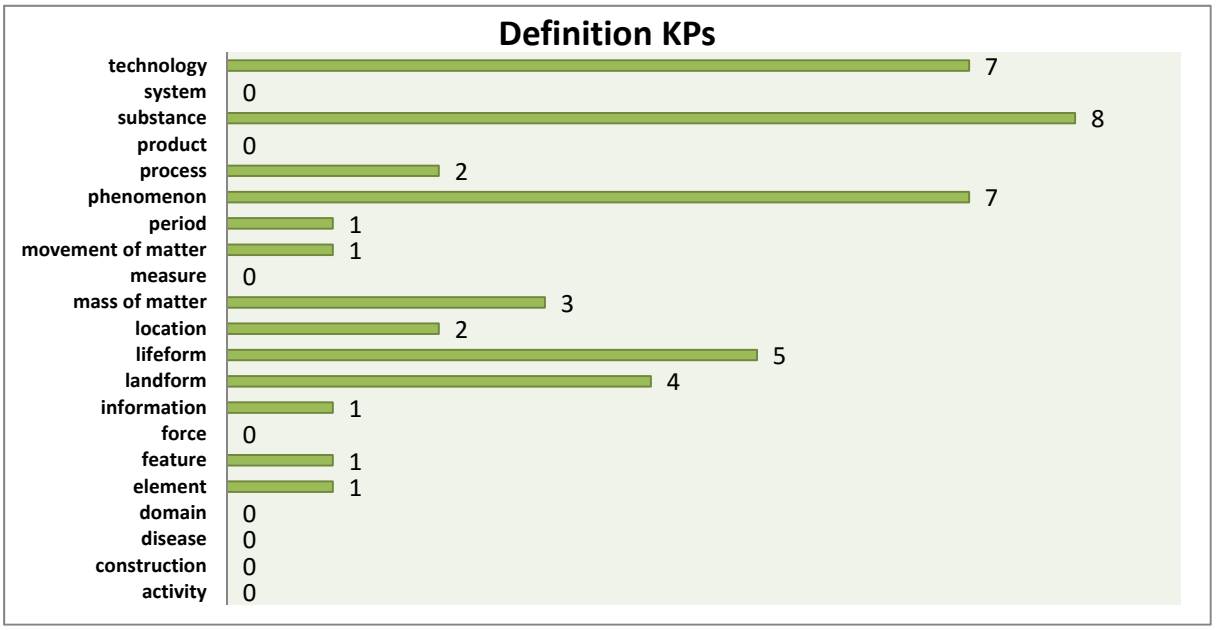


Figure 14: Definition KPs per semantic category

As for range KPs (Figure 15), a different semantic category held first position. The nature of this hyponymic KP makes it ideal for expressing time periods, scales, and degrees (HYPER ranging from HYPO to HYPO). Not surprisingly, the semantic category, measure, which had little or no relevance in the other patterns, frequently occurred in range KPs.

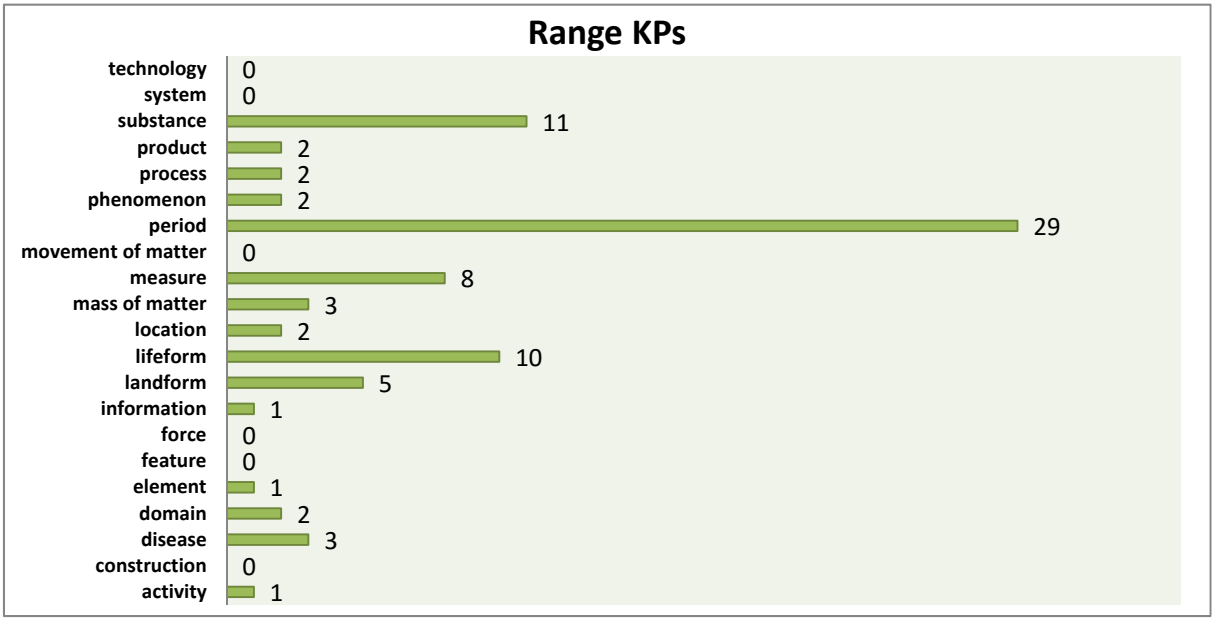


Figure 15: Range KPs per semantic category

Finally, in the case of enumeration KPs (Figure 16) and classification KPs (Figure 17), it was not possible to extract any specific correlation pattern. Our results showed that enumeration KPs, in the same way as exemplification KPs, were applicable to any concept type. Furthermore, the data for classification KPs was insufficient to draw any conclusions.

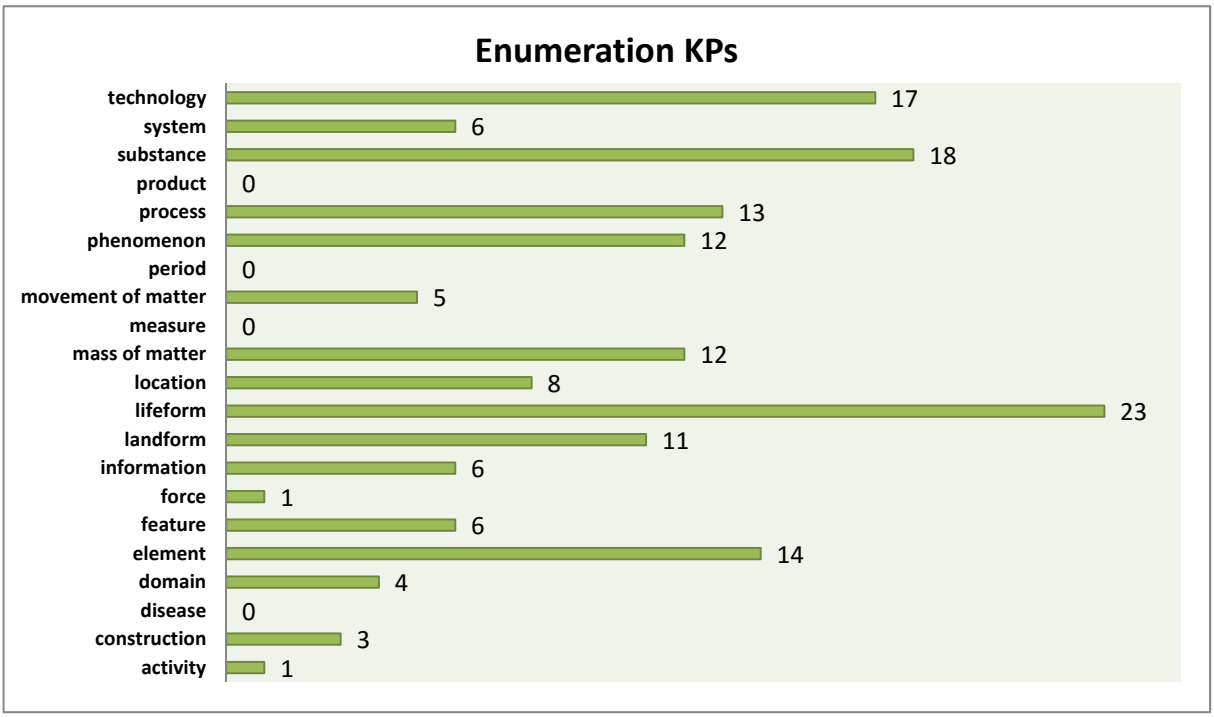


Figure 16: Enumeration KPs per semantic category

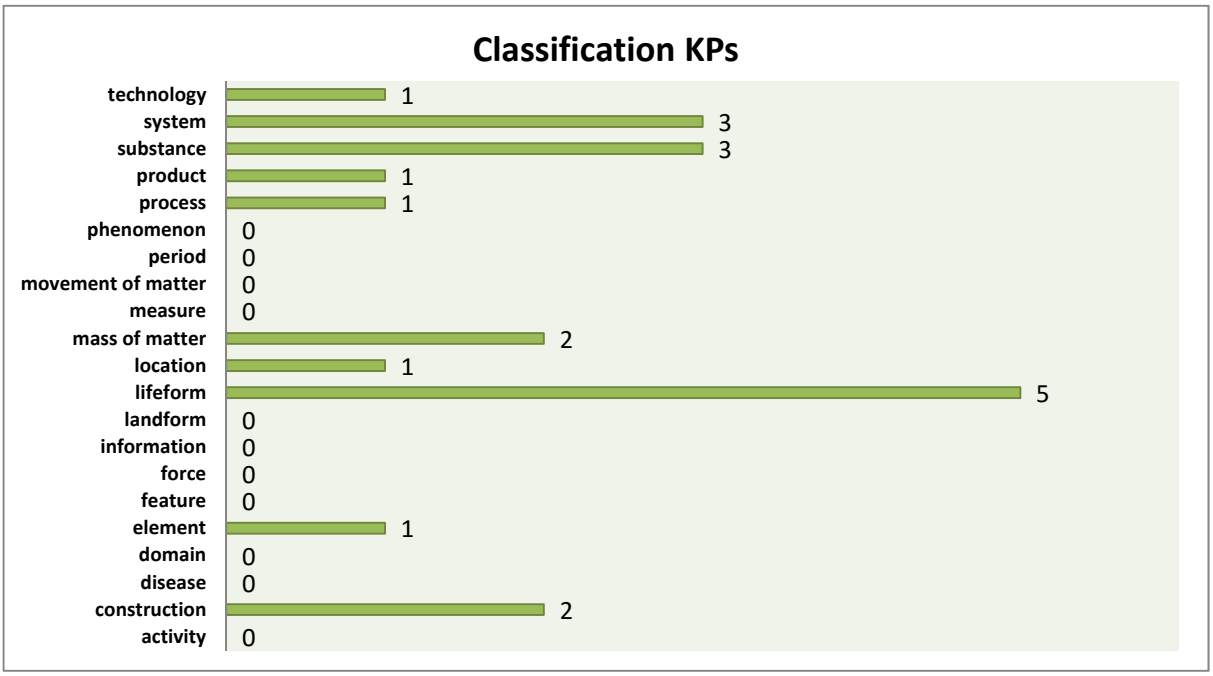


Figure 17: Classification KPs per semantic category

4.2 Hyponymy subtypes analysis: general results

Figure 18 shows the distribution of the 1,912 hyponyms in 13 semantic categories.

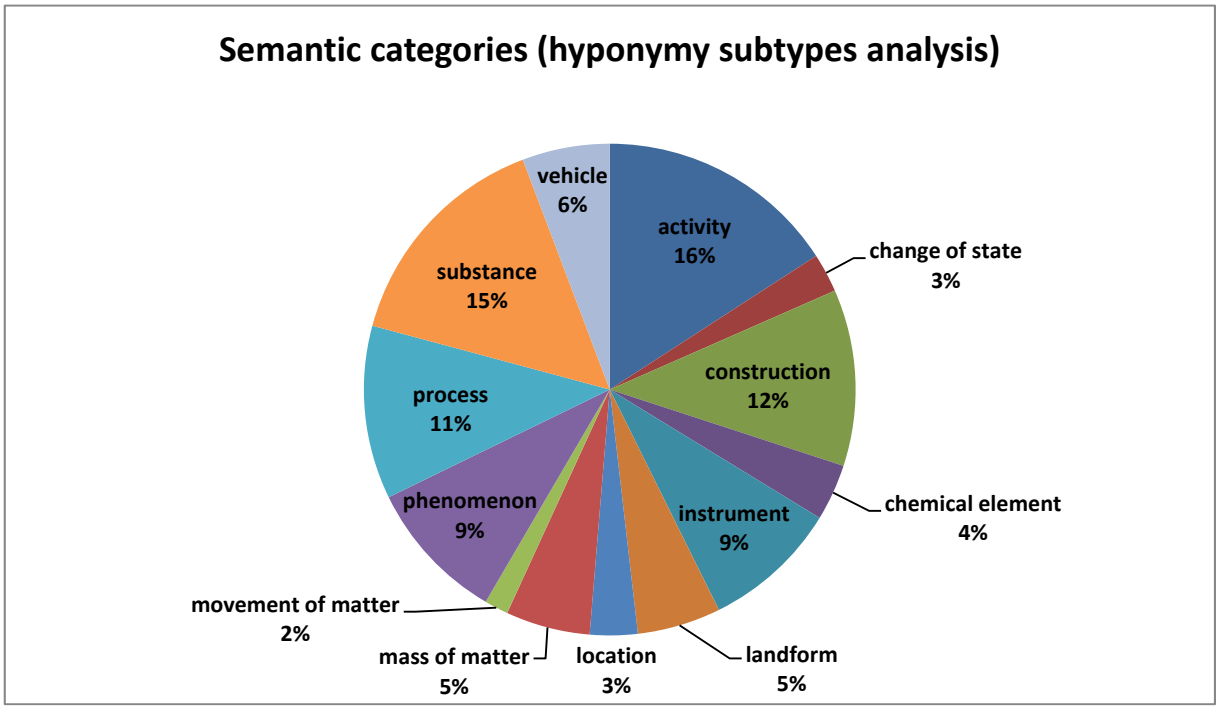


Figure 18: Semantic categories of the concepts of the hyponymy subtypes analysis

Although most of the semantic categories identified during this analysis coincide with those of the hyponymic KP analysis, the categories of *disease, domain, feature, force, information, lifeform, measure, period, product, system* and *technology* do not appear. This was due to the manual selection process. On the other hand, because of the higher frequency of other concept types, it was possible to identify three more semantic categories that are exclusive to the hyponymy subtype analysis: *instrument*, *vehicle*, and *change of state* (Appendix A).

The decomposition of the generic-specific relation was based on common features in the cases analyzed. This led to the identification of 32 different subtypes in the 1,912 hypernym-hyponym pairs (Figure 19). Appendix C describes and exemplifies the full inventory of hyponymy subtypes. In this inventory, a distinction can be made between relational hyponymy subtypes (those specifying a relation between the components of hyponym-hypernym pairs) and attributional hyponymy subtypes (those specifying an intrinsic feature of the hyponym).

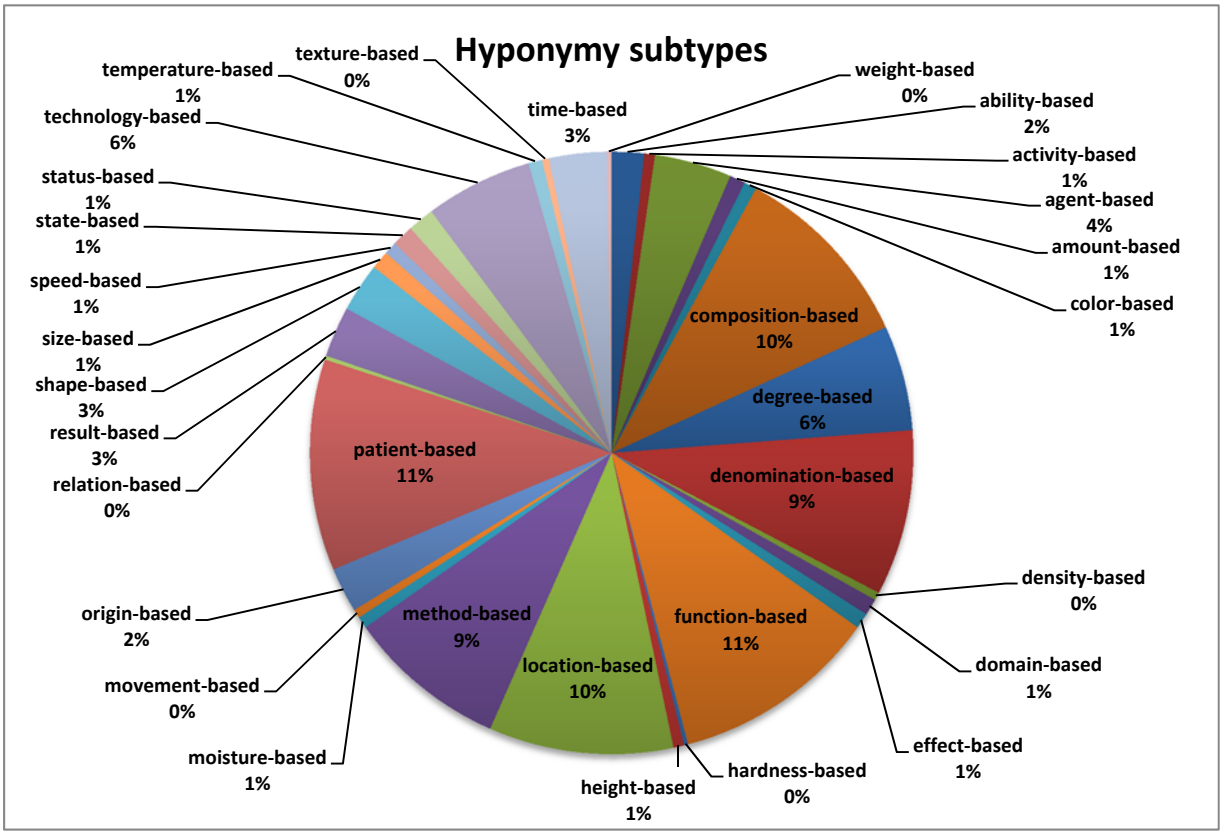


Figure 19: Hyponymy subtypes

As can be observed in Figure 19, the most frequently activated hyponymy subtypes were relational, particularly *patient-based*, *function-based*, *composition-based* and *location-based* hyponymy. On the contrary, attributional hyponymy subtypes (such as *degree-based*, *shape-based*, *ability-based* or *size-based*) were found to be less representative. This seems to indicate that when environmental knowledge is categorized into subtypes, there is a greater emphasis on how concepts interact with each other, rather than on the intrinsic characteristics of individual concepts.

4.2.1 Correlations between hyponymy subtypes and semantic categories

For the sake of conciseness, this section focuses on the 12 most recurrent hyponymy subtypes, derived from 1,582 hypernym-hyponym pairs (83% of the sample). These are *patient-based*, *function-based*, *composition-based*, *location-based*, *denomination-based*, *method-based*, *technology-based*, *degree-based*, *agent-based*, *time-based*, *result-based*, and *shape-based* hyponymy.

In both *patient-based* hyponymy (Figure 20) and *method-based* hyponymy (Figure 21), there was a predominance of the categories of activity, process, phenomenon, and change of state. There were no entity-related semantic categories because these two subtypes of hyponymy are exclusive to process-related semantic categories.

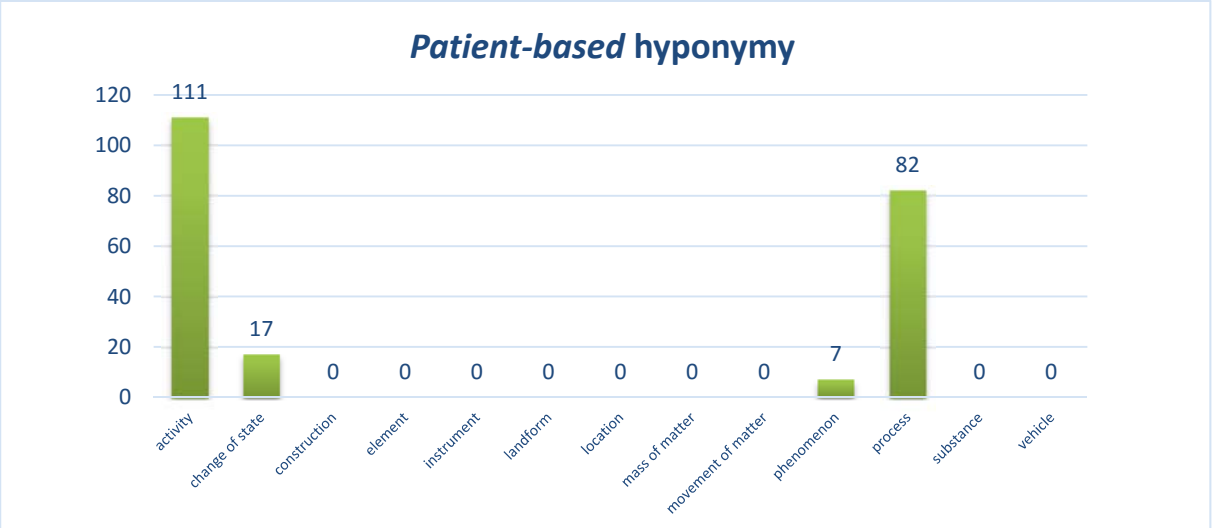


Figure 20: Patient-based hyponymy subtypes per semantic category

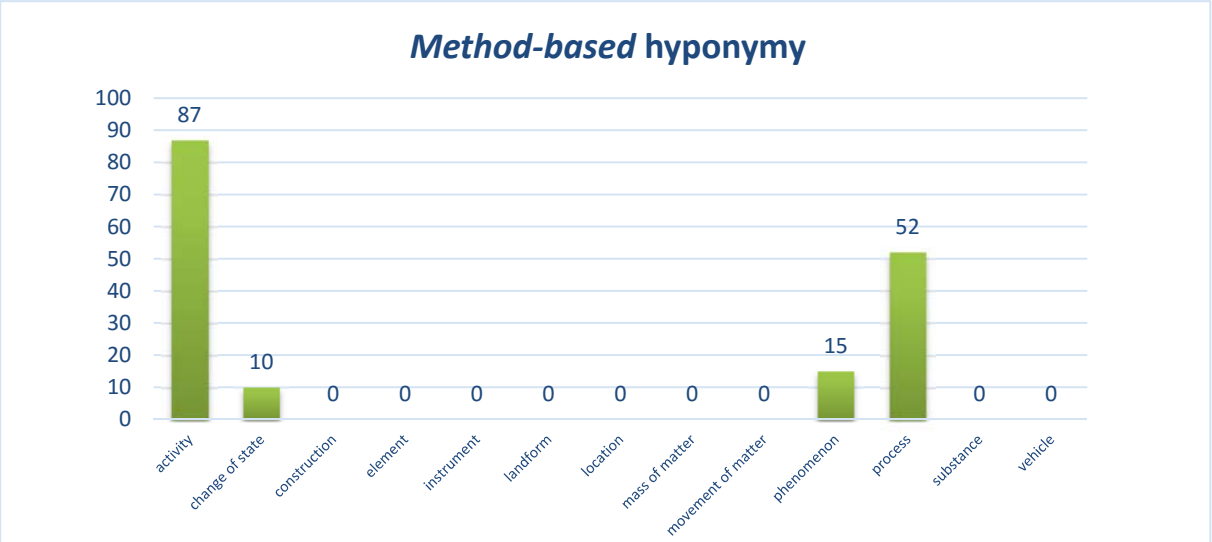


Figure 21: Method-based hyponymy subtypes per semantic category

As can be observed, the most frequent semantic categories were found to be activity and process, which are mostly composed of artificial or deliberate actions and processes. This sharply contrasted with the categories of phenomenon and change of state, which were mostly composed of natural processes. This could indicate that patient and method are what distinguish artificial processes from natural processes, since a natural change is not purposeful or deliberate.

As for *agent-based* hyponymy (Figure 22) and *result-based* hyponymy (Figure 23), once again most of the examples refer to process-related semantic categories, namely activity, process, and phenomenon.

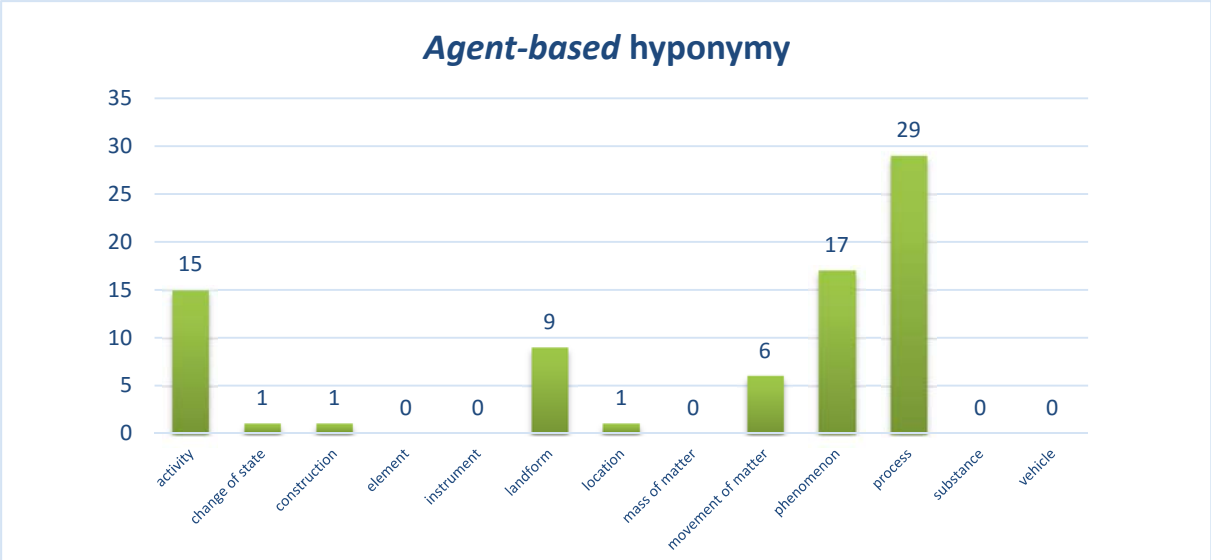


Figure 22: Agent-based hyponymy subtypes per semantic category

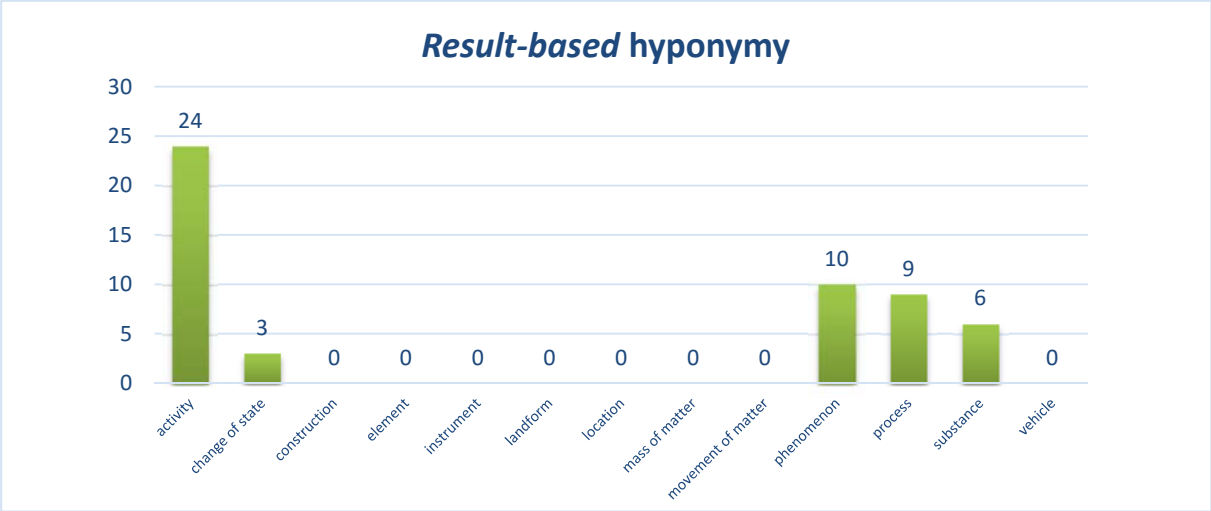


Figure 23: Result-based hyponymy subtypes per semantic category

Interestingly, these hyponymy subtypes also include two entity-related categories: (i) landform in the case of *agent-based* hyponymy, since there are some landforms characterized by the agent that has created them (e.g. GLACIAL LANDFORM, FLUVIAL LANDFORM, VOLCANIC ISLAND); (ii) substance in the case of *result-based* hyponymy, since substances can sometimes be characterized as the result of a process (e.g. DEGRADATION PRODUCT, OXIDATION PRODUCT, FISSION PRODUCT).

Similarly, *degree-based* hyponymy (Figure 24) is also mostly exclusive to process-related semantic categories, such as phenomenon, activity, process, and change of state. Furthermore, and in contrast to the previous results, the category of phenomenon is mostly characterized by degree (e.g. CATACLYSMIC ERUPTION, LOW-MAGNITUDE EARTHQUAKE, KILLER TORNADO, etc.).

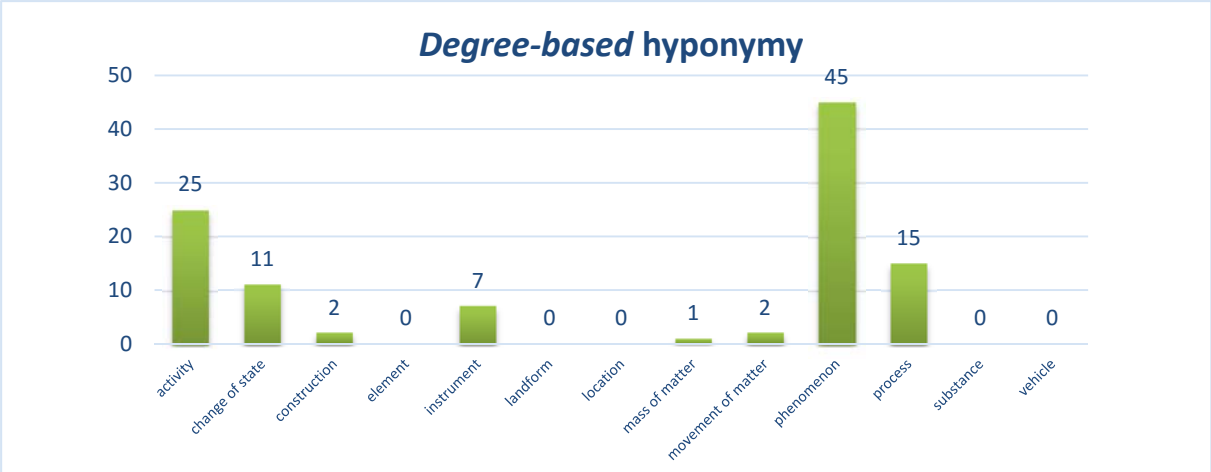


Figure 24: Degree-based hyponymy subtypes per semantic category

Composition-based hyponymy (Figure 25) shows that the most recurrent semantic categories are those involving natural entities, namely substance and chemical element. These are followed by the category of construction, which is composed of artificial entities that can be characterized by their components or their material (e.g. WOODEN BUILDING, RUBBLE MOUND BREAKWATER, CONCRETE DAM, etc.).

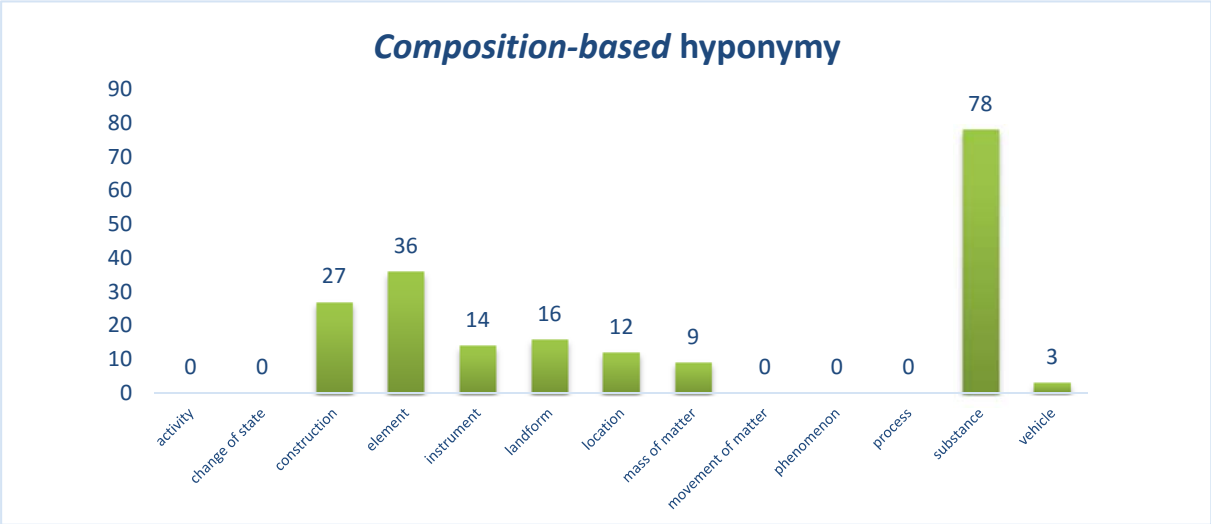


Figure 25: Composition-based hyponymy subtypes per semantic category

Location-based hyponymy (Figure 26) typically occurs with entity-related categories such as substance, construction, mass of matter, and landform. However, the category of phenomenon is also significant because natural processes are also characterized by the location where they occur (e.g. SUBMARINE EARTHQUAKE, MOUNTAIN CYCLOGENESIS, FOREST FIRE, etc.).

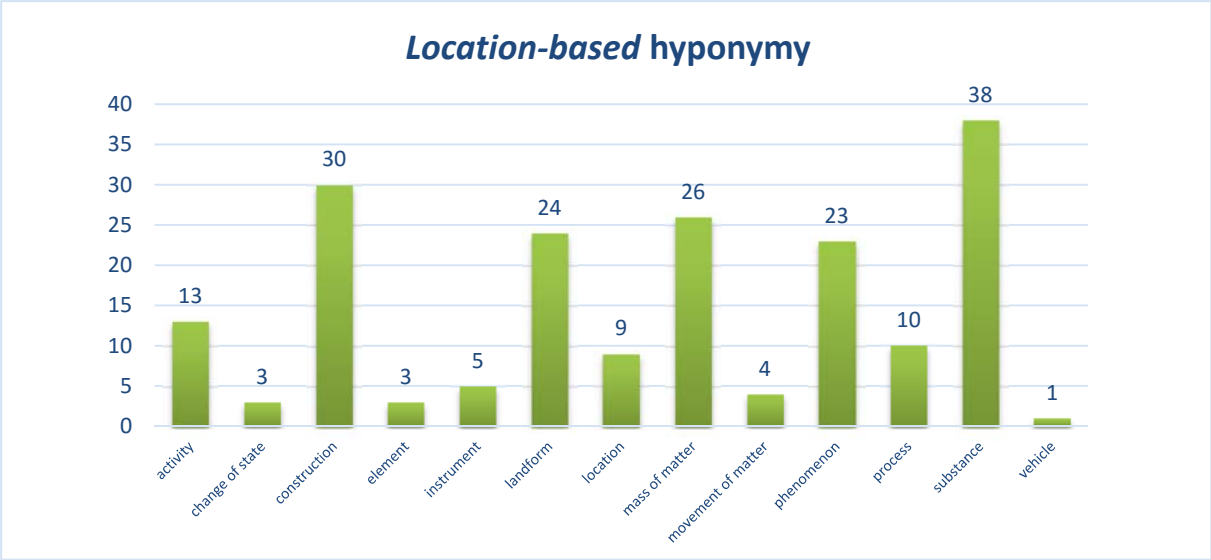


Figure 26: Location-based hyponymy subtypes per semantic category

In the case of *function-based* hyponymy (Figure 27) and *technology-based* hyponymy (Figure 28), the most frequently-activated semantic categories were those pertaining to artificial entities: instrument, vehicle, and construction. However, rather surprisingly, construction, which is the most recurrent category in *function-based* hyponymy, appeared less frequently in relation to *technology-based* hyponymy. This seems to indicate that the identifying feature of a construction is its purpose (e.g. PROCESSING FACILITY, PROTECTION STRUCTURE, LANDING DOCK), rather than its technology (e.g. NUCLEAR FACILITY, COAL-FIRED STATION, ORGANIC FARM).

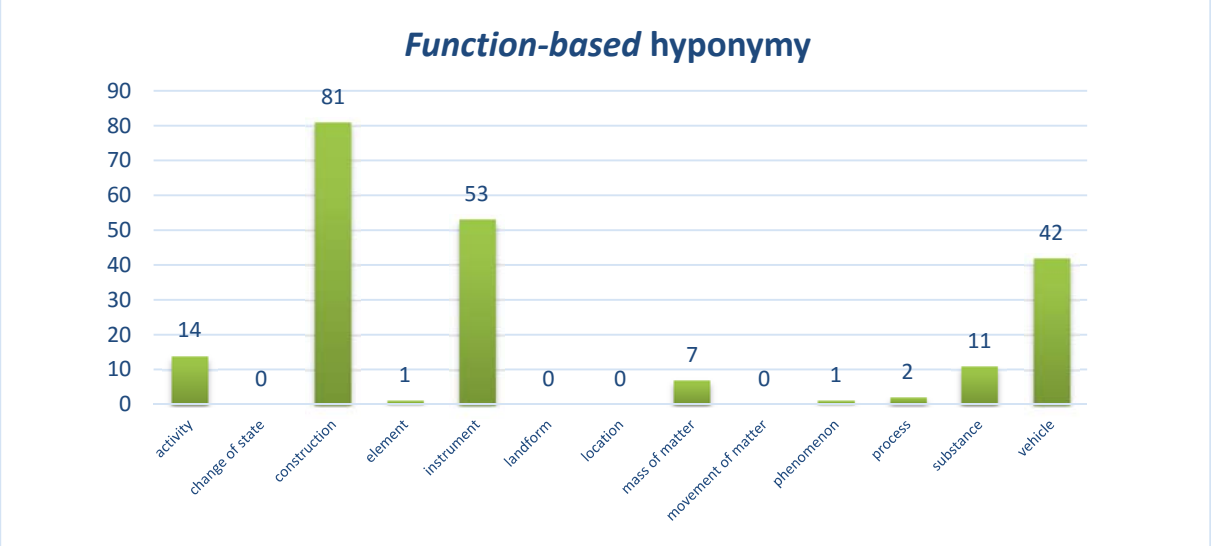


Figure 27: Function-based hyponymy subtypes per semantic category

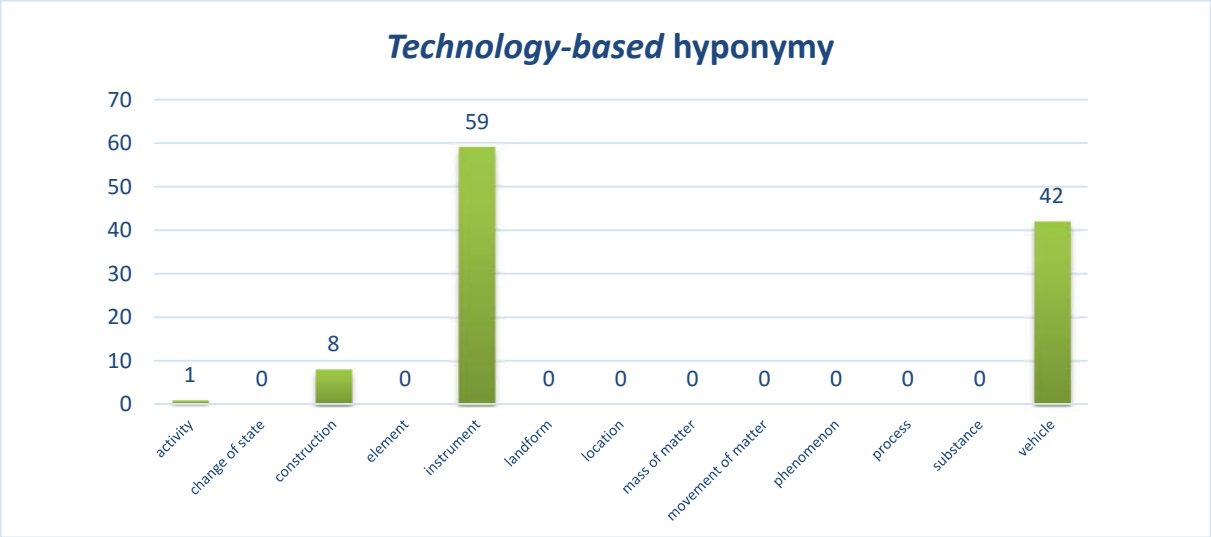


Figure 28: Technology-based hyponymy subtypes per semantic category

Regarding *denomination-based* hyponymy (Figure 29), almost all of the semantic categories activated were entities: landform, location, mass of matter, construction, and instrument. However, the category of phenomenon was in second position along with location, since certain meteorological events tend to receive denominations specifying the location where they occur (e.g. SUMATRA EARTHQUAKE, OKLAHOMA TORNADO, SAHEL DROUGHT).

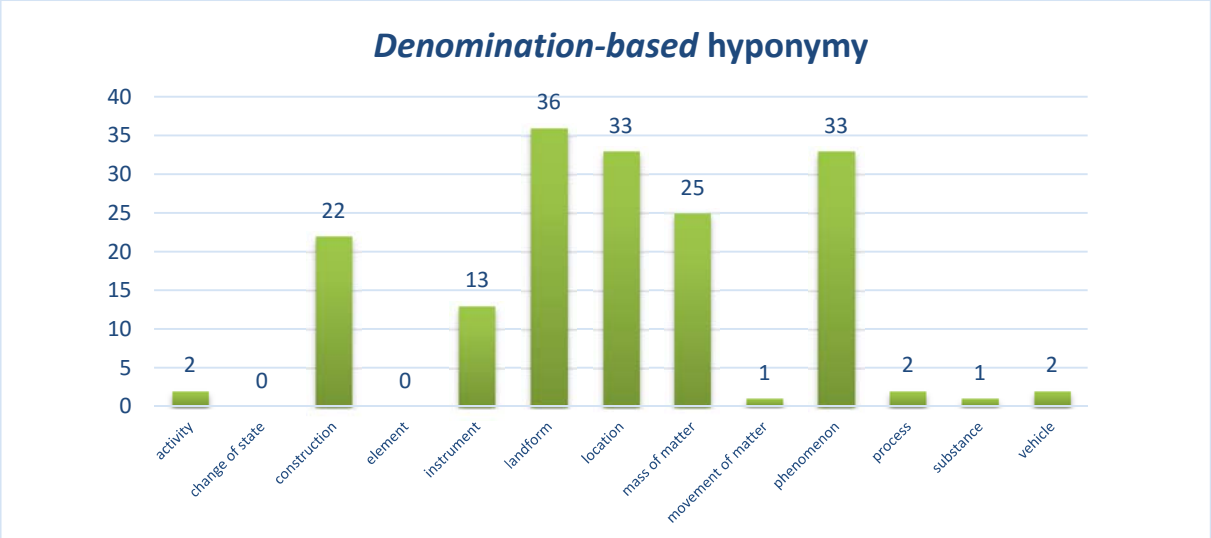


Figure 29: Denomination-based hyponymy subtypes per semantic category

Time-based hyponymy (Figure 30) was related to natural semantic categories, which were both processes (phenomenon and movement of matter) and entities (substance and mass of matter). In fact, time is also a natural factor that affects the environmental domain and phenomena (e.g. SUMMER PRECIPITATION, LATE-SEASON HURRICANE, PERIODIC DROUGHT). However, it rarely occurs with artificial concepts.

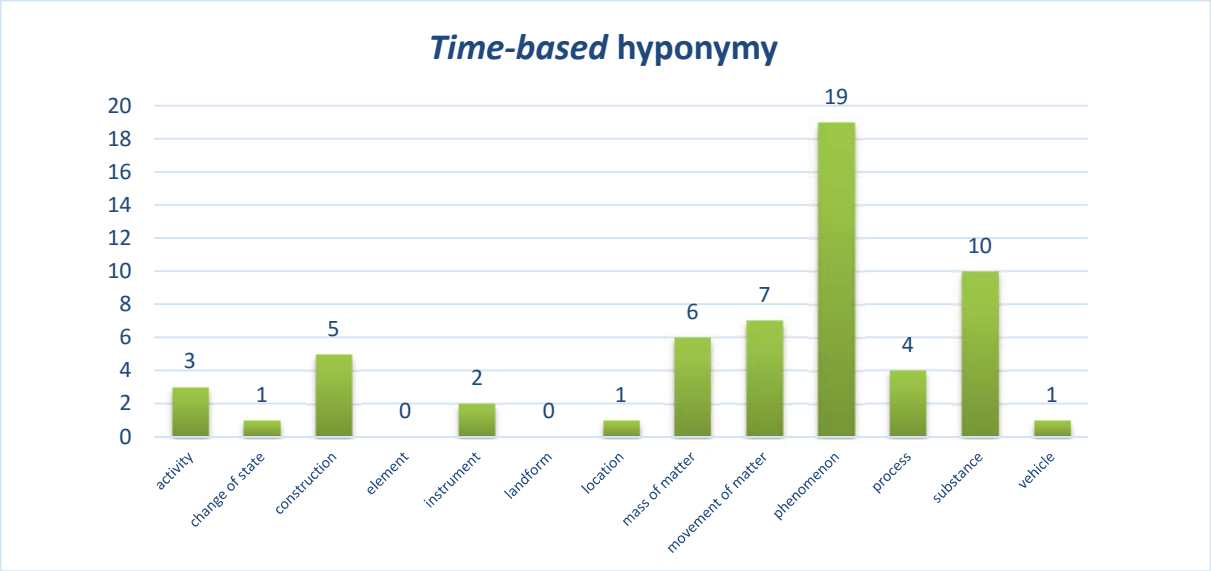


Figure 30: Time-based hyponymy subtypes per semantic category

Finally, with regard to *shape-based* hyponymy (Figure 31), the most recurrent semantic categories were the following artificial and natural entities: construction, landform, and mass of matter. Interestingly, shape occurred most frequently in the case of large formations (e.g. STAR DUNE, RING DIKE, VERTICAL BREAKWATER) than in the case of smaller formations or entities. Furthermore, two process-related semantic categories, movement of matter and phenomenon, are also registered in the table. They include concepts such as WEDGE TORNADO or CROWN FIRE, also characterized by the physical shape acquired by those processes.

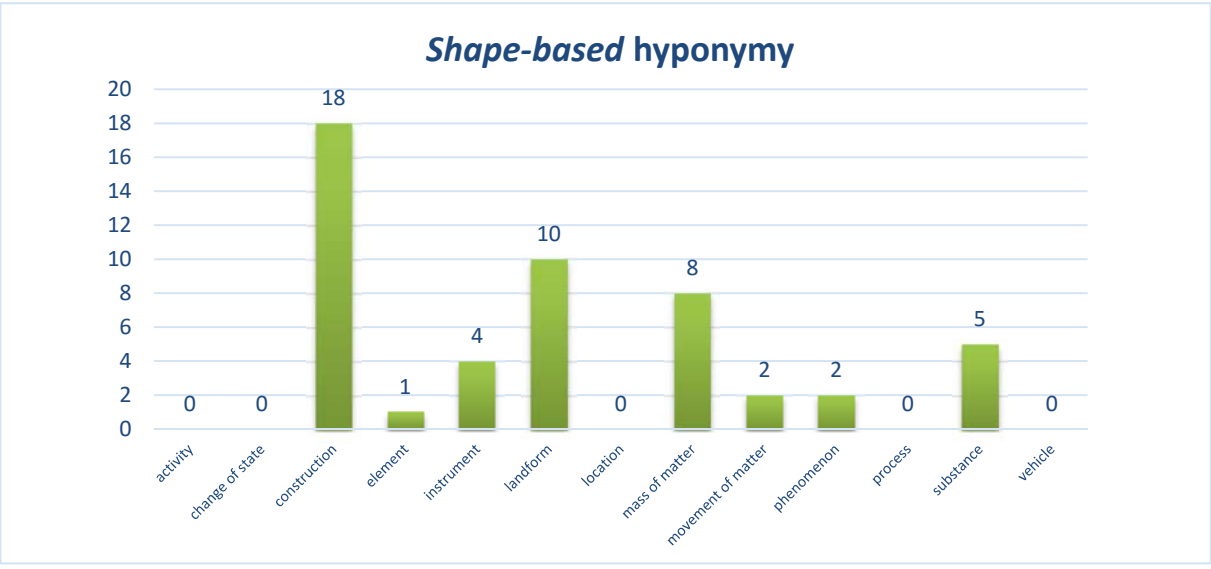


Figure 31: Shape-based hyponymy subtypes per semantic category

5. Conclusion

Hyponymy is a complex semantic relation that can be studied by analyzing concept hierarchies. The results obtained showed that the semantic category of concepts constrained their occurrence in different hyponymy subtypes. By analyzing and classifying hyponymic knowledge patterns and hyponymy subtypes, this study highlights the importance of accounting for semantic categories in the study of the generic-specific relation.

Our results showed that certain KPs (i.e. *exemplification*, *selection*, *itemization*, and *inclusion*) were linked to semantic categories that are the basis of scientific classifications (lifeform and chemical element). Furthermore, other KPs (*identification*, *denomination*, and *definition*) were found to have a more explanatory structure, and were thus most frequently linked to more complex semantic categories involving various participants (phenomenon, process, and technology). They thus invited a more detailed description and/or explanation to facilitate reader understanding. *Range* KPs were mostly associated with time period and measure since these categories are generally composed of values that are characterized by the space/distance between them in terms of time, space, intensity, etc.

The analysis of hyponymy showed that certain subtypes (*agent-based*, *patient-based*, *result-based*, *method-based*, and *degree-based*) closely correlated with process-related semantic categories (activity, phenomenon, process, and change of state). On the other hand, other hyponymy subtypes (*composition-based*, *technology-based*, and *function-based*) were directly linked to entity-related semantic categories (substance, landform, construction, and instruments). In addition, a distinction was made between natural and artificial concepts.

These results open the door to further studies on hyponymy not only in the environmental domain, but also in regard to specialized knowledge in general. In future research, we plan to analyze the whole English EcoLexicon corpus after a previous revision of the customized hyponymic word sketch grammars in order to reduce repetitions and false positives. Regarding hyponymy subtypes, another interesting feature to be explored in future work is the relation between certain subtypes identified (such as *composition-based*, *function-based*, or *origin-based*) and Pustejovsky's (1995) *qualia* structure (with formal, constitutive, telic, and agentive roles).

It would also be necessary to study the distinction between relational and attributional hyponymy subtypes. Another phenomenon to be explored is the correlation between hyponymic KPs and hyponymy subtypes. All of this information related to hyponymy refinement will make it possible to specify a more accurate set of hyponymic relations in the environmental domain.

6. Acknowledgements

This research was carried out as part of project FF2014-52740-P, *Cognitive and Neurological Bases for Terminology-enhanced Translation* (CONTENT), funded by the Spanish Ministry of Economy and Competitiveness.

7. References

Agbago, A. & Barrière, C. (2005). Corpus Construction for Terminology. *Proceedings of the Corpus Linguistics 2005 Conference*, pp. 1–14. Birmingham, United Kingdom.

Barrière, C. (2004a). Knowledge-rich Contexts Discovery. *Proceedings of the 17th Canadian Conference on Artificial Intelligence (AI'2004)*, pp. 187–201. London (Ontario), Canada.

Barrière, C. (2004b). Building a Concept Hierarchy from Corpus Analysis. *Terminology*, 10(2), pp. 241–263.

Bielinskiene, A., Boizou, L., Kovalevskaite, J., & Utka, A. (2012). Towards the Automatic Extraction of Term-defining Contexts in Lithuanian. In A. Tavast, K. Muischnek & M. Koit (Eds.) *Human Language Technologies: The Baltic Perspective*, pp. 18–26. Amsterdam/Berlin/Tokyo/Washington DC: IOS Press.

Cabré, M.T. (1999). *La terminología: representación y comunicación*. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.

Condamines, A. (2002). Corpus Analysis and Conceptual Relation Patterns. *Terminology*, 8(1), pp. 141–162.

Cruse, D.A. (2002). Hyponymy and its Varieties. In R. Green, C.A. Bean, & S.H. Myaeng, (eds.) *The Semantics of Relationships: An Interdisciplinary Perspective*, pp. 3–22. Dordrecht/Boston/London: Kluwer Academic Publishers.

Faber, P. (2009). The Cognitive Shift in Terminology and Specialized Translation. *Monografías de Traducción e Interpretación (MonTI)*, 1, pp. 107–134. Valencia: Universitat de València.

Faber, P. (2015). Frames as a Framework for Terminology. In H.J. Kockaert & F. Steurs (eds.) *Handbook of Terminology*, 1, pp. 14–33. Amsterdam/Philadelphia: John Benjamins.

Faber, P. (ed.) (2012). *A Cognitive Linguistics View of Terminology and Specialized Language*. Berlin/Boston: De Gruyter Mouton.

Faber, P., León Araúz, P., & Reimerink, A. (2014). Representing environmental knowledge in EcoLexicon. *Languages for Specific Purposes in the Digital Era, Educational Linguistics*, 19, pp. 267–301. Springer.

Faber, P., León-Araúz, P., & Reimerink, A. (2016). EcoLexicon: new features and challenges. In I. Kernerman, I. Kosem Trojina, S. Krek, & L. Trap-Jensen, (eds.), *GLOBALEX 2016: Lexicographic Resources for Human Language Technology in conjunction with the 10th edition of the Language Resources and Evaluation Conference*, pp. 73–80. Portorož, Slovenia.

Gil-Berrozpe, J.C. & Faber, P. (2016). Refining Hyponymy in a Terminological Knowledge Base. *Proceedings of the 2nd Joint Workshop on Language and Ontology (LangOnto2) & Terminology and Knowledge Structures (TermiKS) at the 10th edition of the Language Resources and Evaluation Conference (LREC 2016)*, pp. 8–15. Portorož, Slovenia.

Gil-Berrozpe, J.C., León-Araúz, P., & Faber, P. (in press). Subtypes of Hyponymy in the Environmental Domain: Entities and Processes. *Proceedings of the 10th International Conference on Terminology & Ontology: Theories and Applications (TOTh 2016)*. Chambéry, France.

Girju, R., Badulescu, A., & Moldovan, D. (2003). Learning Semantic Constraints for the Automatic Discovery of Part-Whole Relations. *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 1–8.

Jacquemin, C. & Bourigault, D. (2005). Term Extraction and Automatic Indexing. In R. Mitkov (ed.) *The Oxford Handbook of Computational Linguistics*. Oxford: Oxford University Press.

Kilgarriff, A., Rychlý, P., Smrz, P., & Tugwell, D. (2004). The Sketch Engine. In G. Williams & S. Vessier (eds.) *Proceedings of the Eleventh EURALEX International Congress*, pp. 105–116. Lorient: EURALEX.

León-Araúz, P. (2014). Semantic Relations and Local Grammars for the Environment. In S. Joeva, S. Mesfar & M. Silberztein (eds.), *Formalising Natural Languages with NooJ 2013*, pp. 87–102. Newcastle-upon-Tyne: Cambridge Scholars Publishing.

León-Araúz, P., San Martín, A., & Faber, P. (2016). Pattern-based Word Sketches for the Extraction of Semantic Relations. *Proceedings of the 5th International Workshop on Computational Terminology*, pp. 73–82. Osaka, Japan.

Marshman, E. (2002). The Cause Relation in Biopharmaceutical Texts: Some English Knowledge Patterns. *Proceedings of Terminology and Knowledge Engineering (TKE 2002)*, pp. 89–94. Nancy, France.

Meyer, I. (2001). Extracting Knowledge-rich Contexts for Terminography: A Conceptual and Methodological Framework. In D. Bourigault, C. Jacquemin & M. C. L'Homme (eds.) *Recent Advances in Computational Terminology*, pp. 279–302. Amsterdam/Philadelphia: John Benjamins.

Murphy, M.L. (2003). *Semantic Relations and the Lexicon: Antonymy, Synonymy and Other Paradigms*. Cambridge: Cambridge University Press.

Murphy, M.L. (2006). Hyponymy and Hyperonymy. In K. Brown (ed.) *Encyclopedia of Language and Linguistics*, 1, pp. 446–448. New York: Elsevier.

Pustejovsky, J. (1995). *The Generative Lexicon*. Cambridge, MA: MIT Press.

Schumann, A.K. (2012). Knowledge-Rich Context Candidate Extraction and Ranking with KnowPipe. *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, pp. 3626–3630.

Temmerman, R. (2000). *Towards New Ways of Terminology Description: The Sociocognitive Approach*. Amsterdam/Philadelphia: John Benjamins.

Appendix A: Semantic categories: description and examples

SEMANTIC CATEGORY	DESCRIPTION	EXAMPLES
activity	activities, techniques and behaviors	AGRICULTURE REPRODUCTION LAND USE PLANNING
change of state	natural processes involving the change of state of a certain matter	ICE MELTING FLASH EVAPORATION SNOW SUBLIMATION
chemical element	chemical elements and compounds	CHLOROFLUOROCARBON MERCURY NICOTINAMIDE
construction	man-made buildings and structures	TOWER MILL BREAKWATER PIPELINE
disease	illnesses and conditions	BLACK LUNG DISEASE CANCER MALARIA
domain	scientific or knowledge fields	BIOLOGY METEOROLOGY COASTAL ENGINEERING
feature	properties, characteristics and variables	SOIL MOISTURE BODY SIZE DENSITY
force	types of energy	HEAT WAVE SOLAR ENERGY ELECTRICITY
information	documents and data	CLIMOGRAPH BIOLOGICAL CLASSIFICATION BATHYMETRIC CHART
instrument	man-made inventions or creations used as instruments	MONITORING INSTRUMENT DIGITAL BAROMETER SAND FILTER
landform	geographical and geological features	ISLAND KARST MOUNTAIN
lifeform	living beings or organisms	SEABIRD MANGROVE TREE PROTIST
location	spatial environments	MARINE BIOME TROPICAL RAIN FOREST EUROPE
mass of matter	massive entities composed of certain substances	PLANET OCEAN GLACIER
measure	measuring units	CELSIUS HORSEPOWER KILOMETER

movement of matter	types of mass movement	EBBING TIDE LANDSLIDE MUDFLOW
period	time periods or spans	MONTH SEASON HOUR
phenomenon	meteorological and geological phenomena	TSUNAMI RAIN VOLCANIC ERUPTION
process	natural and artificial processes with agents and patients	ABRASION WEATHERING GAS ADSORPTION
product	natural and artificial substances that are the result of a process	GLASSWARE DEODORANT COFFEE
substance	solid, liquid and gaseous substances or materials	GRANITE FOSSIL FUEL WOOD
system	scientific systems and models	THEORY OF RELATIVITY SCIENTIFIC LAW EMPIRICAL METHOD
technology	man-made creations and inventions	GENERATOR AIRCRAFT RADIOSONDE
vehicle	man-made inventions or creations used as vehicles	MOTOR VEHICLE ELECTRIC CAR DELIVERY TRUCK

Appendix B: Hyponymic knowledge patterns: description and examples

HYPONYMIC KP TYPE	DESCRIPTION	EXAMPLES
classification	they classify or divide the hypernym into hyponyms	HYPER is classified into HYPO HYPER is divided into HYPO types of HYPER are classified as HYPO
definition	they introduce the hyponym with a definition where the hypernym is the *genus*	HYPO: a HYPER HYPO: a type of HYPER HYPO, defined as HYPER
denomination	they introduce the hyponyms as particular denominations	a type of HYPER called HYPO a type of HYPER known as HYPO types of HYPER are called HYPO
enumeration	they show an exhaustive and numbered list of hyponyms for the hypernym	# types of HYPER: HYPO # HYPER: HYPO # types of HYPER occur: HYPO
exemplification	they present the hyponyms as examples, types or kinds	HYPER such as HYPO

	of the hypernym	HYPER types such as HYPO HYPER like HYPO
identification	they directly link the hyponym to the hypernym with a copulative verb	HYPO is a HYPER types of HYPER are HYPO a type of HYPER is a HYPO
inclusion	they present the hyponyms as concepts included in the notion of the hypernym	HYPER including HYPO HYPER types include HYPO among HYPER are HYPO
itemization	they introduce a non-exhaustive list of hyponyms for the hypernym	HYPO and other HYPER HYPO and other HYPER types types of HYPER: HYPO
range	they establish a span where several hyponyms can be found for the same hypernym	HYPER ranging from HYPO to HYPO HYPER types ranging from HYPO to HYPO
selection	they highlight main or preferred hyponyms for the hypernym	HYPER, especially HYPO HYPER, mainly HYPO HYPER, usually HYPO

Appendix C: Hyponymy subtypes

HYPONYMY SUBTYPE	DESCRIPTION	EXAMPLES
ability-based	hyponyms characterized by own abilities or characteristics	RENEWABLE RESOURCE HABITABLE PLANET AUTONOMOUS VEHICLE
activity-based	hyponyms characterized by the activity or stability of their composition	RADIOACTIVE SUBSTANCE ALKALI METAL ACTIVE DUNE
agent-based	hyponyms characterized by the agent that causes them	STORM TIDE AIR OXIDATION SPRINKLER IRRIGATION
amount-based	hyponyms characterized by their amount or quantity	TRACE ELEMENT RARE METAL SINGLE STORM
color-based	hyponyms characterized by their color	COLORLESS SOLID RED TIDE YELLOW LIQUID
composition-based	hyponyms characterized by their components or by their material	METALLIC ELEMENT CARBONATE SAND PINE FOREST
degree-based	hyponyms characterized by their degree of intensity, size or consequences	CATACLYSMIC ERUPTION LOW-MAGNITUDE EARTHQUAKE MEGA-SCALE EXTRACTION
denomination-based	hyponyms characterized by having a particular denomination with a proper noun	Pacific Ocean Sahara Desert New York City
density-based	hyponyms characterized by their density or particle concentration	LIGHT ELEMENT DENSE WATER HEAVY METAL

domain-based	hyponyms characterized by the scientific or knowledge field to which they belong	AGRICULTURAL PRODUCT MUSICAL INSTRUMENT CHEMICAL INDUSTRY
effect-based	hyponyms characterized by the effects or consequences that they cause	TOXIC LIQUID HAZARDOUS SUBSTANCE GREENHOUSE GAS
function-based	hyponyms characterized by their function or purpose	DRINKING WATER SURVEILLANCE RADAR MANUFACTURING FACILITY
hardness-based	hyponyms characterized by their hardness level	SOFT WOOD HARD ROCK HARD STRUCTURE
height-based	hyponyms characterized by their height or depth level	SHALLOW WATER DEEP OCEAN HIGH TIDE
location-based	hyponyms characterized by their spatial location or position	OCEAN WATER SURROUNDING AIR TROPICAL STORM
method-based	hyponyms characterized by the method or the process that they involve	AEROBIC OXIDATION DIRECT SUBLIMATION INDUSTRIAL TREATMENT
moisture-based	hyponyms characterized by their moisture level	DRY SOLID SATURATED AIR ARID DESERT
movement-based	hyponyms characterized by their movement or direction	EBB TIDE OCEAN-GOING DREDGE OUTGOING RADIATION
origin-based	hyponyms characterized by their origin, i.e. the place where they come from or where they were created	NATURAL RESOURCE PINE WOOD COUNTRY ROCK
patient-based	hyponyms characterized by the patient that is affected by them	COAST EROSION ICE MELTING WATER TREATMENT
relation-based	hyponyms characterized by being related to other concepts	FOREIGN SUBSTANCE PARENT COMPOUND COVALENT SOLID
result-based	hyponyms characterized by the result that they cause, or by being the result of a process	TSUNAMIGENIC EARTHQUAKE PAPER INDUSTRY UNIMOLECULAR DECOMPOSITION
shape-based	hyponyms characterized by their shape	AMORPHOUS SOLID PARABOLIC DUNE L-SHAPED GROIN
size-based	hyponyms characterized by their size	TINY CRYSTAL GIANT PLANET COMPACT CAR
speed-based	hyponyms characterized by their speed	RAPID EROSION FLASH EVAPORATION SPONTANEOUS DECOMPOSITION
state-based	hyponyms characterized by the state of matter	SOLID SUBSTANCE FLUID ELEMENT

		MOLTEN ROCK
status-based	hyponyms characterized by a particular circumstance or situation	REGULATED SUBSTANCE UNTREATED WOOD CONTAMINATED SOIL
technology-based	hyponyms characterized by the technology that they use	MOTOR VEHICLE GREEN TECHNOLOGY DIGITAL BAROMETER
temperature-based	hyponyms characterized by their temperature	HOT GAS WARM OCEAN COLD AIR
texture-based	hyponyms characterized by their texture	VISCOUS LIQUID FINE SAND SOFT ROCK
time-based	hyponyms characterized by their duration, by their age, or by happening in a particular moment	WINTER ICE OLD ROCK ANNUAL PRECIPITATION
weight-based	hyponyms characterized by their weight	LIGHT-DUTY VEHICLE HEAVY-DUTY TRUCK LIGHT TRUCK

From Translation Equivalents to Synonyms: Creation of a Slovene Thesaurus Using Word co-occurrence Network Analysis

Simon Krek[1], Cyprian Laskowski[2], Marko Robnik-Šikonja[3]

[1] "Jožef Stefan" Institute, Artificial Intelligence Laboratory, Jamova 39, Ljubljana, Slovenia
&
University of Ljubljana, Centre for Language Resources and Technologies, Večna pot 113, Ljubljana, Slovenia
[2] University of Ljubljana, Faculty of Arts, Aškerčeva 2, Ljubljana, Slovenia
[3] University of Ljubljana, Faculty of Computer and Information Science, Večna pot 113, Ljubljana, Slovenia

E-mail: simon.krek@ijs.si, cyprian.laskowski@ff.uni-lj.si, marko.robnik@fri.uni-lj.si

Abstract

We describe an experiment in the semi-automatic creation of a new Slovene thesaurus from Slovene data available in a comprehensive English–Slovenian dictionary, a monolingual dictionary, and a corpus. We used a network analysis on the dictionary word co-occurrence graph. As the additional information, we used the distributional thesaurus data available as part of the Sketch Engine tool and extracted from the 1.2 billion word Gigafida corpus, as well as information on synonyms from a Slovene monolingual dictionary. The resulting database serves as a starting point for manual cleaning of the database with crowdsourcing techniques in a custom-made online visualisation and annotation tool.

Keywords: bilingual dictionary; translation equivalents; thesaurus; automated lexicography; network analysis

1. Introduction

Slovene as a language with approximately two million speakers in the Republic of Slovenia and an additional 0.5 million outside its borders who speak or understand it (cf. Krek, 2012: 44), has been a slow starter in relation to the availability of language reference books. The first, and to date the only, comprehensive monolingual dictionary was compiled and published in five volumes at the end of the 20th century, between 1970 to 1991; and until 2016 no thesaurus or similar reference book describing synonymy in Slovene had been available. In 2016, the Fran Ramovš Institute of Slovene Language, working under the umbrella of the Slovene Academy of Sciences and Arts, published a one-volume thesaurus on a little fewer than 1,300 pages, with its concept and data predominantly based on the already outdated monolingual dictionary. The academic thesaurus project started around 2002; therefore, it took 15 years to compile and publish the dictionary which is currently available only in printed form. The motivation to experiment with bilingual, corpus

and other types of data to create a thesaurus from scratch, originates from two basic deficiencies of the existing academic thesaurus: (1) it is available only in printed form and (2) it describes Slovene that was used in the middle of the 20th century and not the modern language.

In contrast, resources used to compile the thesaurus described in the paper were chosen to reflect primarily what is considered modern Slovene. Major changes in the political and economic system after 1991, when Slovenia became an independent state and abolished the post-WWII single-party system to introduce parliamentary democracy, also had a profound influence on language. Our source data originate from works created in the last 20 years and are explicitly and intentionally corpus-based. In this manner, the used data provide an accurate representation of the current state of the language.

The remainder of the paper is structured as follows. In Section 2, we describe the data sources: bilingual English–Slovene dictionary, the 1.2-billion-word Gigafida corpus of Slovene, and the Slovene monolingual dictionary. In Section 3, we first describe the procedure and algorithms used to automatically create the thesaurus data preprocessing, word co-occurrence graph and extraction of relevant synonyms with the Personal PageRank algorithm. We also present the evaluation of obtained synonymy and the final database. In Section 4, we discuss visualization of the thesaurus data, which we split into three parts: synonyms, collocations and good examples. Our visualization system includes a crowdsourcing component. In Section 5, we conclude the paper.

2. Source Data

2.1 Bilingual dictionary data

As the source dictionary for bilingual data, the Oxford-DZS Comprehensive English–Slovenian Dictionary (ODCESD, 2005–2006; Šorli et al., 2006) was used. Contrary to bi-directional bilingual dictionaries designed to serve for both encoding and decoding purposes for native speakers of languages involved, ODCESD is a mono-directional dictionary intended for Slovene native speakers decoding English texts. Consequently, the headword list is more extensive than usual (120,000), and senses receive a more in-depth treatment (specialised, archaic or rare). Organisation of senses is translation-based, meaning that all the senses which generate the same translation equivalent(s) are joined. While traditional bi-directional dictionaries generally avoid listing (near-)synonymous translations, ODCESD lists an exhaustive list of semantic and stylistic equivalents relying on the native speaker's ability to distinguish between nuances of meaning in translation equivalents. Close synonyms are separated by a comma, while words interchangeable in less than roughly 50% of contexts are separated by a semicolon. We use these strings of Slovene translation equivalents separated by commas or semicolons as a source of data on synonymy in Slovene.

2.2 Corpus data

The second source of information was the Gigafida corpus (Logar et al., 2012); in particular via the Thesaurus module in the Sketch Engine tool (Rychlý, 2016). Gigafida is a 1.2 billion-word corpus of some 40,000 texts of various genres. With its release in 2012, it represents the third iteration of the FIDA family of corpora, which is considered as the reference corpus series for Slovene, starting with the 100 million-word FIDA corpus in the year 2000, followed by the 620 million-word FidaPLUS corpus in 2006. In addition to the Sketch Engine tool, Gigafida is available in a custom-made web concordancer, together with its balanced 100 million-word subcorpus Kres.

2.3 Monolingual dictionary data

The third source of information was a monolingual Slovene dictionary (SSKJ, 2014), the data of which serving as an additional confirmation of associations between words. SSKJ provides the lexicographic description of Slovene from the second half of the 20th century in little more than 92,000 entries. Its first edition was compiled between 1970 and 1991, also representing the first, and to date the only, monolingual dictionary of modern Slovene. In 2014, the second edition was published with some 6,000 new entries, and the dictionary was partly updated. It is available online as part of the Fran dictionary portal and as an independent website.

3. Procedure

3.1 Preparation of data

The first step in building the database was the extraction of translation equivalents from the bilingual dictionary and normalisation of text where truncation devices were applied. The basis of data preparation was an XML version of the ODCESD, which had been stripped of information irrelevant for our purposes (including all English data). The main points of departure were the <tr> tags, which contained the translation(s) of a given headword in a given (sub)sense. For example:

```
<tr>zapustiti; opustiti; odpovedati se, odstopiti od</tr>
```

Here, the particular sense of a headword (abandon, v.) was given four translations, the last two of which are more similar to each other. The first two translations are considered as near synonyms separated by a semicolon, and the last two as core synonyms separated by a comma.

Our first step, however, was to expand two types of truncation devices used inside these tags: brackets and slashes. Brackets indicated shorter and longer versions of a translation, whether just a word or an entire phrase. We handled these by expanding the original text to both versions. For instance, "računalo, abak(us)" expanded to

"računalo, abak, abakus", and "mirovanje; začasni odlog (izvajanja)" became "mirovanje; začasni odlog, začasni odlog izvajanja".

Slashes indicated alternatives. There were two types of devices: if the slash was followed by a dash, it indicated alternative suffixes (e.g., gender variants), the first of which was identified by going back to the first instance of the letter after the dash. For instance, "zaveznik/-ica" became "zaveznik, zaveznica", and "bolj kot/od" expanded to "bolj kot, bolj od". Combinations of brackets and slashes also occurred, in which case the rules were combined. For instance, "(kavno/pšenično) zrno" became "zrno, kavno zrno, pšenično zrno".

Once these expansions were available, the <tr> contents were split at semicolons and commas, to generate a hierarchy of terms with <synch> and <syn> tags, respectively. Each term was placed in a <s> element, possibly with coded attributes which tracked the source truncation devices. Finally, any domain annotations (or labels) within the <tr>, represented by <la> tags, were also copied into the <synch>. For example:

```
<tr><la>knjiž.</la>prilagoditi (se), akomodirati (se);
prirediti; uskladiti</tr>
```

generated:

```
<synch>
  <la>knjiž.</la>
  <syn>
    <s p="1" t="o-vz">prilagoditi se</s>
    <s p="1" t="o-vz">prilagoditi</s>
    <s p="2" t="o-vz">akomodirati se</s>
    <s p="2" t="o-vz">akomodirati</s>
  </syn>
  <syn>
    <s>prirediti</s>
  </syn>
  <syn>
    <s>uskladiti</s>
  </syn>
</synch>
```

At this point, a large list of structured translation equivalents was available. The next key step was finding all the potential synonyms for each term, and generating a reorganised XML file, arranged by headword rather than translation chain (<synch>).

Therefore, a new file was generated with data organised by headwords, with counted frequencies of their co-occurrences with individual candidate synonyms.

For every unique string within the <s> tags, an entry was created with that string as the headword. Then, we generated a synonym candidate list for that headword by looking at all the other <s> strings which co-occurred within a <synch> with the headword. For each such candidate, we tabulated the "core" and "near" counts, by counting all the co-occurrences of the headword and candidate and checking whether they were in the same <syn> or not. In order to detect relationships between the candidates, we calculated these totals for every pair of candidates.

During this phase, we analysed the attributes and labels of the truncation mechanism and used them to filter the data. For instance, since the ODCESD truncation mechanism with a slash-hyphen combination was used mainly for separating female and male translations, we tracked this and filtered out terms with mismatching genders. This way "študent" (male student) and "študentka" (female student) would not be treated as synonyms. Similarly, we removed all the variants that derived from the bracket truncation mechanism, when it resulted in extra words so that only the longest string containing a shorter possible synonym was kept. We also filtered out some labels which were irrelevant for our purposes (e.g., American vs British English).

The result of this phase of data processing was an XML file with 135,073 headwords, organised into entries containing <direct_syns> and <indirect_syns>. The <direct_syns> contained the candidates and their core/near counts with respect to the headword, along with any labels that co-occurred with the combination. The <indirect_syns> contained all the pairs of candidates that co-occurred (with each other and the headword) and their core/near counts with respect to each other. For instance, the word "neobremenjenost" occurred in three <tr> tags in the ODCESD:

```
<tr><la>poet.</la>razpuščenost, zanesenost; sproščenost,
neobremenjenost</tr>
<tr>samozavestnost, neobremenjenost</tr>
<tr>brezskrbnost, neobremenjenost, sproščenost</tr>
```

This resulted in:

```
<entry>
  <head>
    <form>
      <headword>neobremenjenost</headword>
    </form>
  </head>
  <body>
    <syns>
      <direct_syns>
        <direct_syn core="1" near="0">
          <s>brezskrbnost</s>
        </direct_syn>
        <direct_syn core="0" near="1">
          <s>razpuščenost</s>
          <labels>
```

```
            <la>poet.</la>
          </labels>
        </direct_syn>
        <direct_syn core="1" near="0">
          <s>samozavestnost</s>
        </direct_syn>
        <direct_syn core="2" near="0">
          <s>sproščenost</s>
        </direct_syn>
        <direct_syn core="0" near="1">
          <s>zanesenost</s>
          <labels>
            <la>poet.</la>
          </labels>
        </direct_syn>
      </direct_syns>
      <indirect_syns>
        <indirect_syn core="1" near="0">
          <s>brezskrbnost</s>
          <s>sproščenost</s>
        </indirect_syn>
        <indirect_syn core="0" near="1">
          <s>razpuščenost</s>
          <s>sproščenost</s>
        </indirect_syn>
        <indirect_syn core="1" near="0">
          <s>razpuščenost</s>
          <s>zanesenost</s>
        </indirect_syn>
        <indirect_syn core="0" near="1">
          <s>sproščenost</s>
          <s>zanesenost</s>
        </indirect_syn>
      </indirect_syns>
    </syns>
  </body>
</entry>
```

The <indirect_syns> also helped us organise candidate synonyms into groups with a simple rule: if it is possible to reach one candidate from another through a sequence of one or more <indirect_syn> tags, then the candidates belong in the same group. With these data in place, we could set up a co-occurrence graph, as described in the next section.

3.2 Co-occurrence graph

The most important step in organising data according to word associations was the creation of a weighted co-occurrence graph. The graph contains frequencies of co-occurrence of translation equivalents from the whole database. We ran the Personal PageRank algorithm (Page et al., 1999) on this graph to rank the synonym list, separately for each synonym candidate. Having obtained lists of synonyms and near synonyms for each headword, we now pursued three goals: i) determine groups of

words with the same meaning, ii) rank the groups of words with the same meaning according to their semantic similarity with the headword, and iii) rank words within groups according to their frequency of use.

Graphs are a suitable formalism to model semantic relations. We created a word co-occurrence graph G=(V, E), where V is a set of nodes (each node represents one headword with its label if present), and E is a set of connections between nodes. The edge e_{ij} connects nodes i and j and has an associated weight $w_{ij} \in R^+$. The weight models the strength of semantic similarity between words i and j. The larger the weight, the stronger the association between words i and j. Value $w_{ij} = 0$ means that there is no synonymy between words i and j. We organise values w_{ij} into a matrix W, called an adjacency matrix as its values contain degrees of adjacency between nodes. We calculate each cell of the matrix as a weighted sum of synonymy information from our three sources. The primary source of information is the core and near counts for headword-candidate or candidate-candidate combinations (core counts were given twice the weight of near counts). In addition, data from the Thesaurus module in Sketch Engine (Rychlý, 2016) and a monolingual Slovene dictionary, were included as additional information. The Figure 1 below shows a graphical representation of core and near synonyms for the headword *hiša*. Words in rectangles form groups with the same meaning e.g., *bivališče* and *domovanje*. The groups are subgraphs connected with <indirect_syn> tags, as described above. In the actual graph, these words are all connected to the headword but we excluded the connections from the graph to avoid clutter.

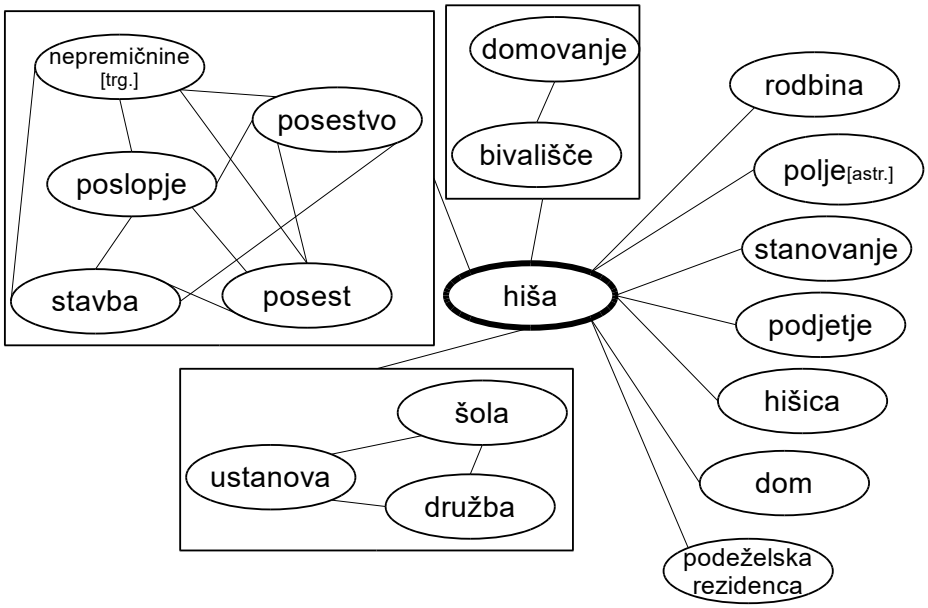


Figure 1: Co-occurrence graph ('hiša' – house)

We set the weight of each connection as the linear combination of contributing factors (core or near synonym, association score from the Sketch Engine tool and confirmation from the monolingual dictionary).

$$W = \text{coreWeight} \cdot \text{coreCount} + \text{nearWeight} \cdot \text{nearCount} + \text{sskjWeight} \cdot \text{sskjScore} + \text{sketchWeight} \cdot \text{sketchScore}$$

Here coreWeight, nearWeight, sskjWeight, and sketchWeight are weights given to each contributing factor. We used a preliminary evaluation on a small number of different headword categories to set sensible default values, namely coreWeight=2, nearWeight=1, sskjWeight=3, and sketchWeight=1. The most important factors are coreCount and nearCount, which are determined as the number of joint occurrences of connected words as core synonyms and near synonyms, respectively. The sskjScore and sketchScore are auxiliary factors with values between 0 and 1 (note that coreCount and nearCount mostly have a far greater range), which strengthen information from the bilingual dictionary. The sskjScore for headword i and connected word j is 0 if the SSKJ dictionary does not contain word j in the description of headword i. If the dictionary description contains the word j then the sskjScore is a value between 0 and 1 depending on the frequency of word j in the Gigafida corpus. If the corpus contains more than 50 instances of word j, the value of sskjScore is 1, if the frequency of a word is less than or equal to 3, sskjScore=0, otherwise it linearly depends on the frequency of j in Gigafida: sskjScore = (frequency - 3) / (50 - 3). The sketchScore is actually the logDice score (Rychlý, 2008) and is reported by the Sketch Engine as the default word association score. It is based on co-occurrence of two words in a corpus of documents, in our case in Gigafida.

Ranking of nodes is one of the frequently used tasks in the analysis of network properties and several so-called node centrality measures exist to assess the influence of a given node in the graph. The objective of ranking is to assess the relevance of a given node either globally (with regard to the whole graph) or locally (relative to some node in the graph). A well-known ranking method is PageRank (Page et al., 1999), which was used in the initial Google search engine. For a given network with the adjacency matrix W, the score of the i-th node returned by the PageRank algorithm is equal to the i-th component of the dominant eigenvector of W'^{T}, where W' is the matrix W with rows normalised so that they sum to 1. This can be interpreted in two ways. The first interpretation is the 'random walker' approach: a random walker starts walking from a random vertex v of the network and in each step walks to one of the neighbouring vertices with a probability proportional to the weight of the edge traversed. The PageRank of a vertex is then the expected proportion of time the walker spends in the vertex, or, equivalently, the probability that the walker is in the particular vertex after a long time. The second interpretation of PageRank is the view of score propagation. The PageRank of a vertex is its score, which it passes to the neighbouring vertices. A vertex v_i with a score PR(i) transfers its score to all its neighbours. Each neighbour receives a share of the score proportional to the strength of the edge between itself and v_i. This view explains the PageRank algorithm with the principle that in order for a vertex to be highly ranked, it must be pointed to by many highly ranked vertices. Other methods for ranking include Personalized-PageRank (Page et al., 1999), frequently abbreviated

as P-PR, that calculates the vertex score locally to a given network vertex, SimRank (Jeh & Widom, 2002), diffusion kernels (Kondor & Laerty, 2002), hubs and authorities (Kleinberg, 1999) and spreading activation (Crestani, 1997).

As we need a locally sensitive ranking of the nodes (i.e., the importance for each headword separately), we applied the P-PR algorithm to each headword (graph node) to rank the synonyms according to the strength of their relationship with the headword.

3.3 Evaluation

We evaluated the obtained results on a set of 50 randomly chosen headwords from different categories by comparing the returned ranks with those manually assigned by two annotators and language experts (denoted as E1 and E2). The selection of headwords by part-of-speech was controlled (25 nouns, 10 verbs, 10 adjectives, 5 adverbs) and some filters were applied, such as the minimum number of synonyms. The overall number of synonyms in 50 entries was 550 (302 core and 248 near synonyms). They were evaluated against two criteria: (1) each synonym was scored according to a three-point numerical scale, with 2 indicating a valid synonym, 1 an acceptable synonym and 0 not a synonym; and (2) the placement of synonyms to groups indicating semantic distinctions was manually evaluated. The main conclusions from the evaluation process were:

1. The results of the procedure are highly useful with regards to both the obtained sets of synonyms and their ranking.

Table 1 below presents the distribution of scores (0–2) for core and near synonyms and both evaluators.

	Evaluator E1						**Evaluator E2**					
score	**core**	%	**near**	%	**sum**	%	**core**	%	**near**	%	**sum**	%
2	164	54.3	58	23.4	222	40.4	107	35.4	20	8.1	127	23.1
1	64	21.2	84	33.9	148	26.9	110	36.4	125	50.4	235	42.7
1+2	228	**75.5**	142	**57.3**	370	**67.3**	217	**71.9**	145	**58.5**	362	**64.0**
0	74	24.5	106	42.7	180	32.7	85	28.1	103	41.5	188	34.2
sum	**302**	100	**248**	100	**550**	100	**302**	100	**248**	100	**550**	100

Table 1: Distribution of scores

The results show that around three quarters of the core synonyms are either valid or acceptable (E1 75.5%, E2 71.9%), as well as more than half of the near synonyms (E1 57.3%, E2 58.5%), which indicates that both core and near synonyms should be taken into account in the final database, although with a clear distinction between the two groups. Together around two-thirds of the obtained results are useful (E1 67.3%, E2 64%).

The agreement between the annotators is shown in Table 2.

Scores (E1-E2)	No. (synonyms)	Sum	%	Sum (1-2)	%
2-2 (valid)	98				
1-1 (acceptable)	99				
0-0 (not synonym)	130	327	59,5		
1-2 (acceptable – valid)	115			442	80,4
0-1 (not synonym – acceptable)	70				
0-2 (not synonym – valid)	38				
Sum	550				

Table 2: Inter-annotator agreement

Annotators completely agreed in 59.5% of cases, or in 80.4% if the difference between valid and acceptable synonyms is ignored, which is less important than between a synonym and non-synonym. The critical disagreement is between valid synonyms and non-synonyms (38 cases). The qualitative analysis shows that in the majority of cases the disagreement originates from a different strategy of evaluators in relation to a limited number of specific situations. For example, a multi-word synonym contains a single-word synonym which is already included in the list (headword: *videti* 'to see', single-word synonym: *srečati* 'to meet', multi-word synonym: *večkrat srečati* 'to meet repeatedly'), or some synonyms are semantically valid but differ in register, style, etc. (headword: *sodnik* 'judge', synonym: *rihtar* 'judge' (old informal expression)). In general, the bias should be towards inclusion: if at least one evaluator found the synonym acceptable, it is likely that some dictionary users would be interested in seeing it in the dictionary. The percentage of synonyms with at least one positive score is 76.4%.

2. Sense groups can only be considered as an indication of sense distribution and cannot be explicitly used as separate dictionary senses.

Table 3 in Appendix 1 shows the results of the evaluation of the sense groups. We used the Adjusted Rand Index (ARI) (Hubert & Arabie, 1985) to calculate agreement between the sense groupings constructed in the automatic procedure (A) or created by the evaluators (E1 or E2). ARI counts pairs of items that are in the same (or different) group in one grouping and at the same time are part of the same (or different) group in the other grouping. It yields a score of 1 for identical groupings and 0 for matching of groups expected by chance (so it is also possible to have negative ARI score). Unsurprisingly, the overlap is high in the case of monosemous headwords with a small number of synonyms, but in the case of polysemous headwords with a higher number of synonyms, the numbers show that the results are not good enough to use groups directly as separate senses. We note that the overlap between the two human evaluators is not very high which shows that the task is fairly difficult and there can be multiple solutions (presumably due to different granularities of categorisation).

Some other important qualitative observations from the evaluation are:

- a selected set of labels should be kept, particularly those indicating terminology (field labels);
- synonym candidates resembling definitions should be removed;
- in (rare) cases when a headword can belong to multiple parts-of-speech, synonyms are mixed and should be separated according to the different parts-of-speech;

We considered the results satisfactory and therefore did not test other network centrality measures (besides P-PR) as this would require another evaluation round. This may be an interesting further work.

3.4 Final database

The result was an automatically created thesaurus with 78,276 headwords where at least one core or near synonym is available. Of them, 41,555 are single-word and 36,721 multi-word headwords. Headwords with the highest number of synonyms are adjectives *hud* (angry, bad), with 110 as the total number synonyms (core + near) and *oster* (sharp) with the highest number of 62 core synonyms. The distribution of synonyms is uneven, with a long tail: 27,136 headwords have only one synonym (core or near), with the next 14,451 headwords having two synonyms. In Figure 2 we show a graphic representation of the number of synonyms per entry/headword.

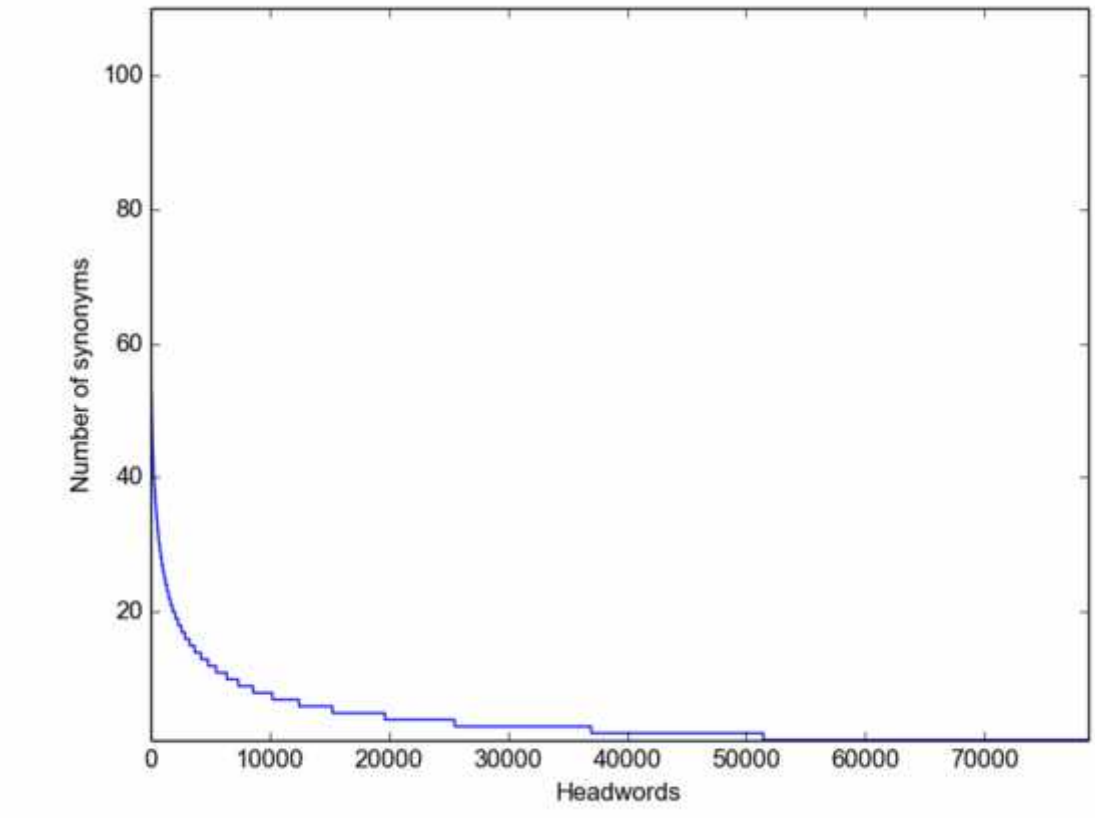


Figure 2: Number of synonyms per entry/headword

Finally, two types of additional information were added to all headwords and synonyms: (1) frequency of occurrence from the Gigafida corpus was added to single-word headwords and synonyms, and (2) part-of-speech category was attributed to headwords. This data set in XML format represents the result of the automatic procedure which was used in the visualisation and crowdsourcing tool described in the next section.

4. Visualisation and Crowdsourcing

From the very beginning of the experiment, automated generation of a thesaurus from the dictionary and corpus data was not the final goal. Crowdsourcing as a language resource creation method is also gaining ground in "traditional" lexicography (Čibej et al., 2015), and the ultimate goal was, in fact, to enable experimenting with crowdsourcing in creating a thesaurus for Slovene.[1] For this purpose, an online visualisation and annotation tool has been developed that enables users of the automatically created data to give feedback on the quality of generated synonyms. The tool was specifically designed to help users in recognising the differences between the two synonyms and enable scoring on the basis of contextual data. For this purpose, we use data extracted from the Gigafida corpus as produced by the Sketch-Diff module in the Sketch Engine tool. Users of the online thesaurus visualisation platform can advance through three consecutive pages when consulting each entry, and through linking to the previously available web concordancer, end up in the Gigafida corpus as the final consultation tool. The three pages are (1) synonyms page, (2) collocations page, and (3) good examples page.

4.1 Synonyms page

The initial page shows sets of synonyms in two groups: core and near synonyms. The two groups can be listed according to four criteria: ranking, alphabet, word length and user scores. We also use frequency data from the Gigafida corpus to show which synonyms are more frequent in the language in general, with a four-degree slider consecutively dimming less frequent synonyms. Finally, the most important features of the initial page are evaluation buttons accompanying each synonym enabling users to score the synonym.

4.2 Collocations page

Provided that relevant data can be found in the Gigafida corpus, we show collocations of both the headword and the synonym on the collocations page. For this purpose, we use data extracted from the Gigafida corpus taking advantage of a combination of Word Sketch and Thesaurus modules in the Sketch-Diff part of the Sketch Engine tool. Sketch-diffs show which collocations appear more frequently with

[1] The crowdsourcing phase has not begun at the time of the submission of the article, therefore we cannot provide any numbers on the success of crowdsourcing.

either the headword or the synonym, or equally with both. Collocations are selected on the basis of the "sketch grammar", a formalism identifying statistically relevant words appearing in specific grammatical relations in the corpus, such as an adjective preceding a noun. The grammar was developed for automatic extraction of data for a new Slovene collocations dictionary (Krek et al., 2016), which was used also in the online thesaurus tool. By clicking on a synonym on the initial page, the user is presented with collocations in four grammatical relations, in three different groups: those that are equally distributed in the context of both the headword and the synonym, and those that appear only with either the headword or the synonym, thus providing the means to the user to understand subtle semantic and contextual differences of synonyms.

4.3 Good examples page

As part of the collocations dictionary project, a tool producing "good dictionary examples" or GDEX tool (cf. Kosem, 2016) was also developed, and relevant dictionary examples were extracted from the Gigafida corpus. If good examples exist in the extracted database for a particular collocation, they are shown on the last page in the visualisation tool. In addition, each combination of the headword/synonym + collocate is equipped with a direct link to the Gigafida web concordancer producing all available concordances for that particular combination.

5. Conclusion

We described the automatic creation of a Slovene thesaurus from bilingual dictionary data, monolingual dictionary data, and corpus data. The methodology that was used in generating the thesaurus is general and can be applied to other dictionaries and languages. The step specific to our case is preprocessing, where the original dictionary data are transformed into XML form, suitable for our purposes. Other steps, such as the creation of word co-occurrence graph or ranking of nodes with P-PR, etc., are more general and can be reused if the input data are available in a suitable format. The methodology is flexible and allows integration of different sources of word association information. Furthermore, we described how we intend to use crowdsourcing in lexicography, which is a new trend, particularly in relation to cleaning the automatically generated data or data extracted from a corpus. Conceptually, visualisation and crowdsourcing solutions are also language independent and, given that similar types of data (synonyms, collocations, good examples) are used, solutions in the interface can be used in other dictionary portals.[2] We are convinced that this trend will grow and that other similar initiatives will follow in the coming years. We believe that our experiment on an automatic generation of a thesaurus represents a step in this direction.

[2] All relevant information will be available on the CJVT website (http://www.cjvt.si) after the lauch of the synonym dicitonary portal which is expected in October 2017.

6. Acknowledgements

We would like to thank DZS, the publisher of the *Oxford-DZS comprehensive English–Slovenian dictionary*, for their permission to conduct the experiment with the Slovene translation equivalents part of the bilingual dictionary database. Jan Kralj kindly provided the Python source code of P-PR algorithm. We would also like to thank the evaluation team: Špela Arhar Holdt, Jaka Čibej, Polona Gantar and Iztok Kosem.

7. References

Čibej, J., Fišer, D. & Kosem, I. (2015). The role of crowdsourcing in lexicography. In I. Kosem, M. Jakubíček, J. Kallas & S. Krek (eds.) *Proceedings of the eLex 2015 conference, 11-13 August 2015, Herstmonceux Castle, United Kingdom*. Ljubljana/Brighton: Trojina, Institute for Applied Slovene Studies/Lexical Computing, pp. 70-83.

Crestani, F. (1997). Application of spreading activation techniques in information retrieval. *Artificial Intelligence Review*, 11(6), pp. 453-482.

Hubert, L. & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2(1), pp. 193–218.

Jeh, G. & Widom, J. (2002). SimRank: A measure of structural-context similarity. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 538-543. ACM.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. (2014). The Sketch Engine: ten years on. In *Lexicography*, vol 1. issue 1. Springer, pp. 7–36.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5), pp. 604-632.

Kondor, R. I. & Laerty, J. D. (2002). Diffusion kernels on graphs and other discrete input spaces. In *Proceedings of the 19th International Conference on Machine Learning*, pp. 315-322.

Kosem, I. (2016). Interrogating a corpus. In P. Durkin (ed.) *The Oxford Handbook of Lexicography*. Oxford: Oxford University Press, pp. 76-93.

Krek, S. (2012). *The Slovene Language in the Digital Age / Slovenski jezik v digitalni dobi*. META-NET White Paper Series "Europe's Languages in the Digital Age", Springer, Heidelberg.

Krek, S., Gantar, P., Kosem, I., Gorjanc, V. & Laskowski, C. (2016). Baza kolokacijskega slovarja slovenskega jezika. *Proceedings of the Conference on Language Technologies & Digital Humanities, Faculty of Arts, University of Ljubljana*. Ljubljana, Slovenia. pp. 101-105.

Logar Berginc, N., Grčar, M., Brakus, M., Erjavec, T., Arhar Holdt, Š. & Krek, S. (2012). *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba*. Ljubljana: Trojina, zavod za uporabno slovenistiko; Fakulteta za družbene vede.

Page, L., Brin, S., Motwani, R. & Winograd, T. (1999). *The PageRank Citation Ranking: Bringing Order to the Web*. Technical Report. Stanford InfoLab.

Rychlý, P. (2016). Evaluation of the Sketch Engine Thesaurus on Analogy Queries. In *Tenth Workshop on Recent Advances in Slavonic Natural Language Processing*, RASLAN. pp. 147–15.

Rychlý, P. (2008). A lexicographer-friendly association score. In *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN*, pp. 6–9.

Šorli, M., Grabnar, K., Krek, S., Košir, T. (2006). Oxford-DZS comprehensive English-Slovenian dictionary. In *Proceedings XII EURALEX international congress*. Edizioni dell`Orso: Universita di Torino: Academia della Crusca, pp. 631-637.

Dictionaries:

ODCESD: *The Oxford®-DZS Comprehensive English-Slovenian Dictionary*. (2005-2006). Ljubljana: DZS.

SSKJ: *Slovar slovenskega knjižnega jezika (Dictionary of Literary Slovene)*. (2014). 2nd edition. Ljubljana: ZRC SAZU.

Appendix 1

Table 3: The comparison of synonym groups produced automatically (A) or by two human evaluators (E1 or E2).ARI score of 0 corresponds to an agreement between randomly assigned groups and 1 to perfectly matching groups. Note that the agreement expressed by ARI score may not be linear.

headword	POS	E1-E2	A-E1	A-E2	groups	synonyms	core	near	SSKJ
pritrditi	v	0.39	0.16	0.06	20	46	21	25	3
ugotoviti	v	0.14	0.04	0.21	20	51	13	38	1
vodilo	n	0.32	0.21	0.28	15	34	16	15	4
prepričati	v	0.21	0.03	0.04	14	28	8	20	1
kraj	n	0.48	0.42	0.27	12	19	12	7	7
preklic	n	0.48	0.12	0	12	26	11	15	5
sodnik	n	0.53	-0.03	-0.05	11	18	6	12	3
zrasti	v	0.34	0.25	0.41	11	25	14	10	10
hudodelstvo	n	0.55	0.17	0	9	12	3	9	1
drgnjenje	n	0.26	0.22	0.20	9	15	6	9	3
miniaturen	adj	0.26	0.14	0.34	9	18	14	4	2
temačno	adv	0.55	0.22	0.42	8	23	10	13	3
videti	v	0.34	0.33	0.42	8	14	4	10	12
analiza	n	0.26	0.09	0.26	8	18	11	7	1
pretepač	n	0.65	0.04	-0.02	7	11	6	5	1
krutost	n	0.37	0.29	0.15	6	18	11	7	1
žvižganje	n	0.01	0.14	0.14	6	10	4	6	5
zmršen	adj	1	0.66	0.66	5	10	4	6	1
kliše	n	1	0.41	0.41	5	7	7	0	2
prileżnica	n	0.64	0.26	0.13	5	6	5	1	1
pretepanje	n	-0.08	0.35	0.24	5	7	5	2	1
odveza	n	1	0.62	0.62	4	4	4	0	2
prasica	n	0.37	0.53	0.24	4	12	10	2	3
grotesken	adj	0.15	0.15	1	4	10	3	7	2
razsipništvo	n	0.12	0.30	0.02	4	7	1	6	1
somišljenik	n	0.63	0.06	-0.10	3	5	2	3	1
izpodbijati	v	0.55	1	0.55	3	5	2	3	2
odmevno	adv	0	0.33	0.57	3	4	3	1	-
spenjati	v	0	0	0	3	3	3	0	2
bivalen	adj	1	1	1	2	4	3	1	1
pravokotnica	n	1	1	1	2	3	1	2	1
zalust	n	1	1	1	2	3	1	2	1
povratno	adv	1	0.33	0.33	2	4	4	0	2
nagajivka	n	0.51	-0.17	-0.18	2	10	10	0	2

tarnanje	n	0.22	-0.06	-0.10	2	15	12	3	1
gensko	adv	0	0	1	2	2	1	1	-
prevohati	v	0	0	1	2	2	2	0	2
despotski	adj	-0.11	0.32	-0.22	2	9	9	0	1
ponazarjati	v	-0.20	0.57	-0.29	2	4	4	0	1
spodjesti	v	1	1	1	1	4	2	2	3
predelan	adj	1	1	1	1	2	1	1	1
ultramoderen	adj	1	1	1	1	2	2	0	1
paleolitik	n	1	1	1	1	2	2	0	1
investicija	n	1	1	1	1	1	1	0	2
konjunkcija	n	1	1	1	1	1	1	0	2
nevedno	adv	1	1	1	1	1	1	0	1
vsemogočnost	n	1	1	1	1	1	1	0	1
naturen	adj	1	1	1	1	1	1	0	1
popoten	adj	1	1	1	1	1	1	0	1
vrvičen	adj	0	0	1	1	2	2	0	1

E1-E2 – ARI between groups of Evaluator 1 and Evaluator 2
A-E1 – ARI between automatically assigned groups and Evaluator 1
A-E2 – ARI between automatically assigned groups and Evaluator 2
groups – number of groups created by the P-PR algorithm
synonyms – number of synonyms
core – number of core synonyms
near – number of near synonyms
SSKJ – number of senses in the monolingual Slovene dictionary

An Ontology-terminology Model for Designing Technical e-dictionaries: Formalisation and Presentation of Variational Data

Laura Giacomini

Hildesheim University/ Heidelberg University
E-mail: laura.giacomini@iued.uni-heidelberg.de

Abstract

This paper presents a model for the description of terms and term variants in technical e-dictionaries designed for professional translators and technical writers. The paper introduces a concept of variation as a phenomenon affecting (quasi)synonymous terms and terminological word combinations with morphological affinity, and provides an overview of the methodological steps involved in ontological/semantic systematisation, in morphosyntactic analysis of terminological variants and in the following data formalisation. The model is based on a multi-layered formalisation procedure that includes the compilation of a coherent domain ontology, the identification of domain-specific frames and frame elements, and the description of term variants along a previously designed morphology-oriented typology. The paper also discusses visualisation options and search query types in the final e-dictionary. Examples are taken from German and English terminology related to thermal insulation products, with the purpose of hinting at the general applicability of the model to other technical subfields.

Keywords: terminology; ontology; variation; technical domain; LSP; e-dictionary

1. Introduction

This paper presents a description model for terms and term variants in technical e-dictionaries. The study is part of a larger project on database representation of terminological variation, in which restricted technical subdomains, belonging to the areas of building and electrical engineering, have been analysed and compared to test the feasibility of the method. Despite clear differences at the level of conceptualisation, standardisation and communicative features between the two domains, the model has proven to be globally efficient, and seems to provide a reliable method for conceptual and terminological representation in other comparable technical subfields. The employed method and the resulting lexicographic presentation are explained via reference to German and English terms belonging to the field of building thermal insulation. First, ontological data are introduced together with their descriptors (Section 2.1). Second, lexical data (terms and variants) are classified along morphosyntactic rules (Section 2.2). In the next step, the method for merging ontological and morphosyntactic formalisation is discussed. It is also shown how conceptual and lexical formalisation can be embedded in NLP procedures for

extracting candidate terms from LSP corpora (Section 2.3). The concluding part of the paper concentrates on visualisation features of the final lexicographic product (Section 3).

2. Systematic analysis and multi-layered formalisation

2.1 Ontological and semantic data analysis

At the core of conceptual formalisation is a domain ontology and its association with a frame-based approach to obtain fine-grained data descriptions. The domain ontology has been built on the grounds of knowledge that was manually retrieved from several specialised texts dealing with the topic of thermal insulation. These texts, which may address different target recipients, belong to the most typical genres in this field, e.g. specialised magazines, handbooks, product descriptions, data sheets, and guides. Due to the complex structure of the ontology and, in particular, its integration with frame elements, a formalisation of ontological knowledge and the corresponding lexical information by means of widespread RDF models (e.g. lemon-OntoLex) has been avoided, at least for the moment. This type of ontological and semantic data has been recorded in a relational database in the same way as terminological data, rather than in a database-external conceptual layer (as is the case of Ontop and similar OntoLex systems, cf. Bosque-Gil et al., 2015).

The domain ontology is structured around a key entity (or class of entities), the THERMAL INSULATION PRODUCT(S), which constitutes the topical focus of a collection of reference texts[1]. It consists of objects and their taxonomic and non-taxonomic relationships (Declerck & Gromann, 2012). Taxonomies may regard both instances and classes of instances, and produce a controlled vocabulary with a hierarchical structure of the kind parent-child or superclass-subclass. Ontological knowledge representation, however, often requires other types of information to express relations between entities as well as properties of entities. As for first-order entities (Lyons, 1977) in the form of inanimate objects, it is useful to approach ontology work by employing a tripartite supercategorisation as a starting point: thermal insulation products can be observed by taking into consideration aspects regarding their MATERIAL, their FORM and their FUNCTION. Each of these macrocategories includes a number of entities that are sometimes taxonomically related to other entities of the ontology. For instance, a category that is connected to the function of thermal insulation products is the BUILDING COMPONENT to which the product is applied, whereby a specific building component is a kind of superordinate entity belonging to a building component class (e.g. a flat roof is a kind of roof, cf. Table 1).

[1] Texts have been selected on the grounds of their relevance for translation (typically translated texts concerning this topic) and for companies (typically published texts concerning manufacturing, selling and application of a specific product).

BUILDING COMPONENT	CLASS OF BUILDING COMPONENTS
flat roof	roof
mono-pitched roof	roof
multi-pitched roof	roof
exterior wall	wall
interior wall	wall

Table 1**:** Example of an ontological category with taxonomic relationships.

Other entities belong to ontological categories in the form of terms that do not build taxonomic relationships to other relevant categories. For instance, PRODUCT USER and PRODUCT FEATURE, both belonging to the macrocategory FUNCTION, show this kind of behaviour (cf. Table 2).

PRODUCT USER
technician/craftsman
handyman

PRODUCT FEATURE
fire-resistance rating
thermal conductivity
heat storage capacity
bulk density

Table 2**:** Examples of ontological categories without taxonomic relationships.

Ontological categorisation is integrated with more specific semantic information in the form of frame elements in terms of the Frame Semantics theory. The key entity, the THERMAL INSULATION PRODUCT, is seen as part of one of the potential frames involving that entity. Frames are typical situational perspectives, in which specific entities (frame elements) play a role. For instance, thermal insulation products (the concept and the corresponding terms) can be considered from the perspective of their production, their selling, or their use. In the preferred frame, in this case, the insulation product is an object with distinctive features that is sold by producers or traders to specific users in order for them to thermally insulate one or more components/areas of a building. The selected frame serves as an interface between the ontology and the lexicon of the subdomain, and provides a relevant tool for semantic categorisation of terms as well as for lexicographical disambiguation of variants.

Each term or term component directly denoting or indirectly referring to a thermal insulation product can be reduced to a frame containing all or some of its typical frame elements. These elements, e.g. MATERIAL, FORMAT, PART OF THE BUILDING, APPLICATION TECHNIQUE, can be understood as potential semes which coincide with the previously identified ontological entities. The relationship between the ontological, semantic and terminological levels of the proposed model can be visualised as follows:

ontological level:	ontological (macro)category	FORM > FORMAT	FUNCTION > APPL. TECHNIQUE
semantic level:	frame element	FORMAT	APPL. TECHNIQUE
term level:	terms/variants	*insulation boards, insulation batts*	*spray foam insulation, blow-in insulation*

The domain ontology should be a general (if not universal) picture of the objects/concepts that compose the domain, whereas the chosen frame is embedded in the description of a specific situation and, accordingly, can match different constellations of ontological entities. At the terminological level, single terms and multiword terms can be subdivided into semantic components that are directly related to the elements of the relevant frame.

2.2 Terminological data analysis

The study concentrates on terminological variation as a key phenomenon in specialised language, which, in recent decades, has been analysed along different theoretical approaches (cf., among others, Auger, 2001; Freixa, 2006). Our description of terminological variants is based on a concept of variation as a phenomenon affecting (quasi)synonymous terms and terminological word combinations with morphological affinity (i.e. they share at least one lexical morpheme). Texts concerning thermal insulation products, belonging to different textual genres and embedded in various communicative situations, often include more or less large clusters of semantically and morphologically homogeneous terms; for instance, *wood fibre insulation boards, wood fiber insulation boards, wood fibre thermal insulation boards, wood fibre boards for acoustic and thermal insulation, wood fibre boards for external wall insulation, wood fibre boards for insulating walls internally and externally*, etc. The generally low degree of standardisation in the subfield of thermal insulation, in which international and national standards provide guidance and specifications only for a part of the involved entities, is one of the main causes of intensive variation. Variant clusters are apparently relevant in technical writing and specialised translation, but language professionals in these fields are often compelled to perform time-consuming queries in parallel and comparable corpora to obtain information on the availability and correctness of potential variants. Lexicographic information tools such as LSP e-dictionaries, glossaries and termbases, in fact, cover only a small fraction of the commonly used variants. They usually record possible variants at different levels of discourse, for instance geographical variants such as *fibre* (BrE) and *fiber* (AmE) or, in general, variants with no morphological affinity (e.g. German *isolieren/ dämmen*), which, however, are relatively infrequent in specialised language. On the contrary, morphologically similar synonyms at the same level of discourse are not taken into consideration, with the exception of rare cases. Variants extracted from texts are

assigned to classes according to a language-independent variation typology. Each synonymous variant of a term is classified along morphological, syntactic and graphical criteria. Graphical variation regards phenomena such as hyphenation, and plays a minor role in variation analysis, not least because these phenomena are scarcely subject to standardisation[2]. Morphological variation may be total, partial, or entirely missing. Figure 1 shows the three most relevant combinations of variation types, which correspond to the light grey areas.

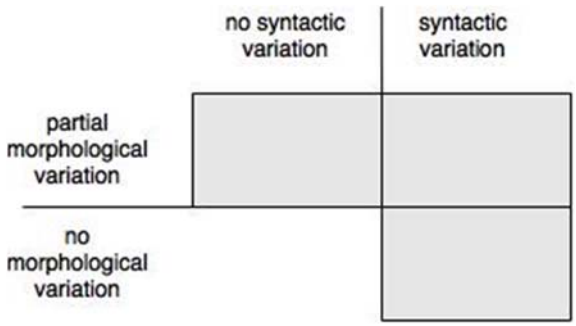


Figure 1**:** Relevant types of term variation

Since the study concentrates on morphologically similar variants, the focus of the study lies on partial morphological variation, independent of its combination with syntactic changes, as well as on syntactic variation without morphological change (light grey areas). Morphological change is missing whenever the variant of a term is made of the same lexical morphemes as the original term.

Variation types indicated in Figure 1 can be illustrated by means of the following examples in German and English[3]:

a) partial morphological variation and no syntactic variation (pMV-nSV)

DE: *Dämmstoff/Isolierstoff*

EN: *thermal insulation/heat insulation*

b) partial morphological variation and syntactic variation (pMV-SV)

DE: *Dämmstoff/wärmeisolierender Stoff*

EN: *polystyrene foam insulation/insulation with styrofoam*

c) no morphological variation and syntactic variation (nMV-SV)

[2] Referring to Figure 1, it can be noted that a case of non-morphological and non-syntactic variation could coincide with mere graphical variation, for instance the absence or presence of hyphenation in the two German terms *Polyurethanschaum*/*Polyurethan-Schaum* (EN *polyurethane foam*).

[3] For simplification reasons, the following abbreviations have been assigned to the relevant variation classes and types: MV = morphological variation; pMV = partial morphological variation; nMV = no morphological variation; SV = syntactic variation; nSV = no syntactic variation.

DE: *Isolierung der Fenster/Fensterisolierung*

EN: *roof insulation/insulation of roofs*

Terms belonging to clusters of morphologically similar synonyms can be analysed on the basis of the presented variation types. Variation can be classified either confronting terms pairwise or, as far as a preferred term can be identified along an existing standard or by conventional use, referring available variants to the preferred term. The following two examples illustrate both approaches to classification:

	pairwise relations:	relations to the preferred term:
(A)		
wood fibre insulation boards (preferred term)		
	} pMV-SV	
wood fiber insulation boards		pMV-nSV
	} pMV-SV	
wood fibre thermal insulation boards		pMV-SV
	} nMV-SV	
wood fibre boards for thermal insulation		pMV-SV
	} pMV-SV+	
wood fibre boards for external wall insulation		pMV-SV+
	} nMV-SV	
wood fibre boards for insulating walls externally		pMV-SV+
(B)		
WDVS-Mineralwolle (preferred term)		
	} nMV-SV	
Mineralwolle als WDVS		nMV-SV
	} pMV-SV	
Mineralwolle zum Dämmen im WDVS		pMV-SV
	} pMV-nSV	
Mineralwolle zum Dämmen im Wärmedämmverbundsystem		pMV-SV
	} nMV-SV	
Mineralwolle… einer der beliebtesten Dämmstoffe für WDVS		pMV-SV

The "+" sign in the first example indicates a semantic expansion: underlined words (cf., for instance, *external wall*) add conceptual information to the contrasted original term, automatically changing both its morphological and semantic nature.

Pairwise classification enables fine-grained interpretation without postulating a hierarchical structure between a preferred term and its variants. In textual analysis, this approach can be useful to follow up variation strategies and motivations inside a specialised text. From a lexicographic perspective, however, classifying variants by comparing them to a preferred term has more advantages, since it may enable a dictionary user to retrieve all variants of a lemma belonging to a specific type with the help of variation-related filters (cf. Section 3).

2.3 A multi-layered formalisation: ontological, semantic and terminological data for the lexicographic database

Data formalisation takes place with the help of ontological, semantic and variational descriptors that are meant to provide lexicographic users with comprehensive information concerning terms and variants of the selected technical domain. Given a term and its synonymous variants, the formalisation process can be illustrated using the examples of *stone wool insulation batts* and *Holzfaserdämmplatten* (*wood fibre insulation boards*), and their synonymous variants (Table 3).

Common terms (i.e. non-proper nouns) indicating a thermal insulation product can be decomposed in semantic unities that refer to specific frame elements and occur in all synonymous variants in order to produce the same term meaning. In the exemplified case, the three frame elements MATERIAL (insulation material of which the product is made), FORMAT (the format in which the product is sold) and GOAL (the purpose with which the product is employed) constitute the semantic profile of the preferred terms *stone wool insulation batts* and *Holzfaserdämmplatten*, and their variants. Variants can combine these frame elements in a different syntactic order:

stone wool insulation batts vs. *insulation batts made of stone wool*

and/or introducing morphological transformations:

stone wool vs. *mineral wool*

insulation vs. *thermal insulation*,

building in this way a large cluster of multiword terms which share the same semantic head, *batts*. When compared with the preferred terms, variants display heterogeneous combinations of morphological and semantic changes.

The lexicographic database, which is structured in a relational form, records in its tables, for each type of thermal insulation product,

- the preferred term indicating this product and

- its terminological variants (synonyms);
- for each variant, the involved variation type;
- the semantic decomposition of the preferred term and its variants[4];
- the frame elements which are realised by the preferred term and its variants.

Frame elements	Terms	Variation	
	[stone wool] [insulation] [batts] (pref. term)	MV	SV
MATERIAL = [stone wool] FORMAT = [batt] GOAL = [insulation]	[mineral wool] [insulation] [batts]	partial	-
	[stone wool] [thermal insulation] [batts]	partial	✓
	[stone wool] [batts] for [thermal insulation]	partial	✓
	[stone wool] [batts] for [insulating] …	-	✓
	[insulation] [batts] made of [stone wool]	-	✓
Frame elements	Terms	Variation	
	[Holzfaser][dämm][platten] (pref. term)	MV	SV
MATERIAL = [Holzfaser] FORMAT = [Platte] GOAL = [Dämmung]	[Holzfaser][platten] zur [Dämmung] …	partial	✓
	[Dämm][platten] aus [Holzfasern]	-	✓
	aus [Holzfasern] hergestellte [Dämm][platten]	-	✓
	[Platten] aus [Holzfasern] zur [Dämmung] …	-	✓

Table 3 – Example of data formalisation by means of frame elements and variation types.

Semantic decomposition, signalled in Table 3 by means of square brackets, is essential for the identification of ontological/semantic differences and similarities between terms that possibly embody other frame element constellations. The frame element FORMAT, for instance, is realised in English by terms such as *slab, board, mattress, rope, foam*, or

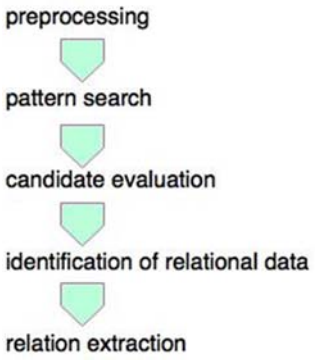
[4] A morphosyntactic decomposition of terms is also provided. However, this topic will not be discussed in this paper.

loose granules, which may belong to larger multiword units together with terms indicating other frame elements. This means that the same frame elements combination can be found in several terms, depending on the language-independent finite number of relevant ontological entities, as well as on the language-specific availability of synonyms (e.g. *panel/ board*), as shown in this example:

	[mineral wool] [mattress]
	[fibreglass wool] [mattress]
	[mineral wool] [batt]
MATERIAL + FORMAT:	[mineral wool] [slab]
	[polystyrene] [panel]
	[polystyrene] [board]
	[polystyrene] [granules]
	[perlite] [granules]

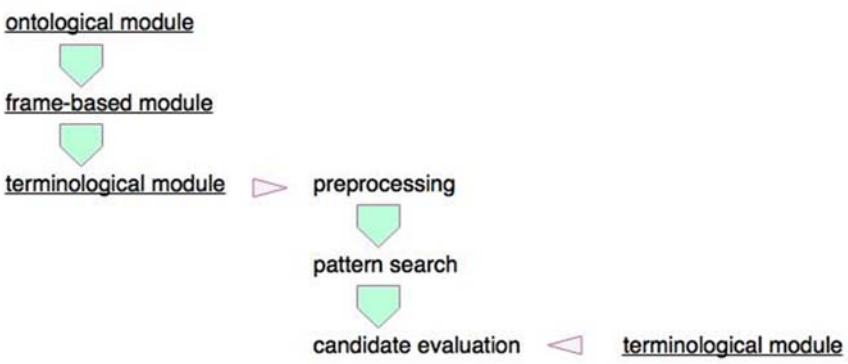
The proposed model of lexical data representation could be combined with NLP techniques to term and relation extraction from LSP corpora to create a semi-automatic pipeline for improving identification of semantically related terms. The formalisation process, as a matter of fact, provides the basis for a consistent rule-based morphosyntactic and semantic analysis, with a direct connection between the two analysis levels. Existing NLP procedures aimed at relation extraction (cf. Rösiger et al., 2016) are based on an inductive method, i.e. on specific instances that lead to generalised statements: relational data obtained by means of corpus preprocessing, pattern search and candidate evaluation are used to extract further relations (a).

(a)



A procedure which integrates the proposed formal model consisting of the ontological, the frame-based and the terminological (morphosyntactic, variational) module allows for the preliminary annotation of terms with new ontological, semantic and morphosyntactic information and thus for a deductive identification of relational data. The new descriptive modules (cf. underlined elements) can interface in different ways with the basic pipeline, depending on the process stage in which their information (sense, morphosyntactic, and variation tags) is most required (b).

(b)



In example (b), terminological formalisation could take place both at a preprocessing and evaluation level, for enhancing both precision and recall of retrieved data. The process could also incorporate bootstrapping techniques in order to iteratively refine extraction results for both terms and relations in a corpus. Extracted candidate terms need to undergo evaluation, preferably in a semi-automatic procedure in which linguistic and conceptual coverage are tested, with the previously identified terminological profiles potentially serving as one of the non-automatic validation tools (for a comparable approach cf. Christensen, 2016).

3. Data presentation in the technical e-dictionary

Terms concerning thermal insulation products are recorded in the technical e-dictionary together with other lexical data (e.g. variants, equivalents in a target language, usage examples) and metadata (e.g. frame elements and relevant entities from the underlying ontology). The main visualisation features in the dictionary will now be shown and discussed (cf. Costa, 2013). These include different presentation modes for conveying user-tailored lexicographic information. Ideally, target users should in fact be able to select specific information, i.e. to group variants along morphosyntactic or conceptual/semantic principles by applying more or less detailed filters.

The addressed user is the technical writer and the professional translator passively translating into the native language. Lexicographically relevant needs arising in specific extralexicographic situations (Tarp, 2008) determine specific dictionary functions to efficiently provide potential users with the required assistance. Technical writers produce functional and user-oriented specialised texts, particularly technical documentation (Göpferich 1998: 1003), whereas professional translators need to produce a native-language target text being tied to a foreign-language source text. Despite this operational and cognitive difference between the two tasks, the presented model aims to serve both target groups by means of a clear text-productive orientation: the main function of the technical e-dictionary is to make variants, and information about variants, available to users in their native language, independent of the qualitative and quantitative features of variation in foreign-language reference

resources used for technical documentation or in a foreign-language source text that has to be translated. On the one hand, dictionary users should be able to perform separate or combined queries involving each data type available in the database. On the other hand, they should be able to customize a search by applying filters to single data types in order to obtain tailored results.

Table 4 displays some of the manifold possibilities of performing search queries by combining different levels of knowledge about variational data. For instance, example (b) given for query type (2) is a combination of query type (2), i.e. the search for a specific variational profile, with query type (1), i.e. the search for a term or part of a term.

Query types, specifically (1)–(3), refer to a terminological layer including terms, variants and their variational and morphosyntactic description. Query types (4)–(5) are related to the ontological/semantic layer of the database, with information concerning frames and the domain ontology. The structure of the database and the multi-layered data formalisation allow for targeted search queries and the combination of query types during a single search act. Users can choose whether to look up a term or to start a query by indicating, for instance, a well-defined set of frame elements, or even if they wish to combine both kinds of information to obtain more fine-grained results (cf. possible query relations in the second column of Table 4).

At the same time, filtering as well as result-widening options in the form of expand/hide commands can be selected during each search query in order to retrieve either more specific or more general results. For instance, the output of the first query example would include by default the term, its classified variants, their morphosyntactic structure, usage examples, as well as the involved frame elements and ontological categories (a). This also constitutes the microstructure of lexicographic entries. However, users can also choose to expand on further results that include additional frame elements (d).

Against the background of the specific user's needs and the relevant microstructural items, it is clear that the technical e-dictionary has both a form-determined and a systematic macrostructure, and that it allows for multiple access paths to the desired data (cf. Giacomini, 2015). Moreover, data representation in the lexicographic database allows for both a monolingual and a bilingual coverage of terms and variants.

Search query type:	Query relations:	Query example (with specific query relation):	Output example:
(1) search for a single or (part of) multiword term		flat roof insulation	(a) preferred term: flat roof insulation variants: flat roof thermal insulation, insulation for flat roofs, insulating flat roofs + *variants types* + *morphosyntactic types* + *information on involved frame elements and ontological entities*
(2) search for a variation profile	and (1)	and (1) nMV-SV	(b) insulation for flat roofs, insulating flat roofs
(3) search for a morphosyntactic structure	and/or (1), and (4)(5)	and (1) NN	(c) preferred term: flat roof insulation variants: flat roof thermal insulation, insulation for flat roofs
(4) search for (a) (combination of) frame element(s)	and/or (1), and (3)(5)	and (1) + MATERIAL + FORMAT	(d) WOOD FIBRE BOARDS for flat roof insulation, EXPANDED POLYSTYRENE SLABS for flat roof insulation
(5) search for an ontological entity or category	and/or (1), and (4)	and (1) MATERIAL	(e) [WOOD FIBRE]/ [POLYSTYRENE]/ [POLYURETHANE] + flat roof insulation

Table 4 – Search query types and visualisation options in the technical e-dictionary.

4. Conclusions

This paper has introduced a description model for technical terms and their variants in an e-dictionary designed for professional translators and technical writers, and covering terminology related to thermal insulation products. The aim of the paper was, on the one hand, to provide an overview of the methodological steps involved in ontological/semantic systematisation, in morphosyntactic analysis of terminological variants and in the following data formalisation. On the other hand, a major goal of this paper was to discuss visualisation options in the final e-dictionary, and to associate them with its overall microstructural, macrostructural and access properties.

As already mentioned in the introduction, results presented in this paper, as well as those obtained in the underlying project concerning this topic, confirm the effectiveness of the method in:

- creating multi-layered, language-independent descriptions for synonymous variation in restricted technical subdomains;

- adapting this description to lexicographic functions of resources that address specific target users; and

- providing formalisation tools to possibly improve NLP procedures for term and variant extraction from specialised corpora.

In the current project, synonymous variation and its morphological peculiarities are at the centre of discussion as one of the most pervasive and, at the same time, underestimated lexical phenomena in terminology. Its relevance for LSP dictionaries addressing professional text producers is indisputable. Special attention is due in the field of electronic lexicography, which can provide the necessary tools (e.g. data formalisation, or database representation strategies) to ensure extensive, modular coverage of the phenomenon, and which can benefit from the availability of data obtained by semi-automatic term and variant extraction. Future work conducted on the language(s) of technology will further investigate these topics and attempt to expand the method of technical (and maybe non-technical) subdomains displaying even larger differences in conceptualisation, standardisation and communicative features.

5. References

Auger, P. (2001). Essai d'élaboration d'un modèle terminologique/terminographique variationniste. In *TradTerm*, 7, 183-224.

Bosque-Gil, J. et al. (2015). Applying the OntoLex model to a multilingual terminological resource. In *European Semantic Web Conference*. Springer International Publishing, 283-294.

Christensen, L. W. (2016). Semi-automatic Evaluation of Terminological Web-crawled Corpora. *Term Bases and Linguistic Linked Open Data*, 64.

Costa, R. (2013). Terminology and Specialised Lexicography: two complementary domains. In *Lexicographica* 29.1/2013, 29-42.

Declerck, T. & Gromann, D. (2012). Combining three ways of conveying knowledge: Modularization of domain, terminological, and linguistic knowledge in ontologies. In *Proceedings of the 6th International Workshop on Modular Ontologies, Graz, Austria, CEUR-WS, Aachen. CEUR Workshop Proceedings* Vol. 875, 28-40.

Fillmore, C. J. (1977). Scenes-and-frames semantics. In A. Zampolli (ed.) *Linguistic Structures Processing*. Amsterdam: North-Holland Publishing Company, 55-81.

Freixa, J. (2006). Causes of denominative variation in terminology: A typology proposal. In *Terminology* 12(1). Amsterdam: John Benjamins, 51-77.

Giacomini, L. (2015). Macrostructural properties and access structures of LSP e-dictionaries for translation. In *Lexicographica* 2015/31. Berlin/Boston: De Gruyter. 90-117.

Göpferich, S. (1998). Schreiben in der Technik/Technical Writing. Fachsprachen: Ein internationales Handbuch zur Fachsprachenforschung und Terminologiewissenschaft. New York, Berlin: de Gruyter, 1003-1014.

Lyons, J. (1977). *Semantics* (vols. i & ii). Cambridge CUP.

Rösiger, I. et al. (2016). Acquisition of semantic relations between terms: how far can we get with standard NLP tools? In *Computerm 2016*, 41.

Tarp, S. (2008). Lexicography in the Borderland between Knowledge and Non-Knowledge. Tübingen: Max Niemeyer Verlag.

The Translation Equivalents Database (Treq) as a Lexicographer's Aid

Michal Škrabal, Martin Vavřín

Institute of the Czech National Corpus, Charles University, Czech Republic
E-mail: michal.skrabal@ff.cuni.cz, martin.vavrin@ff.cuni.cz

Abstract

The aim of this paper is to introduce a tool that has recently been developed at the Institute of the Czech National Corpus, the Treq (**Tr**anslation **Eq**uivalents) database, and to explore its possible uses, especially in the field of lexicography. Equivalent candidates offered by Treq can also be considered as potential equivalents in a bilingual dictionary (we will focus on the Latvian–Czech combination in this paper). Lexicographers instantly receive a list of candidates for target language counterparts and their frequencies (expressed both in absolute numbers and percentages) that suggest the probability that a given candidate is functionally equivalent. A significant advantage is the possibility to click on any one of these candidates and immediately verify their individual occurrences in a given context; and thus more easily distinguish the relevant translation candidates from the misleading ones. This utility, which is based on data stored in the InterCorp parallel corpus, is continually being upgraded and enriched with new functions (the recent integration of multi-word units, adding English as the primary language of the dictionaries, an improved interface, etc.), and the accuracy of the results is growing as the volume of data keeps increasing.

Keywords: InterCorp; Treq; translation equivalents; alignment; Latvian–Czech dictionary

1. Introduction

The aim of this paper is to introduce one of the tools that has been developed recently at the Institute of the Czech National Corpus (ICNC) and which could be especially helpful to lexicographers: namely, the Treq translation equivalents database[1]. It is based on data stored in the InterCorp parallel corpus (always its latest version, currently v9).

2. InterCorp

InterCorp is a large parallel synchronic corpus under continuous construction at the ICNC since 2005. The corpus has been growing systematically every year in the recent past and, since 2013 (version 6), even obsolete versions of the corpus will remain available via our corpus query interface, KonText, in order to preserve the possibility of replicating previous research. InterCorp is composed of several parts, the most

[1] Available online at http://treq.korpus.cz/.

important and valuable of which is arguably the so-called *core*—literary texts with manually corrected OCR and sentence alignment. In addition to the core, there are several *collections*, consisting of texts which were only processed automatically[2], not manually. These include the following types of texts:

- journalistic articles and news published by Project Syndicate and VoxEurop (formerly PressEurop);
- legal texts of the European Union from the Acquis Communautaire corpus;
- proceedings of the European Parliament dated 2007–2011 from the Europarl corpus;
- film subtitles from the Open Subtitles database.

InterCorp v9 contains, besides Czech as the pivot language (for every text in InterCorp, there *must* be a single Czech version, either the original or a translation), another 39 languages that are, however, unevenly represented. You can therefore find languages which have up to 31 million running words in the core (German) and corpora of individual languages can range in size up to 120 million running words (English), but there are also corpora which have no text in the core (i.e., no manually processed texts) and restrict themselves to collections only (e.g., Vietnamese with a total size of nearly 1.5 million words, consisting only of film subtitles, etc.). Texts in more than half of the languages are provided with morphological annotation (23 out of 39) and lemmatized (20 out of 39). The total size of InterCorp v9 is more than 1.2 billion running words / 1.5 billion tokens[3].

3. Data preparation[4]

First, when preparing data for Treq, only sentences that are aligned[5] 1:1 are selected from the entire InterCorp corpus. We restrict ourselves to this simple alignment because it tends to be more reliable; especially in the case of automatically aligned

[2] For the list of used tools, see http://wiki.korpus.cz/doku.php/en:cnk:intercorp:verze9#acknowledgements.

[3] For information about the exact composition of the corpus and the size of its components, see http://wiki.korpus.cz/doku.php/en:cnk:intercorp. For general information about the InterCorp project, see Čermák & Rosen (2012) or Rosen (2016).

[4] Cf., e.g. the process of compiling "statistical translation dictionaries" described in Kovář et al. (2016: 343n).

[5] The core component of InterCorp is aligned with the InterText tool (Vondřička, 2014) and this alignment is subsequently manually checked and corrected, mostly in three stages (for details, see Rosen & Vavřín, 2012: 2448). Collections are aligned only by the Hunalign aligner (Varga et al., 2015; see also http://mokk.bme.hu/en/resources/hunalign/), with no correction following. Basic assessment of the quality of our automatic segmentation and alignment can be found in Rosen & Vavřín (2012: 2450).

texts, potential errors can be prevented[6].

The next step is to perform an automatic word-to-word alignment using the GIZA++ tool (Och & Ney, 2003)[7]. In older versions of Treq, a method called *intersection* was used, creating only such alignments where one word in the source language corresponds to one word in the target language, e.g.:

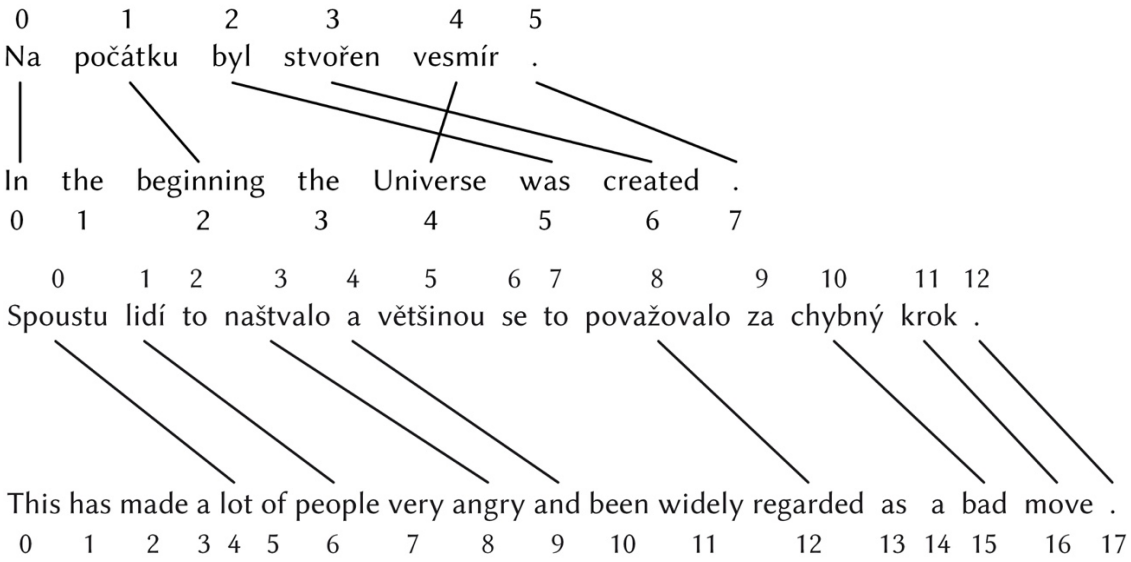


Figure 1: Aligning words using the *intersection* method

That is, the first word in the source language (0) corresponds to the first word in the target language (0), the second word (1) corresponds to the third one (2), etc. (cf. Rosen, Adamová & Vavřín, 2014; Kaczmarska & Rosen, 2015: 164–165).

Starting with release 2.0, apart from this simple alignment method, the so-called *grow-diag-final-and* method has also been used, as it allows the creation of more complicated alignments containing more than one word on both sides of the translation[8]. These multi-word units are not necessarily well-defined entities from a linguistic point of view: some may correspond to what a linguist would analyse as multi-word expressions, some may not.

[6] In the future, however, we would like to experiment also with a non-1:1 alignment (cf. Kovář et al., 2016: 350–351). Other possible plans are outlined in the conclusion of this paper.

[7] For details about our setup, see https://github.com/moses-smt/mgiza/tree/master/mgizapp. An auxiliary script created by Ondřej Bojar (http://www1.cuni.cz/~obo/) was also used.

[8] Individual GIZA++ word alignment methods are described and compared by, e.g. Mareček (2009) or Girgzdis et al. (2014). In both papers, the *grow-diag-final-and* method has been evaluated as the most precise and efficient one, therefore it has been adopted also for our purpose.

Such an alignment may look like this:

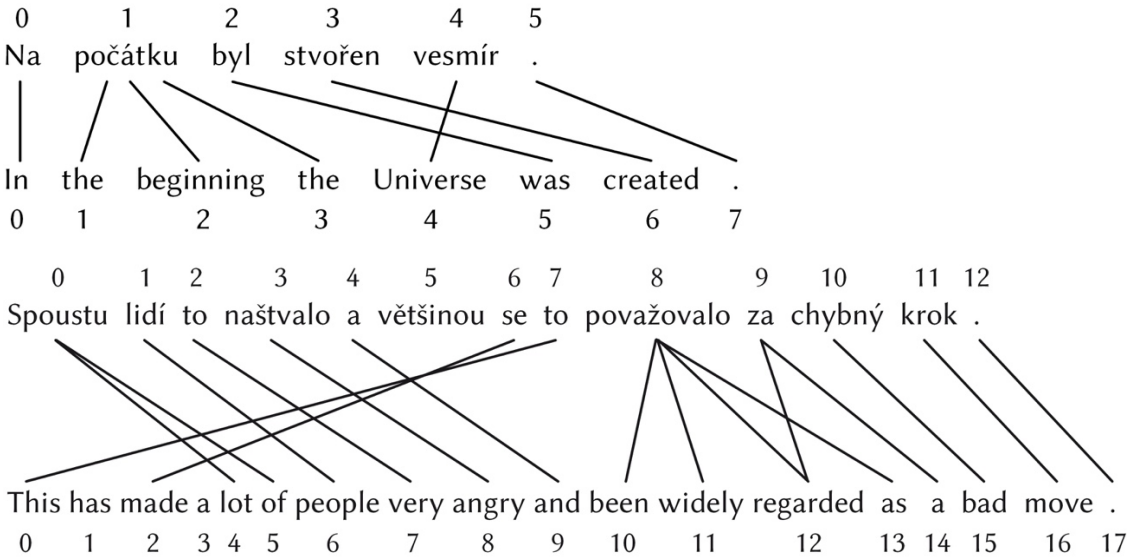


Figure 2: Aligning words using the *grow-diag-final-and* method

(Note the difference: the second word in the target language (0) now corresponds not only to the third (2), but also the second and fourth (1, 3) word in the target language.)

From such an alignment, we choose—using a simple script—the largest possible number of combinations of words that this alignment allows. In both cases, the aligned pairs of words are then sorted and summarized. The result of this automatic excerption is not revised in any way and is provided to users as a list of found equivalents of the given expression, supplemented with absolute and relative frequencies of aligned pairs.

Table 1 indicates in what proportion the frequencies found in KonText are similar to those displayed by Treq. It also specifies the different data types at each stage of their processing for Treq, considering the InterCorp v9 English component (multi-word variant).

Step by step, you can see the gradual loss of data that are used in the resulting dictionary. In the first step, we only use a 1:1 sentence alignment; thus 20.7% of sentences are lost. Subsequently, both one- and multi-word equivalents are selected based on an alignment made by the GIZA++ tool. However, the relationship between the size of the original corpus and the number of extracted equivalents cannot be clearly predicted, especially in multi-word equivalents where various combinations of the same words arise (see bold pairs below). For example, an alphabetical list of Czech–English couples extracted from the second example sentence above would look like this:

a – and

chybný – bad
krok – move
lidí – people
naštvalo – angry
považovalo – been widely regarded as
považovalo za – been widely regarded as
považovalo za – regarded as
se – made
Spoustu – lot of
to – this
to – very
za – regarded as
. – .

Processing phase	Output data		Count (in thousands)						
			Core	**Sub.**	**Acq.**	**Eu.**	**Vox.**	**Synd.**	**Total**
0. Input	Tokens (in English)		25 149	66 790	29 626	17 384	3 123	4 387	146 458
	Sentences (in English)		1 510	9 211	1 426	681	152	190	13 171
1. Sentence alignment (1:1)	Aligned sentences	lemmas	1 267	6 955	1 251	656	127	180	10 437
		word forms	1 267	6 955	1 254	656	127	180	10 440
2. Word alignment	Equivalents identified	lemmas	15 785	41 189	19 344	12 812	1 670	3 352	94 153
		word forms	15 538	41 445	19 656	12 899	1 598	3 344	94 479
3. Dictionary compilation	Dictionary entries	lemmas	3 235	6 697	1 441	1 213	547	550	13 682
		word forms	4 639	9 276	2 056	1 946	670	873	19 460
4. Dictionary cleanup	Dictionary entries	lemmas	2 775	5 375	1 133	1 061	461	458	11 263
		word forms	3 966	7 146	1 722	1 760	566	750	15 909

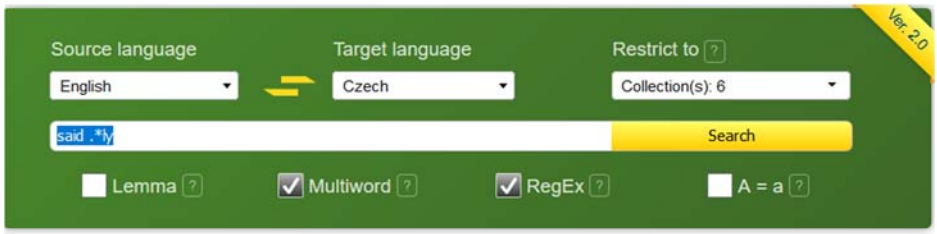
Table 1: Data processing for a Czech-English dictionary (Sub.=Subtitles, Acq.=Acquis Communautaire, Eu.=Europarl, Vox.=VoxEurop, Synd.=Project Syndicate)

In the third step, lines that are the same on both sides of the alignment are added together throughout the text. This will give us the list and the frequency of the equivalents. Finally, in the last step, we exclude all the counterparts containing the punctuation to get the final version of the dictionary. For all language pairs where the

lemmatization is available on both sides of the alignment, we apply the same procedure to the lemmatized form of data (*na počátek být stvořit vesmír – in the beginning the universe be create*).

4. Interface

Access to the extracted data is then mediated by the Treq online search interface.



Frequency	Proportion	English	Czech
13	1.5	said quickly	vyhrkl
8	0.9	said quietly	hlesl
7	0.8	said defensively	bránil
7	0.8	said angrily	rozzlobil
7	0.8	said hoarsely	zachraplal
7	0.8	said angrily	rozzlobil se
6	0.7	said defensively	bránil se
6	0.7	said quickly	vyhrkla
5	0.6	said previously	uvedl
4	0.5	said loudly	nahlas
4	0.5	said indignantly	rozhořčil
4	0.5	said hastily	pospíšil
4	0.5	said angrily	zlobil
4	0.5	said shortly	odsekl
4	0.5	said indignantly	rozhořčil se
4	0.5	said hastily	pospíšil si
4	0.5	said angrily	zlobil se

Figure 3: Advanced searching (via RegEx and multi-word units) in the English–Czech section[9]

By default, found counterparts of the searched expression are ranked in descending order of frequency of these equivalent pairs. Their relative frequency is the user's primary guide: the more often the equivalent of the search term occurred compared to other equivalents, the higher the probability that it is plausible. For large-sized and genre-varied corpora, it is advisable to indicate the frequency of equivalent pairs separately for distinct types of texts (see above Section 1) via the *Restrict to* option.

Starting with version 2, it became possible to enter multi-word expressions into the query window (in both directions, of course), yielding both one- and multi-word units as results, in compliance with user preferences. With non-1:1 word alignments, it is

[9] We have adopted this example from Dr. Lenka Fárová (unpublished presentation). It does a good job of showing a non-symmetric nature of equivalent reporting verbs in English and Czech.

now possible, e.g., in the English–Czech language combination, to search for phrasal verbs, discourse markers, phrases in a general sense, etc. (in the direction from English to Czech); and, in the opposite direction, e.g., reflexive verbs (which are formed in Czech using a separate reflexive morpheme, *se/si*). Moreover, current results more faithfully correspond to the language reality as the equivalence between lexemes in the source and target language cannot, understandably, be limited to an "ideal" 1:1 ratio.

With the implementation of multi-word units, the need to incorporate a query language that would allow the use of wild cards has become urgent[10]: up to now, Treq has only been searching for an exact string of characters. Furthermore, a second primary language (besides Czech), namely English, has been added. And, in addition to the existing bidirectional Czech-X lexicons, bidirectional English-X lexicons have also been generated from the InterCorp data. Thus, the possibility of using Treq is opened up to a much wider audience now as users are no longer limited by the need to master Czech. Theoretically, in the future, the primary language can be extended to any one represented in InterCorp; in this respect, it is necessary to take into account the interests and needs of users.

5. The possible use of Treq in lexicography

(Latvian–Czech dictionary case)

Treq is a relatively new application (its initial version, 0.1 alpha, was released in September 2014[11]), but it is quickly gaining popularity among users, especially for its simplicity and straightforwardness[12]. Possible uses of Treq range from simple, one-shot probes while searching for an equivalent expression for a target language, to more sophisticated and elaborate corpus-assisted translations (Škrabal & Vavřín, 2017: 251–257). However, the equivalents offered by Treq can also be considered as potential dictionary equivalents. This is a handy tool for lexicographers as they instantly get a list of candidates for target language counterparts along with their frequencies (expressed both in absolute numbers and percentages), which suggests the probability that a given candidate is functionally equivalent. A significant advantage is the possibility to click on any of them to immediately verify its individual occurrences in the context, and thus more easily distinguish relevant translation candidates from misleading ones.

---

[10] Treq is based on the database system MySQL, which uses Henry Spencer's regular expression library compliant to the POSIX.2 standard (see e.g., https://garyhouston.github.io/regex/).

[11] Detailed information about individual versions can be found in the Version Info at: https://treq.korpus.cz.

[12] During 2016, over 719 thousand user interactions were registered at the www.korpus.cz portal. The tool used most often was KonText (with more than 85% of the total), followed by the Treq database (more than 70,000 queries, i.e., almost 200 per day, which represents close to 10% of the total number of queries entered).

The extraction of data from parallel corpora for lexicographical purposes is a logical process that is inherent in the very nature of these data. Partial attempts in this regard have also been undertaken in Czech lexicography, e.g., in the case of English (e.g., Čmejrek, 1998; Čmejrek & Cuřín, 2001; Popelka, 2011), Croatian (Jirásek, 2011), or Lithuanian (Skoumalová, 2008). These authors agree that dictionaries automatically extracted from a parallel corpus are merely the starting point for subsequent lexicographical work; nevertheless, they can relieve much of the burden placed on the lexicographer. This is also confirmed by our own experience as Treq is being used—*inter alia*—for the construction of a Latvian–Czech dictionary (Škrabal, 2016a). It is obvious that the extent to which the retrieved data can be utilised in this way depends primarily on the amount of data for the respective language combination[13].

Currently, the Latvian component of InterCorp (release 9) has a total of over 40.6 million words: the initial manually aligned belletristic core (currently 1,666,000 words) was, as for many other languages in InterCorp, extended by a collection of automatically aligned texts from the Acquis Communautaire corpus (24,667,000 words), Europarl corpus (13,895,000 words) and the OpenSubtitles database (381,000 words).

Let us compare these figures to the situation in the early phase of compiling the Latvian–Czech Dictionary, namely to InterCorp version 3.1 (released in May 2011). The Latvian–Czech component then consisted of parallel fiction texts only (20 in the Czech original, 7 in the Latvian original, and 6 in other languages), numbering slightly more than 1 million running words which were neither lemmatized nor tagged. These data were tentatively processed by the NATools workbench[14] (cf. Skoumalová, 2008) and a simple dictionary (or rather glossary) was compiled. We will inspect the lemma *biedrs* (for individual senses, see below)[15].

biedra [Gen.sg.] (13): 0, ***kamarádův***, *všudy*, *uvěřitelný*, ***kamarád***, *oddělení*, *rozchod*

biedram [Dat.sg.] (16): ***kamarád***, ***soudruh***, *čerstvý*, *budižkničemu*, *trmácet*

[13] Cf. Jirásek's (2011: 55) experience from the Croatian–Czech part of InterCorp: "It turned out that if we do not want to stay at the level of pocket dictionaries, we need a parallel corpus of at least 10 million running words. Such a size of corpus allows us to reliably process equivalents for a medium-sized dictionary. For a larger dictionary, however, it can only serve as an orientation aid, not the main source of equivalents." By a medium-sized vocabulary, is meant one containing approximately 20,000 headwords, representing only "typical and predominant meanings in everyday communication". The larger-sized dictionary should contain about 50 thousand headwords (ibid.: 45).

[14] http://linguateca.di.uminho.pt/natools/

[15] Individual grammatical forms are given with their absolute frequencies in the then corpus, followed by Czech equivalent candidates (as lemmas) ordered by plausibility, as estimated by the frequency of aligned pairs. The plausible candidates for dictionary equivalents are in bold, those with limited application (in collocations mostly) are marked by an asterisk (*), and 0 indicates null equivalents.

biedri [Nom.pl./Voc.sg.] (56): ***soudruh***, ***kamarád***

biedriem [Dat./Ins.pl.] (16): ***kamarád***, ***druh***, *trhnout*, ***spolubojovník****, *přeletět*

biedrs [Nom.sg.] (52): ***soudruh***, ***kamarád***, ***člen***, ***společník****

biedru [Acc./Instr.sg./Gen.pl.] (50): ***člen***, ***kamarád***, 0, ***druh***, ***soudruh***

biedrus [Acc.pl.] (13): ***soudruh***, ***kamarád***, *na*, *povzbuzovat*, *brabec*, 0

Nowadays, thanks to the Treq tool, leveraging InterCorp data is as simple for the lexicographer as entering the lemma *biedrs* into the query window, and results can be seen immediately.

člen (483 out of 755, i.e., 64%), ***soudruh*** (81), ***kamarád*** (49), *nečlen* (19), *producent* (16), ***kolega*** (12), ***druh*** (11), ***spolužák**** (10), ***přítel*** (7), *poslanec* (7), ***partner**** (7), ***společník**** (6), *členství* (5), ***spolubojovník**** (4), …

It should be noted that these results are useful only in the advanced phase of the lexicographic work on the relevant headword, preceded by an analysis of the corpus data[16] and, in the case of polysemous headwords, drafting the initial sketch of the sense structure. This can often differ from the existing lexicographical description, especially if it is not corpus-based, which is also the case of the chosen lexeme *biedrs*. Thus, the sense division in the newest Latvian monolingual dictionary (MLVV): 1. 'fellow, friend, colleague'; 2. 'member'; 3. 'comrade' had to be rejected for our purpose. On the basis of a manual analysis of corpus data (776 occurrences of the lemma in LVK2013), an overlooked sense[17] (yet, incidentally more frequent than the third one, historically-marked) was discovered; the rank of the first two senses was adjusted by frequency as well into this resulting semantic framework (cf. also Škrabal, 2016b):

1. 'member' (497 hits in LVK2013, i.e., 64%); 2. 'fellow, friend, colleague' (204 hits, i.e., 26%); 3. 'deputy' (50 hits, i.e., 6%); 4. 'comrade' (25 hits, i.e., 3%).

Only on the background of such a semantic skeleton did we examine the offered translation candidates in terms of the adequacy of the expression in the source language, i.e., we compared the contexts in which the expressions occur in both

[16] A list of corpora used during the work on the Latvian–Czech dictionary includes, besides InterCorp, also the following three:

- representative Latvian corpus *Līdzsvarots mūsdienu latviešu valodas tekstu korpuss 2013* (LVK2013, 5.5 million tokens, lemmatized, tagged)
- *Latviešu valodas tīmekļa korpuss* (LVTK) compiled from Latvian web pages (over 122 million tokens, non-lemmatized, only partially tagged)
- *lvTenTen corpus* as a member of the TenTen corpora family (Jakubíček et al., 2013) accessible via Sketch Engine (658 million tokens, lemmatized, tagged).

[17] This sense is not a new one, just an updated one from the inter-war period.

languages (via the KonText interface).

By simply modifying the query above into *.*biedrs* (and ticking the *RegEx* option) we will get a considerable amount of compounds with the lemma as its component. These can serve in two ways: either as candidates for separate headwords (e.g., *ceļabiedrs*) or, if written separately (as often happens in Latvian, e.g., *ceļa biedrs*), as potential collocations under the respective headword. Regular expressions can thus provide the lexicographer with possible translation equivalents not only for a single word, but even for a word list.

ceļabiedrs ['fellow traveller']: ***spolucestující*** (3), ***společník*** (3), *naštvaný* (1), *průvodní* (1), *spolubojovník* (1), *sužovat* (1), *ušetřený* (1)

cīņasbiedrs ['comrade-in-arms']: ***spolubojovník*** (1)

darbabiedrs ['colleague, co-worker']: ***kolega*** (7), ***spolupracovník*** (5)

domubiedrs ['person who holds the same views']: *podobný* (1), *rodina* (1), ***spojenec*** (1)

dzīvesbiedrs ['spouse, mate']: *manželka* (50), ***manžel*** (43), ***partner*** (5), *manželský* (2), *držitelův* (1), *Lullingové* (1), *pára* (1), *tabule* (1)

galdabiedrs [lit. 'table-mate']: *bodávat* (1), ***kumpán*** (1), *stolní* (1)

karabiedrs ['comrade-in-arms']: ***spolubojovník*** (1), *válečný* (1), *zlíbit* (1)

klasesbiedrs ['classmate']: ***spolužák*** (41)

laikabiedrs ['contemporary']: ***současník*** (6), ***pamětník*** (2), ***vrstevník*** (2), *doba* (1), *Gruzie* (1), *spoluobčan* (1), *vyprávění* (1)

līdzbiedrs [lit. 'co-mate']: *learning* (3), *bližní* (1), ***spolupracovník**** (1), ***vrstevník**** (1), *záhada* (1)

rotaļbiedrs [lit. 'toy-mate']: ***kamarád**** (1)

skolasbiedrs ['schoolmate']: ***spolužák*** (37), *kamarád* (1), *spolužákův* (1), *včerejší* (1)[18]

[18] There were only the following compounds with their translation candidates in the data extracted by the NATools workbench:

darbabiedrs: ***kolega***, *se*, ***spolupracovník***, *známý*

karabiedrs: *vyzvědět*

klasesbiedrs: ***spolužák***, *muset*, *zařídit*

laikabiedrs: *můj*, ***pamětník***, *hodně*, *doba*, *nic*, *místo*, *většina*, *průběh*, ***současník***, *Haškův*

skolasbiedrs: *spolužákův*, *recese*, ***spolužák***, 0, *leccos*, *sejít*, *vůbec*, *kamarád*

Finally, after ticking the box *Multiword*, we can extend our list of multi-word expressions and their counterparts with these relevant non-1:1 pairs.

dzīvesbiedrs: *manžel nebo manželka* (22)

arodbiedrības biedrs: *odborář* (16)

konservatīvās partijas biedrs: *konzervativec* (5)

sarunu biedrs: *společník* (5), *protějšek* (3), *partner* (3)

This probe, as well as others carried out while testing the new Treq version, illustratively indicates that, despite its non-representative nature, size, and composition of texts[19], the Latvian–Czech component of the parallel corpus InterCorp, or its extension Treq, respectively, is a valuable source among the sources used to compile a Latvian–Czech dictionary.[20] This is because it is the only one that directly offers Czech equivalents of Latvian lexemes to such an extent. Unlike other similar projects[21] based on parallel corpus data, InterCorp contains a considerable share of original and translated fiction which has been manually checked and therefore provides more precise results. Another advantage, compared to other tools, is Treq's speed, user-friendliness and direct access to parallel concordances via the KonText interface (with its advanced functionality).

Bilingual word sketches (Kovář et al., 2016) are another tool which could be of significant help in the future, along with the Translate button tool (Baisa et al., 2014); but unfortunately, they are not available now for this language combination.

6. Outlook

Further improvements in the results of Treq yields can be expected along with the increasing volume of data and genre variety of the texts used and a gradual improvement in automatic word-aligning tools. At the moment, InterCorp is the largest parallel corpus available for many Czech-X language combinations, including

[19] More precisely: the minimum proportion of Latvian originals that would be ideal for our purposes. In the belletristic core, there are only four novels, one memoir, one book of fairy tales and one shorter essay with the source language being Latvian, while in the Europarl collection there are 268 transcripts from 16 different authors. The total size of such a subcorpus is 387,544 tokens (incl. punctuation), i.e., less than 1% of the total volume of data in the Latvian part of the InterCorp (in the core, the ratio of Latvian originals is about 20%).

[20] Cf. Nikuļceva's (2006) situation when she was writing her Czech–Latvian dictionary a decade and a half ago: there was no Czech–Latvian parallel corpus at all, not to say a Treq-like tool, just a synchronic corpus of Czech SYN2000 (100 million tokens).

[21] Including, e.g., *Opus* (http://opus.lingfil.uu.se/), *Glosbe* (https://glosbe.com/), *Linguee* (www.linguee.com/), *Europarl* (http://www.statmt.org/europarl/) etc., or a parallel corpus of fiction texts in Slavic and other languages *ParaSol* (http://parasolcorpus.org/).

Czech–Latvian[22]. Generally, this relates to a greater effort in the building of parallel corpora in comparison to monolingual ones.

From the example of the polysemous lexeme above, it is apparent that Treq only offers potential translation equivalents, performing no word sense disambiguation. Therefore, it would be desirable to try to align words while paying attention to morphosyntactic and/or syntactic-semantic categories. We would also like to explore other options of aligning multi-word units, e.g., to start by searching the text for multi-word units using specialized tools and then seek alignment for individual words already within the identified multi-word units.

7. Acknowledgements

This paper resulted from the implementation of the Czech National Corpus project (LM2015044) funded by the Ministry of Education, Youth and Sports of the Czech Republic within the framework of Large Research, Development and Innovation Infrastructures. We would like to thank our colleagues, Michal Křen and Alexander Rosen, for the valuable comments on the article, the latter also for his contribution to the design and implementation of the entire Treq project. In addition, we owe our gratitude to Elżbieta Kaczmarska for giving an impetus to the development of this instrument, to Ondřej Bojar and David Mareček for technical assistance and to Jan Kocek for helping us with the graphic processing of data. Last but not least, we also wish to thank two anonymous reviewers for their thought-provoking comments.

8. References

Baisa, V., Jakubíček, M., Kilgarriff, A., Kovář, V. & Rychlý, P. (2014). *Bilingual word sketches: the translate button.* In A. Abel, C. Vettori & N. Ralli (eds.) *Proceedings of the XVI EURALEX International Congress: The User in Focus.* Bolzano: Institute for Specialised Communication and Multilingualism, pp. 505–513.

Čermák, F. & Rosen, A. (2012). The case of InterCorp, a multilingual parallel corpus, *International Journal of Corpus Linguistics* 13, 3, pp. 411–427.

Čmejrek, M. (1998). *Automatická extrakce dvojjazyčného pravděpodobnostního slovníku z paralelních textů.* Master's thesis. Praha: Ústav formální a aplikované lingvistiky MFF UK.

Čmejrek, M. & Cuřín, J. (2001). Automatic Extraction of Terminological Lexicon from Czech-English Parallel Texts. In *International Journal of Corpus Linguistics Special Issue 2001*, pp. 1–12.

Girgzdis, V., Kale, M., Vaicekauskis, M., Zarina, I. & Skadiņa, I. (2014). Tracing

[22] At least in the traditional sense, as opposed to web-crawled corpora or easily accessible collections of parallel texts online, cf. also Latvian parallel corpora offered via Sketch Engine, including corpus EUR-Lex Latvian 2/2016 with over than 491 million tokens (not tagged yet).

Mistakes and Finding Gaps in Automatic Word Alignments for Latvian-English Translation. In A. Utka et al. (eds.) *Human Language Technologies – The Baltic Perspective. Proceedings of the Sixth International Conference Baltic HLT 2014.* Amsterdam: IOV Press BV, pp. 87–94.

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P. & Suchomel, V. (2013). The TenTen Corpus Family. In *7th International Corpus Linguistics Conference CL 2013*. Lancaster: UCREL, pp. 125–127.

Jirásek, K. (2011). Využití paralelního korpusu InterCorp k získávání ekvivalentů pro chorvatsko-český slovník. In F. Čermák (ed.) *Korpusová lingvistika Praha 2011: 1 – InterCorp.* Praha: NLN, pp. 45–55.

Kaczmarska, E. & Rosen, A. (2015). Jak najít optimální překlad polysémních jednotek – porovnání metod formální analýzy paralelních textů. *Časopis pro moderní filologii* 97, 2, pp. 157–168.

Kovář, V., Baisa, V. & Jakubíček, M. (2016). Sketch Engine for Bilingual Lexicography. *International Journal of Lexicography* 29, 3, pp. 339–352.

Mareček, D. (2009). Using tectogrammatical alignment in phrase-based machine translation. In J. Šafránková & J. Pavlů (eds.) *WDS 2009 Proceedings of Contributed Papers*. Praha: Matfyzpress, pp. 22–27.

MLVV = *Mūsdienu latviešu valodas vārdnīca* [online]. Accessed at: http://www.tezaurs.lv/mlvv/. (23 May 2017)

Nikuļceva, S. (2006). *Česko-lotyšský slovník. Čehu-latviešu vārdnīca.* Praha: Leda.

Och, F. J. & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29, 1, pp. 19–51.

Popelka, J. (2011). *Automatické vytváření slovníků z paralelních korpusů*. Master's thesis. Praha: Ústav formální a aplikované lingvistiky MFF UK.

Rosen, A. (2016). InterCorp – a look behind the façade of a parallel corpus. In E. Gruszczyńska & A. Leńko-Szymańska (eds.) *Polskojęzyczne korpusy równoległe. Polish-language Parallel Corpora*. Warszawa: Instytut Lingwistyki Stosowanej, pp. 21–40.

Rosen, A., Adamová, M. & Vavřín, M. (2014). Extrakce lexikálních ekvivalentů z paralelního korpusu. In *Korpusová lingvistika Praha 2014. 20 let mapování češtiny. Abstrakty.* Praha: Ústav Českého národního korpusu, pp. 177–179.

Rosen, A. & Vavřín, M. (2012). Building a multilingual parallel corpus for human users. In N. Calzolari et al. (eds.) *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12).* Istanbul: European Language Resources Association (ELRA), pp. 2447–2452.

Skoumalová, H. (2008). Extracting dictionaries from parallel corpora. In F. Čermák et al. (eds.) *Proceedings of The Third Baltic Conference on Human Language Technologies.* Kaunas: Vytautas Magnus University, pp. 297–301.

Škrabal, M. (2016a). *Srovnávací aspekty lotyšského a českého lexikonu: Materiály k sestavení lotyšsko-českého slovníku.* PhD. thesis. Praha: Ústav Českého národního korpusu FF UK.

Škrabal, M. (2016b). Straddling the boundaries of traditional and corpus-based

lexicography: A Latvian-Czech dictionary. In T. Margalitadze & G. Meladze (eds.) *Proceedings of the XVII EURALEX International Congress: Lexicography and Linguistic Diversity*. Tbilisi: Ivane Javakhishvili Tbilisi State University, pp. 910–914.

Škrabal, M. & Vavřín, M. (2017). Databáze překladových ekvivalentů Treq. *Časopis pro moderní filologii* 99, 2, pp. 245–260.

Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V. & Nagy, V. (2005). Parallel corpora for medium density languages. In G. Angelova et al. (eds.) *Proceedings of the RANLP 2005*, pp. 590–596.

Vondřička, P. (2014). Aligning parallel texts with InterText. In N. Calzolari et al. (eds.) *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavík: European Language Resources Association (ELRA), pp. 1875–1879.

Cognitive Features in a Corpus-based Dictionary of Commonly Confused Words

Petra Storjohann

Institut für Deutsche Sprache, Mannheim (Germany)
E-mail: storjohann@ids-mannheim.de

Abstract

This paper discusses how cognitive aspects can be incorporated into lexicographic meaning descriptions based on corpus-driven analysis. The new German Online dictionary "Paronyme – Dynamisch im Kontrast" is concerned with easily confused words such as *effektiv/ effizient*, *sensibel/ sensitiv*. It is currently in the process of being developed and it aims at adopting a more conceptual and encyclopaedic approach to meaning. Contrastive entries emphasise usage, comparing conceptual categories and indicating the mapping of knowledge. Adaptable access to lexicographic details offers different perspectives on information, and authentic examples reflect prototypical structures.

Some of the cognitive features are demonstrated with the help of examples. Firstly, I will outline how patterns of usage imply conceptual categories as central ideas instead of sufficiently logical criteria of semantic distinction. In this way, linguistic findings correlate better with how users conceptualise language. Secondly, it is pointed out how collocates are family members and fillers in contexts. Thirdly, I will demonstrate how contextual structure and function are included by summarising referential information. Details are drawn from corpus data; they are usage-based patterns illustrating conversational interaction and semantic negotiation in contemporary public discourse. Finally, I will show flexible consultation routines where the focus on structural knowledge changes.

Keywords: cognitive lexicography; corpus semantics; paronyms, easily confused words;

encyclopaedic-conceptual approach

1. Introduction

Lexicography has undergone dramatic changes over the past two decades. These mainly concern approaches to lexical analysis, the editorial process and the digitation/presentation of data. The relationship between semantic theory and practical lexicography has always been a difficult one (cf. Rundell, 2012). When it comes to employing semantic foundations, it is above all the field of corpus linguistics that has made its mark on dictionary writing. Corpora and their tools have turned lexicography into a more objective and empirical trade which makes use of authentic language data. Lexicographers have also continually taken advantage of hypertextual opportunities to present lexical information in innovative ways, although their full potential has not been exploited, nor have users' needs been extensively considered (Müller-Spitzer, 2014).

Cognitive linguistics, however, has had no major impact on general dictionaries. In particular, the structuring of entries and the definition of senses are two areas where cognitive principles could be used to implement descriptions of conceptual structures and to show how meaning is construed or represented. As Ostermann (2015) points out, novel cognitive theories have been neither recognised nor successfully integrated into general English dictionaries. A few specialised frame-based English systems such as Pattern Dictionary of English Verbs (PDEV) or FrameNet[1] facilitate their meaning descriptions with cognitive foundations (e.g. Fillmore, 1976; 1977). As far as general monolingual dictionaries are concerned, both in English and in German, there is a complete lack of guiding cognitive paradigms (e.g. conceptualisation, construction, categorisation, representation) being incorporated into semantic descriptions with a theoretical foundation.

In this paper, it is argued that cognitive ideas can be successfully implemented in descriptions of meaning and the structuring of entries, and that these provide relevant information which primarily benefits users. In the following, the new German dictionary of commonly confused words "Paronyme-Dynamisch im Kontrast" (Storjohann, 2016) is taken as an example that breaks with tradition by including central conceptual information and by representing both linguistic and encyclopaedic knowledge. Within the German context, it is a first attempt at a more cognitively infused lexicography calling for more realistic documentations of language and the way speakers perceive, conceptualise and linguistically represent the world. For the purpose of illustration, some cognitive features will be demonstrated, particularly those emphasising the interaction of details for more adequate depictions of flexible usage and contextual categorical implications.

2. Where to look for information on commonly confused words

Paronyms are easily confused words which regularly cause problems for both native speakers and language learners. As these lexical items often share morphological roots, they are similar with respect to sound, spelling and/or meaning, e.g. *effektiv/effizient (effective/efficient) sensibel/sensitiv (sensitive/delicate)*, *formell/formal/förmlich (formal/official), Method/Methodologie/Methodik (method/methodology)*, *Elektrik/Elektronik (electrics/electronics)*.[2] Generally, such pairs/sets are not regarded as synonymous (cf. Làzàrescu, 1995; 1999) although corpus analyses suggest that some items undergo meaning change due to the rivalry between the words. Sometimes, they can develop synonymous notions and simply become lexical alternatives (cf. Storjohann, 2015). In other cases, they remain similar in meaning but show subtle differences and restrictions in usage. Inevitably, situations of confusion arise when

[1] A related project in German is the German Frame-Semantic Online Lexicon GFOL (http://coerll.utexas.edu/frames/).

[2] For more examples see Schnörch (2015).

speakers' intuitions contradict information in existing reference works.

The importance of paronyms is based on the assumption that these items play a vital role for users in the process of second language acquisition and foreign language communication in order to avoid misunderstandings. Confusing paronyms is sometimes regarded as a violation of semantic correctness. Prescriptive analysts favour semantic correction and the avoidance of such mishaps (Bolshakov & Gelbukh, 2003). Indeed, the alleged misuse of morphologically and semantically similar words also leads to linguistic uncertainties for native speakers, as numerous language-related Internet blogs show. However, corpus-guided investigations of paronyms partly reveal recent semantic changes, conventionalised overlappings and newly established contexts. Therefore, empirically sound, descriptive documentation is necessary to capture the current use of paronyms. Corpus-assisted investigations of easily confused words and their usage over recent decades can provide valuable insight into principles of semantic shift. It is argued here that such analyses might enable semanticists to integrate the phenomenon into a wider theoretical framework on the one hand and into appropriate lexicographic descriptions on the other hand.

As most general German reference guides still favour a traditional style and structure, recent or new phenomena are hardly captured nor adequately described. Taking a closer look at resources such as Duden online, their lexicographic deficiencies become apparent. Users interested in the differences between *Elektrik/Elektronik* or *effektiv/effizient* find the following facts:

Elektrik	*Elektronik*
Gesamtheit einer elektrischen Ausstattung	Gesamtheit einer elektronischen Anlage oder Ausstattung
effektiv	*effizient*
wirksam, wirkungsvoll lohnend, nutzbringend sich tatsächlich feststellen lassend, wirklich	wirksam und wirtschaftlich

Table 1: Definitions taken from Duden online

The entries of *Elektrik/Elektronik* are circular and "married with content from antiquated dictionaries – the type that define pedantic as 'of, pertaining to, or characteristic of a pedant'" (Rundell, 2012: 74). The entries of *effektiv/effizient* mainly summarise synonyms. Users do not obtain sufficient details concerning their conceptual potential and contextual usage. Quite likely, users will miss information, for example, on semantic reference, relevant conceptual domains or categories, discourse structures and contextual situations. Who/what is specifically characterised

as *effektiv/effizient* and in what kind of contextual circumstances? This question remains open. Similarly, German Wiktionary describes the meaning of *effektiv* as follows:

[1] die Fähigkeit besitzend, eine Aufgabe erfolgreich zu erledigen,
(to have the ability to complete a task successfully),
[2] *ohne Steigerung:* sich tatsächlich feststellen lassend, wirklich,
(*without comparison*: in fact, real).

The adjective *effizient* is described as 'to be able to be productive relative to the invested effort':

[1] fähig, viel Leistung in Relation zum Aufwand zu erbringen.

Again, conceptual details, preferred discourse situations and further encyclopaedic knowledge are not documented. Also, both descriptions suggest a small semantic spectrum for both adjectives.

Today, speakers face a range of consultation options, from traditional print dictionaries to free online resources. As most German e-dictionaries are copied or digitised versions of conventional reference books, unfortunately these often do not offer satisfactory answers to questions about paronym behaviour.

> bezüglich *sensitiv/sensibel*: „Ich hab zwar überall nach einer Definition dieser beiden Wörter gesucht, aber je mehr ich finde, desto irritierender ist es" (aus: http://depriforum.phpbb8.de/diskussionen-f16/sensitiv-sensibel-t1258.html).[3]

Consequently, online forums have turned into widely used social media sources where users consult the community for their linguistic problems (see Figure 1[4]).

Figure 1: Typical Language Question in Internet

[3] Translation: I've looked for a definition of these two words everywhere, but the more I find, the more irritating it becomes.

[4] Example taken from:
http://www.gutefrage.net/frage/was-ist-der-genaue-unterschied-zwischen-effektiv-und-effizient.

In some cases, they explain whole contextual situations in which their uncertainties occur. They seek information on lexical use, prototypical contexts, possible constructions, and conceptual as well as encyclopaedic issues. The answers from the language community are impressively diverse and revealing. As a matter of fact, speakers have good intuitions as to what linguistic and extra-linguistic information is required to form essential parts of authentic communication. In online forums, people share their concerns about easily confused words. It is here, through the study of blogs, that detailed insights into the specific linguistic problems of users, their consultation behaviour and their needs, can be gained. However, it is also here where we see that users do not always obtain satisfactory answers (see Figure 2).

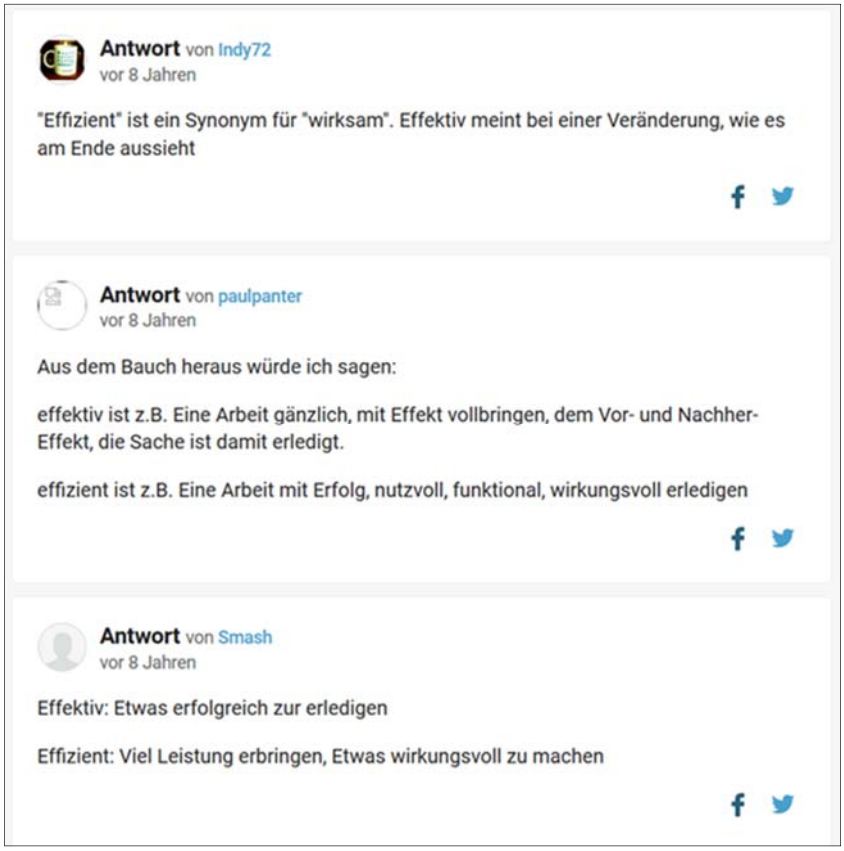


Figure 2: Exemplary Answers in Internet Forum[5]

Undoubtedly, Internet forums are not a reliable source of information. Consultations can be helpful but they are not guaranteed sources of reliable information.

[5] Translation answer 1 “efficient” is a synonym for “productive/effective”. Effective refers to a change of state and what it looks like in the end. Translation answer 2: Intuitively, I would say: effective is, for example, to finish some work with effect, the pre and after effect, the job is then done. Efficient is, for example, to do a work in a useful, functional, effective way. Translation answer 3: effective: to so something successfully. efficient: to work productively, to do something effectively.

3. Dictionaries and the Cognitive Perspective

The subject of paronymy has not been revisited with empirical, data-driven methods either in terms of semantic theory or practical lexicography. Lexicographically, some German paronyms have been documented in printed dictionaries (Müller, 1973; Pollmann & Wolk, 2010), although not systematically. However, there is no corpus-guided reference guide empirically describing paronym sets enabling readers to find the correct usage of such lexical items.

Placing the user in focus, it is essential to strive for conceptual approaches and to document the interplay of lexical, structural and encyclopaedic knowledge in meaning descriptions. While analysing the needs and various interests of users we have come across two prerequisites. On the one hand, it is necessary to implement a semantic structure and network that is closer to actual usage and this requires information on patterns of conceptualisation, on categories, reference and concrete lexical prototypes. For quite some time, there are endeavours to reconcile the branch of lexicography with cognitive semantic theories. As Geeraerts (2007: 1168) has pointed out:

> […] what Cognitive Linguistics seems to offer to lexicography is a conception of semantic structure that is perhaps in a number of respects more realistic than what many other semantic theories (in particular, theories of a structuralist persuasion) can provide.

On the other hand, we need to overcome a rigid, linear ordering of information and strive for a realistic representation of multi-dimensional facets of semantic configurations in language use to be closer to the structure of the mental lexicon (cf. Ostermann, 2015).

3.1 The New German e-Dictionary "Paronyme – Dynamisch im Kontrast"

"Paronyme-Dynamisch im Kontrast" is an electronic dictionary that breaks new ground by adopting a more conceptual and encyclopaedic approach to meaning by incorporating cognitive features. It will be published in the dictionary portal OWID (Online-Wortschatz-Informationssystem Deutsch, www.owid.de) in 2017. It is currently in the process of being developed and includes conceptual, prototypical, and referential categorisation and a flexible structural access to knowledge. This dictionary does not follow sufficiently logical criteria of semantic distinction for its sense disambiguation. Instead, different patterns of usage and their underlying conceptual categories and prototypical realisations function as parameters of contextual distinction. These are then accessed flexibly via menu navigation. As a quick guide, short paraphrases define characteristics of conceptual referential categories. Concerning the adjectival pair *effektiv/effizient*, relevant topic areas (or frame

presentations) are given for each adjective. These are coded as "guide words"[6] together with a synonym (see Figure 3).

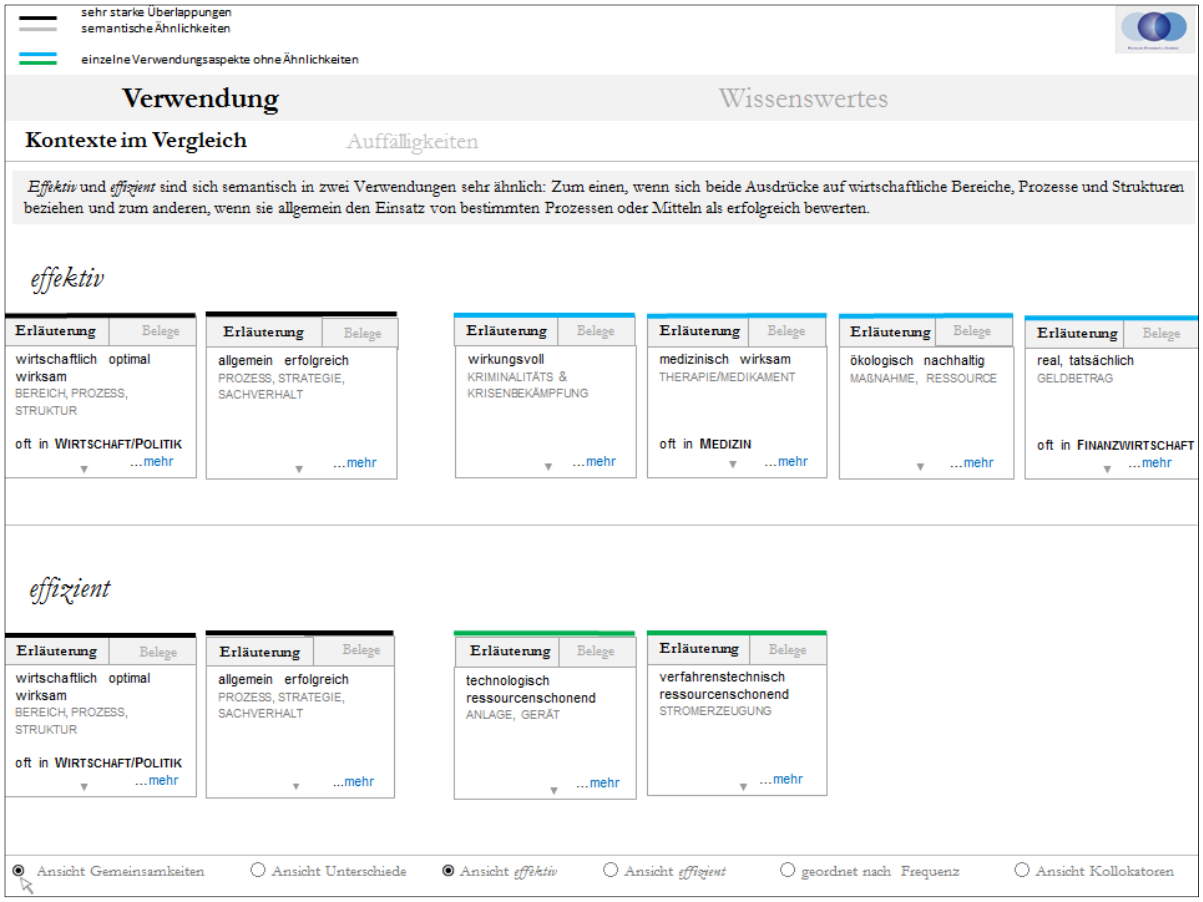


Figure 3: Default Conceptual Navigation Structure

A large amount of knowledge about words, meanings and concepts is derived from experience and from the categories we construct, i.e. mentally represented frames or schemas. It is these categories (e.g. AREA/PROCESS, STRUCTURE, PROCESS/STRATEGY/STATE OF AFFAIRS, CRIME/CRISIS, MEDICINE, MEASURE/RESOURCE, MONEY, TECHNOLOGICAL DEVICE, ENERGY) that justify a distinction of patterns and help to correlate situations of language use to different contexts. Similar ideas of how to use guide words to exemplify contextual frames in which the words are prototypically embedded can be found in Ostermann (2015). In the dictionary, these categories build up a quick contrastive guide and a concept-driven navigation structure (see Figure 1). They are also able to activate corresponding concepts of polysemous words. These also help a user to encode contexts and to identify metonymic and metaphoric mappings (cf. Fillmore & Atkins). Users can more easily relate the adjectives to their meanings and relate these then to the preferred contextual reference (here nouns), e.g.:

[6] Guide words are also used in Cambridge International Dictionary of English.

- *Effektiv* means 'economically optimal' with respect to an AREA, PROCESSES or STRUCTURE and it often occurs in economy or politics.
- *Effektiv* means 'generally successful' with regard to a PROCESS, a STRATEGY or a SITUATION/SPECIFIC MATTERS.
- *Effektiv* means 'working' in terms of FIGHTING CRIME/CRISES.
- *Effektiv* means 'working' in terms of MEDICATION or a THERAPY and it often occurs in medical contexts or contexts describing health issues.
- *Effektiv* means 'ecologically sustainable' in terms of MEASURES or RESOURCES.
- *Effektiv* means 'real' with regard to AMOUNTS OF MONEY and it is often used in contexts describing financial issues.

Compared to traditional dictionaries (see table 1) much more information is provided which can be consulted and then mentally stored together.

Through the more visual explanations, it is possible to answer questions such as *Can German effektiv be used synonymously with effizient in contexts of business to characterise economic methods or structures? Can a motor be described as effektiv or effizient?* or *Is a powerful production of electricity better referred to as being effektiv or effizient? Do I use effizient or effektiv when I want to say that a medical treatment is working well?* With the help of the given synonyms and guidewords in the short paraphrase it is also possible to compare individual contexts of the two paronyms and quickly identify similarities and differences.

3.2 Contextual Fillers as Prototypical Lexical Realisations

Users also have the option of consulting more detailed information on demand. Conceptual reference and encyclopaedic ideas are then explicitly integrated into the longer paraphrase. The relevant ontological category or domain is then specifically illustrated using lexical preferences, i.e. collocates. With a dynamic electronic display at hand, these are shown optionally, as a list of frequent and conventionalised contextual partners, introduced by *such as* underneath the definition (see Figure 4). In this approach, collocates are concrete lexical realisations (or fillers[7]) in specific contexts illustrating the referential category given in the definition.[8]

[7] For verbs, which only make up only a small section of the dictionary, collocates serve as fillers in frame-like constructions. Collocates are then grouped into different sets (argument roles).

[8] The linguistic analysis of corpus-driven collocates is also indicative evidence of distinct usage and senses. They are a primary source for lexicographers for deriving definitions and disambiguating meaning.

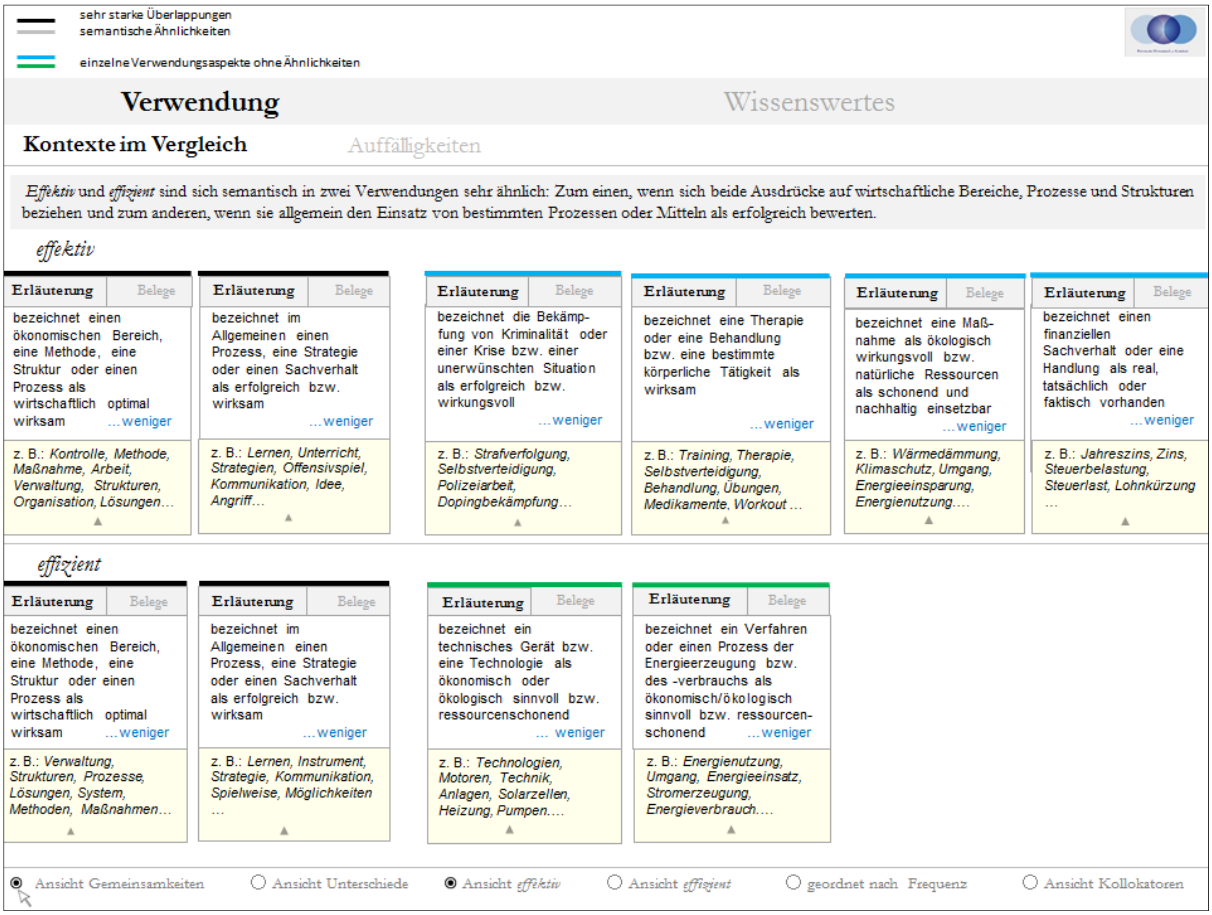


Figure 4: Long Definition and Prototypical Realisation (Fillers)

For example, polysemous *effektiv* prototypically means something like 'economically effective' or 'efficient'. It is the conceptual background where the adjective refers to nouns functioning as non-human subjects or objects and denoting ECONOMIC AREAS, PROCESSES, STRUCTURES or MATTERS OR AFFAIRS such as *control, method, measure, work, administration, structures, organization* or *solutions*. Similarly, German *effizient* also refers to nouns expressing the concepts of ECONOMIC AREAS, PROCESSES, STRUCTURES or MATTERS OR AFFAIRS illustrated by *administration*, *structures*, *processes*, *solutions*, *system*, *methods* and *measures*. Alternatively, both items can refer to STRATEGIES or PROCESSES as 'generally being successful': for *effektiv* these are typically *learning*, *teaching*, *strategy*, *offense*, *communication*, *idea* and *attacks*.

For *effizient* these are *learning, instrument, strategy, ways of playing, communication* or *possibilities*. In other contexts, they differ in terms of their conceptual referents. For example, *effektiv* can be used to describe the successful fight against crime or crises (as exemplified by *self-defense*, *police work*). It refers to the positive results of a therapy or medicine (illustrated by *training*, *therapy*, *exercise*) and the adjective describes measures and natural resources (demonstrated by *insulation*, *climate protection*, *energy saving*) as successful. An adverbial usage is also attested for *effektiv* with

referenced to money or interests, meaning 'real, actual' (indicated *annual interest*, *tax burden*).

Effizient exhibits two further contexts which refer to technological equipment or instruments such as *motors*, *solar cells*, *heating* and *pumps*. It also occurs in contexts were the adjective describes procedures of generation or consumption of power/energy as ecologically sustainable (illustrated by co-occurring *electricity consumption, electricity supply*, *power generation*).

In essence, the lexical representations are prototypical domain elements and structured mental representations of human experience. They shed light on strong affinities to constructions and contextual preferences, and they point to properties correlating with aspects of meaning structure. With prototypical details, we have the possibility of handling polysemous contexts in a way that "more faithfully reflects what corpus data tells us" (Rundell, 2012: 82). For polysemous items, metonymous and metaphorical contexts are listed. These show cognitive processes in which conceptual elements motivate the configuration of another semantically related conceptual entity (cf. Kövecses & Csabi, 2014).

The lexical representations are not intuition-based examples but statistically significant occurrences provided by corpus instances (see Section 4). They are lexicographically analysed, interpreted and classified manually, once automatically retrieved collocation analyses have provided the necessary access to typical contextual structures. Each paraphrased context is illustrated by up to three citations editorially picked from the corpus. The entry as such is not automatically retrieved, corpus tools pre-analyse complex data sets and provide systematic access to significant patterns. These then undergo editorial scrutiny where corpus findings are essential evidence of cognitive entities and categories. In the entries, corpus lexicography meets cognitive lexicography.

3.3 The Organisation of Knowledge

Hypertext dictionaries can break up conventional sequential ordering of information. A granular XML-architecture allows for different data structures and therefore flexible access routes, adaptable presentations and complex searches. As digital data systems can represent their content in a structure that is not dependent on its presentation, it is possible to generate adaptable displays. Tailor-made user-adaptivity is technologically feasible but will only become a realistic option once we know more about the users. Content can be arranged dynamically, changing linguistic focus to "allow users to recreate and re-represent their own dictionary data" (Fuertes-Olivera, 2013: 330).

By focusing on the needs of the user, we have learned that these vary considerably (cf. Storjohann, 2016). Given this, a system of various options has been developed which

enables us to configure different perspectives on the organisation of knowledge. In essence, this dictionary is an XML-based hypermedia resource. Its system is customisable and can adaptively generate and prioritise information for specific user groups. Apart from regular search options, with multi-functional specifications at hand, dictionary data can be individually "reshuffled" by setting different parameters during the consultation process. Consequently, focal points on conceptual structures change and different linguistic aspects are emphasised.

Firstly, as a default, the different instances of usage of each lexical item are established in relation to the individual contexts of the corresponding paronym item, with identical contexts first, followed by similar and dissimilar contexts. Through this, an instant overview of overlapping uses and differences is provided (Figure 1 or 2). Secondly, depending on personal interest, users can also choose parameters for listing the different contexts first. Thirdly, as neither ordering necessarily corresponds to the frequency of occurrence in actual usage, all contexts can be shown according to their distribution in the corpus, so that the predominance or centrality of certain contexts can be seen. Fourthly, it is the user's decision to choose the ordering of paronym items and determine which one appears at the top of the entry. Finally, the menu options also include a visualisation of collocation profiles with behavioural networks and interactive functions (see Figure 5).

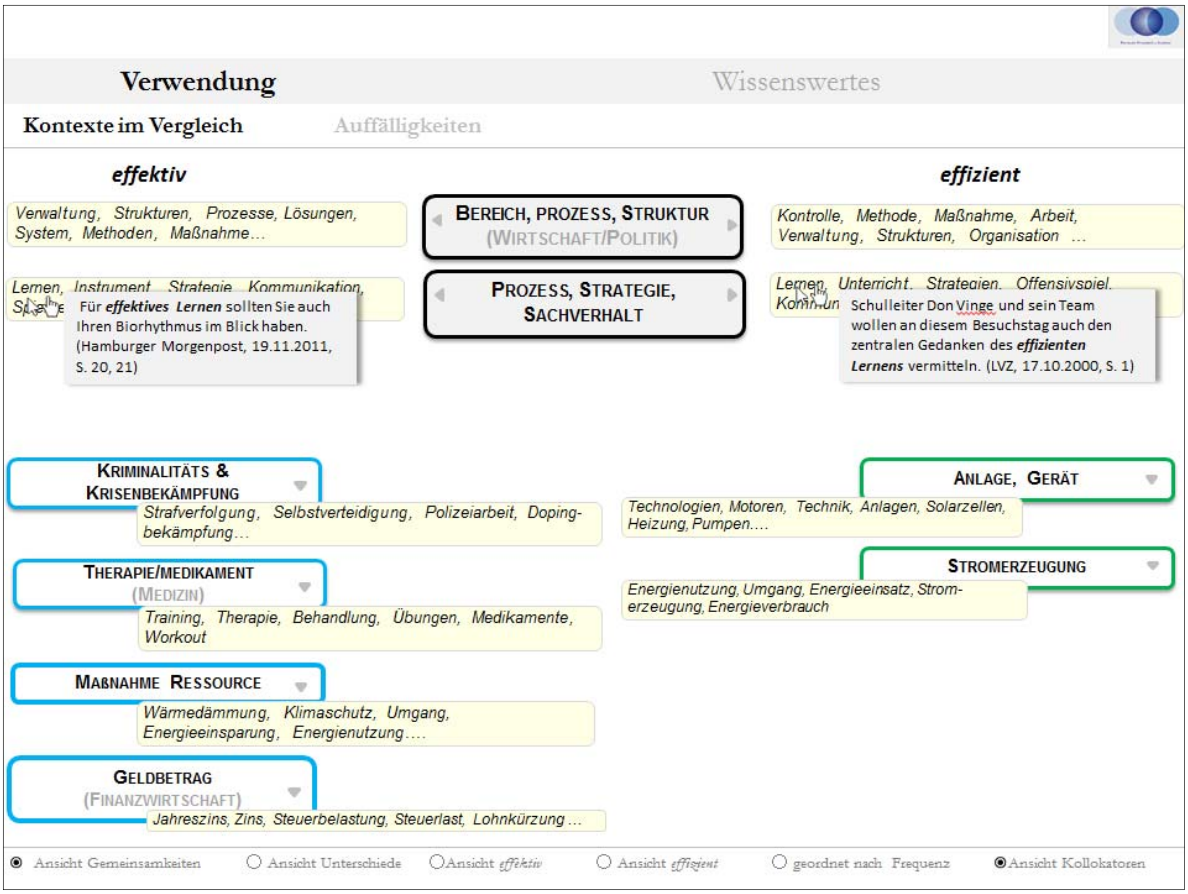


Figure 5: Visualisation of Collocational Profiles and Interactive Functions

Taking the conceptual categories as a starting point, their corresponding collocational representations can be studied contrastively. The denoted concepts that are commonly shared are in the centre, followed by dissimilar concepts arranged separately to the left and right below. Each category, together with its individual lexical realisations, is exemplified by corpus instances. It is a simplified diagram with abstract concepts directly representable in a contrastive conceptual organisation. Overall, this e-dictionary exploits text- and hypertechnological possibilities and offers consultation routines by optionally generating different facets of structural knowledge.

3.4 Corpus-guided dictionaries vs traditional dictionaries

As we can see from Figures 3–5, corpus data strongly suggests that *effektiv* and *effizient* are used synonymously with respect to two contexts. The underlying corpus provides numerous attestations (see Examples 1–3).

(1) Arbeit **effizienter** machen: Mit einem guten Computer-Netz kann jedes Unternehmen **effektiver** arbeiten - gerade, wenn seine Büros auf viele Orte verteilt liegen. (Rhein-Zeitung, 15.03.2002, Das Dekanat soll "Computer-fit" werden.)

(2) Derzeit gilt der Vertrag von Nizza aus dem Jahr 2000. Doch die Strukturen sind nicht mehr **effizient**. Eine Kommission mit 27 Kommissaren kann ebenso wenig **effektiv** arbeiten wie ein Parlament mit fast 800 Abgeordneten. (Braunschweiger Zeitung, 21.06.2007, Fragen und Antworten zum Gipfel.)

(3) Der sture Ablauf, der fast immer eingehalten wird, sei vielmehr das Ergebnis **effektiver** Arbeitsteilung von Spitälern, Bestattern und den Friedhofsbetreibern. Fast alle größeren Bestattungsunternehmen pflegen in Deutschland eine **effiziente** Arbeitsteilung. (Die Zeit, 15.04.2004, Wie man in Deutschland begraben wird)

This entry, demonstrated in Figures 3–5, is a good example to show differences to other existing prescriptive reference books such as Pollmann & Wolk (2010). Its documentation aims at guiding users to the allegedly correct usage and suggests a clear distinction between the items in question (see Figure 6).


effektiv ··· effizient

- Ist etwas gemessen an den Mitteln, die zur Erreichung eines bestimmten Ziels ausgerichtet sind verhältnismäßig, dann ist es effektiv.
- Effizient ist etwas, das besonders wirtschaftlich ist, also die Kosten mit dem Nutzen vergleicht. (Eselsbrücke: -zient / Cent).

Ein Insektenspray vertreibt oder tötet beispielsweise Insekten, somit ist es effektiv. Wesentlich effizienter ist jedoch eine Fliegenklatsche.

Figure 6: Dictionary Entry *effektiv*/*effizient* in Pollmann & Wolk (2010).

Strictly normative language use is also propagated in the German Wiktionary[9], a popular electronic resource which under an explicit headline points out that confusion over the two words *effektiv* and *effizient* should be avoided. Conventional reference guides have so far focussed on the differences between commonly confused words. They entirely fail to explain existing similarities. The usage restrictions that are documented in these reference books cannot be confirmed through corpus data. As is the case for *effektiv/effizient*, strict usage lines cannot be sharply drawn which might have been expected intuitively. The meanings of typically confused words are more freely exposed to semantic negotiation. Following a descriptive empirical view, the semantics of some paronymic lexical items have adopted new semantic aspects and undergone meaning changes that are observable as regular patterns in a corpus and not as single misused occurrences.

Overall, all reference guides mentioned are neither based on semantic examinations of current natural language in use nor on investigations of large data. It is empirical corpus explorations that open up the discrepancies to traditional descriptions. Corpus studies allow for the description of similarities which, on the one hand, might offer a deeper understanding why two words are regularly being confused and, on the other hand, it might indicate ongoing linguistic change worth documenting. Consequently, corpus-driven research on paronymy demands a more differentiated look at the phenomenon than has previously been offered.

4. Corpus Lexicography meets Cognitive Lexicography

The paronym dictionary bases its information on a comprehensive purpose-built corpus comprising 2.3 billion words.[10] The underlying corpus is publicly accessible and provides for transparent lexicographic practices. As the subject of paronymy has not been revisited with empirical, data-driven methods, either in terms of semantic theory or in terms of practical lexicography, suitable corpus methods for contrastive investigation needed to be tested (cf. Storjohann & Schnörch 2014). Currently, complementary software-driven resources facilitating the search for similarity and difference are being exploited, each of which is based on the analysis and interpretation of contextual profiles, collocations and colligations, corresponding semantic roles and syntactic functions.[11] Corpus data reveals how meaning is constantly being negotiated in usage events and how communicative acts can create semantic rivalry or increase vagueness of easily confused words. Accordingly, variation and uncertainties arise from lexical similarity, sometimes leading to the adoption of new conceptual-semantic nuances. It is corpus-guided investigations that uncover discrepancies between conventionalised language use, speakers' intuition and

[9] See Wiktionary entry https://de.wiktionary.org/wiki/effektiv.

[10] See: http://www1.ids-mannheim.de/lexik/paronymwoerterbuch/dasparonymkorpus.html.

[11] For verbs, these would be based on the extraction of complementation patterns.

traditional dictionary entries. They are essential in the tracing of regular, conventionalised or new semantic components. The analysis and interpretation of patterns shows that meaning is conceptualisation, constantly negotiated in usage. Aspects of discourse, domain, reference and ontological categorisation are mentally processed and stored as information on lexical use and meaning.

In the case of the paronym dictionary, linguistic and encyclopaedic details are drawn from corpus data and are included in usage-based linguistic patterns, illustrating conversational interaction and semantic negotiations in contemporary public discourse. Cognitive elements play an essential role when users confuse lexical items. This confusion is often not only related to formal similarities but also to conceptual closeness. Corpus-derived data allow for the search of minimal semantic differences and the integration of necessary encyclopaedic knowledge, information that is complementary to linguistic information and needed by users. While this is not news to cognitivists, lexicographers still have to learn how to integrate this insight into usable tools. Bridging the gap between corpus lexicography and cognitive paradigms is a slow but steady process (Gries, 2006; Rundell, 2012; Hanks, 2013). Writing dictionaries should be informed by theoretical grounding and lexicographers should be linguistically aware corpus analysts. As Lew (2007: 221) points out "let us hope that lexicographers will keep an open mind to developments in linguistics […]".

5. Summary

So far, there is no corpus-assisted German reference guide empirically describing commonly confused words and enabling readers to find the correct contemporary usage.[12] The paronym dictionary is committed to overcoming the discrepancy between traditional practice and insights from language use. This necessarily means finding a way of educating users by showing how linguistic knowledge, encyclopaedic knowledge and human experience are inextricably linked. Given these goals, the dictionary breaks down the binary distinction of dictionary vs. encyclopaedia. Solutions to a number of lexicographical challenges were required. One aim was to bridge the gap between cognitive semantics and corpus lexicography by simultaneously considering user needs. It was argued that cognitive aspects can successfully be incorporated into meaning descriptions based on corpus-driven analysis. Insights into collocational use and the interpretation of contexts can lead to the implementation of more abstract encyclopaedic or conceptual categories as central ideas. Together with concrete prototypical contextual realisation these replace circular definitions and uncommented lists of synonyms. Authentic examples reflect prototypical structures as manifested in discourse and in the mental lexicon.

In contrastive entries, the interaction between lexemes is emphasised. The dictionary

[12] Intuition-based dictionaries include (Müller) 1973 and Pollmann & Wolk (2010).

strives to adequately reflect ideas such as conceptual structure, categorisation and knowledge. While Kövecses & Csábi (2014) argue that employing cognitive linguistics is a profitable theoretical underpinning for lexicographers, we favour the description in terms of cognitive principles as it predominantly embraces user needs.

Only a digital resource is able to solve problems of strict macrostructural ordering. Indeed, "an online dictionary can be adapted to the needs of each dictionary user" (Kwary, 2012: 35). Dynamic look-up options replace rigid structures. An adaptable access to lexicographical information has been suggested, where variable search options enable different foci and perspectives on linguistic information. In addition, the implementation of interactive collocation networks is a more onomasiological approach which offers an alternative access to language and knowledge structures relevant in actual usage events. The Paronymwörterbuch is a dynamic source of information where the interests of different users will hopefully be met.

6. References

Bolshakov, Igor A. & Gelbukh, Alexander (2003). Paronyms for accellerated correction of semantic errors. In *International Journal Information Theories & Applications* 10, pp. 198–204.

Fillmore, Charles J. (1976). Frame semantics and the nature of language. *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech* 280, pp. 20–32.

Fillmore, Charles, J. & Atkins, B. T. Sue (1992). Toward a Frame-based Lexicon: The Semantics of RISK and its Neighbors. In Adrienne Lehrer & Eva Feder Kittay (eds.) *Frames, Fields and Contrast. New Essays in Semantic and Lexical Organization*. Hillsdale & London: Erlbaum, pp. 75–102.

Fillmore, Charles, J. (1977). The need for a frame semantics in linguistics. In Hans Karlgren (ed.) *Statistical Methods in Linguistics*. Stockholm: Scriptor, pp. 5–29.

Fuertes-Olivera, Pedro A. (2013). e-lexicography: The Continuing Challenge of Applying New Technology to Dictionary-Making. In Howard Jackson (ed.) *The Bloomsbury Companion to Lexicography*. London: Bloomsbury, pp. 323–340.

Geeraerts, Dirk (2007). Lexicography. In Dirk Geeraerts & Hubert Cuyckens (eds.) *The Oxford Handbook of Cognitive Linguistics*. Oxford: Oxford university Press, pp. 1160–1174.

Gries, Stefan Th. (2006). Corpus-based methods and cognitive semantics: The many senses of to run. In Stefan Th. Gries & Anatol Stefanowitsch (eds.) *Corpora in Cognitive Linguistics Corpus-Based Approaches to Syntax and Lexis*, Berlin & New York: de Gruyter, pp. 57–100.

Hanks, Patrick (2013). *Lexical Analysis: Norms and Exploitations*. Cambridge (MA): MIT.

Kövecses, Zoltán & Csábi, Szilvia (2014). Lexicography and cognitive linguistics. *Revista Española de Lingüística Aplicada* 27(1), pp. 118–139.

Kwary, Deny Arnos (2012). Adaptive hypermedia and user-oriented data for online

dictionaries: A case study on an English dictionary of finance for Indonesian students. *International Journal of Lexicography* 25(1), pp. 30–49.

Làzàrescu, Ioan. (1995). Deutsche Paronyme. In *Grazer Linguistische Studien* 43, pp. 85–93.

Làzàrescu, Ioan. (1999). *Die Paronymie als lexikalisches Phänomen und die Paronomasie als Stilfigur im Deutschen*. Bukarest: Editura Anima.

Lew, Robert. 2007. Linguistic semantics and lexicography: A troubled relationship. In Małgorzata Fabiszak (ed.) *Language and Meaning. Cognitive and Functional Perpectives*. Frankfurt: Lang, pp. 217–224.

Müller-Spitzer, Carolin (ed.). (2014). *Using Online Dictionaries*. Berlin & New York: de Gruyter.

Ostermann, Carolin (2015). *Cognitive lexicography. A New Approach to Lexicography Making Use of Cognitive Semantics*. Berlin & Boston: de Gruyter

Rundell, Michael (2012). It works in practice but will it work in theory? The unseasy relationship between lexicography and matters theoretical. In Ruth Vatvedt Fjeld, & Julie Matilde Torjusen (eds.) *Proceedings of the 15th EURALEX International Congress,* EURALEX 2012. Oslo: University of Oslo, pp. 47–92. Available at: www.euralex.org/elx_proceedings/.../pp47-92%20Rundell.pdf (10 May 2017).

Schnörch, Ulrich (2015). Wie viele Paronympaare gibt es eigentlich? Das Zusammenspiel aus korpuslinguistischen und redaktionellen Verfahren zur Ermittlung einer Paronymstichwortliste. *Sprachreport* 4 (2015), pp. 16–26.

Storjohann, Petra & Schnörch, Ulrich (2014). Empirical approaches to German Paronyms. In Abel, Andrea & Chiara Vettori & Natascia Ralli (eds.) *Proceedings of the 16th EURALEX International Congress: The User in Focus*. EURALEX 2014, Bolzano/Bozen: EURAC, pp. 463-476. Available at: http://euralex2014.eurac.edu/en/callforpapers/Pages/default.aspx (10 May 2017).

Storjohann, Petra (2015). Was ist der Unterschied zwischen sensitiv und sensibel? *Zeitschrift für Angewandte Linguistik* 62(1), pp. 99–122.

Storjohann, Petra (2016). Vom Interesse am Gebrauch von Paronymen zur Notwendigkeit eines dynamischen Wörterbuchs. *Sprachreport* 4 (2016), pp. 32–43.

Dictionaries & Websites:

Cambridge International Dictionary of English. 1995. Cambridge: Cambridge University Press.

DWDS: *Digitales Wörterbuch der Deutschen Sprache*. Accessed at: http://www.dwds.de. (10-14 March 2015).

Duden Online. Accessed at: http://www.dude-online.de. (10 May 2017).

FrameNet. Accessed at: framenet.icsi.berkeley.edu/fndrupal/. (10 May 2017).

German Frame-Semantic Online Lexicon GFOL. Accessed at: http://coerll.utexas.edu/frames/. (10 May 2017).

Müller, Wolfgang. 1973. *Leicht verwechselbare Wörter.* Duden Taschenwörterbücher Vol. 17. Mannheim: Bibliographisches Institut.

OWID - *Online Wortschatzinformationssystem des Deutschen.* Accessed at: http://www.owid.de. (10 May 2017).

Paronyme - Dynamisch im Kontrast. URL: http://www1.ids-mannheim.de/lexik/paronymwoerterbuch.html. (10 May 2017).

Pollmann, Christoph & Wolk, Ulrike. 2010. *Wörterbuch der verwechselten Wörter. 1000 Zweifelsfälle verständlich erklärt.* Stuttgart: Pons.

PDE - *Pattern Dictionary of English Verbs.* Accessed at: http://www.pdev.org.uk. (10 May 2017).

Wiktionary-*Das freie Wörterbuch*: Accessed at: www.wiktionary.de. (10 May 2017).

From Monolingual to Bilingual Dictionary: The Case of Semi-automated Lexicography on the Example of Estonian–Finnish Dictionary

Margit Langemets[1], Indrek Hein[1], Tarja Heinonen[2], Kristina Koppel[1], Ülle Viks[1]

[1] Institute of the Estonian Language, Roosikrantsi 6, 10119 Tallinn, Estonia

[2] Institute for the Languages of Finland, Hakaniemenranta 6, 00530 Helsinki, Finland

E-mail: margit.langemets@eki.ee, indrek.hein@eki.ee, tarja.heinonen@kotus.fi, kristina.koppel@eki.ee, ylle.viks@eki.ee

Abstract

We describe the semi-automated compilation of the bilingual Estonian–Finnish online dictionary. The compilation process involves different stages: (a) reusing and combining data from two different databases: the monolingual general dictionary of Estonian and the opposite bilingual dictionary (Finnish–Estonian); (b) adjusting the compilation of the entries against time; (c) bilingualizing the monolingual dictionary; (d) deciding about the directionality of the dictionary; (e) searching ways for presenting typical/good L2 usage examples for both languages; (f) achieving the understanding about the necessity of linking of different lexical resources. The lexicographers' tasks are to edit the translation equivalent candidates (selecting and reordering) and to decide whether or not to translate the existing usage examples, i.e. is the translation justified for being too difficult for the user. The project started in 2016 and will last for 18 months due to the unpostponable date: the dictionary is meant to celebrate the 100th anniversary of both states (Finland in December 2017, Estonia in February 2018).

Keywords: bilingual lexicography; corpus linguistics; usage examples; GDEX; Estonian; Finnish

1. Background

The idea of compiling the new Estonian–Finnish dictionary is about 15 years old: it first arose in 2003, after the successful publication of the voluminous Finnish–Estonian dictionary, a project which benefitted from necessary and sufficient financing, adequate time (5 years) and staff (7 lexicographers), and effective management. The dictionary was printed in two 1000-page volumes.

Times tend to change: it was not immediately financially possible to start the vice versa dictionary and the 'electronic' manuscript (text file with XML-like mark-up) was filed away. Then, in late 2015, times changed once more, this time for the better. Finland and Estonia, both approaching their 100th anniversaries of the state (Finland in December 2017, Estonia in February 2018), decided to jointly celebrate their

anniversaries by offering each other, as a gift, two dictionaries: the Finnish–Estonian dictionary (2003) will be made available for free on the web in 2017, and the new and long-awaited Estonian–Finnish dictionary will be compiled and published electronically in February 2018.

The project was initiated in 2016, the agreement of the joint project was signed by the Institute for the Languages of Finland (Helsinki) and the Institute of the Estonian Language (Tallinn) in June 2016, leaving 18 months for the partners to achieve their mission.

2. Generation of the EST-FIN database

The database for the Estonian–Finnish Dictionary (henceforth EST-FIN) was generated combining two databases: the source language part was formed from the database of the monolingual general Dictionary of Estonian (ESTDic, to appear in 2018/2019), and the target language part from the database of the Finnish–Estonian dictionary (2003, 2 vols, 90,000 lemmas; henceforth FIN-EST). The EST-FIN database was created in August 2016 with a list of 80,000 lemmas. Since the Dictionary of Estonian is an ongoing project, we only managed to operate with four-fifths of the complete manuscript (ca 100,000 lemmas). The remaining fifth will be added to the database in June 2017 (after finishing the compilation of ESTDic) following the same principles.

The database structure of the monolingual dictionary was transformed into the structure of the bilingual database thus: we added the elements for the target language information (Finnish translations, grammatical information, e.g. government etc.). Specific tuple- or triple-groups of adverbs and adverbial phrases denoting state, place, direction etc. (functioning in 2-3 internal/external local cases only) were divided into separate entries due to different translation equivalents in the target language. (Technicalities have the bad habit of lasting forever...)

Reversing a bilingual dictionary has already been described in numerous papers (e.g. Maks, 2007; Viks, 2008). However, the otherwise trivial process of extracting lemma-translation pairs, reversing them and re-sorting according to the position of the translation in the article, had some unexpected hurdles.

The FIN-EST, as a typical paper dictionary, used tilde as a replacement for the non-variable start of the lemma (see example a). Restoring full-blown textual representation required, in rare cases, manual editing. Variable parts in Estonian phrases were recursively split into primitives: thus, the pair FIN *aavistuksen verran suolaa* (‘pinch of salt’)–EST *raasuke (natuke, sipake) soola* resulted in three potential pairs (example b).

(a) FIN loik|ata (‘to hop’) – FIN hän ~kasi ojan yli, … FIN ~ata vihollisen puolelle

→ FIN hän loikkasi ojan yli, FIN loikata vihollisen puolelle

(b) FIN aavistuksen verran suolaa ('pinch of salt') – EST raasuke (natuke, sipake) soola

→ FIN aavistuksen verran suolaa – EST raasuke soola

→ FIN aavistuksen verran suolaa – EST natuke soola

→ FIN aavistuksen verran suolaa – EST sipake soola

Estonian morphology subtly differs from Finnish in the usage of verb infinitive forms. The dictionary tradition in Estonian is to present verb lemmas in ma-infinitive, whereas the common da-infinitive is well understood by speakers of both languages and used for Estonian translations in the FIN-EST database. However, the EST-FIN database has lemmas in ma-infinitive and the reversing process had to make use of a morphological analyser for non-phrasal translations. Analysing phrases and collocations was not attempted for several reasons. The bulk of all pairs consist of nouns and noun collocations; verbs are relatively few. Also, many of the phrases would not be used, as they are constructed Estonian translations and not typically used as lemmas, e.g. Finnish *lautailla* ('to surf')–Estonian *rulaga sõita*, *lainelauaga sõita*, *lumelauaga sõita*. Applying morphological analysis to simple collocations without much context would also have produced many meaningless candidates. As the task at hand was not to compile a reverse dictionary, but rather to save as much time as possible by keeping routine tasks to a minimum, usage examples containing verb forms were ignored.

Extracting all translation pairs from the FIN-EST database (90,000 lemmas) resulted in 330,000 pairs, which could be inserted into pre-defined slots in the EST-FIN template. If the template Estonian dictionary entry had the same phrase that was used as translation, the corresponding Finnish phrase was filled in. This worked well for colloquialisms like EST *tere hommikust* ('good morning')–FIN *[hyvää] huomenta*. Most of the phrases (i.e. anything consisting of two or more words) still went unused. Where the EST-FIN database had a matching lemma, all the found Finnish counterparts were added as candidates for the translation equivalents under the first sense subdivision of the first homonym. The best unused candidates (ca. 40,000 words in Estonian) were used to form a complementary dictionary volume with skeleton articles filled in. These data will be used to grow the main dictionary in the future.

Instead of trying to guess the most appropriate homonym and sense from the limited data from the FIN-EST database, a second tool was provided as a first step in compiling the reversed dictionary. Our dictionary management system EELex does not support drag-and-drop editing, and voluminous dictionary entries require scrolling and copy-paste functions when one wants to move part of an entry to another position. To quickly delete, reorder and relocate generated Finnish translation candidates, we made a special tool that only displays enough information in the dictionary entry to indicate if, and under what sense, any of the translations belong (Figure 1). It only lists articles with

several senses (to move) and/or several provided translations (to reorder) and keeps track of finished articles. Essentially it still is just an alternative user interface alongside EELex, giving the opportunity to simultaneously do other tasks the traditional way while automating the tedious task of distributing the translation candidates between senses.

Figure 1: Drag-and-drop editing tool as an alternative user interface alongside EELex

3. The compilation of the dictionary

Due to the lack of time—the unpostponable date of the 100th anniversaries—we have to impose on using automated lexicography as effectively as possible. At the same time, we are trying not to brush aside the substantial principles of dictionary-making.

3.1 Automatically-compiled entries

First, we tried to estimate how many entries might be 'ready' from the very beginning. Inspection of the initial EST-FIN database (80,000 entries) gave us a preliminary picture (Figure 2). Estonian is a predominantly agglutinative language, which, when creating new senses, mostly makes use of morphologic derivation. Polysemy applies to about every tenth Estonian word (Langemets, 2010: 269). Roughly the same is true of Finnish, which belongs to the same group of Uralic languages. The total entries with a single meaning in the EST-FIN database is 73,000. The 'ready'-quality was assumed for the simplest words only, i.e. words with a single meaning, preferably with no subsenses, and with no more than three translation equivalents or examples to be translated. There were 14,389 such 'ready' entries in the EST-FIN database (Figure 2), incudingl 9,784

with one translation equivalent (e.g. EST *hambapasta* 'toothpaste', EST *kuupmeeter* 'cubic meter'); 3,053 with two equivalents (e.g. EST *sisepoliitika* 'home affairs'); and 902 with three equivalents (e.g. EST *hormoon* 'hormone' (see Figure 4), EST *ajaleht* 'newspaper').

The remaining part comprising entries with a single meaning (58,611) still needs further editing: the lexicographer's task is to select the proper equivalents, and select, edit and translate the examples. The most curious case is 69 (!) equivalents for the EST *pritsima* 'to splash' in the initial EST-FIN database. More than half of the entries (41,475) received no translation equivalents at all when reversing the database (e.g. EST *digitelevisioon* 'digital TV', EST *grammatikareegel* 'grammar rule', EST *alatähtsustama* 'understate').

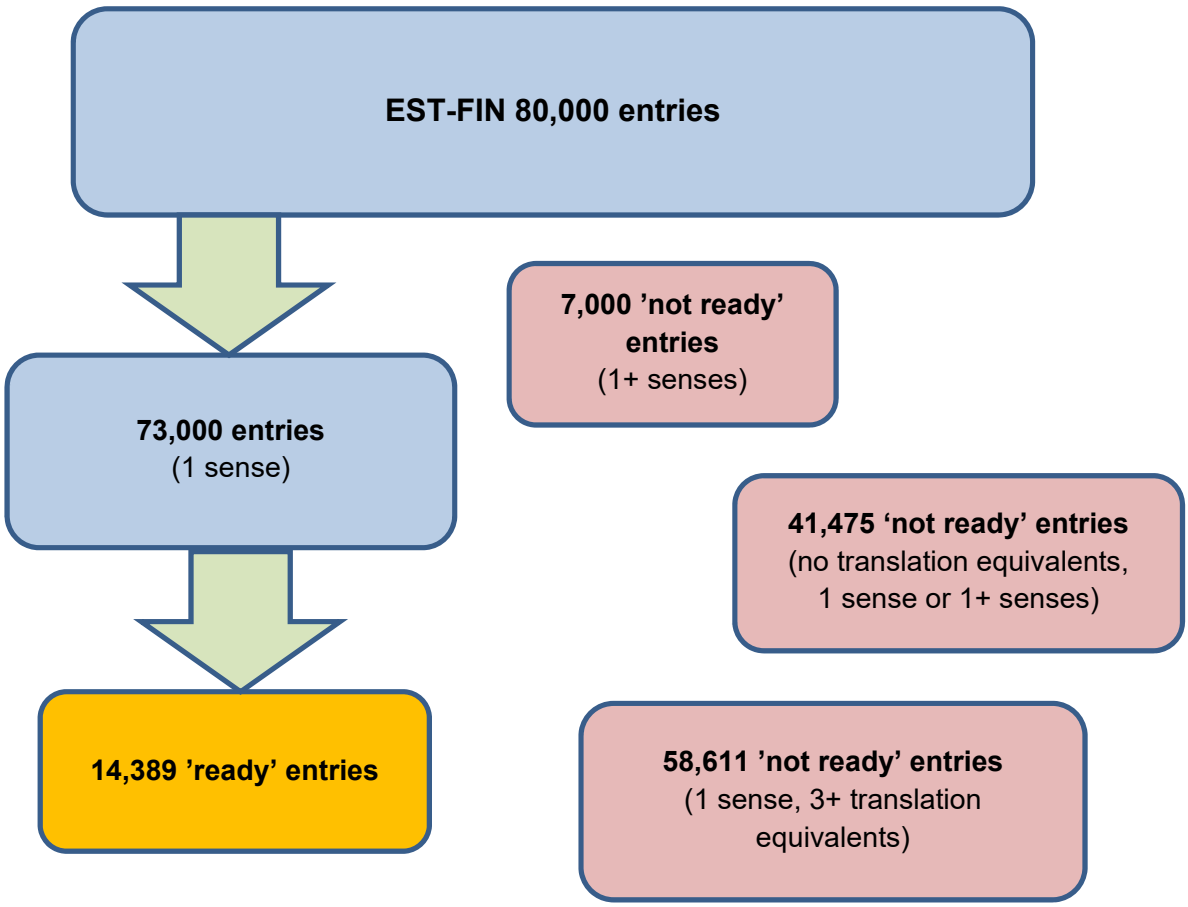


Figure 2: The initial EST-FIN database: automatically completed 'ready' entries vs. 'not ready' entries

Secondly, we have refined the initial 80,000-lemma list by acquiring frequency information from a corpus (5 frequency groups for approx. 50,000 top frequent lemmas). The lexicographers edit the articles along those groups starting with the most frequent words, i.e. the 5,000 lemmas of the Basic Dictionary of Estonian (2014).

The lexicographers' main tasks are to edit the translation equivalent candidates (selecting and reordering) or to find the candidates, if missing, and decide whether or not to translate the existing usage examples, i.e. whether the translation is justified for being too difficult for the user.

3.2 Bilingualizing the monolingual dictionary

Since the source language part of the EST-FIN database comes from the monolingual dictionary, the usual drawbacks of the bilingual dictionary (e.g. insufficient sense discrimination, lack of example sentences, to begin with, mentioned in Adamska-Salaciak & Kernerman, 2016: 276) should be avoided. Unfortunately, drawbacks of another type remain.

The monolingual dictionary of Estonian is not aimed at learners but at native speakers of Estonian. The dictionary describes current Estonian and focuses on sense discrimination. It is being compiled using etTenTen corpus and the Sketch Engine tool (Kilgarriff et al., 2004). The examples are meant to illustrate the senses, they are real as well as natural in Estonian, but in many cases, are definitely not good examples for learners. For another, the Estonian Collocations Dictionary (ECD, to appear in 2018, see Kallas et al., 2015) grammatical constructions (collocations) have been extracted from the corpus. Since the ECD work is in progress, the database is inadequate so far and not usable for the EST-FIN database.

There are 95,000 usage examples in the EST-FIN database. The average number of examples per entry is 1.1. Around half of the entries have no examples in the database. We are aware from the previous studies (Frankenberg-Garcia 2012, 2014, quoted in Lew, 2015: 5) that users find three examples per sense significantly more helpful than just a single example. The problems remain: how could we manage to translate all necessary examples? How could we obtain enough examples for all the entries? How might we obtain good examples for the bilingual dictionary?

Bilingualizing the dictionary involves bilingualizing the metalanguage (domain and style labels, grammatical information etc.) as well as—to some extent—the explanations. Bilingualizing explanation means first and foremost simplifying the definitions: the definitions in a monolingual dictionary are usually much longer and more complex than in a bilingual dictionary. We have decided to preserve the semantic structure of the treatment of the source language, but for better understanding of Estonian (as L2) we have shortened many definitions into glosses.

3.3 Directionality of the dictionary

Traditionally all (paper) dictionaries have been compiled to fulfil the needs of all

conceivable users trying to solve several different tasks. A good bilingual dictionary should enable both understanding L2 and producing L2, though the most important task seems to be the latter—an opinion supported by many researchers (see Adamska-Salaciak & Kernerman 2016). The best dictionary would be 'addressed specifically to the native speakers of one of its two object languages' (Adamska-Salaciak & Kernerman, 2016). The dictionary should somehow selectively separate information for different purposes, i.e. it should be monodirectional.

For Estonian users, the EST-FIN dictionary functions as a L1–L2 dictionary, helping L1 users to talk about specific phenomena of his/her own culture. There are several concepts (senses) that are not lexicalized in L2 (e.g. EST *akadeemiline tund* ('(in the universities:) 45 mins'), EST *prääник* ('a thick soft spicy biscuit')).

For Finnish users, the EST-FIN dictionary functions as the receptive L2–L1 dictionary, helping to render the L2 meanings.

For production of L2 (Estonian or Finnish), it should contain first and foremost collocational information as well as good examples.

Modes of provision of semantic information. In the EST-FIN database, the semantic information explaining the lexical items of the source language is placed into definitions (EST, rarely FIN), equivalents (FIN) and examples (EST, rarely FIN). Estonian seems to dominate over Finnish (Figure 3, Finnish underlined) in the entries for single words (EST *aadress* 'address') but the phrasal verbs and idioms (EST *käsi peseb kätt* 'one hand washes another') are as carefully explained in Finnish as in Estonian. There are about 6,000 multiword units in the EST-FIN database and as part of the headword list they are treated likewise.

aadress 'address' ‹s›
isiku elupaiga või asutuse asukoha andmed
osoite
◦ ***kodune aadress*** kotiosoite
▪ (arvutivõrgus)
www-osoite, osoite
◦ ***koduleheküljе aadress*** kotisivun osoite
▪ [kellegi] **aadressil**
kellegi kohta või pihta
[jotakin] kohtaan, [johonkin] liittyen, osoitettuna [jollekin]
◦ ***kriitika valitsuse aadressil*** kritiikki hallitusta kohtaan

käsi peseb kätt ‘one hand washes another’
ütlus selle kohta, et teenele vastatakse teenega
ilmaus siitä, että palvelukseen vastataan palveluksella
käsi kättä pesee KUV, käsi käden pesee KUV

Figure 3: Providing semantic information in both languages (Finnish translations underlined, other symbols: ◦ usage example; ▪ subsense; FIN definition)

Modes of treating synonyms. If synonymy is presented for the concept (in the Estonian part) and is designated by different terms (lemma and its synonyms, marked with = in the database) and described by the same definition in the particular sense, then the semantic information (incl. Finnish translations) remains the same for all counterparts (Figure 4: EST *hormoon* ‘hormone’ = *sise|nõre* = *inkreet*). We have agreed that the domain label (in Finnish, e.g. FYSIOL) functions as a semantic gloss for the Finnish, so it would not be necessary to translate Estonian definitions.

hormoon ‘hormone’ ‹s›
aine, mis reguleerib inimese ainevahetust ning organismi talitlust
(= sise|nõre, inkreet)
hormoni FYSIOL, sisä+erite FYSIOL, umpi+erite FYSIOL

sise|nõre ‘hormone’ ‹s›
aine, mis reguleerib inimese ainevahetust ning organismi talitlust
(= sise|nõre, inkreet)
hormoni FYSIOL, sisä+erite FYSIOL, umpi+erite FYSIOL

inkreet ‘hormone’ ‹s›
aine, mis reguleerib inimese ainevahetust ning organismi talitlust
(= sise|nõre, inkreet)
hormoni FYSIOL, sisä+erite FYSIOL, umpi+erite FYSIOL

Figure 4: Treating synonyms in the dictionary (Finnish translations underlined)

Collocational information. Collocations and other lexical bundles are not systematically and explicitly treated in the EST-FIN database. As mentioned above, the work on the Estonian Collocations Dictionary is in progress.

Modes of provision usage examples. The examples are essential for all types of learners as well as L2 users. It was attested 20 years ago, that the dominance of bilingual dictionaries is greater for L1–>L2 translation (i.e. for producing L2) than for L2–>L1 translation (Atkins & Varantola, 1997). The examples of the EST-FIN database are meant to illustrate the senses of the monolingual dictionary, i.e. in the EST-FIN database they function for L2–>L1 translation.

Translation of dictionary examples has not always been seen as the best solution for a bilingual dictionary (Adamska-Salaciak, 2006; Hmeljak Sangawa & Erjavec, 2012). A corpus is needed to provide typical L2 examples in the (unidirectional) bilingual dictionary. Adamska-Salaciak (2006) favours presenting L2 examples only, with the exception of difficult cases (2006: 494):

> Naturally, even in dictionaries whose examples are normally left untranslated, exception must be made for sentences or parts thereof which might be too difficult for the average user to interpret on their own.

And another statement from a lexicographer (emailed to one of the authors, January 2017) keeping in mind producing L2:

> Personally, I use bilingual dictionaries pretty rarely. Google is often fine, especially for phrasal expressions: I test what I intend to say against data on the web.

So, would it not be marvellous if our user could have real (authentic) material at hand? Since we cannot give the user access to a large high-quality bilingual (parallel) corpus as well as following Adamska-Salaciak (2006), we should instead provide our users with good L2 examples.

Next, we will discuss the possibilities of obtaining good examples for both languages.

4. Good examples for Estonian

Presenting authentic sentences in dictionaries is a common practice in modern lexicography. One of the possibilities for extracting authentic examples from corpora is to use GDEX (Kilgarriff et al., 2008)—a software part of Sketch Engine (Kilgarriff et al., 2004). GDEX evaluates syntactic and lexical features of sentences and sorts concordances according to how perfectly they meet all the relevant criteria. As a result, GDEX offers a list of sentences: the better candidates are at the top of the list and the not-so-good ones at the bottom. The theoretical framework for GDEX development is proposed in Kilgarriff et al. (2008) and Kosem et al. (2011, 2013).

GDEX was developed as a set of classifiers for specific features and it was first used in the preparation of an electronic version of the Macmillan English Dictionary (Macmillan 2002, 2007). All features are quantifiable, e.g. sentence length, word length, presence or absence of certain words or non-words, the number of pronouns in the sentence etc. Each feature has its own individual value and GDEX counts them in a fixed way. It ranks the sentence with a score from 0 to 1. The specification of measured features and the way in which they are combined is defined in files called GDEX configurations (Figure 5).

```
formula: >
    (50 * is_whole_sentence() * blacklist(words, illegal_chars) * blacklist(lemmas, parsnips)
    + 50 * optimal_interval(length, 10, 14)
    * greylist(words, rare_chars, 0.1)
    * greylist(tags, pronouns, 0.1)
    ) / 100
variables:
    illegal_chars: ([<|\]\[>/\\^@])
    rare_chars: ([A-Z0-9'.,!?](;:-])
    pronouns: PRON.*
    parsnips: ^(tory, whisky, jesus, cowgirl, meth, commie, bacon)$
```

Figure 5: Syntax of GDEX configuration files[1]

While some of the parameters apply to all languages (e.g. sentences start with a capital letter and end with a punctuation mark), some are language-specific (e.g. sentence length, keyword position etc.). Therefore, it is reasonable to modify parameters according to the languages that the GDEX configuration will be applied upon.

First, GDEX configuration for Estonian was developed in the Institute of the Estonian Language in 2014, in collaboration with Lexical Computing Ltd., for the automatic extraction of the Estonian Collocations Dictionary (ECD) database. ECD is a monolingual dictionary aimed at learners of Estonian as a foreign or second language at the upper intermediate and advanced levels. The Estonian configuration is being continuously developed. The latest version was set up in 2016 based on research carried out under ISCH COST Action IS10305 European Network of e-Lexicography during a short term scientific mission in the University of Ljubljana.

The Estonian corpus (560 mio words) contains of two parts: the Estonian Reference Corpus (biased heavily towards newspaper texts), and the web corpus, crawled by SpiderLing in 2013. Syntagmatic relations of content words are described as lexico-grammatical constructions defined by means of morphosyntactic categories (phrase type, part of speech, inflectional categories). The Estonian Sketch Grammar has been worked out by Kallas (2013). The corpus was tagged for sentences, clauses and morphology (POS-tag and inflections) by Filosoft Ltd (ESTMORF).

In developing GDEX for Estonian, the needs of language learners have always been considered. Sentences in learner dictionaries should ideally be short, not syntactically and grammatically complex, include frequent words, help the learner to understand the meaning of an unknown word, and/or show the collocation in its typical context.

[1] https://www.sketchengine.co.uk/syntax-of-gdex-configuration-files/ (25.5.2017).

Parameters for Estonian are as follows (Koppel 2017):

- Starts with a capital letter and ends with a punctuation mark;
- Sentence length is 4 to 20 tokens;
- Optimal sentence length is 6 to 12 tokens;
- Contains a verb;
- Maximum word length is 20 characters;
- Certain characters (e.g. <|\]\[>/\\}{^@•·*#=_~) are prohibited and certain characters (e.g. ;:“„”«»"'×…§-) are penalized;
- Certain words (e.g. *pigem* ‘rather’, *teisisõnu* ‘in other words’, *seetõttu* ‘for that reason’), word pairs (e.g. *seda enam* ‘even more’, *teiste sõnadega* ‘in other words’, *teisest küljest* ‘on the other hand’) and sentence initial tags (e.g. conjunction, abbreviation, interjection) are prohibited from appearing in the beginning of the sentence;
- Words with a frequency of less than 5 are prohibited;
- Lemmas with a frequency of less than 1000 are penalized;
- Keyword repetition is prohibited;
- Sentences including pronouns, words from graylist (e.g sensitive words, profanities), abbreviations, proper names, certain non-finite constructions are penalized;
- Sentences containing more than 2 verbs, more than 1 adverb, more than 1 pronoun, more than 1 conjunction, more than 1 proper name, more than 1 numeral and more than 1 comma are penalized.

Figure 6 shows the GDEX output after the latest version for Estonian is implemented.

Figure 6: GDEX output for Estonian lemma *raamat* ‘book’

5. Good examples for Finnish

In Finnish, the goal is to find good examples of translation equivalents of Estonian headwords with the help of GDEX configuration (Heinonen, 2015). The configuration for Finnish is based on the one developed by Kristina Koppel for Estonian (see above).

In general, the same configuration works well for both languages. We prefer short, simple and context-free examples: conjunctions, sentence-initial connectives and anaphoric elements are unfavourable. At first sight, one might think that there is not much else to do except replace the original Estonian lexical items by Finnish words. For instance, an Estonian connective expression *teiste sõnadega* 'in other words' would be replaced by its Finnish equivalent *toisin sanoin*. However, a considerable part of the Finnish data used in this project represents informal register. Finland's strict copyright legislation is partly to blame for this since it strongly favours the use of freely-accessible web-pages for corpora. In any case, stylistic variation poses a problem which is not as noticeable on the Estonian side. It is possible that Finnish speakers tend to write more informally than Estonians, or that there is a bigger difference between the standard and the vernacular in Finnish than in Estonian. Whichever is the case, it is more confusing than helpful if the examples contain words and expressions that are not even recorded in a standard Finnish dictionary. An ideal solution would be a grammatically augmented dictionary that could deal with variation in words and inflectional affixes.

The Finnish corpus is tagged morphologically but not syntactically, and there is no way to fix its colloquial syntactic patterns. However, what can be done, is to ban most common colloquial word forms by including them in the list of "bad words", which is one of the parameters of GDEX and is originally used for excluding inappropriate words.

In the Finnish GDEX configuration, this is achieved by listing such prevalent items as spoken forms of general verbs, pronouns and some other grammatical words:

- *oon, oot, oo, tuu, paan, sais, vois*, etc. (informal forms of frequent verbs)
- *mä, sä, toi, noi, mulle, sulla, tolle*, etc. (informal forms of pronouns)
- *vaik, nii, niiku, ku, kans*, etc. (informal forms of conjunctions and adpositions)

Since these items tend to co-occur with other informal words and structures, this is in fact a rather efficient way of preventing unwanted example sentences to surface.

The task of obtaining good Finnish examples is complicated also by the fact that many seemingly fine sentences are hampered by morphosyntactic misanalysis. For instance, out of a test sample of 30 sentences intended to illustrate the use of the verb *kaupata* 'to trade', only seven were in fact occurrences of this lemma. The remaining sentences

(23/30) instead contained the noun *kauppa* in its various senses ('a store', 'a deal', 'commerce', etc.). The reason for this is that words *kaupata* and *kauppa* share the same stem, and some frequent inflectional forms are ambiguous. The same situation is also common in English: the form *shops* can be either a noun in the plural or a verb in the third person singular. Since the noun *kauppa* is more common than the verb *kaupata*, its occurrences dominate in the concordance. Furthermore, such ambiguities bring about mismatches in a bilingual context if the intended translation equivalents do not show up but are instead replaced with misanalyzed forms in corpus examples.

Figure 7 displays a list of top examples for the word *kirja* ('a book'). The top score is given to a sentence *Kirjassa kaksi osaa, ei erillisiä lukuja.* This sentence should not score as highly since it does not have a finite verb. Its literal translation is: 'two parts in the book, no separate chapters'. Again, the problem lies in an ambiguous word form: this time the word *osaa*, which has two readings, 'part' in the partitive singular, or 'can, be able' in the third person singular.

Old rank	Rank	Sentence	Old score	Score
1	1	Kirjassa kaksi osaa, ei erillisiä lukuja.	1.00	1.01
2	2	Samoin kirjan kuvauksissa näkyy ahdistus, joka liittyy oman erillisen minuuden löytämiseen.	0.99	1.00
3	3	Sisällön laadun kannalta onkin ikävää, että itse kirja on esineenä kehnosti tehty.	0.98	0.99
7	4	Toinen vastasi, että kyseinen kirja oli juuri kesken.	0.94	0.98
4	5	Mitä Platon on minulle opettanut, kysyy Pietarinen kirjansa nimessä.	0.96	0.97
13	6	Olen lukenut kirjan parikymmentä vuotta sitten, mutta jotkut asiat siitä jäivät mieleen.	0.93	0.96
6	7	Kirja kertoo, miten maahiset selvisivät uudenlaisessa maailmassa.	0.95	0.96

Figure 7: GDEX output for Finnish lemma *kirja* 'book'

However, Figure 7 shows that even with some faulty analyses, the parameters succeed in ordering the sentences in the corpus in a reasonable way. Scrolling down the list, one encounters lengthy, complex, or fragmentary sentences.

The GDEX for Finnish is still being developed. This is being performed in the GDEX editor[2], which is a standalone tool of Sketch Engine developed by Lexical Computing Ltd. software developer Jan Michelfeit. The GDEX editor enables comparisons between two settings of parameters in parallel, and this is used so that the Estonian-based configuration has been taken as a starting point and it is modified step by step towards a configuration that arranges Finnish data in an optimal way. In Figure 7, the results from Estonian-based GDEX (with few lexical additions) are labelled as "Old rank" and "Old score". Once a parameter is modified, the changes to its ranking and scores can be immediately calculated. As can be seen, the tops of the lists are, in any case, very

[2] https://beta.sketchengine.co.uk/gdex_editor

similar. After the screenshot in Figure 7 was taken, the word *toinen* ('other', 'another', 'second') was tentatively added to the list of prohibited words in a sentence-initial position, with the effect of dropping the rank of the sentence number 4 *Toinen vastasi, that...* ('The other replied that...') to number 25.

6. Conclusion

In the semi-automated compilation process of the Estonian-Finnish (EST-FIN) dictionary we have roughly followed the same steps as mentioned by Gantar et al. (2016: 201).

We reused the previous lexical databases. The database was generated combining two existing databases: the source language part was formed from the database of the monolingual general Dictionary of Estonian (DicEst, to appear in 2018/2019), and the target language part from the database of the Finnish–Estonian dictionary (FIN-EST, 2003). It is worth mentioning that the FIN-EST dictionary did not start from scratch: the base for the source language (Finnish) came from another bilingual dictionary, Finnish–Swedish dictionary (1997, Helsinki). The Estonian lexicographers worked with the XML-like 'electronic' manuscript, filling in the slots for translation equivalents as well as translating the usage examples for the FIN-EST dictionary.

We refined the initial 80,000-lemma list by acquiring frequency information from a corpus (5 frequency groups for approx. 50,000 top frequent lemmas).

The best unused candidates from the FIN-EST database (ca 40,000 words in Estonian), which so far constitute the complementary EST-FIN dictionary volume with skeleton articles filled in, will be used to grow the main dictionary in the future.

We will extract L2 example sentences—daydreaming of getting hold of good examples for both languages, Estonian and Finnish. The first GDEX configuration for Estonian was developed in the Institute of the Estonian Language in 2014 in collaboration with Lexical Computing Ltd. Since then it has been under continuous development. The latest version was set up in 2016. The GDEX configuration for Finnish is based on the Estonian one; it still needs developing.

Subsequently, likely after the finalization of both projects (EST-FIN and ESTDic) we will link the collocations database to these dictionaries to fulfil the productive needs of advanced learners of Estonian as well as the needs of L2 users. The overall development plan at the Institute of the Estonian Language concerning dictionaries and (terminological) databases is to change over to the standardized (Unified) Data Model for better presentation and linking of the lexicographic information congregated at our Institute.

Finally, we have thought of using a free/open-source Machine Translation platform Giellatekno Apertium to provide translations for usage examples (Kaalep et al. 2017). Figure 7 displays the translation of the first sentence in the Estonian GDEX output (Figure 6). Those understanding Finnish might get the feeling that the translation system does not work well at all (Figure 7). However, we still have a well-grounded hope, as the rule-based translation system will be complemented by the real 90,000-lemma Finnish–Estonian dictionary (2003, 2 vols) instead of the small 15,000-lemma dictionary compiled automatically via other pairs of languages.

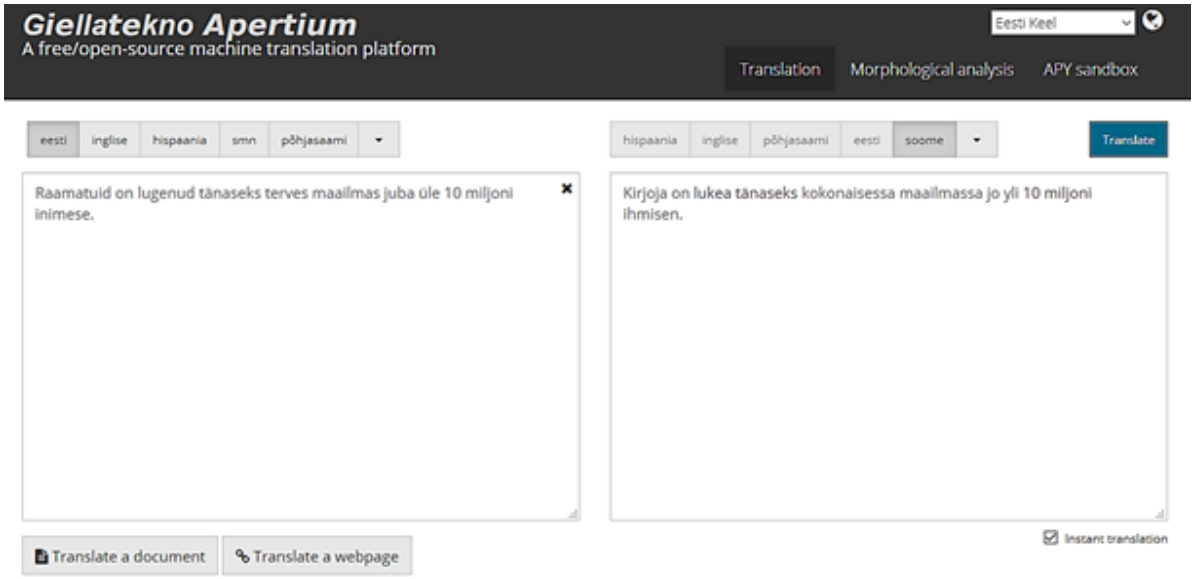


Figure 7: Machine translation platform Giellatekno Apertium: from Estonian to Finnish

7. Acknowledgements

We would like to thank Jelena Kallas for her constructive comments and suggestions. The research was partly funded by ISCH COST action COST-STSM-IS1305-27016 European Network of e-Lexicography.

8. References

Adamska-Salaciak, A. (2006). Translation of Dictionary Examples – Notoriously Unreliable? In *Proceedings of the 12th EURALEX International Congress*. Torino, Italy, 493–501.

Adamska-Salaciak, A. & Kernerman, I. (2016). Introduction: Towards better dictionaries for learners. In *International Journal of Lexicography*, 29(3), 271–278. doi: 10.1093/ijl/ecw033

Atkins, S. & Varantola, K. (1997). Monitoring dictionary use. In *International Journal of Lexicography*, 10(1), 1–45.

DicEst = The Dictionary of Estonian (to appear in 2018/2019). Institute of the Estonian Language.

ECD = The Estonian Collocations Dictionary (to appear in 2018). Institute of the Estonian Language.

EST-FIN = Estonian-Finnish database (to appear in 2018). Institute of the Estonian Language.

ESTMORF = Eesti keele morfoloogiline analüsaator [Morphological Analyzer of Estonian]. Filosoft OÜ. http://www.filosoft.ee/html_morf_et/morfoutinfo.html (3.7.2017).

FIN-EST = Soome-eesti suursõnaraamat. Suomi–viro-suursanakirja (2003). [Finnish-Estonian dictionary.] 2 vols. Tallinn: Eesti Keele Sihtasutus.

Gantar, P. & Kosem, I. & Krek, S. (2016). Discovering automated lexicography: the case of the Slovene database. In *International Journal of Lexicography*, 29 (2), 200–225. doi: 10.1093/ijl/ecw014

Giellatekno Apertium = Giellatekno Apertium: A free/open-source machine translation platform http://gtweb.uit.no/mt-testing/index.est.html?dir=est-fin#translation

Heinonen, T. (2015). Development of Sketch Grammar and GDEX (Good Dictionary Example) for Finnish. Scientific Report of Short Term Scientific Mission, COST STSM.

Hmeljak Sangawa, K. & Erjavec, T. (2012). JaSlo: Integration of a Japanese-Slovene bilingual dictionary with a corpus search system. In *Acta Linguistica Asiatica*, 10 (3). http://revije.ff.uni-lj.si/ala/article/view/223

Kaalep, H.-J. & Tyers F. M. & Trosterud, T. (2017). Soome-eesti reeglipõhise masintõlkesüsteemi DEMO. [Finnish-Estonian rule-based machine translation DEMO.] In Abstracts of EAAL 16 th Annual Conference, April 20-21, 2017, Tallinn, Estonia. Accessed at: https://www.rakenduslingvistika.ee/wp-content/uploads/
2016/04/Teesid-2017-3.pdf (18.5.2017)

Kallas, J. (2013). Syntagmatic relationships of Estonian content words in corpus and pedagogical lexicography. PhD thesis.

Kallas, J. & Koppel, K. & Tuulik, M. (2015). Korpusleksikograafia uued võimalused eesti keele kollokatsioonisõnastiku näitel. [New possibilities in corpus lexicography based on the examples of the Estonian Collocation Dictionary.] In *Eesti Rakenduslingvistika Ühingu aastaraamat*, 11, 75–94. doi: 10.5128/ERYa11.05

Kilgarriff, A. & Rychly, P. & Smrž, P. & Tugwell, D. (2004). The Sketch Engine. In G. Williams, S. Vessier (eds.) *Proceedings of the XI Euralex International Congress.* Lorient: Université de Bretagne Sud, 105–116.

Kilgarriff, A. & Husák, M. & McAdam, K. & Rundell, M. & Rychlý, P. (2008). GDEX:

Automatically finding good dictionary examples in a corpus. In E. Bernal, J. DeCesaris (eds.) *Proceedings of the 13th EURALEX International Congress*. Barcelona: Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra, 425–432.

Koppel, K. (2017). Heade näitelausete automaattuvastamine eesti keele õppesõnastike jaoks. [Automatic detection of good dictionary examples in Estonian learner's dictionaries.] In *Eesti Rakenduslingvistika aastaraamat*, 13, 53−71.

Kosem, I. & Husák, M. & McCarthy, D. (2011). GDEX for Slovene. In *Proceedings of eLex 2011*, 151–159. http://elex2011.trojina.si/Vsebine/proceedings/eLex2011-19.pdf

Kosem, I. & Gantar, P. & Krek, S. (2013). Automation of lexicographic work: an opportunity for both lexicographers and crowd-sourcing. In I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets, & M. Tuulik (eds.) *Electronic lexicography in the 21st century: thinking outside the paper*. Proceedings of the eLex 2013 conference, 17–19 October 2013, Tallinn, Estonia, 17–19. http://eki.ee/elex2013/proceedings/eLex2013_03_Kosem+Gantar+Krek.pdf

Kovář, V. & Baisa, V. & Jakubíček, M. (2016). Bilingual Word Sketches. In *International Journal of Lexicography*, 29 (3), 339–352. doi: 10.1093/ijl/ecw029

Langemets, M. (2010). *Nimisõna süstemaatiline polüseemia eesti keeles ja selle esitus eesti keelevaras.* [Systematic polysemy of nouns in Estonian and its lexicographic treatment in Estonian language resources.] Tallinn: Eesti Keele Sihtasutus.

Lew, R. (2015). Dictionaries and their users. In *International Handbook of Modern Lexis and Lexicography.* Berlin-Heidelberg: Springer-Verlag, 1–9. doi: 10.1007/978-3-642-45369-4_11-1

Macmillan 2002, 2007 = *Macmillan English Dictionary for Advanced Learners* (First and Second editions). 2002, 2007. Rundell, ed. Macmillan, London.

Maks, E. (2007). OMBI: the practice of reversing dictionaries. In *International Journal of Lexicography*, 20(3), 259–274. doi: 10.1093/ijl/ecm028

Viks, Ü. (2008). Eesti-X-keele sõnastik ja grammatika. [Estonian-X dictionary and grammar.] In *Eesti Rakenduslingvistika aastaraamat*, 4, 247−261.

The Croatian Web Dictionary Project – Mrežnik

Lana Hudeček and Milica Mihaljević

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

Abstract

Croatia is still one of the countries which do not have a publically-available online dictionary of their national language compiled according to the rules of contemporary e-lexicography. This paper presents the current, as well as the planned, activities of the Croatian Web Dictionary—MREŽNIK project. The aim of the MREŽNIK project is to compile a free, monolingual, corpus-based, hypertext, easily searchable, online dictionary of Croatian standard language with three modules (for adult native speakers: 10,000 entries, for school children: 3000 entries, and for foreigners: 1000 entries). The dictionary entries will contain links to repositories which will be created as a part of this project (Linguistic Advice Repository: 300 entries, Conjunction Repository: all conjunctions, The Idiom Etymology Repository: 50 idioms, The Repository of Ethnics and Ktetics (place names, feminine and masculine names of the inhabitants and corresponding adjectives): 300 entries) as well as repositories which have already been compiled within other projects at the Institute of Croatian Language and Linguistics: The Verb Valence Repository, The Collocation Repository, The Croatian Terminology Repository (Struna), The Croatian Metaphor Repository, and the website Better in Croatian. The dictionary will be based on these two corpora: the Croatian Web Repository and the Croatian Web Corpus. The dictionary will be compiled using TLex. SketchEngine, a corpus manager and analysis program, and Tickbox Lexicography will be used to search the corpora and extract data from it. As a part of the project, a reverse dictionary will be compiled.

Keywords: e-lexicography; web dictionary; corpus-based dictionary; Croatian language; dictionary grammar

1. Introduction

The fact that Croatia is still one of the countries which do not have a publically-available online corpus-based dictionary of their national language compiled according to the rules of contemporary e-lexicography, or systematic research on e-lexicography, was the reason for starting a new project: Croatian Web Dictionary—MREŽNIK. The project started on the 1st March 2017 and the duration of the project is four years. The result of the MREŽNIK project will be a free, monolingual, hypertext, easily searchable, online dictionary of the Croatian standard language. This dictionary has three different modules: a dictionary for adult native speakers of Croatian, a dictionary for elementary school children, and a dictionary for foreigners. As we are still in the first half of the first year of the project, in this paper we will present the dictionary grammar and style manual for three different modules, which are being compiled at the moment, as well as connected databases and computer tools. Some of the connected databases have already been compiled while others will be compiled at the same time with the dictionary. Some of test definitions and lists of labels have also already been compiled, as well as the pilot reverse dictionary based on the pilot word-list. We will also compare MREŽNIK with *Wrječnik* and *Hrvatski jezični portal* and explain why MREŽNIK will

not be connected with these dictionaries.

2. Foreign E-dictionaries

In modeling the Croatian Web Dictionary many similar foreign dictionaries have been consulted, e.g. *elexiko* (http://www.owid.de/wb/elexiko/start.html) of the Institute of German Language, *Wielki słownik języka polskiego* (http://www.wsjp.pl/) of the Institute of Polish Language, *Swedish online dictionary* (http://spraakbanken.gu.se/karp), *Das Wortauskunftssystem zur deutschen Sprache in Geschichte und Gegenwart* (https://www.dwds.de/), *Algemeen Nederlands Woordenboek* (http://anw.inl.nl/), *Online-Bildwörterbuch* (http://www.bildwoerterbuch.com/), etc. In particular, *elexiko* was used as an inspiration for modeling some of the dictionary fields presented below.

3. Croatian E-dictionaries

While planning the dictionary grammar and fields, all existing Croatian printed dictionaries and e-dictionaries were consulted.[1] However, MREŽNIK will not be in any way connected with *Wječnik* or with *Hrvatski jezični portal* (Croatian Language Portal; HJP). The reasons for this are numerous: 1. While these two dictionaries are descriptive dictionaries of the Croatian language including dialects, MREŽNIK is a descriptive and prescriptive dictionary of Standard Croatian. 2. While MREŽNIK is corpus-based (see below) *Wječnik* and *Hrvatski jezični portal* are not. 4. HJP presents an online dictionary, which is the result of the collaboration between Novi Liber and Srce (http://hjp.znanje.hr/), today owned by the publishing house Znanje. It is an online version of *Hrvatski enciklopedijski rječnik* (*Croatian Encyclopedic Dictionary*), which was not compiled as an online dictionary, but is a printed dictionary published by the publishing house Novi Liber, and sold in the printed version for the last 15 years. This fact is the reason for many drawbacks of this dictionary. This online dictionary has relatively inefficiently interconnected entries, i.e. only links to other headwords in the etymological part of the entry, and is out of date as it has not been revised for a long time. 5. *Wječnik* is a *Wiktionary* project, a collaborative project, based on crowdsourcing to produce a free-content multilingual dictionary. It is a lexical project based on Wikipedia software (Wikimedia).

MREŽNIK, on the other hand, is a scientific project; the collaborators of MREŽNIK are experienced scientists and lexicographers. The dictionary is compiled taking into account semantic relations and the systematic nature of language. We will illustrate this using one simple example, comparing the first definition of seasons in all three dictionaries:

[1] For more on Croatian e-lexicography see Jermen et al. (2015); Štrkalj Despot & Möhrs (2015).

	Wječnik	**Croatian Language Portal**	**MREŽNIK**
zima winter	jedno od četiriju godišnjih doba, kalendarski (na sjevernoj Zemljinoj polutci) traje od 21. odnosno 22. prosinca do 21. ožujka, dolazi između jeseni i proljeća.	**a.** kalendarsko doba od 22. prosinca do 21. ožujka **b.** jedno od četiriju godišnjih doba, između jeseni i proljeća	godišnje doba koje najčešće na sjevernoj hemisferi počinje 22. prosinca i traje do 21. ožujka
proljeće spring	Jedno od četiriju godišnjih doba. Kalendarski traje od 22. ožujka do 22. lipnja	**a.** kalendarsko doba od 21. ožujka do 21. lipnja **b.** jedno od četiri godišnja doba (između zime i ljeta)	godišnje doba koje najčešće na sjevernoj hemisferi počinje 22. ožujka i traje do 21. lipnja
ljeto summer	Jedno od četiriju godišnjih doba. Kalendarski traje od 21. lipnja do 22. rujna.	kalendarsko doba od 21. lipnja do 22. rujna, jedno od četiri godišnja doba	godišnje doba koje najčešće na sjevernoj hemisferi počinje 22. lipnja i traje do 22. rujna
jesen autumn	godišnje doba koje traje od 23. rujna do 21. prosinca	**a.** kalendarsko doba između 23. rujna i 21. prosinca **b.** jedno od četiri godišnja doba	godišnje doba koje najčešće na sjevernoj hemisferi počinje 23. rujna i traje do 21. prosinca

Table 1: Comparison of the definitions of four seasons in three dictionaries

From Table 1, it is obvious that only MREŽNIK has all definitions structured in the same way. All definitions start with the same hypernym (godišnje doba – season) written in lowercase letters; definitions have the same syntactic structure and consist only of one sentence; and they give the same data, the date of the beginning and the end of the season. The analysis of semantic fields having more members would show even greater differences between dictionary definitions.

4. Corpus

The Croatian web dictionary MREŽNIK is based on two Croatian corpora: *Croatian Web Repository*, the corpus of the Institute of the Croatian Language and Linguistics

(http://riznica.ihjj.hr/index.hr.html) and the *Croatian Web Corpus* (http://nlp.ffzg.hr/resources/corpora/hrwac). These corpora are managed by the corpus tool Sketch Engine. The corpus is always checked by the lexicographer, as the dictionary is corpus-based and not corpus-driven. One of the reasons for this is: "For most native-speaker dictionaries, corpora are still inadequate in size and are also slightly out of date by the time they are available to lexicographers. So 'reading-and-marking' of the latest newspapers and magazines, and attentive listening to radio and television broadcasts, are still a necessity." (Brown 2006: 250–254). Although this may be less and less true for big languages like English, it is still true for Croatian.

The corpus-based principle was used in *elexiko* as Klosa points out: *elexiko* is basically corpus-based, i.e. there are no lexicographic entries in *elexiko* which do not come from the *elexiko*-Korpus and there is no information that is simply taken over from other dictionaries. (Klosa 2011: 16.) However, she also points out some negative aspects of this principle while some topics appear more often in the newspaper corpus. (Klosa 2011: 58.) and *elexiko*-Korpus is basically a newspaper corpus. Another reason for using the corpus-based and not corpus-driven approach is the normative aspect of MREŽNIK. The corpus will provide bases for creating the list of headwords, differentiating meanings, selecting derivatives, compounds and collocations, composing definitions, selecting or creating examples (depending on the module). Three different modules have three different approaches to the corpus as will be shown below.

5. Three Modules

MREŽNIK is a scientific dictionary which is also user-friendly and fulfils different needs of different user groups. Thus, the dictionary gives as the information the user needs and is connected to many other databases. A similar idea when speaking about *elexico* has been stated in Haß (2005: 3) as she points out that *elexico* can fulfill different user needs and interests and this approach would not be possible in a printed dictionary.

MODULES
module for adult native Croatian speakers – 10 000 entries
module for school children – 3000 entries
module for foreigners – 1000 entries

Table 2: Three modules of MREŽNIK with the number of entries

MREŽNIK consists of three separate modules which are connected by the fact that all given data is coordinated and synchronized. However, each module functions as a separate dictionary compiled for a different target group of users. The first module is a dictionary for adult native speakers of Croatian consisting of 10,000 entries. The second module is a dictionary for elementary school children consisting of 3000 entries,

and the third is a dictionary for foreigners consisting of 1000 entries. Each dictionary module has different dictionary grammar which is based on the specific needs of the dictionary user.

Different modules have a different approach to examples from the corpus. In the dictionary for adult users, the lexicographer will select the examples from the corpus. Each meaning and definition will have examples from the corpus. The approach to the corpus of three different modules is shown in Table 3:

Three modules	**Module for adult native Croatian speakers**	**Module for elementary school children**	**Module for foreigners**
Explanation	The headword will be a direct link to the corpus. For each meaning, examples will be taken from the corpus. These examples will be selected by the lexicographer. In addition, each headword will have a link to the corpus.	For each meaning very simple examples will be devised by the lexicographer.	For each meaning the example will be taken from the corpus and simplified by the lexicographer.
Example	Djelatnici Inspektorata Ministarstva zaštite okoliša i kriminalistička policija na zgarištu su u Parku prirode Kopački rit proveli više od deset sati kako bi se utvrdio uzrok požara koji je u nedjelju navečer poharao ovaj baranjski biser i to njegov najvredniji dio - poseban ZOO rezervat u kojemu se gnijezde rijetke i zaštićene vrste ptica.http://riznica.ihjj.hr/philocgi-bin/search3t?dbname=Cijelihr&word=ptica&OUTPUT	Ptica leži na jajima u gnijezdu.	U parku se gnijezde rijetke vrste ptica.

Table 3: Approach to the corpus in three different modules

6. Word List and the Corpus

Three different modules will have three different lists of headwords. The starting point for the word list for adult native speakers is the corpus from which the 10,000 most-frequent lemmas will be extracted. The words extracted from the corpus will be manually checked by dictionary editors, compared with the word list which has been compiled manually by the authors and editors of the dictionary and supplemented using the criteria of word formation and semantic fields. As the compilation of the corpus

extracted word list is as yet in progress, this will be illustrated from the word list that has been compiled manually.

autobus	autoportret	bacač
autobusni	autoput	bacil
autocesta	autor	baciti
autogram	autostop	bačva
automat	avantura	badem
automatski	avion	badminton
automehaničar	b	Badnjak
automobil	baba	badnji
automobilizam	babaroga	

Table 4: Extract from the pilot wordlist of the MREŽNIK project

This word list will be supplemented for example by the words *automehaničarka*, *automobilistički*, *bacačica*.

As there are no specialized corpora for elementary school children and foreigners, the list of headwords for these users has to be derived manually. Fortunately, some of the members of the MREŽNIK project have experience in writing lexicographic works for school children as some of them are authors of *Prvi školski rječnik hrvatskoga jezika* (The First Dictionary of the Croatian Language—Čilaš Šimpraga, Jojić & Lewis, 2008), *Školski rječnik hrvatskoga jezika* (School dictionary of the Croatian language—Birtić at al, 2012), and *Prvi školski pravopis* (First Orthographic Manual Hudeček, Jozić, Hudeček, Lewis & Mihaljević, 2016) and are the editors of *School portal*, which is one of the elements which will be connected with MREŽNIK (see below). A member of the MREŽNIK team works with foreigners learning Croatian in Croaticum at the Faculty of Humanities and Social Sciences and, on the basis of her experience and textbooks for foreigners learning Croatian, 1000 words for foreigners will be selected.

7. Dictionary Grammar

A dictionary is a highly structured document. The 'dictionary grammar' is at the center of the project. It names the different fields of information and says how they are to be nested and ordered, and which are obligatory and which are optional. (*Encyclopedia of Language & Linguistics* 2006: 783–793). Simultaneously with the extraction of the headwords, the editors of the dictionary are working on the 'dictionary grammar' of MREŽNIK. The idea is to have a three-module and three-dimensional dictionary. The

representation of the dictionary in each module will consist of fields giving basic data for each entry and links giving additional information which the user can see by clicking on the links. Dictionary grammar will be analyzed for each module separately.

7.1 Module for Adult Speakers

The dictionary grammar consists of these elements: accentuated headword (direct link to the type in the corpus), homonym mark, grammatical information, accentuated inflectional forms, link to inflectional forms, masculine/feminine pairs, perfect/imperfect pairs, cross-references to other entries, accentuated sub-entry, grammatical label, stylistic label, usage label, field label, differentiation of meaning, grammatical restriction, definition, examples from the corpus, link to collocations, link to pragmatic comments, link to semantic relations, phrase, idiom, word formation analysis of the headword, link to derivatives and compounds from the corpus. Semantic relations are divided into synonyms, antonyms, hyponyms and, co-hyponyms. Some elements re-occur as many times as needed, e.g. differentiation of meaning, definition, semantic relations, link to pragmatic explanation are given for the headword, phrase, and idiom. All the fields are optional, except the headword and the grammatical information. Dictionary grammar is shown in Table 10 in the Appendix.

The fields in the table will be illustrated by examples from different entries as it is impossible to find an entry containing all these elements: accentuated headword: **kùćа** (house), grammatical information: *im. m.*, accentuated forms <G kùćē; *mn.* N kùće, G k ćā>.

In the four-year duration of the project from each headword a link to all forms that will be automatically derived and manually checked will be attached, e.g.

	Singular	**Plural**
N	kuća	kuće
G	kuće	kuća
D	kući	kućama
A	kuću	kuće
V	kućo	kuće
L	kući	kućama
I	kućom	kućama

Table 5: Example of word forms for the headword *kuća* (house).

In the future the plan is to have all these forms accentuated as well, e.g.:

	Singular	**Plural**
N	kȕća	kȕće
G	kȕće	k ćā
D	kȕći	kȕćama
A	kȕću	kȕće
V	kȕćo	kȕće
L	kȕći	kȕćama
I	kȕćom	kȕćama

Table 6: Example of accentuated word forms for the headword *kuća* (house).

This is not a problem with nouns, but there are very many verbal forms, and this will probably not be possible within the four-year duration of the project.

An important part of the dictionary are the masculine-feminine pairs, e.g. the masculine noun *učitelj* and the feminine noun *učiteljica* (teacher) will be interconnected:

ùčitelj *im. m.*

učitèljica *im. ž.*

The criterion for such interconnection is not only word-formation, but also semantics. So e.g. nouns *jelen* (deer) and *košuta* (doe) will also be connected:

jȅlēn *im. m.*

kòšuta *im. ž*

Such interconnection is also possible for phrases, e.g.:

medicinska sestra (nurse feminine)

medicinski tehničar (nurse masculine)

The connection to the pair will be a link if the pair has its own entry in the dictionary.

In the same way, perfect and imperfect verbs will be connected, e.g.:

> **ocijéniti** *gl. svrš. prijel.* (evaluate)
>
> **ocjenjívati** *gl. nesvrš. prijel.*

Headwords will be cross-referenced with the label *v.* (vidi – see) which directs a headword not belonging to the standard language to its standard equivalent. These words will usually also have a usage *label* and/or will be connected with linguistic advice.

Sub-entry is used for reflexive verbs which are analyzed under the main verbal entry, i.e.:

> **sèliti** gl. *prijel./neprijel.* (move)
>
> - **sèliti se** *povr.*

Grammatical restriction is used in the cases when a specific grammatical description applies only to a specific meaning, e.g.:

> **mȉš** *im. m.* <G mȉša, A mȉša/mȉš, L mȉšu/mìšu; mn. N mȉševi, G mȉšēvā> 1. <A mȉša, L mȉšu> *zool.* 2. <A mȉš, L mìšu> *inform., tehn.* (mouse)

Each meaning has a definition. Definitions usually start with hypernyms which are links to the dictionary entry, e.g.:

> **mȁnastīr** *im. m.* <G mȁnastīra, I mȁnastīrom; mn. N mȁnastīri, G mȁnastīrā> crkva … (monastery)
>
> **cȓkva** im. *ž.* <G cȓkvē; mn. N cȓkve, G cȓkāvā/c kvā/cȓkvī> 1. Građevina … (church)
>
> **gràđevina** *im. ž.* <G gràđevinē; mn. N gràđevine, G gràđevīnā> grad. 1. objekt … (building)

This is the reason why hyperonyms are not stated among semantic relations in the dictionary grammar.

Examples follow each meaning and are selected from the corpus as described above.

Special attention in MREŽNIK will be paid to collocations. Collocations are differentiated from phrases and idioms. They will be analyzed according to the model from *elexiko* and derived from the corpus via Word Sketches. In *elexiko* e.g. headwords *girl* and *boy* are analyzed according to these questions: *What are the characteristics of a X?, What does X do?, What happens to X?, Which themes are used with X?* (http://www.owid.de/wb/elexiko/gruppen/maedchen-junge.html).

Some entries will have a link to the pragmatic comment. Pragmatic comment will be given, e.g. with pronouns *ti* and *Vi* (you), greetings *dobar dan*, *dobro večer*, *dobro jutro*, *zdravo*, *bok,* etc.

Special attention will also be paid to semantic relations which will be attached to each meaning: synonyms, antonyms, and hyponyms, e.g.:

> **dȍbar** 1. koji ima pozitivne osobine ili poželjna svojstva ***sin.*** valjan; ***ant.*** loš 2. koji je onakav kakav treba biti, koji ispunjava očekivanja ***ant.*** loš 3. koji čini i želi dobro ***ant.*** zao, zločest 4. koji je ispravno utemeljen i logičan ***sin.*** pravi *razg.*, valjan; ***ant.*** loš 5. <neodr.; u im. funkciji> srednja školska ocjena označena s 3; ***sin.*** trojka razg. 6. <sup.> koji ima najpozitivnije osobine i najpoželjnija svojstva]; ***sin.*** (optimalan)

Three different classes of labels will be used as shown in Table 7.

Grammatical labels	**Field labels**	**Usage labels**
m. – muški rod male	anat. – anatomija anatomy	razg. razgovorno – colloquial
s. – srednji rod neuter	astr. – astronomija astronomy	reg. regionalizam regional
ž. – ženski rod female	astrol. – astrologija astrology	žarg. – žargonizam – jargon
pl. tantum – *pluralia tantum*	biol. – biologija biology	
sg. tantum – *singularia tantum*	bot. – botanika bothanics	
neprijel. – neprijelazni glagol intransitive verb	el. – elektrotehnika electrical engineering	
povr. – povratni glagol reflexive verb	farm. – farmacija pharmacy	
prijel. – prijelazni glagol – transitive verb…	fil. – filozofija philosophy	
	fiz. – fizika physics	
	fiziol. – fiziologija physiology	
	geol. – geologija geology…	

Table 7: Three different classes of labels in MREŽNIK

The dictionary entries contain links to repositories which have already been compiled within other projects conducted at the Institute of Croatian Language and Linguistics: Valence database e-Glava, Repository of Metaphors, Terminology database STRUNA , Better in Croatian and to databases which are created as a part of this project and compiled simultaneously with the dictionary. These repositories are: Linguistic Advice Repository, Conjunction Repository, Repository of Idioms, Repository of Ethnics and Ktetics, Male/female repository, and Pragmalinguistic repository.

7.2 Module for Elementary School Children

The second module of the project is the module for school children. The aim is to contribute to Croatian language learning in schools as it is evident that e-dictionaries have a much greater chance of being accepted in schools than classical dictionaries.[2] The dictionary grammar is shown in Table 11 in the Appendix.

This dictionary will consist of the headword with marked accentuation place (djev**o**jčica), some grammatical information (part of speech, gender), syllable marking (e.g. dje-voj-či-ca), simple definitions, examples written by the lexicographer, very few synonyms, collocations, idioms. In this module some of the entries will contain illustrations. The principles of using appropriate illustrations and their role in the understanding of semantic relations in language manuals for children have been given in Hudeček & Mihaljević (2015).[3]

Some dictionary entries contain links to simple language advice and explanation of idioms for school children in the repository Croatian in School (http://hrvatski.hr/).

The block of fields starting with differentiation of meanings re-occurs as many times as needed. All the fields are optional except the headword and the grammatical information.

7.3 Module for Foreigners

The third module of the project is the module for foreigners. This module will contain audio recording of the pronunciation of each headword. It will also provide information (pragmatic, cultural, collocations) useful to foreigners learning Croatian.[4] The dictionary grammar is shown in Table 12 in the Appendix.

The block of fields starting with differentiation of meanings, re-occurs as many times as needed. All the fields are optional except the headword and the grammatical information. Some entries will be linked to simplified language advice and explanation of idioms.

8. Reverse Dictionary

In the fourth year of the project, a reverse dictionary in which entries are alphabetized from the end is planned. This dictionary is very important for the analysis of word-formation. This dictionary will be based on the completed list of 10,000 words. At this phase of the project a pilot reverse dictionary based on the manually compiled word-list has already been compiled and is available to team members.

2 The mariginal role of dictionaries in teaching German as a mother tongue in German schools is also mentioned by Töpel (2014: 291).

3 Illustrations have also an important function in *elexiko*. Müller-Spitzer (2005: 212).

4 Similar ideas appear in Möhrs (2014: 322).

baba	oporaba	skladba
visibaba	zloporaba	obradba
tužibaba	uporaba	preradba
koraba	zlouporaba	doradba
poraba	žaba	razradba

Table 8: Extract for the pilot reversed wordlist of the MREŽNIK project

9. Computer Tools

The two basic computer tools for the compilation of this three-module dictionary are SketchEngine, a corpus query system (loaded with corpora) to support the analysis of the language, and TLex, a dictionary writing system to support the preparation of the dictionary text. The dictionary will be compiled using TLex, a professional software application for compiling dictionaries. SketchEngine, a corpus manager and analysis program, will be used to search the corpora and extract data. SketchEngine can be used to retrieve the context in which a word is usually found using word sketches, grouping the strongest collocations into syntactic categories and finding adequate examples of lexemes and collocations. Data are selected from the corpus by simply ticking boxes inside a lexicographic interface (Tickbox Lexicography). Data selected in this way are automatically saved in TLex and other lexicographers can continue describing the lexemes on all other levels. For the morphological description, the morphological lexicon hrLex (http://nlp.ffzg.hr/resources/lexicons/hrlex/) will be used, the content of which will be adapted and connected with MREŽNIK. The data acquired from hrLex will be checked by the lexicographer and accentuated by the accentologist.

After completion of the dictionary entries, the data will be exported from TLex, in order to be used in the Web application, which will be developed for the dictionary and the CLARIN repository; a European research network working in the field of archiving and processing of language-related resources (https://www.clarin.eu/). MREŽNIK will in this way become available for use through web application. It will also be available for various purposes at the CLARIN repository.

10. Conclusion

In the paper, the dictionary grammar of a three-module three-dimensional corpus-based dictionary is presented. As the project only began in March 2017, many elements are still being developed and programmed in TLex so the fields are presented in tables and not as screenshots of the program. Table 9 sums up the connection of three dictionary modules with special databases.

Dictionary module	**Connected databases**
dictionary for adult native speakers 10,000 entries	Linguistic Advice Repository http://jezicni-savjetnik.hr/
	Conjunction Repository
	Repository of Idioms
	Repository of Ethnics and Ktetics
	Portal Bolje je hrvatski Better in Croatian http://bolje.hr/
	Male/female portal
	Pragmalinguistic portal
	Repository of Metapors http://ihjj.hr/metafore/
	Valency database e-Glava: Baza hrvatskih glagolskih valencija http://ihjj.hr/projekt/baza-hrvatskih-glagolskih-valencija/27/
	Croatian Special Field Terminology http://struna.ihjj.hr/
dictionary for elementary school children 3000 entries	Croatian in schools – explanation of idioms and language advice for children http://hrvatski.hr/
dictionary for foreigners	explanation of idioms explanation of usage simple language advice

Table 9: Connection of three dictionary modules with special databases

11. Acknowledgements

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

12. References

Baza hrvatskih glagolskih valencija – GLAVA. Accessed at http://ihjj.hr/projekt/baza-hrvatskih-glagolskih-valencija/27/ (18 May 2017)

Birtić, M. & et al. (2012). *Školski rječnik hrvatskoga jezika*. Zagreb: Školska knjiga – Institut za hrvatski jezik i jezikoslovlje.

Bolje je hrvatski. Accessed at http://bolje.hr/ (18 May 2017)

Brown, K. (ed.) (2006). *Encyclopedia of Language & Linguistics* (second edition) Elsevier Ltd.

Čilaš Šimpraga, A.; Jojić, Lj. & Lewis, K. (2008). *Prvi školski rječnik hrvatskoga jezika*. Zagreb: Školska knjiga – Institut za hrvatski jezik i jezikoslovlje.

Haß, U. (2005). Das Bedeutungsspectrum. In U. Hass (ed.) Grundfragen der elektronischen Lexikographie. elexiko – das Online-Informationssystem zum deutschen Wortschatz. Berlin/New York: De Gruyter, pp163-181.

Hrvatski jezični portal. Accessed at http://hjp.znanje.hr/ (30 April 2017)

Hrvatsko strukovno nazivlje – STRUNA. Accessed at http://struna.ihjj.hr/ (18 May 2017)

Hudeček, L. & Mihaljević, M. (2015). Dječji jezik i jezični priručnici za djecu in In L. Cvikić; B. Filipan-Žignić, I. Gruić, B. Majhut & L. Zergollern-Miletić (eds.) *Istraživanje paradigma djetinjstva i obrazovanja.* Zagreb: Učiteljski fakultet, pp. 46-67 Available at http://hrvatski.ihjj.hr/content/hr-nastava-1.pdf

Hudeček, L., Jozić, Ž., Lewis, K. & Mihaljević, M. (2016). *Prvi školski pravopis*. Zagreb: Institut za hrvatski jezik i jezikoslovlje.

Jermen, N., Kraus, C. & Starčević Stančić, I. (2015). Lexicography and Encyclopaedistics in the Digital Environment in K. Anderson, L. Duranti, R.Jaworski, H. Stančić, S. Seljan & V. Mateljan. *Infuture 2015. The Future of Information Sciences. E-Institutions Openness, Accessibility, and Preservation,* Zagreb: Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences, pp. 65-75.

Jezični savjetnik Accessed at http://jezicni-savjetnik.hr/ (18 May 2017)

Klosa, A. (ed.). (2011). elexiko. *Erfahrungsberichte aus der lexikographischen Praxis eines Internetwörterbuchs*. Tübingen: Narr. Verlag.

Möhrs, Ch. (2014): Landeskundliche Wortschatzübungen auf der Basis von Kollokationen. Zur Nutzung von *elexiko* für Deutschlehrende. In A. Klosa (ed): *Themenheft "Dateninterpretation und -präsentation in Onlinewörterbüchern am Beispiel von elexiko"*. Deutsche Sprache 4/2014, pp. 309-324.

Müller-Spitzer, C. (2005). Illustrationen in elexiko. In U. Haß (ed.). *Grundfragen der elektronischen Lexikographie. elexiko - das Online-Informationssystem zum deutschen Wortschatz*. Berlin/New York: De Gruyter, pp. 204-226.

OWID. Accessed at http://www.owid.de/wb/elexiko/gruppen/maedchen-junge.html (15 May 2017)

Repozitorij metafora. Accessed at http://ihjj.hr/metafore/ (18 May 2017)

Štrkalj Despot, K. & Möhrs, C. (2015). Pogled u e-leksikografiju. *Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje*, 41(2), pp. 329-353.

Štrkalj Despot, K. (2014). Croatian Metaphor Repository. *Proceedings of the 2nd COST ENeL Working Group 3 („Innovative e-dictionaries“)* Meeting: „Workflow of Corpus-based Lexicography“. Available at http://www.elexicography.eu/wp-content/uploads/2014/07/Strkalj-Despot_2014_COST_Bolzano.pdf

Töpel, A. (2014). Wörterbücher im muttersprachlichen Deutschunterricht der Sekundarstufe II – Zur Nutzbarkeit der Bedeutungserläuterungen von *elexiko*. In: A. Klosa (ed.), *Themenheft "Dateninterpretation und -präsentation in Onlinewörterbüchern am Beispiel von elexiko*. Deutsche Sprache 4/2014, pp. 291-308.

Wječnik. Accessed at https://hr.wiktionary.org/wiki/Glavna_stranica (30 April 2017)

13. Appendix

Field	**Sub-field and/or comment**
accentuated headword	headword is a direct link to the type in the corpus
homonym mark	
grammatical information (part of speech code, e.g. im. m. noun, masculine)	
accentuated inflectional forms (selected forms)	
link to the inflectional forms (all forms)	
masculine/feminine pair	
perfect/imperfect pair	
cross-reference to another entry (*v.* see)	
accentuated sub-entry - reflexive verb	
grammatical label *povr.* - reflexive	
style and usage label	
field label	
differentiation of meanings 1., 2., 3. …	
stylistic label	
field label	
grammatical restriction	
definition	
examples from the corpus	
link to collocations	
link to pragmatic comment	
link to semantic relations	synonyms

	antonyms
	hyponyms
	co-hyponyms
phrase	
style label	
field label	
definition	
examples from the corpus	
link to collocations	
link to pragmatic comments	
link to semantic relations	synonyms
	antonyms
	hyponyms
	co-hyponyms
idiom	
link to the explanation of the idiom	
definition	
examples from the corpus	
link to semantic relations	synonyms, antonyms
link to pragmatic information	
word formation analysis of the headword	
link to derivatives and compounds from the corpus	

Table 10: Dictionary fields in the module for adult native speakers

Field	**Sub-field and/or comment**
headword with marked place of the accent	link to the audio recording of the pronunciation
homonym mark	
grammatical information (full words)	
masculine/feminine pair	
cross-reference to another entry	
inflectional forms	
differentiation of meanings 1., 2., 3. …	
stylistic label	
field label	
grammatical restriction	
definition	
examples	
collocations	
usage	
semantic relations	synonyms
	antonyms
phrase	
style label	
field label	
definition of meaning	
collocations	list of most frequent collocations
link to pragmatic comments	

link to semantic relations	synonyms
	antonyms
Idiom	link to the explanation of idioms
definition	
semantic relations	synonyms
	antonyms
headword with marked place of the accent	link to the audio recording of the pronunciation
homonym mark	
grammatical information (full words)	
division into syllables	
masculine/feminine pair	
cross-reference to another entry	
inflectional forms	
subentry reflexive verb	
differentiation of meanings	
grammatical restriction	
definition	
examples	most common collocations
synonyms – introduced into the definition by the formula: the same meaning have the words…	
antonyms – introduced into the definition by the formula: the opposite meaning have the words	
phrase	
definition	definition contains information on style and field (in spoken language, when speaking to friends, in mathematics…) as well as on

	synonyms and antonyms
examples	
synonyms – introduced into the definition by the formula: the same meaning have the words…	
antonyms – introduced into the definition by the formula: the opposite meaning have the words	
idiom	link to the explanation of the meaning of idioms for children
definition	
synonyms – introduced by the formula: the same meaning have the words…	
antonyms – introduced into the definition by the formula: the opposite meaning have the words	

Table 11: Dictionary fields in the module for school children

Field	**Sub-field and/or comment**
headword with marked place of the accent	link to the audio recording of the pronunciation
homonym mark	
grammatical information (full words)	
masculine/feminine pair	
cross-reference to another entry	
inflectional forms	
differentiation of meanings 1., 2., 3. …	
stylistic label	
field label	
grammatical restriction	
definition	

examples	
collocations	
usage	
semantic relations	synonyms
	antonyms
phrase	
style label	
field label	
definition	
collocations	list of most frequent collocations
link to pragmatic comments	
link to semantic relations	synonyms
	antonyms
idiom	link to the explanation of idioms
definition	
semantic relations	synonyms
	antonyms

Table 12: Dictionary fields in the module for foreigners

Dicționariul Limbei Române (LM) by A. T. Laurian and I. C. Massim – the Digital Form of the First Romanian Academic Dictionary

Marius-Radu Clim, Mădălin-Ionel Patrașcu, Elena Isabelle Tamba

"A. Philippide" Institute of Romanian Philology, Romanian Academy, Iasi, Romania
E-mail: marius.clim@gmail.com, madalin.patrascu@gmail.com, isabelle.tamba@gmail.com

Abstract

We aim to present a project called DICTIONE, the digitalization of the first Academic Dictionary written in the Romanian language, the *Dicționariul limbei române* (LM) by A. T. Laurian and I. C. Massim. LM was published in three volumes between 1873-1877, has 3600 pages and includes 70,000 headwords out of which over 20,000 words are the personal creation of the authors. This dictionary is unique in Romanian cultural history, due to the fact that the two lexicographers did not aim to illustrate the language vocabulary in a particular moment of its history, but instead intended to impose a certain direction to the language, beyond its use at that time. A novelty is the fact that the authors proposed new words which they attempted to popularize through this dictionary. The impact of this work on Romanian culture is a significant one: from more than 20,000 newly-created words, based on terms taken from Latin, most stayed in use and became neologisms. This fact led to the enrichment of Romanian terminology in many domains and to the modernization of the Romanian language at the same time as that of other European cultures.

Keywords: digitization; academic; neologism; cultural heritage

1. European context

European cultures were, and still are, preoccupied with the recovery and valuation of their own lexicographical thesauruses, within which the cultural stages of a language are stored. The current digital means permit not only the recovery of these "cultural databases", but their promotion by making this lexical richness available to the public. We mention several examples, such as:

a) *Trésor de la langue française* (1971–1994, first printed edition), http://atilf.atilf.fr;

b) *Dictionnaire de l'Académie française.* La 9 édition en ligne (1694, first printed

edition), http://atilf.atilf.fr/academie9.htm;

c) *Diccionario de la lengua española de la Real Academia Española* (DRAE) (1780, first printed edition), http://buscon.rae.es/draeI/;

d) *Tesoro della lingua italiana delle origini* (TLIO); http://tlio.ovi.cnr.it/TLIO/index2.html;

e) *Deutsches Worterbuch der Grimm* (DWB) (1838-1961, first printed edition), http://germazope.uni-trier.de/Projects/DWB;

f) *Oxford English Dictionary* (1928, first printed edition), http://www.oed.com.

In 2016 we celebrated 150 years since the establishment of the Romanian Academy, whose original purpose was to create a dictionary and a grammar of the Romanian language, to solve the orthographic problem and to write a book of Romanian history. In this context, what we intend to do through the project DICTIONE, namely the digitalization of the first Academic Dictionary written in the Romanian language, is to both recover the Romanian cultural heritage and make profitable the activity, ideas and erudition of the first scholars who realized this academic thesaurus. The project DICTIONE fits perfectly into the European trend and aims to capitalize on the Romanian lexicographical heritage by digitizing the first academic dictionary to have been printed in the Romanian language.

2. Related Romanian research projects

This project had several related and relevant predecessors concerning the digitization of various lexicographic works. We would like to mention the most representative. The project *CLRE* concerned the *Corpus lexicografic românesc esenţial. 100 de dicţionare din bibliografia DLR aliniate la nivel de intrare* [Essential Romanian Lexicographic Corpus. 100 dictionaries from DLR Bibliography aligned by entries][1] received national financing between 2010 and 2013, and was conceived as a linked database between 100 dictionaries from the DLR Bibliography aligned by entry level. Today, this project continues as part of the research plans of the Romanian Academy – Iasi Branch, by the "A. Philippide" Institute of Romanian Philology. This constitutes a great resource for lexicographers, providing fast access to dictionaries and helping to present a better historical perspective of the Romanian lexicography (Clim, 2015).

Another Romanian lexicographic resource available online is *Lexiconul de la Buda* [The Lexicon of Buda], the electronic edition[2] of the first etymological and explanatory dictionary of the Romanian language, a benchmark for modern Romanian

[1] The CLRE project will be accessible at http://clre.philippide.ro at the end of 2017. More about this project in Clim et al. (2016).

[2] http://www.bcucluj.ro/lexiconuldelabuda/site/login.php

lexicography. The current electronic edition restores the volume that was printed at the Buda press, in 1825, under its full title, *Lesicon romanescu-latinescu-ungurescu-nemtescu quare de mai mulţi autori, in cursul a trideci, si mai multoru ani s'au lucrat. Seu Lexicon valachico-latino-hungarico-germanicum quod a pluribus auctoribus decursu triginta et amplius annorum elaboratum est.* This is a multilingual dictionary, targeted more towards the equivalence of terms in four languages (Romanian, Latin, Hungarian and German) and less towards the semantic description of lemmas (Patrașcu et al., 2016).

3. DICTIONE project

3.1 DICTIONE – highlights of the first Romanian academic dictionary

In contrast to this dictionary, the one implicated in the DICTIONE project is a dictionary exclusively dedicated to Romanian vocabulary, etymology and semantics. *Dicționariul limbei române* (LM) [Romanian Language Dictionary] by A. T. Laurian and I. C. Massim was published in three volumes between 1873-1877, has 3600 pages and includes 70,000 headwords, of which over 20,000 words are the personal creation of the authors, based on latinized terms. In this dictionary, the authors were interested in achieving two main outcomes: creating Romanian terminology, capable of expressing the concepts of modern culture; and proving by lexical means the latinity of the Romanian language. However, in order to facilitate the process of consulting this lexicographical work by the foreign specialists, the Romanian lemmas are accompanied by their Latin equivalent.

The preface contains three chapters describing the principles used in elaborating the dictionary, the conception of the Romanian language and how it can be modernized. The authors, who are representatives of the latinizing direction in Romanian linguistics, proclaim the purity of the language as a basic principle of the entire work and argue for maintaining the latinity of the Romanian language and eliminating foreign words, using the French Academy as a model. The last chapter promotes the orthography according to the etymological principle, the authors aim to admitting only the letters that correspond to the primitive sounds preserved by the Romanian language from Latin. Laurian was a supporter of etymological writing, to the detriment of phonetic writing, and attempted to demonstrate as clearly as possible the Latin form and origin of Romanian words (Clim, 2012). Thus, access to understanding this writing was possible only for those who knew Latin. Also, the authors tried to assimilate the current forms of Romanian words – forms that resulted from a long historical evolution – with the appropriate forms from Latin, and the difference between the words inherited from Latin and the Latin-Romance neologisms was ignored.

Laurian and Massim divide the words in this dictionary in two categories, considering this historical criterion:

a) words in use before 1830;

b) words in use after 1830.

The words of the first category are numerous in regions inhabited by Romanian people, while those from the second category had not yet spread until the dictionary was published and were thus marked in this work with an asterisk.

The two authors declared themselves against defining the terms through synonyms, as, in their opinion, this creates more confusion regarding the meaning of the headwords. Therefore, they adopted definition by periphrasis. Broad definitions are followed by illustrative phrases. Another lexicographical novelty is the fact that LM includes detailed orthoepic explanation at the beginning of each letter, taken from the work of Laurian, *Tentamen criticum in originem, derivationem et formam linguae romanae in utraque Dacia Vigentis vulgo valachicae*, Vienna, 1840. In general, the dictionary article is structured as follows: the entry word, bearing no accent indication, followed sometimes by the explicit indication of pronunciation (e.g. *coctoriu*, *pronuntiatu coptoriu*), and then details about the word flexion and about the grammatical category. The equivalence of the Romanian term with another from a well-known foreign language like Latin should be appreciated. Also, the fact that the etymology of the entry word is indicated before its explanation is a novelty in the lexicographical technique. It also has a solid scientific argumentation in the preface of the work, many of the etymological solutions proposed by the authors are used even today. Furthermore, the adoption of multiple etymologies as a way of explaining the origin of Romanian terms is notable.

However, the greatest value of this dictionary is that it tried to impose a large number of neologisms into the Romanian language, over 20,000 of them. Although unnatural, this effort to fill the gaps of Romanian vocabulary remains quite impressive. Here are some of the neologisms proposed by the authors of this dictionary: *accelerator, adjunct, admirativ, adversitate, aerofagie, austeritate, anxietate, benign, bibliologie, biochimie, biotic, calvar, fabricabil, fabulație, factură, fastuozitate, felin, feroce, ferocitate, figurativ, fluență, formativ, fotogenic, fracționa, frazeologic, genetic, genuin, germina, ginecologic, giratoriu, gnoseologie, gnostic, grandilocvență, imersiune, imixtiune, imobiliar, imobiliza, imortaliza, matador, mercantil, meteoric, metronom, micrometru, miligram*, etc. These terms exist in the current Romanian language, many of them being (re)borrowed from Romance languages. In order to illustrate the peculiarity of this dictionary we present the definition of the term *factură* [invoice]:

> factură: *factura*, s. f., **factura**, resultatu allu facerei, opera; 2. in commerciu, (it. **fattura**, fr, **facture**), statulu care arréta in detaliu speci'a, cantitatea, calitatea si pretiulu merciloru ce tramitte unu fabricante sau unu negotiatoriu la veri-unulu dintre confratii sau associatii sei, la veri-unu commissionariu, etc. (dar și ceilalți termeni din familia de cuvinte: *factura*, *facturariu*, *facturat*).

Laurian and Massim did not aim to highlight the state of the language in their time, but instead intended to set a certain direction, outside the use of the period, and thus proposed new words (an impressive novelty in the lexicographical works) that they tried to popularize through this dictionary. In other words, they pointed out to the maximum the regulatory role of this Academic dictionary. It is recognized that the two lexicographers have sought to provide modern definitions, both for the new terminology, and for old words. This is the reason for which the work has been appreciated by subsequent lexicographers who have treated it as a valuable source of information and documentation. Through the efforts of the authors to enrich the vocabulary of the Romanian language, many neologisms that have been preserved in the language were put into circulation. Taking into consideration the large number of registered neologisms, this dictionary is unprecedented in Romanian culture. This dictionary is mentioned in prestigious lexicographical works and represents a documentary base exploited by Romanian language researchers. Regarding the problem of etymology, and especially the primary, not only direct, etymology problem studied by European researchers interested in the migration of words, this dictionary is a valuable bibliographic resource. The conversion of this first dictionary of the Romanian Academy in a digital format, easy to read and to use in the documentation process and in linguistic research, would facilitate access to the information gathered by Laurian and Massim to Romanist specialists.

In the current European context, there is interest in collating linguistic resources for determining a correct etymology of a new term for any given language, of finding the primary etymon and also the transition of the term from the source language to the host language. Thus, the digitization of this work will help Romanian researchers solve etymological problems, while also assisting foreign researchers who want to study the filiation of the terms or meanings. Therefore, this dictionary is proposed for inclusion on the list of prestigious European dictionaries to be used by all the people interested in the study of the Romanian language.[3]

3.2 DICTIONE goals

Through its digital version, this dictionary will have a significant impact both on lexicographic Romanian works, and on research of Romanian language history.

The objectives of this project are not new and they reflect the usual objectives of digitizing an old dictionary. The project DICTIONE aims to digitize the first academic dictionary printed in the Romanian language. The proposed objectives are the following:

[3] This dictionary will be included in the European Dictionary portal www.dictionaryportal.com created in the COST project *ENeL European Network of e-Lexicography.*

1. transforming the digital form into an editable format and correcting the text of the dictionary;

2. recognizing the entries of the dictionary;

3. creating a site for the project that would contain, among other information, the scanned and the editable versions of the dictionary. The written version will be annotated at the levels of morphology, etymology, and so on. This will permit various types of search.

4. aligning the dictionary to other digital versions of Romanian dictionaries. In the primary stage, this will be done at headword level. After this step, annotated information will be considered to be linked with similar data from other dictionaries.

5. aligning the dictionary to other Romanian dictionaries and integrating it into the list of representative dictionaries of European languages;

6. creating an exhaustive list of neologisms defined into the dictionary and developing a study regarding their circulation during the period 1862-1927.

3.3 DICTIONE project steps

This project is meant to last two years and comprises a number of steps that we will present briefly here. The first step is administrative and presumes the preparation of the lexicographic material and the establishment of the working stages for the entire team. That means, first of all, to check the current status. There are already scans of this dictionary (some made in Romania: for example one made by the Bucharest Metropolitan Library, available on the site www.digibuc.ro, or in CLRE project and another made by Google Books) and it is necessary to verify and compare the existing scans and to select the most appropriate scan for the project DICTIONE. Before using the OCR program, it is necessary to process the 3600 scanned images to eliminate page noise and any typographical points and to optimize them for the process of recognizing characters. This process will be followed by a OCR testing phase. It is possible for the program to recognize most letters, but it will not be able to recognize many words because of the latinizing orthography: vowels, doubled consonants, the lack of diacritical marks and others. After recognizing the characters, a program will be created in order to allow the validity of each term separately. Then the text will be corrected for keeping the exact orthographic version of the dictionary, as it was drafted. This step, to the text correcting arising from the character recognition process, will be assisted by software specifically developed for this purpose. This computer program should identify types of corrections made to the text by the specialist and propagate similar changes throughout the text. We rely on the fact that we can achieve this desideratum, because from our experience with OCR-ized texts we

have noticed that these software tools introduce inaccuracies towards printed text which can be resolved systematically. In the first stage, the program will identify possible errors by statistical analysis of the text (for example: words with one or two occurrences). Secondly, the program will assist a human specialist and propose words which are similar to the modified words operated on the text. The similar terms will be generated by considering the OCR score of each recognized sign and an analysis based on n-grams. This is the approach by which we hope to reduce the length of this stage and at the same time to develop a software instrument that will later be used for similar texts. Another goal of the project would be that the electronic text obtained from the correction of the OCR stage be similar to the printed one in both content and graphic form.

After the text is revised the next step is to recognize the dictionary entries. In order to validate the delimitations of the dictionary entries, a program created through the CLRE project will be used. This program will be adapted to insert the modern orthographic form for each headword, to align this dictionary to the other Romanian works. Subsequently, a parser will be created to delimit the information (the fields) of a drafted article: entry word, morphological information, etymology, the translation of the term, examples etc. Once the text from the dictionary is parsed, it is necessary to make the correlations between the terms from the dictionary and those mentioned by the authors in the preface in order to put at the disposal of the users not only the definitions, but also the commentaries and analyses made by Laurian and Massim [for example, a term as *federatione* [federation] will be found according to the modern orthography „federație/federațiune" both in the text dictionary and also in the preface. In addition, parsing the text will enable searching the Latin etymon of the terms if, for example, a term borrowed from Latin is searched.

The last step involves the creation of a website for the dictionary, aligning the project DICTIONE to other digital Romanian dictionaries, such as CLRE, and including this dictionary on the list of representative dictionaries at a European level (www.dictionaryportal.eu). After the dictionary is fully digitized we intend to extract an exhaustive list of new terms included in Laurian and Massim's work, terms that did not exist in the language at the time, and to conduct research into their adaptation in the Romanian language.

The digitization of this dictionary comes with some risks, due to its uniqueness in Romanian culture. The greatest risk is the potential incorrect automatic recognition of words, because the latinizing orthography requires that the user should mentally transpose the terms into a modern orthographical version. Also, this dictionary can cause problems to public users because of the Latinist spelling. However, transposing the scanned text into its latinizing version reflects the novelty of this dictionary and for this reason researchers will employ automatic or semi-automatic electronic means, but also using manual verification to maintain the accuracy of the dictionary text. In addition, all headwords will appear in their modern orthographic form in order to

allow easy access for all users familiar with the Romanian language. If the headwords are commented in the Preface, the user will be able to search according to the current form of the terms. Because the entire text will be corrected, researchers can search, for example, for a Latin term and verify if it has been borrowed into the Romanian language and determine under which form it is mentioned in LM. This will allow correlations to other languages (Romanic or not) which have also borrowed that Latin term.

This dictionary remains a valuable source for the lexicographers involved in drafting *The Dictionary of the Romanian Language (DLRi),* the current academic thesaurus, but also for the dialectologists and philologists interested in the etymology of old words or their semantic evolution.

4. Acknowledgements

This work was supported by a grant of the Romanian National Authority for Scientific Research and Innovation, CNCS – UEFISCDI, project number PN-II-RU-TE-2014-4-0195.

5. References

Catană-Spenchiu, A.-V. & Clim, M.-R. & Patrașcu, M. & Tamba, E. (2015). CLRE. Corpus lexicographique roumain essentiel. Résultats et perspectives in *Integrare europeană/identitate națională; plurilingvism/multiculturalitate – limba și cultura română: evaluări, perspective (European Integration/ National Identity; Plurilingualism/Multiculturality – Romanian Language and Culture: Evaluation, Perspectives)*, Luminița Botoșineanu, Ofelia Ichim (eds). Roma, Italia, ARACNE Editrice, Colecția „Danubiana", p. 323-332.

Clim, M.-R. & Tamba, E. & Catană-Spenchiu A.-V. & Patrașcu, M. (2016). CLRE. Corpus lexicographique roumain essentiel. 100 dictionnaires de la langue roumaine alignés au niveau de l'entrée et, partiellement, au niveau du sens, in vol. Buchi, Éva, Chauveau, Jean-Paul & Pierrel, Jean-Marie (éd.) : *Actes du XXVII^e Congrès international de linguistique et de philologie romanes (Nancy, 15-20 juillet 2013)*, Strasbourg, ÉLiPhi (Editions de linguistique et de philologie), vol. 2, Section 16, p.1611-1622, (and on-line in vol. Trotter, David/Bozzi, Andrea/Fairon, Cédric (éd.): *Actes du XXVII^e Congrès international de linguistique et de philologie romanes (Nancy, 15-20 juillet 2013). Section 16 : Projets en cours; ressources et outils nouveaux.* Nancy, ATILF: http://www.atilf.fr/cilpr2013/actes/section-16.html).

Clim, M.-R. (2012). *Neologismul în lexicografia românească*, Iaşi, Editura Universităţii „Alexandru Ioan Cuza", 355 p.

Clim, M.-R. (2015). La lexicografía rumana informatizada: tendencias, obstáculos y logros in vol. *Lexicografia de las lenguas románicas. Aproximaciones a la lexicografía moderna y contrastiva,* vol. II, coords.: María Dolores Sánchez

Palomino y María José Domínguez Vázquez, eds.: María José Domínguez Vázquez, Xavier Gómez Guinovart y Carlos Valcárcel Riveiro, Editura De Gruyter, pp. 95-110.

Dănilă E. & Haja, G. (2005) Neologismul din perspectivă lexicografică, în „Studii şi cercetări lingvistice", LVI; nr. 1–2, ianuarie–decembrie, Bucureşti, p. 71–78.

De Schryver, G.-M. & Joffe, D. (2004). On How Electronic Dictionaries are Really Used. In G. Williams & S. Vessier (eds.) *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004. Lorient: Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud*, pp. 187–196.

Krek, S. & Kilgarriff, A. (2006). Slovene Word Sketches. In T. Erjavec & J. Žganec Gros (eds.) *Proceedings of the 5th Slovenian and 1st International Language Technologies Conference.* Ljubljana, Slovenia. Available at: http://nl.ijs.si/is-ltc06/proc/12_Krek.pdf.

Patrașcu, M.-I. & Clim, M.-R. & Haja, G. & Tamba, E. (2016). Romanian Dictionaries. Projects of Digitization and Linked Data in Diana Trandabăț, Daniela Gîfu (eds.) *Linguistic Linked Open Data.* 12th EUROLAN 2015 Summer School and RUMOUR 2015 Workshop Sibiu, Romania, July 13–25, 2015. Revised Selected Papers, Springer, p. 110-123.

Pruvost, J. & Sablayrolles, J.-F. (2003). *Les Neologismes*, Paris, Presses Universitaires de France, 128 p.

Seche, M. (1966). *Schiţă de istorie a lexicografiei române*, vol. I: *De la origini până la 1880*, Bucureşti, Editura Ştiinţifică, 192 p.

Tamba, E. & Clim, M.-R. & Catană-Spenchiu, A.-V. & Patrașcu, M. (2012). Situația lexicografiei românești în context european, in „Philologica Jassyensia", An VIII, Nr. 2 (16), 2012, p. 259-268 and on-line, http://www.philologica-jassyensia.ro/upload/VIII_2_Tamba_Clim.pdf

Tamba, E. & Clim, M.-R. & Catană-Spenchiu, A.-V. (2012). The Evolution of the Romanian Digitalized Lexicography. The Essential Romanian Lexicographic Corpus in *Proceedings of the 15 th EURALEX International Congress,* 7–11 august 2012, Oslo, eds. Ruth Vatvedt Fjeld, Julie Matilde Torjusen, Press Reprosentrales, UiO, ISBN: 978-82-303-2095-2, p. 225; the text can be accessed on-line at: http://www.euralex.org/proceedings-toc/euralex_2012/.

What Do Users of General Electronic Monolingual Dictionaries Search for? The Most Popular Entries in the Polish Academy of Sciences Great Dictionary of Polish

Ewa Kozioł-Chrzanowska

[1] Institute of the Polish Language, Polish Academy of Sciences, al. Mickiewicza 31, 31-120 Cracow, Poland
E-mail: ewa.koziol-chrzanowska@ijp-pan.pl

Abstract

This article reports on an analysis of the most popular entries in the online general monolingual dictionary, based on the Polish Academy of Sciences Great Dictionary of Polish (GDP). The GDP was created from scratch over 12 years. The given survey aims to present an overview of its users' needs after the completion of the first stage of work (which was the 15,000 most frequently used lexemes) and to draw conclusions which may become useful for other lexicographers facing similar challenges. The analyzed data consist of 500 most popular entries in a four-year time period. The majority (80%) constitutes multi-word expressions: phraseological units (50%), proverbs (29%) and terms (1%). All of the subgroups are varied in style, meaning and form. The remaining 20% of the most popular entries are made up by single lexemes, mostly (15%) by the ones with a low subjective probability factor. Additionally, possible reasons for such results are addressed, considering school needs as well as the content of other online dictionaries.

Keywords: dictionary use; online dictionary; monolingual dictionary

1. Introduction

For the last few decades, there has been a growing interest in the needs and habits of dictionary users. This interest has resulted in most experts now appreciating the necessity to compile dictionaries with the user needs foremost in mind (Lew, 2011: 1). Undoubtedly, intuitions and predictions cannot expand the empirical data to achieve the goal. The given article contributes to the growing literature which seeks to inform lexicographers regarding user needs and expectations.

1.1 Research on Monolingual Electronic Dictionary Use

As electronic dictionaries have been replacing printed counterparts (Lew, 2012: 243) it seems obvious that research into dictionary use is increasingly focussing on the former. However, Töpel (2014), in her paper which was part of the first volume of Lew (2015:

232), focussed on the use of online dictionaries and claims that the "description of the current state of research into electronic dictionaries makes it clear that in several areas there remains much to investigate. On the content side, both research into online dictionaries, in this case particularly monolingual dictionaries, and issues of user-friendly presentation of content have been investigated only a little or not at all" (Töpel 2014: 48). A similar statement was made a year later by Gromann, arguing that most studies empirically evaluate specific learners' dictionaries or specialised translation dictionaries (Gromann, 2015: 55). In 2012, Müller-Spitzer et al. mentioned only three studies that focus solely on monolingual electronic dictionaries. The authors stressed also the fact that most studies focus on multilingual, mainly bilingual, dictionaries or on comparing bilingual with monolingual ones (Müller-Spitzer et al., 2012: 426).

1.2 The "Polish Academy of Sciences Great Dictionary of Polish" Project

The Polish Academy of Sciences Great Dictionary of Polish (GDP) is a general dictionary of the Polish language, published online at http://wsjp.pl/. Access is open and free of charge.

The project has been underway for over 12 years. The initial idea for a new dictionary was presented in 2005. The actual lexicographical work on the dictionary began in January 2008 and continued until the end of 2012. During this period, 15,000 entries were prepared (describing the most frequently used words in the Polish language collected from corpora available from that period). The current stage of lexicographical work began in September 2013 and is expected to continue until 2018. Its aims are to enrich the previously compiled entries as well as to compile an additional 35,000 entries. The latter goal consists of preparing:

- lexemes that were already included in the dictionary (per the rule of compilation) in the meaning relations with previously compiled words;
- formative derivatives from the words already described;
- the most recent vocabulary items, which have not yet been recorded in any general Polish language dictionary.

The current stage of the project is not the last one. After its completion, the dictionary is expected to attain 50,000 main headwords (aside from entries describing idioms and proverbs) (Żmigrodzki, 2014: 37–40).

To provide a background to the current paper it seems indispensable to present the general characteristics of the dictionary:

- Two corpora (the National Corpus of Polish – www.nkjp.pl – and an auxiliary corpus created to serve the needs of the emerging dictionary) and Internet websites

constitute the sources of linguistic data for the dictionary. While preparing specific parts, other lexicographical sources are also used, e.g. the Grammatical Dictionary of Polish Language [Słownik gramatyczny języka polskiego] provides inflectional paradigms.

- The entries are compiled based on contemporary texts only: they include the sources that came into use after 1945.

- The dictionary is, in principle, descriptive, which means that the authors do not exclude from the description any lexicographical facts deemed incorrect. However, the normative unacceptability of a given fact (as per the Normative Dictionary of Polish [WSPP: Wielki słownik poprawnej polszczyzny PWN]) is highlighted.

- The dictionary employs wherever possible the achievements of Polish 20th Century linguistics, especially in the field of semantic, inflectional and syntactic descriptions of lexical units. However, the description is created in a way which is accessible to a very varied group of Polish language users[1].

- The macrostructure consists of single lexemes, idiomatic expressions and so-called functional words (prepositions, conjunctions etc.) as well as the most frequently used proverbs, abbreviations, acronyms and proper names.

- The microstructure covers a headword form (with variants); information about pronunciation (so far only for the words with unpredictable pronunciation, especially recent borrowings); chronology; etymology; description of meaning (in other words definition and, in polysemous entries, an additional guideword); thematic classification; superordinates, synonyms and antonyms of the entry word in the specific meaning; inflection (especially the full paradigm of the word's inflection, its affiliation to a part of speech); syntactic requirements (especially for verbs); collocations; full sentence quotations; abbreviations (if any); normative information (pertaining to some incorrect uses of the word); notes on usage (any other information pertaining to the usage of the word in texts). This set of information is used in the description of the two most numerous types of language units: single lexemes and idiomatic expressions (Żmigrodzki, 2014: 41–43).

1.3 Objectives – Survey Design – Tools

The aim of this article is to identify the types of the most popular entries in the GDP and to share these experiences. The paper outlines one aspect of GDP user behaviour

[1] According to Polish literature regarding this topic, the solutions considered as user-friendly are, for e.g.: the lack of abbreviations and symbols, and grouping all the information about the particular word in one place (like not using the references to inflectional information but joining it with the entry) (Żmigrodzki, 2005: 42).

(what they do, what entries they look up) and draws some conclusions regarding their needs and expectations (what they want, what kinds of entries they are prone to look up, and what are the reasons for such choices). The result of the analysis may indicate solutions for lexicographers facing the challenge of compiling monolingual general dictionaries from scratch, as well as for those continuing such work (including the GDP project itself). Undoubtedly, a dictionary should contain the entries which are needed by its users. This paper attempts to define these needs and identify their possible motivations. The latter issue is important if the conclusions are supposed to be useful for lexicographers participating in projects similar to the GDP, who should thus be able to compare GDP user motivation to that of their own users, as different motivations are reflected through different needs. The paper outlines the general interest in particular groups of entries available in the monolingual general dictionary. In other words, considering the behaviour of different users (e.g. foreign vs native speakers, children vs adults, professionals vs non-professionals) is beyond the scope of the survey. It can be assumed that all these mentioned groups (and many others) have some representation in the large set of those who entered the GDP in the analyzed period. Unfortunately, this approach may lead to an overrepresentation of the needs of those groups of people who use dictionaries more often than the others, e.g. editors, proofreaders, teachers or translators. This problem is well-known and some researchers try to solve it by devising profiles of users (Arhar Holdt et al., 2016). However, if the given analysis is to be useful for lexicographers working on a dictionary from scratch, it seems more effective to provide them with general conclusions. Meeting the expectations of different types of users is, seemingly, the next step in compiling a dictionary. This paper also provides a preliminary attempt to analyze the behaviour of GDP users – this is another reason for choosing a more general perspective for its starting point. Undoubtedly, it is going to be more detailed in the future.

The analyzed data were gathered by Google Analytics. The analysis covers 500 entries which were the most popular between 01.01.2013 and 31.12.2016. This period was chosen since 2013 was the first year of use of the dictionary after the completion of its first stage of preparation. It lasted four years and ended just before the beginning of the data collection which is presented in the current paper.

1.4 Obtaining the Data

Regarding the method of data collection, a few approaches can be distinguished: questionnaires, providing the participants with the task (e.g. a translation of the text), following user behaviour in online dictionaries and via eye-tracking. Collecting unobtrusive[2] data is more reliable in this type of research – that is why the Google Analytics tool was chosen. By "type of research" I consider the above mentioned goal:

[2] "In general, an unobtrusive method can be understood as a method of data collection without the knowledge of the participant [...]" (Müller-Spitzer et al., 2012: 427).

identifying the most popular entry types. According to many authors, analysing log files is not an ideal method of research into dictionary use; its disadvantages are considered by e.g. Müller-Spitzer et al. (2012), Müller-Spitzer et al. (2015), and de Schryver & Joffe (2004). It is probable that some of the problems raised by these authors regarding log files also apply to Google Analytics. However, as has already been mentioned, Google Analytics is a good choice for compiling a list of the most popular entries. Still, log files are used more commonly for studying user behaviour and it can be claimed that they have dominated empirical research in recent years (Lew, 2015: 235). Nonetheless, some researchers (e.g. Lorentzen & Theilgaard, 2012) have also used Google Analytics.

The list of the 500 most popular entries was compiled by making a report using: Behaviour – Site Content – All Pages and adjusting the Date Range (from 01.01.2013 to 31.12.2016). This part was executed automatically by Google Analytics. The most popular pages were ranked by measuring their page view[3] rate. The next step was to exclude those pages which did not refer to the entries, e.g. the search engine of the dictionary, the history of the dictionary, the instruction for users, the page presenting the authors; instead, this stage was completed manually. The ways of entering the sub-sites (e.g. the search engine of the GDP vs the search engine of Google, typing the whole headwords vs typing their parts only) do not fall within the scope of the survey.

2. The Study

It should be emphasised that no assumptions relating to the division of groups have been made in advance, before gathering data. In other words, the criteria for distinguishing groups of entries were prepared after ranking on the basis of character. In the study, two factors are considered: popularity of the given group of entries and its strength. The "popularity" is understood as the percentage of occurrences of the group in question in the ranking of 500 entries. The "strength" is measured in terms of the number of page views.

2.1 Remarkable groups of entries and their popularity

The first conclusion drawn from the observation of the 500 most-popular entries in the GDP is a domination of multi-word expressions over single lexemes. The latter group consists of 99 entries, whereas the former totals 401 entries. Additionally, the popularity of single lexemes is strongly correlated with their position in the rank. Among the 100 most popular entries (i.e. from 1st to 100th position) there are only 12 single lexemes, whereas among the 100 least popular ones (i.e. from 400th to 500th

[3] "A *pageview* is defined as a view of a page on your site that is being tracked by the Analytics tracking code. If a user clicks reload after reaching the page, this is counted as an additional pageview. If a user navigates to a different page and then returns to the original page, a second pageview is recorded as well." (Analytics Help, access: 13.05.2017).

position) there are 33. In the remaining hundreds, the number of single lexemes remains the same amounting to 18.

Having created the two categories (multi-word expressions and single lexemes), we face the problem of dividing them into subcategories as the abovementioned conclusion is far too general. There are many possibilities, e.g. singling out the types of multi-word expressions (MWEs) (verbal, noun, adjectival), contrasting polysemic and monosemic entries, distinguishing the loanwords. As previously mentioned, no criteria for division were given in advance. The analysis of the list led to two surprising conclusions: the proverbs appear to be extremely popular and words that seem to be part of the basic vocabulary scope are rare. The distinction of subcategories was based on these two statements.

2.1.1 Multi-word Expressions (MWEs)

Since among the MWEs proverbs are a distinctive group, the principle of looking for other subcategories was to check if there are any other types of MWEs (e.g. slogans, wing words, phraseological units). Those found in the ranking were: phraseological units and terms. These three subcategories are different in number: the subcategory of proverbs comprises the most popular entries (29%), phraseological units (50%) and terms (1.2%) (all numerical data are provided in Figure 1).

Among terms there is no regularity in form or meaning; they consist of full words as well as abbreviations (one example: *ABS* 1. 'Anti-lock Braking System', 2. 'Avalanche Airbag System'[4]) and concern different topics, e.g. *capital letter* [*drukowana litera*][5], *sign language* [*język migowy*], *collective responsibility* [*odpowiedzialność zbiorowa*].

A similar situation of ambiguousness can be found in two other groups: proverbs and phraseological units. Among proverbs there are examples of old units, *Guest at home, God at home* [*Gość w dom, Bóg w dom*], as well as quite contemporary ones, *The finger and the head are school excuses* [*Paluszek i główka to szkolna wymówka*]. The former is a proverb which encourages the warm and hospitable welcome of guests. In Polish texts, this proverb was noted for the first time in the 17th century (Krzyżanowski, 1969: 717). The latter example is used to deride the complaint of a minor ailment. This proverb has been noted since the end of 19th century (Krzyżanowski, 1972: 803). Both examples, as well as others, highlight the differentiation of mentioned topics: *Better to be safe than sorry* [*Gdyby / Żeby kózka nie skakała, toby nóżki nie złamała*], *He who is born to be hanged shall never be drowned* [*Co ma wisieć, nie utonie*], *One swallow does not make a summer* [*Jedna jaskółka wiosny nie czyni*], *After the New Year the days become longer very fast* [*Na Nowy Rok przybywa dnia na barani skok*]. It can also be

[4] The numbers mark polysemic entries.

[5] In brackets Polish equivalents are provided.

stated that GDP users were interested in popular, well-known proverbs as well as in those which are rarely used. The information regarding the popularity of proverbs is drawn from the research that established a paremiological minimum of the Polish language.[6] The latest one was completed in 2013 (Szpila, 2014) and it contains, for example, *Where two are fighting the third wins* [*Gdzie dwóch się bije, tam trzeci korzysta*], *What goes around, comes around* [*Co się odwlecze to nie uciecze*], *It is a mixed blessing* [*Każdy kij ma dwa końce*]. These proverbs are present in the paremiological minimum as well as in the ranking of the 500 most popular entries in the GDP. However, there are also units absent from the minimum list, even in its extended version from 2013[7]. Here are some examples: *Corruption starts at the top* [*Ryba psuje się od głowy*], *Humility gets you everywhere* [*Pokorne cielę dwie matki ssie*], *A nobleman at the farm is equal to a palatine* [*Szlachcic na zagrodzie równy wojewodzie*].

The differentiation in origins, meanings, forms and stylistic features is also characteristic for the most popular phraseological units in the GDP. Some of them originate from the Bible, mythology or literature, e.g. *Aesopian language* [*język ezopowy*], *Balzakian age* [*wiek balzakowski*], *in the arms of Morpheus* [*w objęciach Morfeusza*], *thorn in the side* [*cierń w oku*]; whereas others are quite new and originate from colloquial language: e.g., humorous equivalent of alcoholic drink [*napój wyskokowy*], units that can be translated literally as *a warmed up chop* [*odgrzewany kotlet*] 'sth that was known in the past, but then was forgotten and is currently presented falsely as sth new' and *sth gobbled so well and then it croaked* [*tak dobrze żarło i zdechło*] 'sth was going well, but the difficulties occurred'. Some traditional phraseological units referring to the world of nature or traditions passing by can be indicated here as well. One such unit can be literally translated as *a spoon of tar* [*łyżka dziegciu*]. The word used here refers to a kind of tar which is made in a process of distillation of wood. It has antiseptic and antifungal characteristics. The meaning of the unit is 'sth unpleasant in a generally good situation'. The phraseological units are also different in their forms – clause, noun, verb, adjective, adverb, exclamation: *hit the bull's-eye* [*strzał w dziesiątkę*], *abc* [*abc*], *the scales fell from sb's eyes* [*łuski spadają z oczu komuś*], *to loosen sb's tongue* [*język rozwiązał się komuś*], *pitch dark* [*choć/że oko wykol*], *my word* [*masz babo placek*]. A wide variety in style can be observed. The gathered units are bookish, *sb takes sb for a ride* [*ktoś gra znaczonymi kartami*] as well as neutral *light sleep* [*lekki sen*] and informal: *you pay your money and you take your* [*do wyboru, do koloru*], *bullshit* [*o dupie Maryni*].

[6] Paremiological minimum is „a set of proverbs that all members of society know or an average adult is expected to know" (Ďurčo, 2015: 183).

[7] The results of the survey were divided into two parts: the proverbs which were indicated by at least 8% of informants and fewer up to two informants; the latter version includes 254 examples (Szpila, 2014: 91-93).

2.1.2 Single Lexemes

As per previous observations, one preliminary conclusion was that few ranked entries seemed to be part of the basic vocabulary range. This statement was a starting point for distinguishing subcategories of single lexemes: lexemes with low and high frequency.

As the correlation between the corpus frequency of a word and the frequency of look-ups in online dictionaries is a subject of analyses[8], applying this method should probably be the natural choice for a given study. However, extracting the list of frequently used lexemes from the National Corpus of the Polish Language [NKJP] did not seem the best solution because of the dominance of the written texts. The National Corpus of the Polish Language contains about 10% of speech data. However, most are media speech data (transcriptions of TV and radio programs) and transcriptions of parliamentary speeches and discussions. The conversational speech data (transcriptions of dialogues of people of different ages, different education levels and descent) amount to about 900,000 tokens (Pęzik, 2012: 38-39). The problem of overrepresentation of data of the written language was also raised by Janusz Imiołczyk who addressed this problem with reference to frequency dictionaries (1987: 24). One of the aims of the basic frequency dictionary completed by the author (Imiołczyk, 1987) was to cope with this problem. To achieve the goal, Imiołczyk conducted a psychometric experiment in which he prepared a list of about 5,000 lexemes ordered according to the rule of subjective probability. He asked informants to fill in the questionnaires by labelling the given lexemes with numbers from 1 to 7 (where 7 means the word is used constantly and 1 means that the word is unknown or never used). For statistical analysis, he provided each lexeme with a rank (Imiołczyk, 1987: 34–39). The author claims that the frequency of words was not the only criterion used by informants; other important issues were: ordinariness, abstraction, connotative meaning (Imiołczyk, 1987: 48). This approach constitutes the next argument for using Imiołczyk's list instead of the rank list extracted from the corpora. Of course, the fact that the list was prepared 30 years ago cannot be ignored. However, I assume that due to the abovementioned reasons using this list is still the most trustworthy reference point. Additionally, in the list of the most popular single lexemes from the GDP there was no word created or borrowed during last 30 years (of course this fact does not exclude the possibility of new meanings)[9]. Thus, the term "frequency" should be

[8] E.g. (de Schryver & Joffe, 2004), (de Schryver et al., 2006), (Verlinde &Binon, 2010), (Koplenig, Meyer, Müller-Spitzer, 2014), (Müller-Spitzer et al., 2015).

[9] This situation is probably the result of the fact that the most recent vocabulary items are currently being added to the GDP, during the second stage of the project (started in September 2013). At the moment of collecting the analyzed data the entries describing the most frequently used words of the Polish language were available for users (Żmigrodzki, 2014: 39). Therefore, the entries being the newest vocabulary are still being prepared and cannot be fully represented in queries (their popularity can be checked later, after completing the current stage of the project).

abandoned and replaced with "subjective probability".

Comparing the list of the most popular single lexemes in the GDP and the list prepared by Imiołczyk confirms this intuitive assumption: units which are absent from his list dominate in the GDP look-ups. They amount to 14% of all analyzed entries (see Figure 1 – Single Lexemes: Low Subjective Probability), whereas lexemes included in the Imiołczyk list form only 5% of all analyzed entries (see Figure 1 – Single Lexemes: High Subjective Probability). The lexeme which has the highest[10] subjective probability and is present on the list of the most popular the GDP entries is *house/home* [*dom*], with rank 22. Other words from the first thousand entries of Imiołczyk's list include a perfective form of *to slice* [*ukroić*], *patience* [*cierpliwość*], *love* [*miłość*], *problem* [*problem*], *youth* [*młodzież*]. The last 21 words hold different places in Imiołczyk's list (from 1256 to 4808). The words which were not included in the list, which is equal to having low subjective probability, are, to name but a few: *abortively* [*aborcyjnie*], *stocky* [*krępy*], *gully* [*źleb*], *optimal* [*optymalny*], *liberalization* [*liberalizacja*], *absorption* [*absorbcja*], *submission* [*uległość*], *empirical* [*empiryczny*], *to whisper* [*szeptać*].

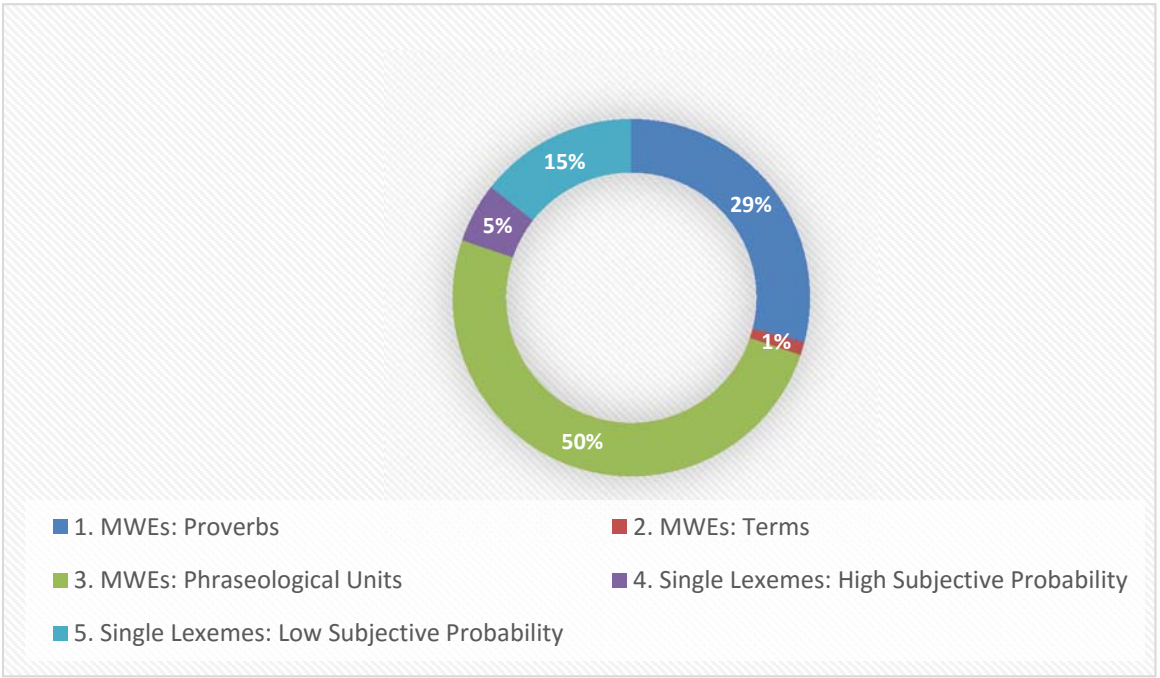


Figure 1: Popularity of Entries

2.2 Remarkable groups of entries and their strength

The attention of GDP users can be measured not only in the number of units representing remarkable groups, but also in their strength (represented by the number

[10] The lower the rank, the higher the subjective probability.

of page views, Figure 2). Its presence on the list of the 500 most popular entries is a distinctive factor. However, it is also important how many times the single entry was viewed.

The analysis of strength of entries (Figure 2) leads to conclusions similar to those drawn regarding their popularity (Figure 1) with reference to the single lexemes and being slightly different in the case of MWEs. The former group differs from the rate of popularity only in 1% with reference to the group of single lexemes with high subjective probability. The latter diverged from popularity in about 10% in both the most numerous subgroups – proverbs and phraseological units. As a matter of fact, the measurement of strength supports the thesis regarding GDP user interests in proverbs. When considering the number of page views, proverbs and phraseological units are almost equal and both constitute groups of entries drawing the most attention from users, despite the fact that the group of phraseological units consists of 250 units, whereas proverbs amount to 145 units.

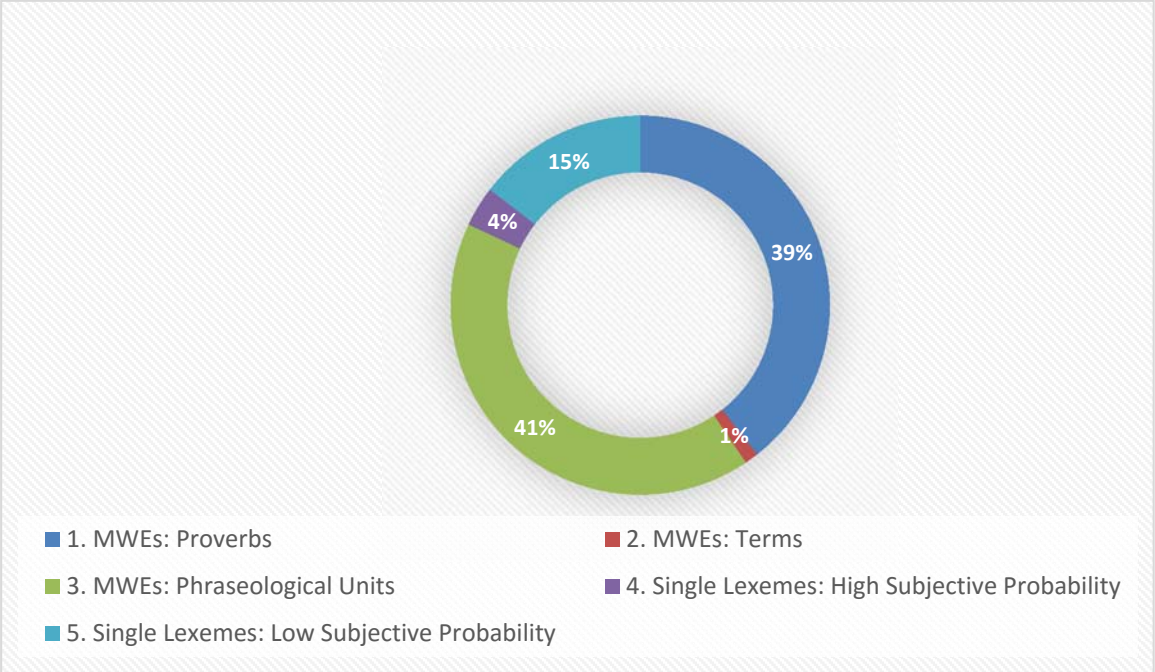


Figure 2: Strength of Entries

3. Findings

The gathered data reveal the following:

- GDP users are mostly interested in multi-word expressions, which constituted 401 of the 500 most popular entries, meaning that single lexemes cover only about 19.8% of the most popular entries.

- Single lexemes are represented in queries mostly by those with a low subjective

probability rate (15% of all most popular entries, 75% of the single lexemes), whereas single lexemes with a high subjective probability rate amount to only 5% of the most popular entries (25% of single lexemes). The popularity (percentage of the group in the rank of 500 most popular entries) and strength (measured in terms of the number of page views) for both subgroups are almost equal (Figure 1 and 2).

- In the MWEs three subgroups can be distinguished: proverbs, phraseological units and terms. Their popularity and strength is usually not equal (Figures 1 and 2) (except for terms). The popularity rank shows that phraseological units account for 50%, whereas proverbs account for 29% of all most popular entries (Figure 1). Considering the number of page views (strength, Figure 2) leads to the conclusion that the two subgroups are almost equal.

- All three subgroups of MWEs are varied in their origins, meanings, forms and stylistic features. No patterns in user needs can be indicated here.

4. Discussion

Knowledge of proverbs in the Polish language is decreasing, according to some authors, for the last 30 years (Buttler, 1989; Szpila, 2000). This observation is supported by empirical research on informants to establish the paremiological minimum of the Polish language. The research conducted in 1998 showed that the minimum consisted of 72 proverbs, whereas the survey from 2013 (using the same method and including as minimum only the units which were indicated by at least 8% of informants) identified only 39 proverbs (Szpila, 2014: 91–93). At the same time, GDP users are mostly interested in MWEs, particularly in proverbs. This surprising fact requires an additional comment.

One possible explanation for the interest in proverbs is school needs. This statement has been raised a few times during discussions among members of the GDP project. What would be the effect of confronting this assumption with school reality? The easiest way to check this is via school textbooks and other widely available sources. The term *proverb* is mentioned only once in the official document, which is currently in force and constitutes the basis of the syllabuses and textbooks of Polish schools (with regards to the subject "Polish language", devoted both to Polish language and literature). The document recommends that pupils from primary school years 4-6 should be able to recognize proverbs as well as stories, legends, novels and so on.[11] A little more attention is given to phraseological units. Pupils from junior high schools should be able to use phraseological dictionaries, understand phraseological units and use them. However, the exercises referring to phraseological units and proverbs often appear in textbooks for Polish language in primary schools and junior high schools.

[11] The document is called *the programme basis* and it is announced by the Minister of Education. The current one has been in force since December 2008.

To check how often Polish pupils face MWEs, four Polish language textbooks were analysed; three from primary schools (in accordance to the previously mentioned document this is the only stage of education which pays attention to proverbs): one chosen at random for each year from the second level of education (i.e. years 4, 5 and 6); and one from junior high school chosen at random from year 2. The scope of the analysis covered only textbooks (without workbooks or any other additional sources) and only those exercises in which pupils were obliged to work with MWEs. The explanations, definitions and texts regarding MWEs were not considered since it was assumed that pupils were encouraged to use dictionaries (e.g. the GDP) only when performing the task. The analysis showed that in the first part of the year 4 textbook there are seven exercises related to MWEs (Michałkiewicz & Mucha, 2011). In year 5 there are 10 exercises (Horwath & Żegleń, 2013), in year 6 19 exercises (Dobrowolska & Dobrowolska, 2014) and in year 2 of junior high school there are 15 exercises (Horwath & Kiełb, 2016). In each textbook, the tasks were mainly related to phraseological units (proverbs were in minority). The exercises comprise tasks such as: explain the meaning of MWE, check the meaning of MWE, create a sentence with MWE, find in a dictionary examples of MWEs containing a particular word and so on. The popularity of the topic is visible not only in textbooks but also in online educational webpages, e.g. *It is a mixed blessing* [*Każdy kij ma dwa końce*] present in www.sciaga.pl (in the part prepared by the website authors), www.zaliczaj.pl, www.zapytaj.onet.pl (as user questions).

On the other hand, exercises in which pupils were obliged to deal with single words were rarer. This fact partially confirms the assumption about the impact of school needs on GDP queries. Table 1 presents the exact MWEs and single words used in exercises for one of the analyzed school years, the fourth year of primary school. The table provides information on the presence or absence of the given MWEs and single words on the list of the 500 most popular entries of the GDP. It should be stated that the number of exercises (mentioned in the last paragraph) is not equal to the number of MWEs and single words. This is because a lot of exercises concern more than one lexical unit. Pupils are also obliged to find some MWEs and single words on their own (instructions such as: find the examples of MWEs containing the given word, give the examples of MWEs linked to the given topic, and so on).

Table 1 shows that in the fourth year of primary school MWEs were more numerous in the exercises than single words (29 MWEs vs 18 single words). Most are not present on the list of the 500 most popular entries from the GDP. The situation was similar in other analyzed textbooks – most lexical units in the exercises requiring meaning checks, finding synonyms or antonyms, or using the units in sentences were MWEs and not single words. Few of these lexical units were present on the 500 most common list.

MWEs	The presence of the MWE on the 500 most popular list	Single words	The presence of the single word on the 500 most popular list
to have a sour look on one's face [*ma skwaszoną minę*]	no	*popular* [*popularny*]	No
to put on a brave face [*nadrabia miną*]	no	*famous* [*sławny*]	No
his face fell [*zrzedła jej mina*]	yes	*scallywag* [*ziółko*]	No
looks askance at sb [*patrzy krzywym okiem*]	no	*fairytale* [*baśniowy*]	No
looks at sb piercingly [*przeszywa kogoś wzrokiem*]	no	vocabulary connected with theatre (chosen by pupils from the given text)	
looks on with a fixed stare [*postawiła oczy w słup*]	no	*tradition* [*tradycja*]	No
truth hurts [*prawda w oczy kole*]	no	*scholar* [*uczony*]	No
'very distant relative' [*dziesiąta woda po kisielu*]	no	*doctor* [*doktor*]	No
'a complete stranger' [*ani brat, ani swat*]	no	*associate professor* [*docent*]	No
as alike as two peas in a pod [*kubek w kubek podobny do*]	no	*house* [*dom*]	Yes
as alike as two peas in a pod [*kropka w kropkę podobny do*]	no	*cottage* [*chałupa*]	No

as alike as two peas in a pod [*podobni jak dwie krople wody*]	no	*small hut* [*chatka*]	No
'the spitting image of one's father/mother' [*wykapany tata, wykapana mama*]	no	*flat* [*mieszkanie*]	no
'talk man to man' [*porozmawiać z kimś po męsku*]	no	*apartment* [*apartament*]	No
'make a quick and firm decision' [*podjąć męską decyzję*]	no	*ruin* [*rudera*]	No
'severe rules' [*ojcowska ręka*]	no	*tenement* [*kamienica*]	No
'done in a way a woman would do' [*znać w czymś kobiecą rękę*]	no	*villa* [*willa*]	No
'woman's intuition' [*mieć kobiecą intuicję*]	no	*residence* [*rezydencja*]	No
'motherly heart' [*matczyne serce*]	no		
radiant smile [*promienny uśmiech*]	no		
glimmer of hope [*promyk nadziei*]	no		
glimmer of joy [*promyk radości*]	no		
glimmer of happiness [*promyk szczęścia*]	no		
feel at home [*czuć się jak u siebie w domu*]	no		
host [*pan/pani domu*]	no		

do the honours [*czynić honory domu*]	no		
'establish a family' [*założyć dom*]	no		
a friend of the family [*przyjaciel domu*]	no		
live out of a suitcase [*życie na walizkach*]	no		

Table 1: The MWEs and Single Words Used in Exercises from a Chosen Textbook for Polish Language for 4th Year Primary School Children

The conclusions of the given analysis are ambiguous. On one hand, the MWEs undoubtedly constitute an important part of school practice. On the other hand, it is clear that most MWEs (as well as single words) found in the textbook exercises were not present on the list of the 500 most popular entries in the GDP. Additionally, other relevant factors can be indicated here. One has already been mentioned: a lot of exercises oblige pupils to find MWEs not mentioned in the exercises. This fact excludes the possibility of preparing the list of MWEs (or single words) taught at school and checking their popularity in the GDP. Unfortunately, it is also impossible to combine the school activities related to the GDP queries with time periods. For example, at the moment of preparing the article there are five textbooks series which can be used for the Polish language subject in schools: in 2012 there were 10 (for years 4–6 and for junior high school)[12]. Additionally, teachers are not obliged to work through all textbook chapters nor to complete all exercises, but instead might set different exercises. Therefore, although this method would likely provide the most convincing evidence of the relation between the growing interest in MWEs and school needs, it is not a feasible analysis.

To sum up, it can be stated that pupils' needs are at least partially responsible for a big popularity of MWEs, especially proverbs. However, it is not the only reason. It is evident that some of the aforementioned examples of entries are not part of the school teaching program (e.g. colloquialisms). In seeking other reasons for the phenomena, the scope of other online dictionaries should be considered. It seems probable that users cannot find answers to their questions elsewhere and therefore turn to the GDP which results in the overrepresentation of the MWE queries.

[12] According to the official website of the Ministry of National Education related to textbooks (www.podreczniki.men.gov.pl).

When looking for sources like the GDP, the website www.sjp.pwn.pl should be considered. This is the source shared by one of the biggest Polish publishing houses, PWN. Under this address, one search engine enables the look-up of words and expressions in two general dictionaries, a spelling dictionary, a corpus and the answers given to questions which have been sent in by users over the past few years. Although this resource is vast, the overwhelming majority of the MWEs which were popular in the GDP cannot be found in dictionaries (some however appear in the user questions). Only 10 of 250 phraseological units which were most popular in the GDP are present in dictionaries provided by PWN publishing house, e.g.: *Aesopian language* [*język ezopowy*], *Balzakian age* [*wiek balzakowski*], *sb leads the way* [*ktoś wiedzie prym*]. Additionally, some are a part of the spelling dictionary, which means that the only available information is regarding their spelling.

5. Concluding Remarks

It has been shown that research on dictionary user behaviour should concern their typology. If not, results will over-represent the needs of the groups which use dictionaries more often than others (Arhar Holdt et al., 2016). The current study on GDP users does not overcome this obstacle; however, even when assuming that the gathered data are not fully representative, the study clearly shows that users are very interested in MWEs. This statement sheds new light on the previous analysis focused mainly on single lexemes.

Generally, the most important answer to the question regarding popular entries in the general monolingual dictionary (on the basis of the GDP) is that users look for MWEs, especially phraseological units and proverbs, and for single lexemes which are not well-known to them (i.e. having low subjective probability). Of course, this statement is not an absolute truth. When considering candidates for inclusion in the dictionary, one should think about additional circumstances which may influence user behaviour. The study demonstrates that this may be school needs or the content of the other dictionaries.

6. Acknowledgements

This scientific work was financed under the programme of the Ministry of Science and Higher Education entitled "National Programme for the Development of the Humanities" in the years 2013-2018, Project No.: 0016/NPRH2/H11/81/2013.

Praca naukowa finansowana w ramach programu Ministra Nauki i Szkolnictwa Wyższego pod nazwą „Narodowy Program Rozwoju Humanistyki" w latach 2013-2018, nr projektu: 0016/NPRH2/H11/81/2013.

7. References

Arhar Holdt, A., Kosem, I. & Gantar, P. (2016). Dictionary User Typology: The Slovenian Case. In Margalitadze, T. & Meladze, G. (eds.) *Proceedings of the XVII EURALEX International Congress: Lexicography and Linguistic Diversity.* Tbilisi: Ivane Javakhishvili Tbilisi State University, pp. 179–187.

Buttler, D. (1989). Dlaczego zanikają przysłowia w dwudziestowiecznej polszczyźnie?. Poradnik Językowy (5), pp. 332–337.

De Schryver, G.-M. & Joffe, D. (2004). On How Electronic Dictionaries are Really Used. In G. Williams & S. Vessier (eds.) *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004. Lorient: Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud*, pp. 187–196.

Dobrowolska, H. & Dobrowolska, U. (2014). *Jutro pójdę w świat*, Warszawa: WSiP.

Ďurčo, P. (2015). Empirical Research and Paremiological Minimum. In H. Hrisztova-Gotthardt & M. A. Varga (eds.) *Introduction to Paremiology: A Comprehensive Guide to Proverb Studies.* Warsaw/Berlin: De Gruyter Open Ltd., pp.183–205.

Horwath, E. & Kiełb, G. (2016). *Bliżej słowa (podręcznik do gimnazjum, kl. II)*, Warszawa: WSiP.

Horwath, E. & Żegleń, A. (2013). *Słowa z uśmiechem. Literatura i kultura.* Warszawa: WSiP.

Imiołczyk, J. (1987). *Prawdopodobieństwo subiektywne wyrazów. Podstawowy słownik frekwencyjny języka polskiego.* Warszawa: Państwowe Wydawnictwo Naukowe.

Kernerman, L. (1996). English Learners' Dictionaries: How Much do we Know about their Use? In Margalitadze, T. & Meladze, G. (eds.) *Proceedings of the XVII EURALEX International Congress: Lexicography and Linguistic Diversity.* Tbilisi: Ivane Javakhishvili Tbilisi State University, pp. 405–411.

Koplenig, A., Meyer, P. & Müller-Spitzer, C. (2014). Dictionary users do look up frequent words. A log file analysis. In: C. Müller-Spitzer (eds.) *Using Online Dictionaries.* Mannheim: De Gruyter Mouton, pp. 229-250.

Krzyżanowski, J. (eds.), (1969). *Nowa księga przysłów i wyrażeń przysłowiowych polskich.* Warszawa: Państwowy Instytut Wydawniczy.

Krzyżanowski, J. (eds.), (1972). *Nowa księga przysłów i wyrażeń przysłowiowych polskich.* Warszawa: Państwowy Instytut Wydawniczy.

Lew, R. (2011). Studies in Dictionary Use: Recent Developments. *International Journal of Lexicography*, 24 (1), pp. 1–4.

Lew, R. (2012). How can we make electronic dictionaries more effective? In S. Granger & M. Paquot (eds.) *Electronic Lexicography.* Oxford: Oxford University Press, pp. 343-362.

Lew, R. (2015). Research into the Use of Online Dictionaries. *International Journal of Lexicography*, 28 (2), pp. 232–253.

Michałkiewicz, T. & Mucha, K. (2011). *O to chodzi!*, vol. 1. Warszawa: Wydawnictwo

Stentor.

Müller-Spitzer, C., Koplenig, A. & Töpel, A. (2015). Online dictionary use: Key findings from an empirical research project. In *Electronic Lexicography*. Oxford: Oxford University Press, pp. 425–457.

Müller-Spitzer, C., Wolfer, S. & Koplenig, A. (2015). Observing Online Dictionary Users: Studies Using Wiktionary Log Files. *International Journal of Lexicography*, 28 (1), pp. 1–26.

Pęzik, P. (2012). Język mówiony w NKJP. In A. Przepiórkowski & M. Bańko & R.L. Górski & B. Lewandowska-Tomaszczyk (eds.) *Narodowy Korpus Języka Polskiego*. Warszawa: PWN, pp. 37–49.

Schryver de, G.M. & Joffe, D. & Joffe, P. & Hillewaert S. (2006). Do Dictionary Users Really Look Up Frequent Words? – On the Overestimation of the Value of Corpus-based Lexicography. *Lexikos* (16), pp. 67–83.

Szpila, G. (2000). Skamielina czy żywy organizm – przysłowie w prasie polskiej. In G. Szpila (eds.) Język trzeciego tysiąclecia: zbiór referatów z konferencji Kraków, 2–4 marca 2000. Kraków: Krakowskie Towarzystwo Popularyzowania Wiedzy o Komunikacji Językowej "Tertium", pp. 215–224.

Szpila, G. (2014). Znajomość przysłów wśród polskich studentów: minimum paremiologiczne. *Literatura Ludowa*, 58 (4-5), pp. 87–101.

Töpel, A. (2014). Review of research into the use of electronic dictionaries. In: C. Müller-Spitzer (eds.) *Using Online Dictionaries*. Mannheim: De Gruyter Mouton, pp. 13-54.

Verlinde, S. & Binon, J. (2010). Monitoring Dictionary Use in the Electronic Age. In A. Dykstra & T. Schoonheim (eds.) *Proceedings of the XIV EURALEX International Congress*. Leeuwarden/Ljouwert: Fryske Akademy – Afûk, pp. 1144–1151.

Żmigrodzki, P. (2005). *Wprowadzenie do leksykografii polskiej*. Katowice: Wydawnictwo Uniwersytetu Śląskiego.

Żmigrodzki, P. (2014). Polish Academy of Sciences Great Dictionary of Polish [Wielki słownik języka polskiego PAN]. *Slovenščina 2.0*, 2 (2), pp. 37-52.

Dictionaries & Websites:

Analytics Help. Accessed at: Analytics Help; https://support.google.com/analytics/answer/1257084#pageviews_vs_unique_views. (13 May 2017)

nkjp.pl. Accessed at: www.nkjp.pl. (24 May 2017)

podreczniki.men.gov.pl. Accessed at: https://podreczniki.men.gov.pl/dopuszczone_lista5.php?file=szko%C5%82a%20podstawowa%20(kl.%204-8) ; https://podreczniki.men.gov.pl/dopuszczone_lista3.php?file=szko%C5%82a%20podstawowa%20(kl.%20IV-VI) ; https://podreczniki.men.gov.pl/dopuszczone_lista3.php?file=gimnazjum. (4 July 2017)

sciaga.pl. Accessed at: www.sciaga.pl. (2 May 2017)
SGJP: *Słownik gramatyczny języka polskiego.* (2017). [Grammatical Dictionary of Polish Language], available at: www.sgjp.pl. (24 May 2017)
sjp.pwn.pl. Accessed at: www.sjp.pwn.pl. (14 May 2017)
wsjp.pl. Accessed at: www.wsjp.pl. (24 May 2017)
WSPP: *Wielki słownik poprawnej polszczyzny PWN.* (2010). Warszawa: PWN. [Normative Dictionary of Polish]
zaliczaj.pl. Accessed at: www.zaliczaj.pl. (2 May 2017)
zaliczaj.pl. Accessed at: www.zapytaj.onet.pl. (2 May 2017)

Pictorial Illustrations in Encyclopaedias and in Dictionaries – a Comparison

Monika Biesaga

Institute of the Polish Language at the Polish Academy of Sciences
E-mail: monika.biesaga@interia.pl

Abstract

One of the most important differences between an encyclopaedia and a dictionary, which is very often underlined in scientific papers, is the relatively common occurrence of pictorial illustration in encyclopaedias compared to dictionaries. Theoreticians indicate different goals of these two types of reference works. While an encyclopaedia describes objects using scientific knowledge, a dictionary presents words with linguistic arguments. Since the presumed differences are so crucial in their nature, the two types of reference works should not have much in common. On the other hand, a pictorial technique in dictionaries is relatively young and non-omnipresent, and furthermore, undoubtedly arose in a predominantly encyclopaedic surrounding. Therefore, in this paper, I have focused on this graphical distinction: do visual facilities in an encyclopaedia vary from their counterparts in general dictionaries? As a result of this analysis, it can be stated that, apart from general differences (aim of description, types of units, function of caption), an encyclopaedia and a dictionary have surprisingly much in common regarding the visual mode.

Keywords: multimodality; encyclopaedia; dictionary; theory of lexicography

1. Dictionaries and Encyclopaedias

Consuetudo altera natura est – custom is second nature. These famous words, written by Cicero, serve as a good starting point for this text. It goes without saying that in contemporary lexicographical practice, authors of reference works, along with lexicography theoreticians, distinguish between two types of reference works, namely encyclopaedia and dictionary (Lara 1989; Hartmann & James, 1998: 48–49; Burada & Sinu, 2016: 61–62). While encyclopaedias describe things (actual referents) from a scientific standpoint, dictionaries characterize senses of words and discontinuous units from a linguistic point of view (in fact they usually mix linguistic knowledge with common-sense and scientific knowledge—because of the immense impact of the latter on our everyday lives). It has also been mentioned in these papers that encyclopaedic works tend to present historical personalities and well-known places (logical definite descriptions). Therefore, the linguist will surely notice the overrepresentation of proper names, usually not included in general dictionaries (unless unabridged, compare, for example, Merriam-Webster Dictionary).

Apart from these aforementioned crucial distinctions, one of the most important differences between these two types of reference works, very often underlined in such

papers, is the presence of pictorial illustration: it is assumed "relatively common" in encyclopaedias; whereas in dictionaries it is perceived as "relatively rare" (Hartmann & James, 1998: 49). During the last 30 years, a significant number of papers concerning the pictorial technique[1] in general and bilingual dictionaries have been written (see: Hupka, 1989; Nesi, 1989; Stein, 1991; Langridge, 1998; Gangla, 2001; Jones, 2004; Müller-Spitzer, 2005; Gumkowska, 2008; Lew & Doroszewska, 2009; Kemmer, 2014; Klosa, 2016; Biesaga, 2017). Although the amount of obtained knowledge should be described as quite rich (see theoretical summary in Biesaga, 2016), in the case of encyclopaedic visual facilities we remain laypersons. There is likely to be a practical connection between these types of reference works. First, there is a century-long tradition of using graphical elements in encyclopaedias. As semioticians describe it, when a visual code of a certain type becomes standardized, its features tend to vanish in the eyes of the users. What is more, the conventionality of images is less recognized by the audience than the conventionality of verbal communication (Chandler 2007: 68, 77). On the other hand, readers expect from a new product the standard to which they have become accustomed. Did non-specialized lexicography perhaps adopt encyclopaedic pictorial practice without even realizing it? Second, in the case of many long-standing publishing houses, dictionaries grew along with encyclopaedias (e.g. MW and Encyclopedia Britannica, Larousse Dictionary and Encyclopaedia) further complicating this connection between crafts.

2. Scientific Procedure

In this paper, I would like to focus on this pictorial distinction: do visual facilities in encyclopaedias vary from their counterparts in general dictionaries? I will analyze a group of entries from an established Polish encyclopaedia in which graphical illustrations were used—namely *Encyklopedia PWN* (all entries starting with the letter A will be taken into account). Subsequently, a division of the illustrated meanings, according to the criteria of lexical semantics, will be made (common and proper names, language level, and thematic classification in the case of common names[2]). I will subsequently highlight the most typical visual techniques used by the authors of encyclopaedias (relation between the type of entry and the illustration). Next, the procedure will be partially repeated and implemented on the entries taken from the two well-established general and monolingual dictionaries[3] (one printed and one

[1] Pictorial technique is defined here as a special lexicographical (not artistic) approach related to the inclusion of illustrations in reference works. To avoid constant repetitions, the pictorial technique will be also called visual practice, graphical technique etc.

[2] Since proper names do not have meaning (sense, denotation, intension) they will not be subjected to this sort of thematic classification. More about the pictorial thematic classification can be found in Biesaga (2017); see also Batko-Tokarz (2008) for strictly verbal communication.

[3] In the case of analyzed general dictionaries, the relation between the verbal mode and the graphical mode will not be taken into account. This issue should be considered as a separate topic for a completely different paper or papers.

published on the Internet): *Ilustrowany słownik języka polskiego* (Illustrated Dictionary of Polish (IDP)—entries starting with the letters A and B) and *Merriam-Webster Dictionary (MW)*—entries starting with the letter A).[4] Accordingly, a preliminary pictorial distinction between encyclopaedias and dictionaries will be drawn.[5] Such a distinction will help lexicographers in the future to shape this visual technique more consciously and to increase the number and the types of entries which are graphically illustrated.

3. Encyclopaedias

3.1 Proper Names

Probably the most striking difference, connected with the illustrated encyclopaedic entries, is the overrepresentation of proper names, among them especially geographical names (toponyms). They form the majority of all definite descriptions subjected to this graphical process (71 entries). When it comes to the referential typology of entries, most inform their readers about places located outside their native country or beyond the borders of their native continent (e.g. *Abisyńska Wyżyna* [Ethiopian Highlands][6], *Abu Simbel*, *Aconcagua*, *Agra*, *Ahtamar*, *Algier* [Algiers], *Amir*, *Angkor*, *Antarktyda* [*Antarctica*], *Antyliban* [Anti-Lebanon Mountains], *Aso-San*, *Asuan* [Aswan], *Ayers Rock*). More rarely illustrated proper names are connected with places in Europe (*Aix-en-Provence*, *Aletsch*, *Akmaar*, *Apulia*, *Ateny* [Athens], *Atreusza Skarbiec* [Treasury of Atreus], *Avebury*). Exceptionally definite descriptions, related to Polish lands, are mentioned (*Augustowski Kanał* [Augustów Canal], *Antonin*).

I would like to emphasize the role of perceived 'exoticism' as a criterion for inclusion, which is not necessarily done for educational purposes only. It seems that it was inherited from an old encyclopaedic and thesaural tradition according to which the task of the author was to somehow awe his reader with unfamiliar places, animals or objects. This special exotic touch will be present, as we shall see, in our encyclopedic entries, including the illustrations (compare Picture 1).

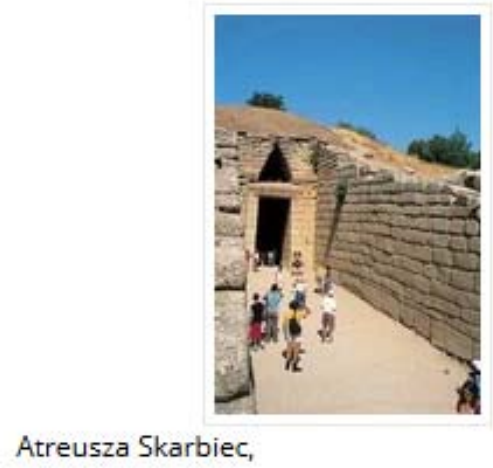
---

[4] Because of the declared number of entries (description balance) in the case of *Encyklopedia PWN* (122.000 entries) and MW (165.000 entries) only the letter A entries will be analysed, in the case of IDP (40.000 entries) both letters A and B will be subjected to scrutiny.

[5] In the next step, more reference works could be added to balance the results. However, one established encyclopedia with a long-standing tradition and two different, also well-established dictionaries should give quite reliable results of comparison.

[6] Translated versions of the entry labels will be used whenever such equivalents exist. If the label will be the same in Polish and in English only one head of such an entry will be provided for the reader.

Picture 1: An illustration from the entry *Adamawa* (Adamawa Plateau), caption: *Adamawa, village on the plateau*



Picture 2: An illustration from the entry *Atreusza Skarbiec* (Treasury of Atreus), caption: *Treasury of Atreus*

Authors illustrate different geographical entries that refer to natural objects (mountains, plateaus, rivers, volcanos etc.) or man-made places (cities, monuments etc.). What I would like to focus upon here is the problem of prototypical images, perceived by a certain culture as the most important for the place in question. Sometimes, while analyzing geographical entries, there was an impression that this certain proper name formed an independent entry just to present such a typical sort of monument or other similar object (e.g. *Agra* – images of Taj Mahal, *Alberobello – trullo* buildings, *Amritsar* – Golden Temple, *Andżar* [Anjar] – ruins of the Umayyad palace, *Awinion* [Avignon] – cathedral, see also Picture 3).

Typically, the role of these encyclopaedic visual aids is to either feature culturally important places within a city or some other location, or, on the other hand, to make the cultural image of the world standardized and, to some extent, flattened.

Akwizgran, katedra

Picture 3: An illustration from the entry *Akwizgran* (Aachen), caption: *Aachen, cathedra*

Apart from geographical proper names, *Encyklopedia PWN* describes multimodally historical personalities, ethnic groups and non-authentic personalities (category of anthroponyms in onomastics). For biblical and mythological personalities (7 entries), authors published their images taken from world-renowned works of art (e.g. *Adam i Ewa* [Adam and Eve] - see Picture 4, *Ahura Mazda*, *Amaterasu*, *Anubis*, *Apollo*, *Atena* [Athena]).

Cranach Lucas (st.), *Adam i Ewa w raju,*

Michał Anioł, *Grzech pierworodny i Wygnanie z raju, fragment fresków na sklepieniu Kaplicy Sykstyńskiej w Watykanie, 1509–12*

Biblia z Grandal,

Adam i Ewa, malowidło wczesnochrześcijańskie

Dürer Albrecht, *Adam i Ewa,*

Michał Anioł Buonarroti, *Stworzenie Adama,*

Picture 4: Illustrations from the entry *Adam i Ewa* (Adam and Eve) – different works of art were chosen

Regarding authentic personalities (30 entries), the pictorial technique is rich. First, we will consider portraits of these described personalities (paintings, photographs, etc.). Of interest is that this visual strategy is often connected with the personality's profession according to his or her typical work context (clothes, tools) and is displayed to the user (see Picture 5).

Aszantowie, farbowanie tkanin *(Benin)*

Picture 5: An illustration from the entry *Abraham Roman* (Polish general)

Secondly, the user is far more often offered the image of the personality's work result (painter: painting, architect: building, writer: image of the book [see Picture 6], scientist: his invention, famous ruler: a battle etc.). It seems that such a method, naturally, has something in common with deep semantic relations (compare the basic meaning shifts, typical for systematic and regular polysemy, see: Apresjan, 2000). This issue should be further analyzed with more visual evidence from different reference works (e.g. broad typologies of dictionary pictorial facilities, compare: Hupka, 1989; Stein, 1991).

fot. Archiwum Ilustracji WN PWN SA © Wydawnictwo Naukowe PWN

Picture 6: An illustration from the entry *Awicenna* (Avicenna) – image of his scientific tract

The last group of encyclopedic entries, connected with anthroponyms, refer to the tribes and other similar ethnic groups of the users (4 entries): e.g. *Aborygeni* [Indigenous Australians], *Ajmarowie* [Aymara people], *Aszantowie* [Ashanti people]. It is also worth pointing out that the use of 'exoticism` as a feature is particularly

prominent here (see Picture 7).

Awicenna, *Al-Kanun fi at-tibb,*

Picture 7: An illustration from the entry *Aszantowie* (Ashanti people), caption: *Ashanti people, dyeing fabric* (Benin)

Finally, there are two more entries that cannot be classified in any other previously listed group. The first entry, *Al-Fatiha,* informs the user about a famous surah from the Qur`an, and the second, *Apollo,* is devoted to NASA spaceships (see Picture 8).

Misja Apollo 11,

Apollo 2,

Apollo 11, załoga:

Picture 8: Illustrations from the entry *Apollo*, captions: *Apollo 11 mission*; *Apollo 2*; *Apollo 11, crew*

Entries related to proper names are highly standardized. The use of visual elements is limited to geographical and personal proper names; however, one could point out many visual possibilities (famous, contemporary non-commercial industrial products, works of art, monuments etc.). Furthermore, the ways in which objects and people are presented ought to be seen as conventional.

3.2 Common Names

As aforementioned, proper names that constitute the majority of all illustrated encyclopedic entries are usually not included in dictionaries. The case is completely different with common names, single words and discontinuous units, which are primarily incorporated into linguistic reference works. Of interest is that we will similarly find a significant number of illustrated common names in an encyclopedia. They will serve as a basis for a later comparison with multimodal dictionary entries.

Regarding the illustrated discontinuous units, all (11 entries) are scientific terms (e.g. *aberracja chromatyczna* [chromatic aberration, see Picture 9], *aberracja chromosomów* [chromosome abnormality], *agama kołnierzasta* [frill-necked lizard], *accelerator plazmy* [plasma accelerator], *algorytm Euklidesa* [Euclidean algorithm], *arnica górska* [arnica montana/wolf`s bane], *autonomiczny układ nerwowy* [autonomic nervous system]). Most of the referred terms belong to natural sciences (biology, physics, maths).

Picture 9: An illustration from the entry *aberracja chromatyczna* [chromatic aberration]

The situation is more complicated in the case of single words (30 entries). We encounter illustrated units connected with more professional areas of language (e.g. *aksonometria* [axonometry], *aktinidia* [actinidia], *apadana*, *azeotropia* [azeotropy]) along with entries related to everyday, basic language (e.g. *akordeon* [accordion], *ananas* [pineapple], *autostrada* [highway]). It is worth highlighting the special focus on entries describing exotic, non-native reality (*agawa* [agave], *aloes* [aloe], *alpaka* [alpaca], *atol* [atoll]) and historical items (*akwedukt* [aqueduct], *alabastron,* see Picture 10, *antyfonarz* [antiphonary] *astrolabium* [astrolabe]).

Alabastrony korynckie,

Picture 10: An illustration from the entry *alabastron*, caption: *Corinthian alabastrons*

Regarding thematic classification, the majority of illustrated entries relate to plants (11). Other popular categories include: animals, mathematical phenomena, musical instruments, architectural elements and optical phenomena.

As seen in the previous pictures, an encyclopaedia displays a wide use of captions. Verbal support serves not only to indicate the entry headword (semantic recognition, see Picture 2) but also to present additional knowledge. Since an encyclopaedia is focused on the transmission of accurate and precise information, a caption is often used to clarify the object in the picture (a certain type of thing generally described in the encyclopedia entry, elements of its context or scene presented, etc.). For example, almost all single word entries related to plants have associated pictures within which the caption points out the exact species presented to the user (see Picture 11).

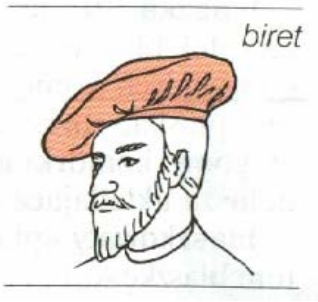
Aloes drzewiasty,

Picture 11: An illustration from the entry *aloes* (aloe), caption: *Krantz aloe*

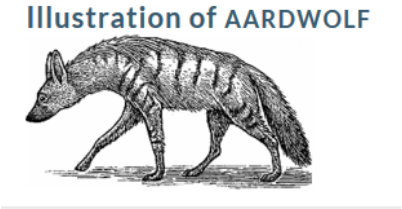
4. Dictionary

Regarding the two analyzed general dictionaries, we observe differences related to the area of vocabulary connected to the illustrated entries. IDP is considered mixed in its

visual approach: specialized and historical vocabulary (e.g. akant [acanthus], *bałałajka* [balalaika], *biret* - see Picture 12 [biretta], *bodziszek* [geranium], *buzdygan* [mace]) is subjected to this illustrative technique along with basic vocabulary (e.g. *autobus* [bus], *bocian* [stork], *brokuł* [broccoli], *budzik* [alarm clock]). On the other hand, MW tends to present more elaborated words (e.g. *aardwolf*, see Picture 13, *abelia*, *alpenhorn*, *aneurysm*, *ankh*, *anvil*, *arteriole*).



Picture 12: An illustration from the entry *biret* [biretta] (IDP)



Picture 13: An illustration from the entry *aardwolf* (MW)

Regarding the opposition between proper and common names, crucial for encyclopedia visual facilities, this does not exist in the dictionaries analyzed. They describe only the referents of common names. Additionally, in comparison to the *Encyklopedia PWN,* both general reference works illustrate only a few discontinuous units (IDP: *bez czarny* [elder/sambucus nigra]; MW: *angora goat*, *arctic fox*). Their presence is deemed accidental. The remaining entries with visual facilities represent the category of single words.

Captions also differ between an encyclopedia and a dictionary. While encyclopedic works use verbal support to identify a meaning or to clarify the content of a picture, in comparison to more general verbal information, dictionaries often do not use captions or, like IDP and MW, use it mainly (though not only) to identify the entry headword with the picture (see Picture 12).

Picture 14: A dictionary entry with a headword identifying caption (entry: *akacja* [acacia], IDP)

As well as these general differences between encyclopedias and dictionaries (type of linguistic objects which are illustrated; function of caption), there also exist similarities, especially when we consider the thematic division of meanings.

Similar to encyclopaedias, IDP and MW dictionaries tend to illustrate the meanings of words related to plants. This group of entries represents the majority of all dictionary units with visual facilities, like in the *Encyklopedia PWN*. In the IDP we will encounter, for example, entries such as: *akacja* (acacia), *aksamitka* (tagetes), *arbuz* (watermelon), *awokado* (avocado), *baobab*, *batat* (sweet potato), *bluszcz* (ivy), *bonsai*, *bulwa* (bulb). The authors of the MW dictionary offer to the user dictionary units like: *abelia*, *acorn*, *agave*, *almond*, *ash*, *asparagus*, *aster*. These special thematic characteristics are most probably inherited in general lexicography from encyclopaedic descriptions.

Both dictionaries offer a significant number of illustrative entries related to animals (e.g. in IDP: *albatross* [albatross], *anaconda* [anaconda], *batalion* [ruff], *bawół* [buffalo], *bażant* [pheasant], *bison* [bison]; in MW: *aardvark*, *addax*, *agouti*, *amoeba*, *arctic fox*, *armadillo*). Simultaneously, encyclopaedias offer just a few such units (e.g. *alpaka* [alpaca], *ara* [scarlet macaw], *archeopteryks* [archaeopteryx]).

A similar relation can be observed in the field of artistic activity (Musical Instruments and Fine arts). Both dictionaries tend to include such illustrations (e.g. in IDP: *akant* [acanthus], *arabeska* [arabesque]; *altówka* [viola], *banjo*, *bębenek* [tambour]; in MW: *accordion*, *alpenhorn*). We also encounter them in encyclopaedias: *akordeon* (accordion), *altówka* (viola).

Furthermore, the thematic field of architecture is shared by both types of reference works (e.g. in *Encyklopedia PWN*: *apadana*, *akwedukt* [aqueduct]; in IDP: *absyda* [apse], *architraw* [architrave], *arkada* [arcade], *bazylika* [basilica]; in MW: *alcazar*, *anta*, *arbor*). The same situation exists in the case of machines and devices (e.g. in *Encyklopedia PWN*: *akcelerator plazmy* [plasma accelerator], *arytmometr* [arithmometer], *astrolabium* [astrolabe]; in IDP: *brona* [harrow]; in MW: *abacus*, *anvil*). Other thematic fields, such as Transport, are present in both resources, albeit to a

lesser extent (e.g. in *Encyklopedia PWN*: *autostrada* [highway]; in IDP: *amfibia* [amphibia], *balon* [balloon], *bryczka* [chaise]; in MW: *airplane*, *anchor*).

Although some thematic fields activated in both dictionaries were not found in the analyzed encyclopaedia entries (e.g. Army and War, Closest Human Environment, Clothing, Sport and Leisure time, Body Parts and Body Functions, Diseases and Treatment), it is evident how they could be included, in accordance with previous encyclopedic experiences. Therefore, the thematic difference between illustrated entries from encyclopaedias and their counterparts from dictionaries is not qualitative, but rather quantitative. The only semantic fields that come to mind and could be underrepresented pictorially in encyclopaedias are those connected with humans' closest environment (elements are widely known to users, too obvious to be scientifically described) and human social and psyche life (subjective and abstract meanings, which cannot be easily presented in a picture and are hardly ever included in dictionary, see Biesaga, 2017).

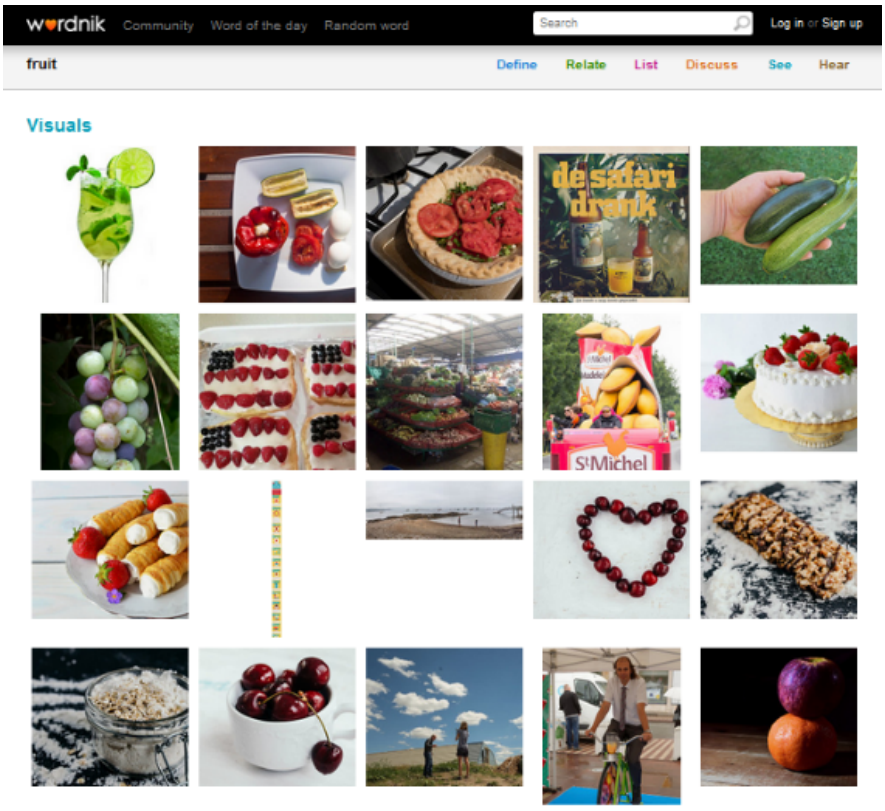
5. Similarities and Differences in Pictorial Technique

The table below gathers the most important differences and similarities between an encyclopaedia and a dictionary, each of them related to visual technique.

Criterion	Encyclopaedia	Dictionary
The presence of proper names	prevalent	none (except for unabridged dictionaries)
The presence of discontinuous unit meanings	significant (scientific terms)	very few
The presence of single word meanings	moderate	prevalent
Area of vocabulary	mostly advanced and specialized vocabulary	mostly basic and advanced vocabulary
Caption	very important, often supplementary knowledge or details given	limited
Most activated thematic fields	Plants	Plants
Moderately activated thematic fields	Animals, Artistic Activity, Architecture, Machines and Devices, Transport	Animals, Army and War, Artistic Activity, Architecture, Machines and Devices, Transport

Table 1: Pictorial techniques in encyclopaedias and dictionaries – comparison.

To summarize, this analysis leads us to different kinds of conclusions. On the one hand, illustrations in encyclopaedias and dictionaries reflect the general purpose of the reference work. This explains why we will find so many pictorial entries connected with certain places, authentic personalities and scientific terms in an encyclopaedia. Their mission is to transmit scientific and cultural knowledge which is presumed important for the user living in a certain society. These characteristics relate to risky decisions because the authors point out what does not belong to a native culture. This could create a temptation to underline this "being-foreign" category which could lead to an abusive usage of graphical materials (political correctness issues). On the other hand, if the author points out culturally and scientifically important graphical objects, he is automatically leaving other images outside the descriptive scope. That helps in information selection (we cannot know everything about everything pictorially) but standardizes and narrows the image of the world. Therefore, in comparison to dictionaries, encyclopaedic graphical descriptions wield a much greater responsibility.



Picture 15: Illustrations from the entry *fruit* (WordNik)

Aside from their differences (general aim of description, types of units, function of captions), encyclopaedias and dictionaries have much in common when it comes to the pictorial technique. Basically, they illustrate similar thematic categories of single word units (plants, animals, architecture, machines and devices etc.). What dictionary authors could learn from the encylopaedic craft is the incorporation of a wider selection of discontinuous units that are subjected to visual description. This would be especially helpful in relation to scientific terms. They are strongly resistant to accurate

description with verbal units that belong to natural, nonspecialized language.

As a final note, I would like to mention a new trend that is developing in the field of Internet lexicography, a tendency connected with the broadly understood encyclopaedic paradigm in illustrated reference works. There are several projects (e.g. BabelNet, eLexiko, WordNik) in which certain entries are illustrated with a set of different illustrations, sometimes automatically taken from multimodal corpora (see Picture 15).

Such a strategy enables us to solve many of the previously indicated pictorial problems, those typical for the printed lexicographical era (e.g. lack of prototypical example, generic meanings, etc., compare Hupka, 1989: 711; Stein, 1991: 119-120). Like with an encyclopaedia, the user is offered additional knowledge which boldly exceeds the dictionary paradigm. Furthermore, webpages seem to be the perfect "spacious" media for this kind of technique. On the other hand, however, in comparison to website capacities, user attention is under constraint. It is not clear which thematic fields are apt for illustrating (complete list), which are the crucial and additional visual features for addressing meaning (exhaustive typology connected with the types of senses is a must), how many illustrations are required for the meaning recognition process, etc. Therefore, a further theoretical and experimental analysis is advocated. One could also imagine in the future an open access pictorial repository for lexicographers (similar to WordNet).

6. Acknowledgements

This scientific work was financed under the auspices of the Ministry of Science and Higher Education entitled "National Programme for the Development of Humanities" in the years 2013-2018, Project No.: 0016/NPRH2/H11/81/2013.

Praca naukowa finansowana w ramach programu Ministra Nauki i Szkolnictwa Wyższego pod nazwą „Narodowy Program Rozwoju Humanistyki" w latach 2013-2018, nr projektu: 0016/NPRH2/H11/81/2013.

7. References

Apresjan, J.D. (2000). Semantyka leksykalna, Wrocław: Zakład Narodowy im. Ossolińskich. Wydawnictwo.

Batko-Tokarz, B. (2008). Tematyczny podział słownictwa w Wielkim słowniku języka polskiego. In R. Przybylska & P. Żmigrodzki (eds.) *Nowe studia leksykograficzne 2*, Kraków: Lexis, pp. 31-48.

Biesaga, M. (2016). Pictorial Illustration in Dictionaries. The State of Theoretical Art. In T. Margalitadze, G. Meladze (eds.) *Proceedings of the XVII EURALEX International Congress. Lexicography and Linguistic Diversity*, Tbilisi: Ivane Javakhishvili Tbilisi State University, pp. 99-108.

Biesaga, M. (2017). Which Vocabulary Thematic Fields Should be Ilustrated? Dictionary Tradition versus Pictorial Cor*pora*,TBA (in print).

Burada, M. & Sinu, R. (2016). *Research and Practice in Lexicography*. Braşov: Transilvania University of Braşov.

Chandler, D. (2007). *Semiotics. The Basics*. Second Edition, London-New York: Routledge.

Gangla, L.A. (2001). Pictorial Illustrations in Dictionaries, MA Thesis, supervisor: Daniel Prinsloo, Pretoria: University of Pretoria. Accessed at: repository.up.ac.za/bitstream/handle/2263/22862/Complete.pdf (19 May 2017).

Gumkowska, A. (2008). The Role of Dictionary Illustrations in the Acquisition of Concrete Nouns by Primary School Learners and College Students of English, BA Thesis, supervisor: Robert Lew, Szczecin: Collegium Balticum.

Hartmann, R.R.K. & James, G. (1998). *Dictionary of Lexicography*. London/New York: Routledge.

Hupka, W. (1989). Die Bebilderung und sonstige Formen der Veranschaulichung im allgemeinen einsprachigen Wörterbuch. In F. J. Hausmann, O. Reichmann, H.E. Wiegand, L. Zgusta (eds.) *Dictionaries. An International Encyclopedia of Lexicography*, vol. 1. Berlin/New York: Walter de Gruyter, pp. 704-726.

Jones, L. (2004). Testing L2 Vocabulary Recognition and Recall Using Pictorial and Written Text Items. *Language Learning&Technology*, 8(3), pp. 122-143.

Kemmer, K. (2014). Rezeption der Illustration, jedoch Vernachlössigung der Paraphrase? In C. Müller-Spitzer (ed.) *Using Online Dictionaries*, Berlin/Boston: de Gruyter, pp. 251-278.

Klosa, A. (2016). Illustrations in Dictionaries; Encyclopeadic and Cultural Information in Dictionaries, In P. Durkin (ed.) *The Oxford Handbook of Lexicography*, Oxford: OUP, pp. 515-531.

Kwaśnicka-Janowicz, A. (2007). Problemy związane z wprowadzaniem ilustracji do haseł słownikowych. In P. Żmigrodzki, R. Przybylska (eds.) *Nowe studia leksykograficzne*, vol. 1. Kraków: Lexis, pp. 163-174.

Langridge, S. (1998). The Genesis and Development of Dictionary Illustrations. In R. de Beaugrande, M. Grosman, B. Seidlhofer (eds.) *Language Policy and Language Education in Emerging Nations: Focus on Slovenia and Croatia and with Contributions from Britain, Austria, Spain and Italy*. Stamford-London: Ablex Publishing Corporation, pp. 69-76.

Lara, L.F. (1989). Dictionnaire de langue, encyclop**é**die et dictionnaire encyclop**é**diqe: le sens de leur distinction. In F.J. Hausmann et al. *W**ö**rterb**ü**cher. Dictionaries. Dictionnaires. Ein internationales Handbuch zur Lexikographie*, vol. 1, Berlin-New York: de Gruyter, pp. 280-287.

Lew, R. & Doroszewska, J. (2009). Electronic Dictionary Entries with Animated Pictures: Lookup Preferences and Word Retention. *International Journal of Lexicography*, 22(3), pp. 239-257.

Liu, X. (2015). Multimodal Definition: The Multiplication of Meaning in Electronic Dictionaries. *Lexikos*, 25, pp. 210-232.

Müller Spitzer, C. (2005). Vorüberlegungen zu Illustrationen in elexiko. In U. Haß (ed.) *Grundfragen der elektronischen Leixikographie. Elexiko – das Online-Informationssystem zum deutschen Wortschatz*. Berlin-New York: de Gruyter, pp. 204-226.

Nesi, H. (1989). How many words is a picture worth? A review of illustrations in dictionaries. In M.L. Tickoo (ed.) *Learners` Dictionaries: state of the art*, Singapore: SEAMEO, pp. 124-134.

Stein, G. (1991). Illustrations in Dictionaries. *International Journal of Lexicography*, 2, pp. 100-127.

Dictionaries, Encyclopaedias and Translation tools:

BabelNet: Accessed at: http://babelnet.org/ (5 July 2017).

eLexiko: Online-Wörterbuch zur deutschen Gegenwartssprache, Accessed at: www.elexiko.de (5 July 2017).

Encyklopedia PWN, Acessed at: http://encyklopedia.pwn.pl/ (20 May 2017).

IDP: Ilustrowany słownik języka polskiego (English translation: The Illustrated Dictionary of Polish), ed. E. Sobol, Wydawnictwo Naukowe PWN, Warszawa 1999.

MW: Merriam-Webster Dictionary. Accessed at: www.merriam-webster.com (1-27 January 2017).

Wordnik: Accessed at: www.wordnik.com (5 July 2017).

A *lemon* Model for the ANW Dictionary

Carole Tiberius[1], Thierry Declerck[2]

[1] Instituut voor de Nederlandse Taal, Matthias de Vrieshof 2,
2311 BZ Leiden, the Netherlands
[2] DFKI GmbH – Multilingual Technologies, 3 Stuhlsatzenhausweg,
D-66123 Saarbrücken, Germany
E-mail: carole.tiberius@ivdnt.org, declerck@dfki.de

Abstract

In this paper, we explore how we can reuse data from the ANW – an online corpus-based, scholarly dictionary of contemporary standard –, improve and optimise it by porting some of its elements into modules of the *lexicon model for ontologies (lemon)*. For the current study, the focus was set on the application of the ontolex and decomp modules, together with the associated LexInfo vocabulary in order to model the semantic and morphosyntactic features of nominal entries in the ANW.
We observe that encoding the ANW information in *lemon* has a number of advantages, including a better modularisation of the data, linking to other (lexical) data and data access using the standardised SPARQL query language.

Keywords: *lemon* model; lexical entry; semagram

1. Introduction

The Algemeen Nederlands Woordenboek (ANW) is a comprehensive online scholarly dictionary of contemporary standard Dutch, which is being compiled at the Dutch Language Institute.[1] It was set up as on online dictionary from the start and, as such, it truly represents a new generation of electronic dictionaries in the sector of academic and scientific lexicography. The dictionary focuses on written Dutch and covers the period from 1970 onwards. For a general introduction to the ANW and its features, the reader is referred to Schoonheim and Tempelaars (2010).

In this paper, we explore how we can reuse ANW data, and improve and optimise its internal formal representation by porting some of its elements into modules of the LExicon Model for ONtologies (*lemon*), using the version published as the result of the W3C Ontology-Lexica Community Group.[2] The original aim of *lemon* was to provide rich linguistic grounding for ontologies. This grounding includes the formal representation of morphological and syntactic properties of lexical entries as well as the syntax-semantics interface, i.e., the meaning of these lexical entries with respect to

[1] See http://ivdnt.org/the-dutch-language-institute [accessed 18.05.2017]

[2] See https://www.w3.org/2016/05/ontolex/ [accessed 18.05.2017]

a descriptive vocabulary or an ontology (McCrae, 2012).

The main modules of *lemon* are:

- Ontology-lexicon Interface (ontolex)
- Syntax and Semantics (synsem)
- Decomposition (decomp)
- Variation and Translation (vartrans)
- Linguistic Metadata (lime)

For the current study the focus was set on the application of the ontolex and decomp modules, and we used the associated LexInfo vocabulary.[3] Ontolex is, in fact, the core module of *lemon*, describing in detail the interface between elements of a lexical entry and the conceptual or world knowledge encoded in lexicon external knowledge bases. Decomp is the module depicting how to encode elements that are a part of a multi-word or compound lexical entry. Both modules are graphically displayed in Figure 1 and Figure 3. LexInfo, building in part on the ISOcat vocabulary[4], is an ontology that was defined to provide data categories (e.g., to denote gender, number, part of speech, etc.) for the *lemon* model.

Our starting point for the study is given by a small set of representative examples of ANW lexical entries encoded in an internal XML format. Our work consisted of proposing a mapping of this XML format onto the *lemon* vocabulary, which makes use of OWL, RDF(s) and RDF constructs.[5] The objective is to investigate if the ANW data can be encoded in an improved modular manner, supporting a higher level of re-usability within the ANW dictionary environment and an improved interoperability with other data sources, especially in the context of the Linked Open Data framework.[6] At the same time, the ANW data offer an excellent source for testing the validity of the *lemon* approach for comprehensive lexicographic resources (similar to the work by El Maarouf et al. (2014), Bosque-Gil et al. (2016), Kahn et al. (2017) or Stolk (2017)) and for suggesting potential extensions.

2. Data Modelling

We started our study with the description of nominal entries in the ANW dataset, considering in the first instance a description of the semantic and morphosyntactic

[3] See http://lov.okfn.org/dataset/lov/vocabs/lexinfo for more details.

[4] See http://www.isocat.org/ for more details.

[5] See respectively https://www.w3.org/OWL/, https://www.w3.org/TR/rdf-schema/ and https://www.w3.org/RDF/

[6] See http://linkeddata.org/ and also the Linguistic Linked Open Data cloud (http://linguistic-lod.org/llod-cloud).

features of these entries. As mentioned in the introduction, the ANW is a scholarly dictionary, providing a detailed description of each lexical entry. In the dictionary, special attention is paid to words in context (combinations, collocations, idioms, proverbs), relations with other words (lexical relations like synonymy, antonymy, hypernymy, hyponymy), semantic relations within the entry (metaphor, metonymy, generalisation, specialisation) and morphological patterns, the word structure of derivations and compounds. This means that the ANW has a rich microstructure.

To model the ANW microstructure with *lemon*, we start with its core module, ontolex (ontology-lexicon interface), as depicted in black in Figure 1[7]. In red we mark the additional elements, either taken from the LexInfo vocabulary or our suggestions, for extending LexInfo in order to account for ANW features.

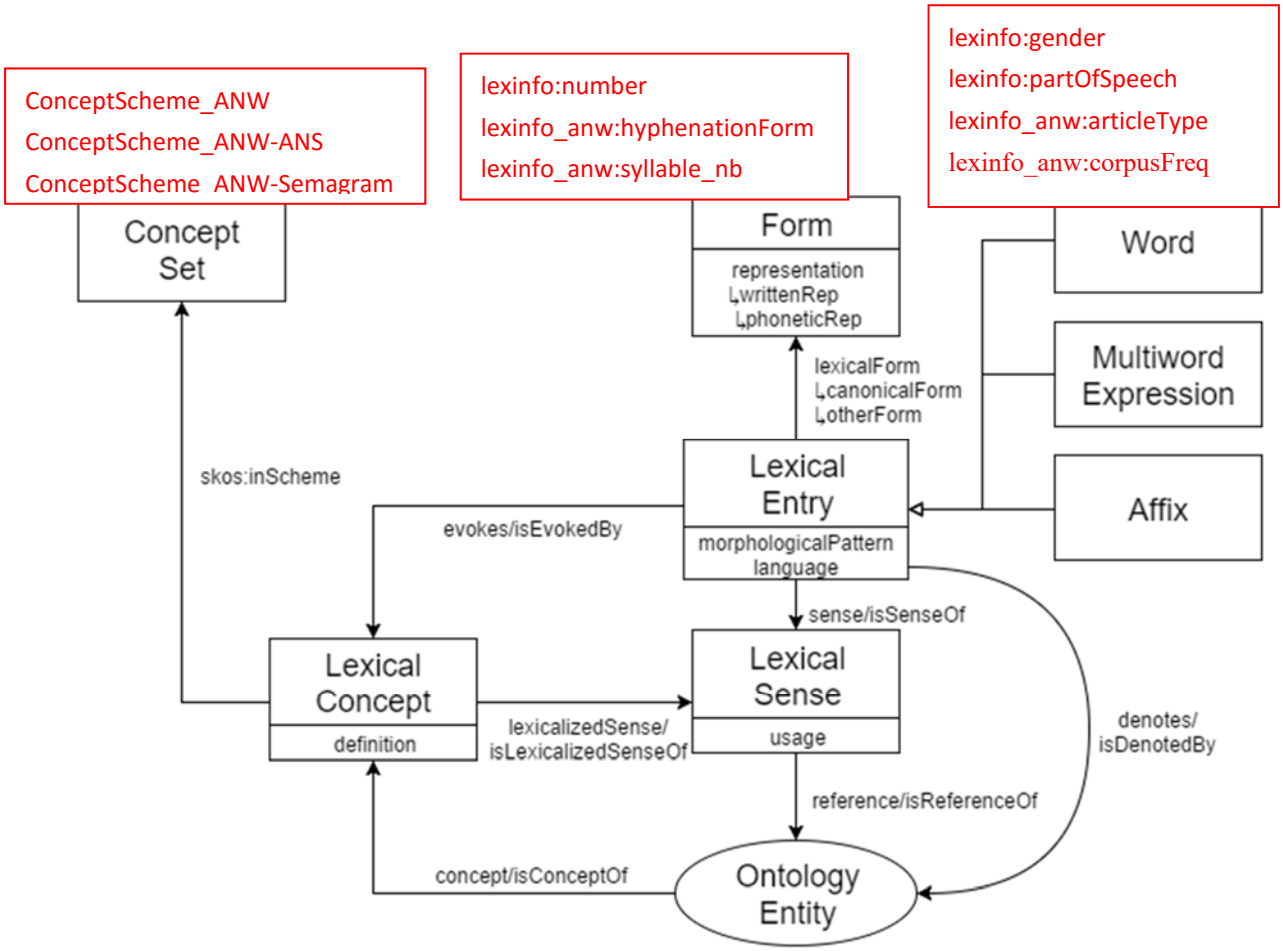


Figure 1: The ANW data model based on ontolex, the core module of *lemon*

In the remainder of this section, we discuss the *lemon* model for the ANW on the basis of an example entry, i.e. *wijn ('wine')*[8]. First we discuss the morphosyntactic encoding, then the semantic encoding and we conclude with the modelling of compounds.

[7] https://www.w3.org/2016/05/ontolex/. Figure created by John P. McCrae for the W3C Ontolex Community Group.

[8] http://anw.inl.nl/article/wijn.

2.1 Encoding of morphosyntactic information

Table 1 lists the morphosyntactic features for the entry *wijn* ('wine') in the ANW. The corresponding *lemon* encoding is given in the last column. Our suggested extensions to the LexInfo vocabulary include the sub-string `anw` and are marked in red.

ANW Features	**ANW Data**	***lemon* encoding**
Lemma		
	Lemma form: wijn	:form_wijn ... ontolex:writtenRep
	Lemma type: woord	rdf:type ontolex:Word
Syntactic Category		
	Type: noun	lexinfo:partOfSpeech lexinfo:noun
	Name type: soortnaam	lexinfo:partOfSpeech lexinfo:commonNoun
	Gender: mannelijk	lexinfo:gender lexinfo:masculine
	Article: de	lexinfo_anw:articleType
	Meaning class : stofnaam ('substance noun')	:ConceptSchema_ANW-ANS :Concept_SubstanceNoun
Spelling and Flexion		
	Forms	:form_wijn_singular ; :form_wijnen_plural
	Singular form: wijn	lexinfo:number lexinfo:singular
	Singular hyphenation: wijn	lexinfo_anw :hyphenationForm
	Plural form: wijnen	lexinfo:number lexinfo:plural; ontolex:writtenRep
	Plural hyphenation: wijnen	lexinfo_anw:hyphenationForm
Pronunciation		
	Number of syllables: 1	lexinfo_anw:syllable_nb
	Phonetic transcription: *w ɛɪ n	ontolex:phoneticRep
Morphology		
	Type: ongeleed ('simplex')	We have no mapping for this as the ontolex class "Word" is disjoint with the class "MultiWordExpression" and therefore has as instances only "non-compound" words
Usage information	Frequency: 6970	lexinfo_anw:corpusFreq

Table 1: Details of the ANW morphosyntactic features for the entry *wijn*

Table 1 shows that most of the morphosyntactic information encoded in the ANW can be coded in *lemon* using the ontolex module and the associated LexInfo vocabulary. Only a few extensions were introduced; for instance, the number of syllables of a word. The encoding of this information is currently not foreseen in LexInfo. However, we feel that this property may also be useful to other lexical resources, therefore we added `lexinfo_anw:syllable_nb`. The same applies to the features hyphenation, frequency and (morphological) type, which do not seem to be language-specific.

An example of a necessary extension that seems to be specific to Dutch, is the feature `lexinfo_anw:articleType,` which contains information on the type of definite article

that is required by the nominal lexical entry. This information is encoded in the ANW because in Dutch it is important to know with which definite article a noun can be used. Dutch has two definite articles; some nouns can only be used with the definite article *de*, some can only be used with the definite article *het*, some cannot have a definite article, and some can be used with either definite articles. In some instances, where both articles are possible, there is a preference for either *de* or *het*. We were unsure how to encode this preference information in *lemon*. This issue also applies to labels which mark that a word or meaning is *mostly* used in singular (or in plural) or in a particular language variety or region, etc.

2.2 Encoding of semantic information

Table 2 shows the information structure for the main sense of the lexical entry *wijn*, the sense of an 'alcoholic drink of fermented grape juice'.

ANW lexical features	**ANW Data**	***lemon* encoding**
		ontolex:LexicalSense
Lemma	[see above]	
Syntactic Category	[see above +]	
	Number: no plural	ontolex:usage lexinfo:massNoun[9] ontolex:usage lexinfo:singular
Pronunciation	[see above]	
Spelling and Flexion	Forms:	
	Singular form: wijn	lexinfo:number lexinfo:singular
	Singular hyphenation: wijn	lexinfo_anw:hyphenationForm
Usage Information	[see above]	
Meaning:	alcoholhoudende drank, verkregen door gisting van het sap van druiven of van andere vruchten, met een middelmatig alcoholgehalte van doorgaans ongeveer 12 procent; alcoholhoudende drank van gegist druivensap	:ConceptScheme_ANW skos:definition
Minidefinition	alcoholhoudende drank van gegist druivensap	:minidefinition
Word Relations		
	Hypernym: drank	:lexinfo hypernym
Semagram		:ConceptSchema_ANW-Semagram
	Top category: is stof	:Semagram_Stof
	Upper category: is vloeistof	:Semagram_Vloeistof
	Category: is drank	:Semagram_Drank
Example sentences	[...]	Not focus of current study
Combinations		Not focus of current study
	Combination type*: as subject of a verb	
	Realisation: gisten, rijpen	

[9] Here, we use the LexInfo element `massNoun`, since such a noun is typically uncountable. But we could also introduce a new element `uncountable`, to be more precise and explicit on this feature.

	Example sentences: […]	
Fixed Expressions		Not focus of current study
	Form*: nieuwe wijn in oude zakken (with definition and example sentences)	
Proverbs		Not focus of current study
	Form*: Wijn op bier is plezier en bier op wijn is venijn (definition and example sentences)	
	Form variant: Wijn na bier is plezier en bier na wijn is venijn; ... (including meaning description)	
Word family		Not focus of current study
	Right-headed compounds: abdijwijn; alsemwijn; ...	
	Left-headed compounds: wijnaanbod; wijnacademie; …	
	Derivational compounds: wijnkleurig; wijnmakerij; ...	

Table 2: Details of information for the main sense of the ANW entry *wijn,* sense 1.0

As can be seen in Table 2, the ANW contains semantic information about the lemma in various information categories within the entry, i.e., within the definitions, within the semagrams (an innovative feature of the ANW, described below in Section 2.2.3) and for nouns also in the so-called meaning classes.

2.2.1 Definitions

As any traditional monolingual dictionary, the ANW contains definitions that explain the meaning of the entry. In addition, the ANW provides mini definitions, i.e., short definitions that are used in sense menus to give the user a quick impression of the different senses of a word.

2.2.2 ANS Meaning classes

For nouns, the ANW also classifies the different senses of an entry in so-called meaning classes, a semantic classification of nouns which is based on the *Algemene Nederlandse Spraakkunst (ANS;* Haeseryn et al., 1997*).*

On the basis of Table 3, the following values are distinguished in the ANW: human nouns, animal nouns, object nouns, substance nouns, collective nouns, abstract nouns, proper nouns and plant nouns (an additional value in the ANW). The advantage of having these meaning classes is that it enables lexicographers to provide a global labelling for the sense distinctions. More precise sense information is given in the semagrams in the ANW.

Nouns			**common**	**proper**
concrete	**individual** voorwerpsnamen	**human nouns** persoonsnamen **animal nouns** diernamen **object nouns** zaaknamen	*man 'man', meisje 'girl', huis 'house'*	*Jan, Minou, Amsterdam*
	substance stofnamen		*water 'water', bier 'beer', goud 'gold'*	
	collective verzamelnamen		*vee 'cattle', kroost 'offspring', gebergte 'mountains'*	*Alpen 'Alps', Antillen 'Antilles'*
Abstract			*maand 'month', voetbalclub 'football club', goedheid 'kindness'*	*april 'April', Vitesse, romantiek 'romantics'*

Table 3: Semantic classification of Nouns according to the *Algemene Nederlandse Spraakkunst*

2.2.3 Semagrams

Semagrams are an innovative feature of the ANW, which were introduced by Moerdijk (2008), the first editor-in-chief of the ANW. A semagram is the representation of knowledge associated with a word in a frame of 'slots' and 'fillers'. 'Slots' are conceptual structure elements which characterise the properties and relations of the semantic class of a word (e.g. COLOUR, SMELL, TASTE, COMPOSITION, INGREDIENTS, PREPARATION for the class of beverages). On the basis of these slots specific data are stored ('fillers') for the word in question.

The ANW adopted its own method for defining the semantic classes and the corresponding frames, as it wanted a classification geared towards lexicographic description and based as far as possible on linguistic foundations rather than on a division of words over various social domains. In addition, it wanted a classification which was relatively transparent such that it could also be used in the dictionary's search function going from content to form. The need to include semagrams in addition to definitions in dictionary entries stems in the first instance from the consideration that definitions alone cannot explain meaning. There is often a lot more semantically relevant knowledge associated with a word than can be shown in a definition. Figure 2 shows the semagram for *wijn* ('wine'), translated into English for the purpose of this paper.[10] At the moment, only the classification information is

[10] For more information on semagrams, see Moerdijk (2008); Tiberius and Schooheim 2015).

encoded in *lemon.* However, the ontolex model can also be used to encode all additional semantic information, taking advantage of the linkage to the SKOS[11] vocabulary, as can be seen in Figure 1. The work to be done here consists of mapping the ANW semagram into the SKOS structure and then to link the whole SKOS construct to the lexical entry by means of the property `isEvokedBy` and to the corresponding sense of *wijn* with the property `isLexicalizedSenseOf`. The advantage of this approach is that all information from "both" sides of the properties are available using the same representation languages.

Wine: beverage; liquid; substance

- [**Smell**] has depending on the developed aroma bouquet, the odour of earth, red fruit, white flowers, forest scents etc.
- [**Colour**] is mainly red, rose, transparent colourless or yellowish
- [**Taste**] is mildly acidic in the case of red or dry white wine but can depending on the grape variety and fermentation also be semi-sweet, semi-dry or sweet
- [**Transparency**] is generally clear
- [**Ingredient**] is a brew based on fermented juice of fruit, especially of grapes, and contains alcohol, acids, unfermented residual sugar and tannin
- [**Function**] serves to enjoy gastronomically, whether or not during a meal, or is to be drunk for pleasure
- [**Preparation**] is prepared by pressing fruit and allowing the juice to ferment
- [**Raw materials**] is made from the juice of grapes or other fruits
- [**Place of Origin**] is produced worldwide in areas with sufficient sunshine for ripening grapes or other fruit
- [**Container**] is in a bottle, carafe, jar or pack, or is being drained from a barrel
- [**Age**] can be young or old, if suitable as a storage wine
- [**Temperature**] is being drunk cold, cool, at room temperature or warm depending on the type [**Property**] usually has a moderate alcohol percentage, often around 12 percent
- [**Mode of use**] is drunk from a goblet or cup
- [**Working**] can make someone happy, rosy or drunk
- [**Occasion**] is being drunk at meals and during meetings with a certain atmosphere such as parties, ceremonies, a celebration, cosy gathering etc.

Figure 2: Semagram for the lemma *wijn* in the ANW

We have chosen to model the semantic information in the definitions, the semagrams and the meaning classes in the ANW into three SKOS concept sets, i.e.:

:ConceptScheme_ANW (for the definitions)

:ConceptScheme_ANW-ANS (for the ANS meaning classes)

:ConceptSchema_ANW-Semagram (for the semagram)

[11] SKOS stands for "Simple Knowledge Organization System". See also https://www.w3.org/2004/02/skos/ for more details.

In addition, some entries also contain domain information. For instance, the sixth and seventh senses of the entry *kat* 'cat' are marked as belonging to the domain of military history. To model the domain information, we propose to use `dct:subject` from the Dublin Core[12] vocabulary.

On the basis of the above information, the semantic information for the ANW entry for *wijn* is modelled as a skos:Concept which has five lexicalised senses: the main sense and four subsenses. This concept is evoked by the lexical entry for wijn, i.e., `lex_wijn_182155`[13], and the lexical entry for *wijnfles*, i.e., `lex_wijnfles_182210`.

"wijn" lexical entry in *lemon*

```
:Concept_325624
  rdf:type skos:Concept ;
  rdf:type ontolex:LexicalConcept ;
  rdfs:comment "Kernbetekennis for lex_wijn_182155" ;
  skos:inScheme :ConceptScheme_ANW ;
  skos:topConceptOf :ConceptScheme_ANW ;
  ontolex:isEvokedBy :lex_wijn_182155 ;
  ontolex:isEvokedBy :lex_wijnfles_182210 ;
  ontolex:lexicalizedSense <http://tutorial-topbraid.com/anw#sense_wijn1.0> ;
  ontolex:lexicalizedSense <http://tutorial-topbraid.com/anw#sense_wijn1.1> ;
  ontolex:lexicalizedSense <http://tutorial-topbraid.com/anw#sense_wijn1.2> ;
  ontolex:lexicalizedSense <http://tutorial-topbraid.com/anw#sense_wijn1.3> ;
  ontolex:lexicalizedSense <http://tutorial-topbraid.com/anw#sense_wijn1.4> ;
.

:lex_wijn_182155
  rdf:type ontolex:Word ;
  lexinfo_anw:articleType "\"de\"" ;
  lexinfo:gender lexinfo:masculine ;
  lexinfo:partOfSpeech lexinfo:commonNoun ;
  lexinfo:partOfSpeech lexinfo:noun ;
  ontolex:canonicalForm :form_wijn_singular ;
  ontolex:otherForm :form_wijnen_plural ;
  ontolex:sense <http://tutorial-topbraid.com/anw-entry#sense_wijn1.0> ;
.
```

"form" information for the lexical entry "wijn" in *lemon*

```
:form_wijn_singular
  rdf:type ontolex:Form ;
  <http://lemon-model.net/lexinfo_anw:hyphenationForm> "\"wijn\"" ;
  <http://lemon-model.net/lexinfo_anw:syllable_nb> 1 ;
  dct:language <http://www.lexvo.org/page/iso639-3/nld> ;
  lexinfo:number lexinfo:singular ;
  ontolex:phoneticRep "* w ɛɪ n"@nl-ReadSpeaker-fonipa ;
  ontolex:writtenRep "wijn"@nl ;
.
:form_wijnen_plural
```

[12] See http://dublincore.org/ for more details

[13] The number refers to the PID of the ANW entry. ANW entries have a PID at the entry level and at the sense level.

```
  rdf:type ontolex:Form ;
  <http://lemon-model.net/lexinfo_anw:hyphenationForm> "\"wij.nen\"" ;
  <http://lemon-model.net/lexinfo_anw:syllable_nb> 2 ;
  dct:language <http://www.lexvo.org/page/iso639-3/nld> ;
  lexinfo:number lexinfo:plural ;
  ontolex:writtenRep "wijnen"@nl ;
.
```

main sense information associated to the lexical entry "wijn" in *lemon*

```
<http://tutorial-topbraid.com/anw#sense_wijn1.0>
  rdf:type ontolex:LexicalSense ;
  skos:definition "alcoholhoudende drank, verkregen door gisting van het sap van
druiven of van andere vruchten, met een middelmatig alcoholgehalte van doorgaans
ongeveer 12 procent; alcoholhoudende drank van gegist druivensap" ;
  ontolex:isLexicalizedSenseOf :Concept_325624 ;
  ontolex:isLexicalizedSenseOf :Concept_Stofnaam ;
  ontolex:isLexicalizedSenseOf :Concept_mass ;
  ontolex:isLexicalizedSenseOf :Semagram_drank ;
  ontolex:isSenseOf :lex_wijn_182155 ;
  ontolex:reference <https://www.wikidata.org/wiki/Q282> ;
  ontolex:usage lexinfo:massNoun ;
  ontolex:usage lexinfo:singular ;
.
```

subssenses originally associated to the entry "wijn", here in the *lemon* encoding

```
<http://tutorial-topbraid.com/anw#sense_wijn1.1>
  rdf:type ontolex:LexicalSense ;
  skos:definition "wijnsoort of wijnmerk" ;
  ontolex:isLexicalizedSenseOf :Concept_Zaaknaam ;
.
<http://tutorial-topbraid.com/anw#sense_wijn1.2>
  rdf:type ontolex:LexicalSense ;
  skos:definition "druiven gekweekt als gewas voor de wijnproductie; wijndruiven als
gewas" ;
  ontolex:isLexicalizedSenseOf :Concept_Zaaknaam ;
.
<http://tutorial-topbraid.com/anw#sense_wijn1.3>
  rdf:type ontolex:LexicalSense ;
  lemon:broader <http://tutorial-topbraid.com/anw#sense_fles1.0> ;
  rdfs:td_is_container_of <http://tutorial-topbraid.com/anw#sense_wijn1.0> ;
  skos:definition "fles wijn" ;
  ontolex:isLexicalizedSenseOf :Concept_325624 ;
  ontolex:isLexicalizedSenseOf :Concept_Zaaknaam ;
  ontolex:reference <https://www.wikidata.org/wiki/Q23490> ;
.
<http://tutorial-topbraid.com/anw#sense_wijn1.4>
  rdf:type ontolex:LexicalSense ;
  lexinfo:partMeronym <http://tutorial-topbraid.com/anw-entry#sense_wijn1.0> ;
  skos:definition "portie of hoeveelheid wijn; glas wijn" ;
  ontolex:isLexicalizedSenseOf :Concept_Zaaknaam ;
  ontolex:reference
<https://commons.wikimedia.org/wiki/File:Glass_wine_white_background.jpg> ;
.
```

2.3 Encoding of compounds

To represent ANW compounds in *lemon*, we make use of the decomposition module, which is depicted in Figure 3 below. An important point being that at this stage we consider compounds as an instance of the MultiWordExpression class of ontolex (see the graphical representation of the ontolex module further above).

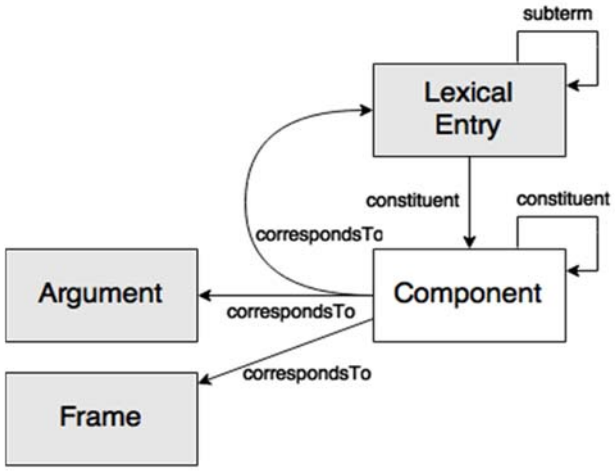


Figure 3: The decomposition module of lemon[14]

In the following *lemon* code below, we can see how the word *wijnfles* ('wine bottle') is decomposed in both its surface form elements (via the property `constituent`) and its compounding lexical entries (via the property `subterm`). The ordering of the elements of the compound is marked with the rdf construct `rdf_1`, etc. The whole compound entry is listed as having the sense `sense_wijn1.3,` which itself is one of the senses for the entry *wijn*. This example shows the potential of *lemon* for sharing and re-using elements of the lexicon across the whole dictionary, and also for linking to other data sources, as every element is encoded internally as a unique resource identifier (URI), including its location on the web.

```
:lex_wijnfles_182210
  rdf:type ontolex:MultiWordExpression ;
  lexinfo_anw:articleType "\"de\"" ;
  lexinfo:gender lexinfo:feminine ;
  lexinfo:gender lexinfo:masculine ;
  lexinfo:partOfSpeech lexinfo:commonNoun ;
  lexinfo:partOfSpeech lexinfo:noun ;
  rdf:_1 :comp_wijn_1 ;
  rdf:_2 :comp_fles_1 ;
  <http://www.w3.org/ns/lemon/decomp#constituent> :comp_fles_1 ;
  <http://www.w3.org/ns/lemon/decomp#constituent> :comp_wijn_1 ;
  <http://www.w3.org/ns/lemon/decomp#subterm>
<http://dictionary_lemon/anw#lex_wijn_182155> ;
  <http://www.w3.org/ns/lemon/decomp#subterm> :lex_fles_18089 ;
  ontolex:sense <http://tutorial-topbraid.com/anw#sense_wijn1.3> ;
.
```

[14] Figure created by John P. McCrae for the W3C Ontolex Community Group.

3. Concluding remarks

In this paper, we have presented a *lemon* model for the morphosyntactic and semantic information in the ANW, a comprehensive scholarly dictionary of Dutch. Encoding the information in *lemon* has a number of advantages:

- **Modularization of the data**

As we could observe especially in the case of the representation of compounds, the *lemon* model implements a strong modular approach to the encoding of lexicon data, and therefore strongly supports the re-use of such elements. This is also true when we look at the internal XML encoding of the ANW, in which for every sense of an entry the whole morphosyntactic information—with some local variations—has to be repeated. This can be avoided in the *lemon* model, as all the different elements of an entry are modularly encoded and interlinked by specific interpretation. There is no redundancy in the graph-based *lemon* model.

- **Linking**

As the lemon model is making use of W3C standards for encoding its elements, linking is the major way to express relations between such elements within one dictionary, but also for external data sources that are encoded as an URI (with a valid location). In the case of the *wijn* entry, we are for example linking the sense 1.0 to a wikidata[15] entry and to a DBpedia[16] entry:

```
<http://tutorial-topbraid.com/anw#sense_wijn1.0>
  rdf:type ontolex:LexicalSense ;
  …
  ontolex:isSenseOf :lex_wijn_182155 ;
  ontolex:reference <https://www.wikidata.org/wiki/Q282> ;
  ontolex:reference http://nl.dbpedia.org/page/Wijn ;
  …
.
```

Accessing then the wikidata or the DBpedia location, one can gain additional information, for example a relevant number of translations of the word *wijn* in this particular sense, as the screenshot of the (partial) page of DBpedia shows in Figure 4.

[15] See https://www.wikidata.org/wiki/Wikidata:Main_Page.

[16] See http://wiki.dbpedia.org/.

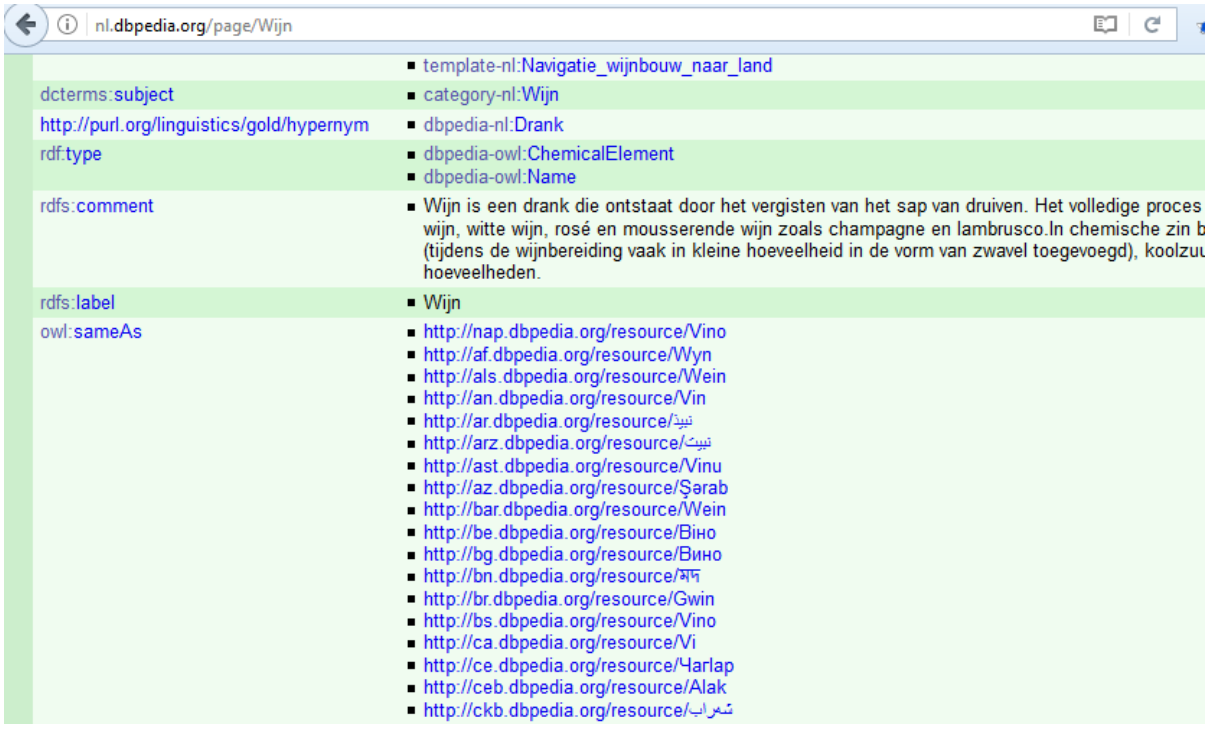
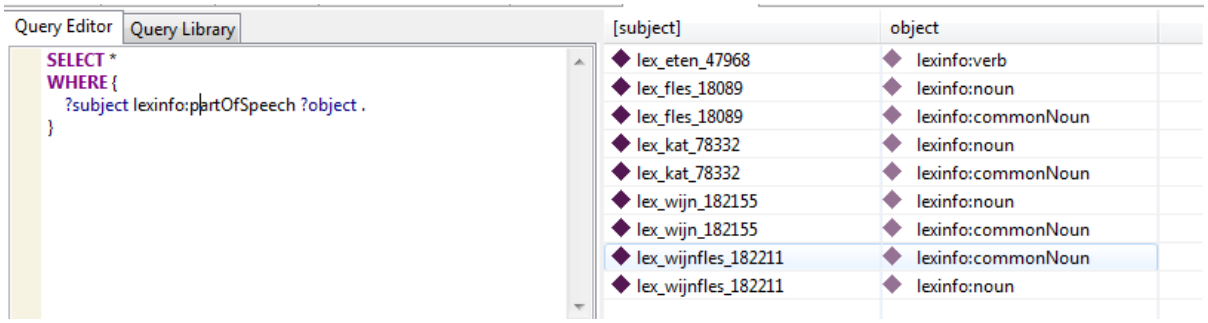


Figure 4: The DBpedia page on 'wijn'

- **Query and access to the data**

The dictionary data encoded in *lemon* are stored in so-called triple stores and thus can be queried and are accessible by the use of the standardised SPARQL query language[17]. It is worth mentioning here, that SPARQL can also be used for augmenting the original data set. The main point is the fact that the ANW data can, in this way, be made available for processing engines, since it is now in a fully machine-readable format. Below we show an example of a simple query we performed with the TopBraid composer[18]. On the left is the query and on the right the results. In this example, the query asks for all entries that have a part-of-speech, while also querying for information about the part-of-speech.



[17] https://www.w3.org/TR/rdf-sparql-query/.

[18] http://www.topquadrant.com/tools/ide-topbraid-composer-maestro-edition/.

In general, we can state that the ontolex and the decomposition modules of *lemon* could be used as they are, while the modifications needed for being compliant with the richness of the ANW data can be addressed in the context of the LexInfo vocabulary, and our ongoing work is to make sure that the inclusion of those ANW features are either made part of LexInfo, or are made available within a similar ontology.

4. Acknowledgements

The DFKI contribution to this paper was partly supported by the H2020 project QT21 with agreement number 645452.

5. References

Bosque-Gil, J., Gracia, J., Montiel-Ponsoda, E. & Aguado-de-Cea, G. (2016). Modelling multilingual lexicographic resources for the Web of Data: The K Dictionaries case. In *GLOBALEX 2016 Lexicographic Resources for Human Language Technology Workshop*, pp. 65-73.

Declerck, Th., Wandl-Vogt, E., Krek, S. & Tiberius, C. (2015). Towards Multilingual eLexicography by Means of Linked (Open) Data. *MSW@ESWC 2015*, pp. 51-58.

Declerck, Th. & Mörth, K. (2016). 'Towards a Sense-based Access to Related Online Lexical Resources'. In: T. Margalitadze & G. Meladze (eds.) *Proceedings of the XVII EURALEX International Congress, Tbilissi, Georgia*, pp. 660-667.

Khan, F., Bellandi, A., Boschetti, F. & Monachini, M. (2017). The Challenges of Converting Legacy Lexical Resources to Linked Open Data using Ontolex-Lemon: The Case of the Intermediate Liddell-Scott Lexicon. In: *Proceedings of the 1st Workshop on the OntoLex Model (OntoLex-2017).*

Haeseryn, W., Romijn, K., Geerts, G., de Rooij, J., ven den Toorn, M. C. (1997). *Algemene Nederlandse spraakkunst.* Groningen Nijhoff.

El Maarouf, I., Jane Bradbury, J., & Hanks, P. (2014). PDEV-lemon: a Linked Data implementation of the Pattern Dictionary of English Verbs based on the Lemon model. In: *Proceedings of the 3rd Workshop on Linked Data in Linguistics,* pp-87-93.

McCrae, J. P., Cimiano, P., Buitelaar, P. & Bordea, G. (2016a). 'Representing Multiword Expressions on the Web with the OntoLex-Lemon model'. In *PARSEME/ENeL workshop on MWE e-lexicons.*

McCrae, J. P., Chiarcos, C., Bond, F., Cimiano, P., Declerck, Th., de Melo, G., Gracia, K., Hellmann, S., Klimek, B., Moran, S., Osenova, P., Pareja-Lora, A. & Pool, J. (2016b), 'The Open Linguistics Working Group: Developing the Linguistic Linked Open Data Cloud'. In *Proceedings of the 10th Language Resource and Evaluation Conference (LREC).*

McCrae, J. P., Aguado-de-Cea, G., Buitelaar, P., Cimiano, P., Declerck, Th., Gómez-Pérez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D. &

Wunner, T. (2012). 'Interchanging lexical resources on the Semantic Web'. In *Language Resources and Evaluation*, 46(6), pp. 701-709.

Moerdijk, F. (2008). 'Frames and Semagrams. Meaning Description in the General Dutch Dictionary'. In: E. Berndal & J. De Cesaris (eds.) *Proceedings of the XIII EURALEX International Congress, Barcelona*, pp. 561-569.

Schoonheim, T. & Tempelaars, R. (2010). 'Dutch Lexicography in Progress, The Algemeen Nederlands Woordenboek (ANW)'. In: A. Dykstra & T. Schoonheim (eds.) *Proceedings of the XIV Euralex International Congress. Leeuwarden*, pp. 718-725.

Stolk, S. (2017). OntoLex and Onomasiological Ordering: Supporting Topical Thesauri. In: *Proceedings of the 1st Workshop on the OntoLex Model (OntoLex-2017).*

Tiberius, C. & Schoonheim, T. (2015). Semagrams, Another Way to Capture Lexical Meaning in Dictionaries'. In *Journal of Cognitive Science*, 16(4), pp. 379-400.

Precise Annotation of Questionnaires for Dialect Research: The Bavarian Dictionary and its Digitization

Manuel Raaf

Bavarian Academy of Sciences and Humanities, Munich, Germany
E-mail: raaf@badw.de

Abstract

The goal of this paper is to present a new web-based application for the analysis of scanned paper-based questionnaires that serve as the basis of the Bavarian Dictionary, by offering project-specific semi-automatic functions which relate and assign image snippets to lemmas. The main requirement was to create software capable of handling several hundred thousand pages containing examples of dialect expressions in several million single parts of images. Additionally, in order to underline the expandability of the software, the requirements of the Franconian Dictionary are briefly described.
Since standard techniques for elaborating dictionary articles or, at least, the preparatory steps to that end, did not fit the particular needs of the project, after digitalization of the questionnaire the web-application LexHelfer was developed. The focus of the application is to aid the editors in the process of gathering relevant information for the lemma entries.
The short history of the application's development and use to date are fully described, so that readers have a complete overview and understanding both of the special type of data and the workflow for creating the entries of the dialect dictionary.

Keywords: annotation; dialect research; image-text-relation; lexicography; questionnaires

1. Introduction

In this paper, in order to provide a complete overview of the digitization of the Bavarian Dictionary and the development of the application, all of the stages are described, beginning with a brief history of the project. The main part of Section 2 then focuses on the application that has been developed for viewing and analyzing the collected material and assisting in compiling the lemma entries. There follow notes on sustainability and basic technical details.

1.1 The Bavarian Dictionary

In 1816, Johann Andreas Schmeller (Frommann: 1872–1877) began work on the first edition of the Bavarian Dictionary, completed in 1837. It was the first scientific work about the Bavarian Language and received extraordinarily high praise by Jacob Grimm (1854: XVII) and others. The second edition, edited by Georg Karl Frommann and published between 1872 and 1877, remains a standard reference work on Bavarian.

The collection of data for the new Bavarian[1] Dictionary has been ongoing since 1913 in order to provide a dictionary that extends beyond Schmeller. The new project also aimed, in cooperation with the Austrian Academy of Sciences, to include the Bavarian varieties spoken in Austria. At that time, today's technical possibilities were unthinkable; there was no thought of a digital strategy, the plan was to create a classical printed dictionary.

In 1961, the Austrian and Bavarian projects, both still ongoing, parted ways. From the mid-80's, the analysis of the material and the writing of lemma entries began, parallel to a survey using questionnaires to collect further information. Up until April 2016, the questionnaires remained almost entirely paper-based and filled out by hand. They were also analyzed by hand, i.e. by choosing the particular individual questionnaire archived in one of dozens of boxes (Figure 1). This meant that the linguists had to review several hundred questionnaires manually in order to determine the variants of the lemma in question. The lemma entries (i.e. the dictionary articles) were subsequently written in MS Word, summarizing all the information gathered on notepads, and thus also paper-based. There now are more than 450,000 paper pages from 240 questionnaires, containing in all over 6,000,000 individual documentations of dialectal expressions. These were digitized to 1.2 Terabytes of images.

Figure 1: paper-based questionnaires (© 2017 Bavarian Academy of Sciences and Humanities)

1 Bavarian is a German dialect that consists of three main variants. It is spoken mainly in the State of Bavaria, Germany, and in Austria, by about 13 million speakers. The project "Bavarian Dictionary" at the Bavarian Academy of Sciences and Humanities focuses on the varieties spoken in Germany. Despite its current relatively large number of speakers, Bavarian is, according to UNESCO, a "vulnerable" language (Moseley: www.unesco.org/culture/languages-atlas/en/atlasmap/language-id-1019.html).

To directly access this large amount of image data for analysis and then use it to write the dictionary, the web-based application *LexHelfer* has been developed, which, among other features, is able to semi-automatically generate relations between questions in the questionnaires and the particular dialect expressions given as answers, enabling quick manual assignment of lemmas. Due to the specific procedure of drafting entries for the Bavarian Dictionary by referring to and analyzing handwritten questionnaires, its digitalization is very different from other methods of drafting digital and/or digitally-based dictionaries—both from the outset and in the concrete steps taken in working with the material.

Editors will soon be able to not only acribicly compile all occurrences of dialect words but also to automatically reduce this information to only the most important cases needed to create the print-version of the entry. In comparison to the manual practices previously used, both actions take only about one eighth of the time.

1.2 Digitization

After preparation of the questionnaires by student assistants, e.g. removal of attachments like nails, photographs or flowers contributed by the informant as clarifications for the dialect word, a service contractor[2] scanned the 450,000 A4 format pages to JPEG files at 300dpi. Because of the semantic structure of the file names (i.e. wordlist, place, region, informant, page number), which was set by hand by a service contractor in Vietnam with a very low error rate, each scan could be linked to the informant. More importantly, the information about place and administrative region could be directly extracted, entered into the database and used by the program to restrict the search to specific regions and locations. The creation of maps showing the distribution of a lemma with the aid of *SprachGIS* (Schmidt et al., 2008) is also based on this structure. A script parsed through the scanned images, inserting their information into the database.

Because the majority of the questionnaires were filled in by hand by hundreds of informants (i.e. too many different handwritings), OCR-techniques would not have been feasible for machine-aided text extraction of the scanned images.

As of April 2017, the informants have the option of filling the questionnaires in online, so that the material is digitally native and can be directly imported into the application. Though most informants are quite aged, this new method of gathering the desired information has so far been very well accepted; by almost 20%.

[2] The material was scanned by MFM Hofmaier (http://www.mfm.de); the file names were set by double-keying in Vietnam on behalf of MFM Hofmaier.

1.3 The Franconian Dictionary

The Franconian Dictionary was also founded, and is hosted by, the Bavarian Academy of Sciences and Humanities. It too is based upon questionnaires that contain examples of dialect expressions. However, the workflow is different: every example is gathered in an Excel file that contains only examples of one specific base lemma. Because of the precision in gathering all the information, each row refers to an image file showing the scanned original questionnaire. By importing the Excel contents into the database and adding table handling to the application, the researchers of the Franconian Dictionary are also able use *LexHelfer* and save time in preparing the material for dissemination to the scientific community as well as to the public.[3] Thus, the Franconian Dictionary and its specific needs will also be part of the following description.

2. The Application: LexHelfer

LexHelfer is designed as a writing assistant for dictionary entries and as a search tool for researchers working on the dialect. Furthermore, the public is also able to search, for example, for expressions documented in their home region. The public view of the application is an automatically created restricted version of the researchers' full version, so that there is no need for the scientist to adjust contents or deactivate program features.

The application consists of different components for preparing digitized questionnaires for analysis and performing the analysis in order to gather information for creating dictionary entries and/or for acribically collecting and online-publishing language examples. From the start, it was developed in close cooperation with linguists, and with consultation about their requirements. Thus, the application fully respects the editors' scientific needs and demands on usability. Furthermore, it was assumed from the beginning that public users should also be able to search the database intuitively and therefore without the need to read (long) instructions.

2.1 From Image Files to Data Relations

As noted in Section 1.2, the names of the image files have a simple semantic structure: wordlist number, place, administrative region, informant ID, and page number. By creating a database table that fits this structure and reading all files into it, the application has all the information needed to point directly to the desired questionnaire and also to restrict searches to specific areas. The particular select-boxes of the HTML-formula for searching were created by a simple and grouped SQL-request

[3] The Franconian Dictionary will not result in a classical dictionary, but rather in a collection of examples linked to standard lemmas enriched with grammatical information and comments.

on the rows containing the information about place and region. Thus, the process can be automated, avoiding the programmer having to type hundreds of place names. Furthermore, it permits the setting of a one-to-one relationship between a lemma and the specific source of examples after performing the actions described in the following sub-section.

2.2 Semi-Automatic Creation of Snippets

Each of the 240 questionnaires contains 60 questions on four pages. This does not result in just 240*60 coordinates of questions/answers, because each questionnaire was distributed to some hundred places all over the State of Bavaria and completed in each place by between one and a dozen informants. Hence, the amount of potential relevant material totals more than 6,000,000 coordinates. Handling so many examples by hand is not feasible in an acceptable period. However, since on each of the 240 questionnaires there is a fixed position for every question and its answer, a simple tool could be created that would allow student assistants to capture these positions and store their coordinates together with the number of the particular question in the database. It was not deemed necessary to choose a specific best-suited copy of a questionnaire to represent all other copies, so the decision of simply taking the first one in the directory of the file system came very quickly. In the final analysis of the results, this decision was spot-on. A script subsequently used the gathered coordinates and created database entries for every single occurrence. Later, the coordinates could be adjusted in the application by researchers in cases where the informant entered information above, below or right next to the designated answer area.[4]

Coding and using the tool for the semi-automatic snippet-creation took about four days, which is a modicum of time compared to manually capturing over six million coordinates. Another day was needed to capture the personal information of the informants, i.e. the address dates, with due respect for data privacy. These areas were then overwritten for the public view.

2.3 Status of a Snippet: None, Irrelevant, Finished

2.3.1 None

The inherent status of each snippet is “none”, i.e. the snippet—or better: its content—is not yet reviewed. Snippets having the status “none” will always occur in the search results.

[4] Please see the appendix for examples of a questionnaire and its semi-automatically created snippets.

2.3.2 Irrelevant

Fortunately, there are many valuable examples. Nonetheless, there are also many impractical ones, i.e. empty answers, which slow the researchers down when they are analysing the given examples because they interrupt the browsing and analysis of the material. To avoid this, each snippet can be set to "irrelevant" and will not then appear in the list of search results unless irrelevant snippets are explicitly demanded on loading. Filtering out impractical snippets can be undertaken quite quickly by student assistants, so that costs are low.

2.3.3 Finished

After finishing the work on a snippet by relating at least one lemma to it, the status "finished" excludes the snippet from (re-)occurring in the search results unless explicitly so defined. Hence, reviewing the contents of the snippets will result in fewer search results after each round, each of which is usually concentrated to a specific region.

2.4 Displaying Evidences and Relating them to Lemmas

The workflow of the Bavarian Dictionary generally starts by choosing a specific question from a wordlist questionnaire, with or without filtering of the search results by place name or administrative region. The snippets then appear ten to the page, surrounded by several action handlers for changing their status, assigning a lemma, or loading the full scan of the corresponding page. In addition, the clip of the snippet can be scrolled up or down by mouse in order to look at information that is potentially written beyond the designated position of the question.

Viewing the full page and choosing a lemma opens a new internal window in which the content in question appears. In this view, one is able to modify the snippet as well as the image file itself, e.g., in the few cases where the scan was made upside-down, by rotating the image. For quick assignment to an existing lemma within the same questionnaire, a drop down menu (<select>) is presented in order to relate the snippet to a lemma in only two clicks.

2.5 Table Handling for the Franconian Dictionary

Table handling is currently used only by the Franconian Dictionary. It allows researchers to modify the content of columns in order to perform error correction faster than in Excel, since the server is more powerful than a typical workstation computer. Sorting 5000 rows or changing their content in Excel might even lead to crashing the system, whereas the database server executes both actions with ease.

Moreover, the table contents can be worked on collaboratively and there is no need for colleagues to first finish and save their changes to the previously used single Excel file.

However, there still are limits relative to the workstation used in accessing the application: the lower the performance, the higher the risk of receiving a blank HTML page as a result of a request for several thousand rows. Fortunately, this only affects reception of the output and not the execution of a server-side command, e.g. changing the content of a field based on that of another field for thousands of entries.

2.6 Export of Search Results

For the Bavarian Dictionary, the search results (list of examples for a particular chosen lemma or list of search results of the concrete search requested by user input) can be downloaded as a PDF file. This contains the snippets and the information about the related lemma (if present) as well as the information extracted from the semantic structure of the particular file name. This function was requested in order to provide a printable and storable list of lemmas and their examples. The same is true of a complete list of examples for a particular chosen lemma.

The Franconian Dictionary yet does not need this functionality; instead, it offers a download of the table of search results as a CSV file.

2.7 Subject Categories (Basic Type of Ontology)

Many dictionary users are interested not only in the meaning and distribution of a (dialect) expression but also in superordinate categories. For instance, before official holidays and celebrations of local traditions, people came to the researchers with questions like “are there any, maybe little-known, dialect words for Christmas in the region of …?”. From time to time, journalists of local newspapers have similar interests. As answering such requests must not take too much of the lexicographer’s time, the solution could be hierarchical categorization of the examples in semantic groups for speedy reference so that a quick search and potential additional examples would satisfy all needs.

A positive side effect of a reliable link between a lemma and a semantic group would be the possibility of adding images automatically to the dictionary. Bearing in mind the widely held opinion that younger generations have less and less knowledge about their dialect, this could be exploited by using new media in preschools to spark off children’s interest and qualification. Moreover, such a link would be a good basis for enabling complete accessibility to the content of the dictionary by enriching the database with additionally linked sound files, which would help disabled people to have full access to the evidences by audio- and image-based indication.

Because there is little scientific work on German ontology and none regarding dialect expressions, this task was resolved using the "subject groups" from the "Badisches Wörterbuch" (Dictionary of Baden Dialects) as well as in the "Pfälzisches Wörterbuch" (Rhenish Palatinate Dictionary), which base their groups upon those of Hallig and von Wartburg (1963). This material is not yet printed or published online. We thankfully received a legal copy of the groups and corresponding lemmas from the Rhenish Palatinate Dictionary by courtesy of Rudolf Post and matched the material with the basic form and the meaning of the dialect examples of the Franconian Dictionary. Where no direct hit could be achieved, the Levenshtein algorithm was used to calculate the shortest distance between two strings, where one is the Franconian evidence (in two rounds: first the basic form, second the meaning) and the other the entry of the aforementioned list of subject groups from the Rhenish Palatinate Dictionary. With the aid of some basic normalizations on the strings (i.e. deletion of articles, pronouns, conjunctions, punctuation, spaces, converting all to lower-case) and accounting for the possibility of attributive metathesis (e.g. *young beautiful woman* vs. *beautiful young woman* as meaning of the example), the error-rate could be reduced a little. However, there are still a very great number of erroneous results, because the dataset available for this task is not yet sufficiently large.

Controlling and correcting the results of the automatic assignment by one student assistant is ongoing and will probably be completed by the end of 2017. The method in its two variants[5] and the particular results are also part of a Bachelor thesis at the Chair of Computer Science and Modern German Literary History at the University of Würzburg, Germany, in cooperation with the Chair of German Linguistics at the University of Erlangen-Nuremberg, Germany. The issue is which method is better, or if both are equally good or bad depending upon the present dataset.

With time, the task will be completed with fewer incorrect results, because the amount of data will increase with each involved dictionary and thus the script will have more clear hits. The long term aim is to create a widespread list of German subject groups that is based on dialects spoken in the State of Bavaria, not on Standard High German. The list could also be used for other dictionaries.

2.8 Lists

LexHelfer offers the possibility of listing all entries by different options in order to produce an overview of what has already been gathered. The options are: listing in

[5] Basically, using Levenshtein as described above. However, there are some possibilities for preferring the strings for comparison: the dialect expression always has the field "meaning" and "basic form" (i.e. basic lemma). For the first option, the meaning serves as first comparative value. The second option can then be either every meaning within the loop through all the meanings or the loop of only the search results for the meaning or (logical *or* in SQL) the basic form in the list of the Rhenish Palatinate Dictionary and its subject groups.

alphabetical order (lemma or basic form; soon: subject group), wordlist and question number, grammatical annotation (for the Franconian Dictionary only). For example, the list of grammatical annotations helps researchers to verify whether the annotation is already present, and how it is written. Every entry in a list links to the page containing the particular search results (e.g. grammatical annotation for "Art AkkSgN partitive"[6] lists all entries for the article (*Art*) in case accusative (*Akk*) singular (*Sg*) neuter (*N*)). Rare constructions and/or rare uses of grammatical structures can thereby be found quite easily. Linguists can thus describe uncommon cases by browsing the list. This is also a good method to recognize typos, e.g. "AdjSgM" instead of the correct notation "Adj SgM" (i.e. adjective singular masculine).

2.9 Upcoming: Compiling Fascicles

To date, researchers write the lemma entries in MS Word. Changing the current workflow to an XML-base is time-consuming because the structure needs to be created in a precise and user-friendly fashion, so that linguists only need a short time for the migration. The plan is to integrate this process into *LexHelfer*, thereby allowing the researchers to prepare the material and directly import it into the function for compiling lemma entries. After reducing the information, the entry can be finished and directly published online or printed (alone or with other entries) to PDF for further processes by the publisher. As of May 2017, the XML-structure is being created in close consultation with the researchers.

3. Sustainability

The questionnaires are present in high-resolution JPEG files, one scan per page, and saved on a highly professional long-time archive system at the governmental Leibniz-Rechenzentrum[7], Garching, Germany, which is part of the Bavarian Academy of Sciences and Humanities. The decision about what format to use was easily taken in favour of compressed JPEG instead of (lossless) TIFF or PNG because of the comparatively low quality of the sources: the questionnaires consist of yellowish paper containing black typewriter font and answers written in either blue, grey (lead pencil), or black. Hence, important visible image information loss by choosing a lossy format

[6] "Art AkkSgN partitive" stands for: article (Art) in the case accusative (Akk) partitive (partitive), number singular (Sg), gender neuter (N). The case *partitive* is not present in German nor in Germanic languages as morphological case. However, it is in the semantics, since atelic actions might force the object of the action to be in a partitive state (depending on the context). For example, the phrase "I'm drinking tea" in the sense of "I'm drinking tea sip by sip" is *partitive* from semantic point of view, whereas "I'm drinking tea" in the sense of "I will completely gulp this cup of tea right now" is not.

[7] www.lrz.de

was not to be expected. The possibility of losing not only visible information but the whole file due to by bit-errors[8] was considered, but neglected for two reasons: 1) The total amount of disk capacity needed for storing around 450,000 files in high-quality JPEG is 1.2 Terabytes, whereas in an uncompressed format it would been at least ten times more. The service contractor also recommended JPEG, with regard to the large amount of data. It is true that high quality scans stored to an uncompressed format would be best for long-term archiving and for reducing the risk of information loss caused by bit errors. However, the compromise with JPEG was regarded as the best choice for the project. 2) The original questionnaires could be re-scanned without much effort, if necessary.

The database containing all the data of the application and thereby the digital base of the dictionary is stored in a MySQL-database. Since MySQL is an open and widely used database management system that can be converted to another (standard) format (e.g. XML), and since JPEG can be read on any device and be converted to any other image format, the data will be available and convertible for a long time. Although technological changes usually happen quickly, the technological base behind the changes does not. Therefore, the current formats will guarantee technical sustainability better than proprietary or less common ones.

4. Technique

LexHelfer runs under Apache 2.4 on a virtual machine running a modern Linux distribution. On the server, PHP 5.6 is the primary language in the scripts handling requests and creating the HTML-output for the client. The latter then uses pure JavaScript (i.e. no libraries) for modifying the output and preparing the requests. The scripts for importing the information about the scanned image files and for the creation of the snippet coordinates for each single occurrence are written in Perl 5. As database, the latest free open-source version of MySQL 5 is used.

On the server, PHP and Perl were used because of the higher power of Perl in processing text information, especially with the aid of regular expressions, which are built-in in Perl. Nevertheless, all actions performed by Perl-scripts could also be performed by PHP-scripts.

Following the open-access-policy of the Bavarian Academy, the source code is licensed under ALv2.[9] After expanding the application in order to have a single one that can be configured by authorized users via the integrated administration panel, e.g. having the

[8] In compressed formats, even a single error-bit can lead to a total and unrecoverable loss of all the information of the file. In uncompressed formats, on the other hand, a destroyed bit would not lead to other bits also being defective. As yet, there is no such case known in the project.

[9] Apache License Version 2, see http://www.apache.org/licenses/LICENSE-2.0.

focus set to the tables (Franconian) or lemma-snippet-assignment (Bavarian), the code of *LexHelfer* will be available via Git. Hence, the direct usage as well as forking the program will be possible and appreciated, since this would transform the application from its status as an in-house solution to a more generic one. The community of users would create synergies and hopefully benefit greatly from it.

5. Public Effectiveness

Government funded projects must always serve not only the scientific community, but also the public. Since a public version of *LexHelfer* is available on-the-fly without researchers having to unlock the contents, anybody is able to access the material on the website.[10] From February 1st until May 16st 2017, around 13,000 public requests where recorded[11]—for a project that is not yet commonly known and that targets a relatively small number of prospective users, this is a relatively high rate of requests by the public.

In addition, the Franconian Dictionary is used by students at the Universities of Erlangen-Nuremberg and Würzburg and has registered about 45,000 cases of public access in the period from January 1st to May 16st 2017.

6. Summary

The schedule of digitizing the Bavarian and Franconian Dictionaries was maintained throughout the whole process. Despite some minor problems, which are to be expected in any project, all went well, thanks to the constant close contact with researchers. Without this, development of the application to meet the scientists' needs would have taken much longer. Scanning the paper-based examples and developing the application have both had a timesaving effect on the linguists' work. The additional financial expense compared to zero expense options, i.e. not going digital, was completely worth the effort, and the additional expenditure was mostly for the scanning of several hundred thousand questionnaires. The speed of gathering all available information for one specific question/lemma has increased about eight fold compared to the analogue manual work.

The options regarding further development[12] are promising and will gradually be instigated in coming years.

---

[10] Database of the Bavarian Dictionary: http://www.bwb.badw.de/bwb-digital/datenbank. Database of the Franconian Dictionary: http://www.wbf.badw.de/wbf-digital/wörterbuch.

[11] In respect of data privacy, only the action of requesting data is recorded/counted, not the request itself or any user-specific data.

[12] Among others: compiling fascicles in XML, improving the automation on relating examples to subject groups, adding images as well as sound files spoken by native speakers to the subject groups as described in Section 2.7

7. References

Christmann, Ernst & Krämer, Julius (ed.) (1965 – 1997): *Pfälzisches Wörterbuch*. Steiner. Wiesbaden.

Fromman, Georg Karl (ed.) (1872 – 1877): *Bayerisches Wörterbuch von J. Andreas Schmeller*. Oldenbourg. Munich. Accessed at http://daten.digitale-sammlungen.de/~db/bsb00005027/images/ (19.05.2017)

Grimm, Jacob (1854): *Deutsches Wörterbuch. Band 1. A – Biermolke*. Hirzel. Leipzig

Hallig, Rudolf & von Wartburg, Walther (1963): *Begriffssystem als Grundlage für die Lexikographie*. Akademie-Verlag. Berlin

Moseley, Christopher (ed.) (2010): *Atlas of the World's Languages in Danger*, 3rd edition. UNESCO Publishing. Paris. Accessed at http://www.unesco.org/culture/en/endangeredlanguages/atlas (18.01.2017)

Ochs, Ernst & Müller, Karl Friedrich & Post, Rudolf et al. (eds.) (1925–today): *Badisches Wörterbuch*. Oldenbourg. Lahr/Munich

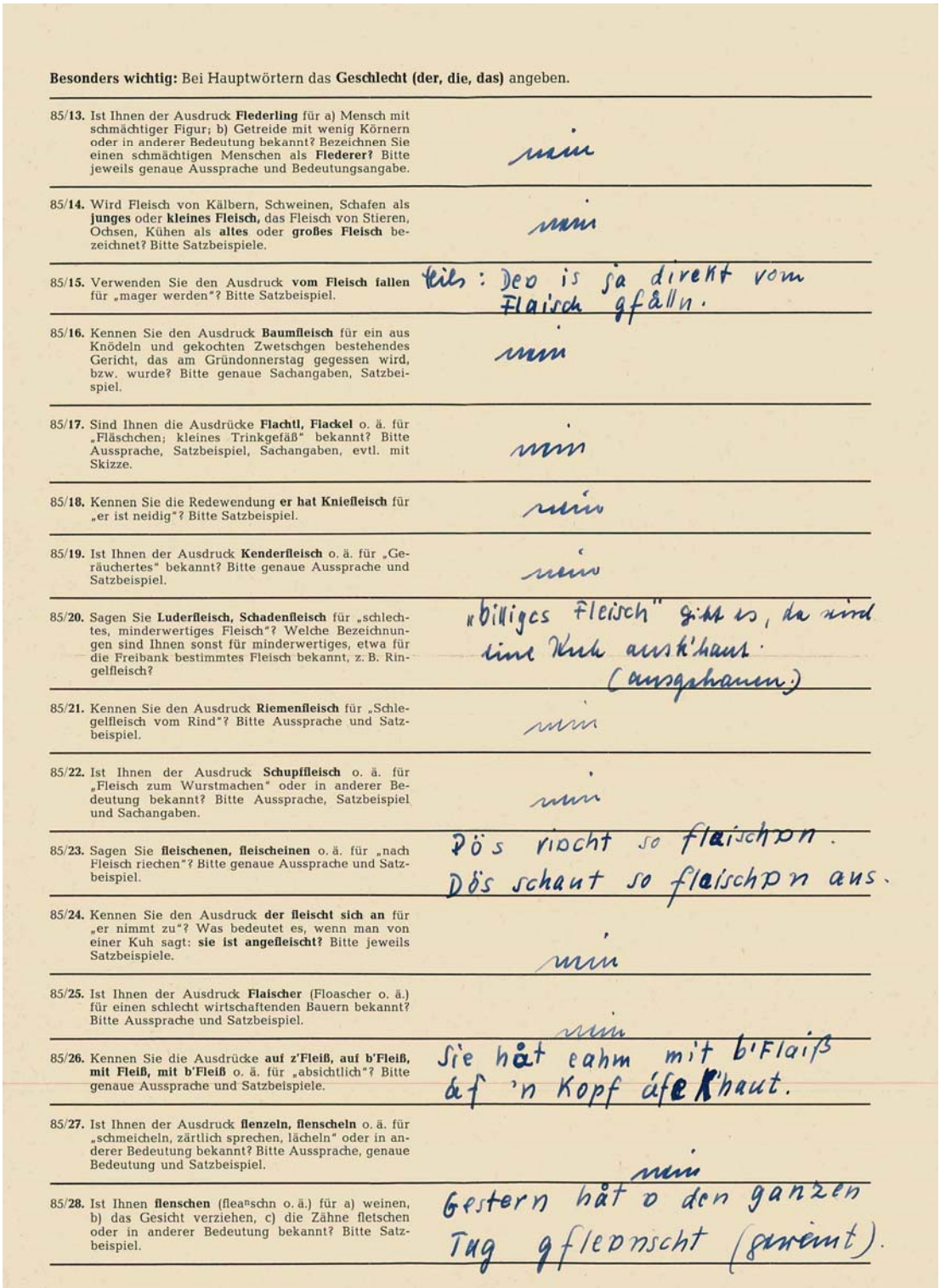
Raaf, Manuel (2016–today): *LexHelfer BWB*. Bavarian Academy of Sciences and Humanities. Munich. Accessed at http://www.bwb.badw.de/en/digital-platform/database (16.05.2017)

Rowley, Anthony et al. (eds.) (1995–today): *Bayerisches Wörterbuch*. Oldenbourg. Munich.

Schmidt, Jürgen Erich & Herrgen, Joachim & Kehrein, Roland (eds.) (2008 ff): *Regionalsprache.de (REDE). Forschungsplattform zu den modernen Regionalsprachen des Deutschen. Teil 6: REDE SprachGIS – Das forschungszentrierte sprachgeographische Informationssystem von Regionalsprache.de*. edited by Dennis Bock et al. Forschungszentrum Deutscher Sprachatlas. Marburg. Accessed at https://www.regionalsprache.de (18.01.2017)

8. Appendix

8.1 Example of a Questionnaire and its Snippets



Besonders wichtig: Bei Hauptwörtern das **Geschlecht (der, die, das)** angeben.

85/13. Ist Ihnen der Ausdruck **Flederling** für a) Mensch mit schmächtiger Figur; b) Getreide mit wenig Körnern oder in anderer Bedeutung bekannt? Bezeichnen Sie einen schmächtigen Menschen als **Flederer?** Bitte jeweils genaue Aussprache und Bedeutungsangabe.

mir

85/14. Wird Fleisch von Kälbern, Schweinen, Schafen als **junges** oder **kleines Fleisch,** das Fleisch von Stieren, Ochsen, Kühen als **altes** oder **großes Fleisch** bezeichnet? Bitte Satzbeispiele.

mir

85/15. Verwenden Sie den Ausdruck **vom Fleisch fallen** für „mager werden"? Bitte Satzbeispiel.

Kils: Deo is ja direkt vom Flaisch gfalln.

85/16. Kennen Sie den Ausdruck **Baumfleisch** für ein aus Knödeln und gekochten Zwetschgen bestehendes Gericht, das am Gründonnerstag gegessen wird, bzw. wurde? Bitte genaue Sachangaben, Satzbeispiel.

mir

85/17. Sind Ihnen die Ausdrücke **Flachtl, Flackel** o. ä. für „Fläschchen; kleines Trinkgefäß" bekannt? Bitte Aussprache, Satzbeispiel, Sachangaben, evtl. mit Skizze.

mir

85/18. Kennen Sie die Redewendung **er hat Kniefleisch** für „er ist neidig"? Bitte Satzbeispiel.

mir

85/19. Ist Ihnen der Ausdruck **Kenderfleisch** o. ä. für „Geräuchertes" bekannt? Bitte genaue Aussprache und Satzbeispiel.

mir

85/20. Sagen Sie **Luderfleisch, Schadenfleisch** für „schlechtes, minderwertiges Fleisch"? Welche Bezeichnungen sind Ihnen sonst für minderwertiges, etwa für die Freibank bestimmtes Fleisch bekannt, z. B. Ringelfleisch?

„billiges Fleisch" gibt es, da wird eine Kuh ausg'haut. (ausgehaut.)

85/21. Kennen Sie den Ausdruck **Riemenfleisch** für „Schlegelfleisch vom Rind"? Bitte Aussprache und Satzbeispiel.

mir

85/22. Ist Ihnen der Ausdruck **Schupfleisch** o. ä. für „Fleisch zum Wurstmachen" oder in anderer Bedeutung bekannt? Bitte Aussprache, Satzbeispiel und Sachangaben.

mir

85/23. Sagen Sie **fleischenen, fleischeinen** o. ä. für „nach Fleisch riechen"? Bitte genaue Aussprache und Satzbeispiel.

Dös riocht so flaischpn.
Dös schaut so flaischpn aus.

85/24. Kennen Sie den Ausdruck **der fleischt sich an** für „er nimmt zu"? Was bedeutet es, wenn man von einer Kuh sagt: **sie ist angefleischt?** Bitte jeweils Satzbeispiele.

mir

85/25. Ist Ihnen der Ausdruck **Flaischer** (Floascher o. ä.) für einen schlecht wirtschaftenden Bauern bekannt? Bitte Aussprache und Satzbeispiel.

mir

85/26. Kennen Sie die Ausdrücke **auf z'Fleiß, auf b'Fleiß, mit Fleiß, mit b'Fleiß** o. ä. für „absichtlich"? Bitte genaue Aussprache und Satzbeispiele.

Sie hat eahm mit b'Flaiß af 'n Kopf afe K'haut.

85/27. Ist Ihnen der Ausdruck **flenzeln, flenscheln** o. ä. für „schmeicheln, zärtlich sprechen, lächeln" oder in anderer Bedeutung bekannt? Bitte Aussprache, genaue Bedeutung und Satzbeispiel.

mir

85/28. Ist Ihnen **flenschen** (fleaⁿschn o. ä.) für a) weinen, b) das Gesicht verziehen, c) die Zähne fletschen oder in anderer Bedeutung bekannt? Bitte Satzbeispiel.

Gestern hat o den ganzen Tag gflennscht (gereint).

Figure 2: second page of one copy of questionnaire 85

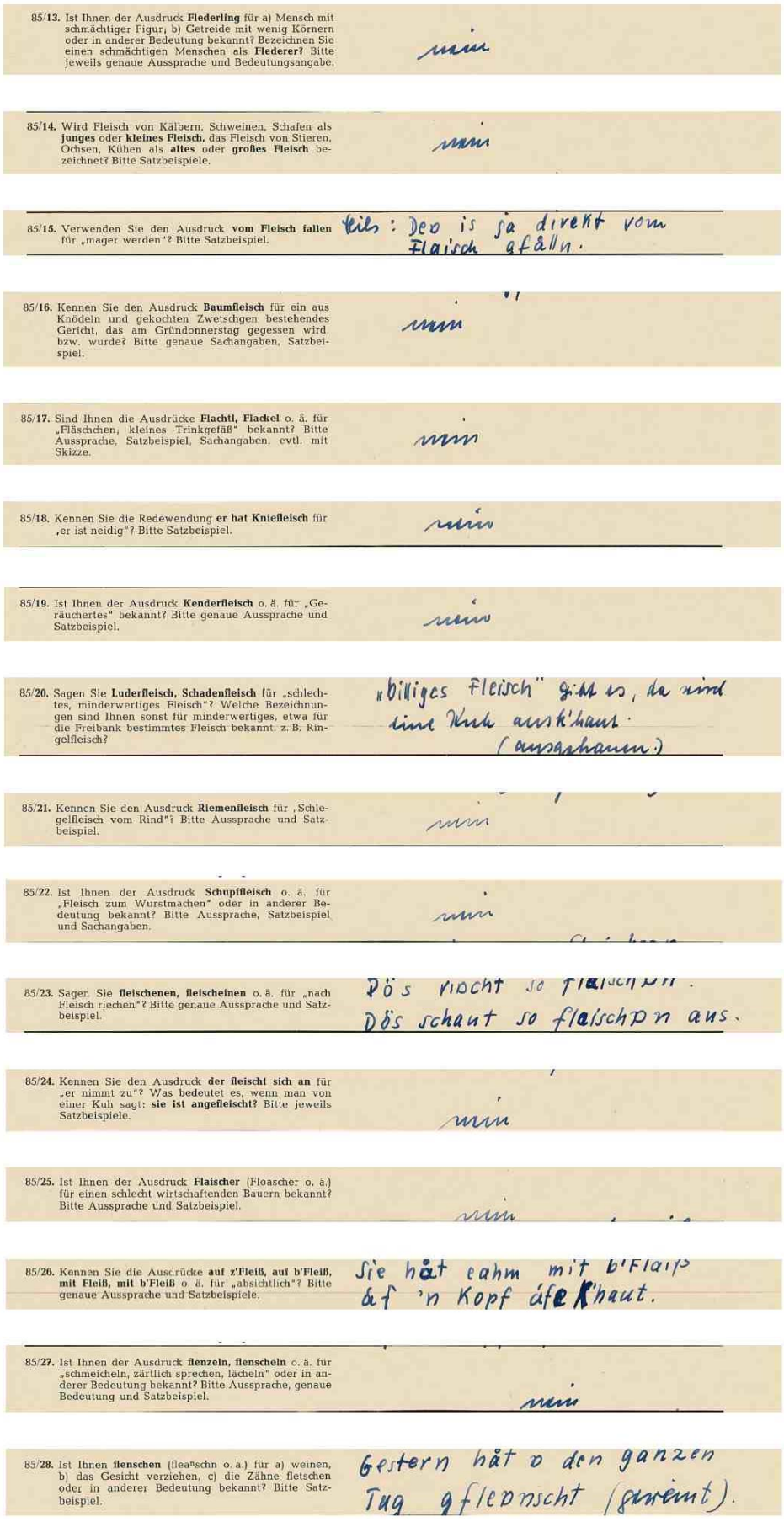


Figure 3: semi-automatically created snippets of one copy of questionnaire 85

8.2 Screenshot of Search Results (Bavarian Dictionary)

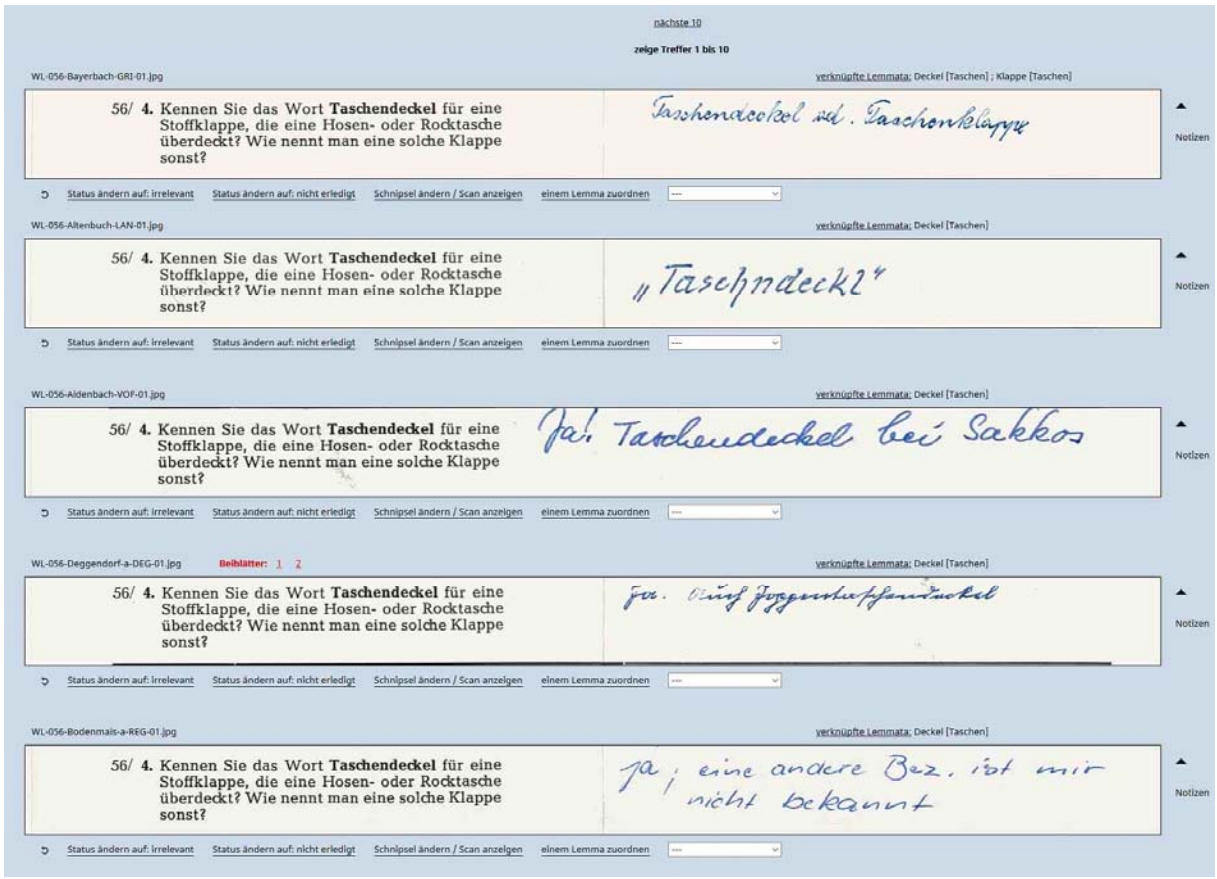


Figure 4: main view of the search results for questionnaire 54. Due to limited space, only half of it is visible in this screenshot

8.3 Screenshot of Search Results (Franconian Dictionary, Table View)

Ihre Suche nach Grundform *Haus** ergab 55 Treffer.

Stamm	Grundform	Bedeutung	Grammatik	Originaltext	Umschrift	Ort	Planquadrat	Kommentar Gewährsperson	Kommentar Bearbeiter	GP	Bogen	Frage	Bild
Arrest	Hausarrest	Freiplatz beim Fangenspiel	S NomSg	Hausarrest	Hausarrest	Bühler	V26,7			1	5	31	
Bock	Hausbock	Insekt, Holzschädling	Sm NomSg	Hausbock	Hausbock	Flachslanden	d30,1	Holzschädling		2	114	7	
Ecke	Hauseck	Außenkante des Hauses (in Redensart)	Sn NomSg	der zünd's Hauseck o	der zündet das Hauseck an	Forth	a34,8		Gesamtbedeutung: urinieren	1	69	35	
Geld	Hausherausgeld	Geldzahlung des Erben an Elnern bei Hofübergabe	Sn NomSg	das Hausraus geld	das Hausherausgeld	Heinrichsthal	V23,1			1	7	47	
halten	Haushalten	führen eines Haushalts	Sn AkkSg	versorcht dös Haushälta	versorgt das Haushalten	Edlendorf	T37,2	jedoch ohne das Erbe im Auge zu haben		1	8	34	
halten	haushalten	sparen, Geld nicht verschwenden	Vst PP	haushalta haouta und haouts za wos bracht	haushalten hat er und hat es zu was gebracht	Vordorf	V38,5		#evtl verschr	1	56	33	
halten	haushälterisch	sparsam, geizig	Adj präd	haushälterisch	haushälterisch	Rothenburg o.d.T.	d28,4			1	69	2	
halten	haushälterlich	sparsam, geizig	Adj präd	haushä(e)lterli	haushälterlich	Hürbel a. R.	d29,9			1	69	2	
hauen	Hauspan	Fällkerbe in Baumstamm schlagen	Sm NomSg	Ha>uspan	Hauspan	Röttenbach	g33,1		zweites a ist unterstrichen	1	1	50	
Haus	Haus	Anrede unter Verliebten	Sn NomSg	aldds Haus	altes Haus	Schwemmelsbach	V27,1			1	116	11	

Figure 5: search results for the base form “Haus” (house), extended by the asterisk * to encompass compounds in the results

Word Sense Frequency Estimation for Russian: Verbs, Adjectives and Different Dictionaries

Anastasiya Lopukhina[1], Konstantin Lopukhin[2]

[1] National Research University, Higher School of Economics;
Russian Language Institute of the Russian Academy of Sciences, Moscow, Russia
[2] Scrapinghub, Moscow, Russia
E-mail: alopukhina@hse.ru, kostia.lopuhin@gmail.com

Abstract

In this paper we investigate several extensions to our prior work on sense frequency estimation for Russian. Our method is based on semantic vectors and is able to achieve good accuracy for sense frequency estimation trained on dictionary entries from the Active Dictionary of Russian and unannotated corpora. We apply our method to verbs and adjectives to obtain sense frequencies for 329 verbs and 256 adjectives in an academic corpus and a web-based corpus. We compare frequency distributions against dictionary sense ordering and between two corpora and find that the first dictionary sense is not the most frequent for almost half of the words we studied. Evaluation of verbs and adjectives shows that frequency estimation error is lower than 15%. We investigate the effect of sense granularity, evaluating how the accuracy of our method changes when applied to more coarse-grained senses. We also investigate if our method can be applied to other dictionaries with less elaborate sense descriptions, by evaluating its accuracy when training on dictionary entries from two other dictionaries.

Keywords: frequency; sense frequency; word sense disambiguation; semantic vectors; sense granularity

1. Introduction

When words have several senses, it is important that dictionaries describe them properly and exhaustively (see e.g. Pustejovsky, 1996; Apresjan, 2000; Iomdin, 2014). One of the properties of word senses is their frequency in a language, as the different senses are not distributed evenly. However, this information is not represented in dictionaries. We cannot rely on the ordering of word senses in a dictionary to obtain this information, as it is not always consistent with real sense distribution in a language. In the Russian lexicographic tradition the ordering of senses follows etymological principles: the first sense of a polysemous word is usually the original, non-figurative meaning (Kruglikova, 2012). For example, the Russian word *veha* can be described as having two distinctly different senses: (1) 'boundary-mark' and (2) 'a milestone in smb's life' (Apresjan, 2014). Although native speakers might agree that the first sense of the word *veha* is rare, we cannot quickly check this assumption; instead, relative frequency is assessed subjectively by intuition.

The lack of word sense frequency information becomes a problem in language learning and teaching. Nesi and Haill (2002) stress the problem of learners being satisfied with the first sense listed in a dictionary, even if the meaning does not fit the context, which often leads to incorrect interpretations. The information about sense frequency is especially necessary if a dictionary is going to be used for text production (Lew, 2013). Discussing the question of word lists for teaching a language, Beck et al. (2013) state that there is no way to obtain the relative frequency of one meaning or sense of a word from the general frequency of this word. It evokes the problem of selecting the appropriate meaning that should be studied first. The same problem can be illustrated for Russian. For example, the first dictionary sense of the Russian word *bremya*—'heavy load'—is perceived as rare in comparison with its second sense—'burden' (according to the *Large Explanatory Dictionary of Russian* (Kuznetsov, 2014)). So, the information about word sense frequency could help students prioritise learning the most relevant sense of a word.

Word sense frequencies can be useful for theoretical studies of the meaning structures of polysemous words. The information about relative sense frequencies can be a basis for comparing the cross-linguistic meaning structure of cognate words in two languages (like *base—basa*, *clay—klej* in English and Russian) and translation equivalents (like *thing—veshch'* in English and Russian). Iomdin and colleagues (2016) described three cases of cognates in Russian and English whose meaning structures are dissimilar: words with senses that have no match in the other language (*vagon—wagon*, *gradus—grade*); words with one or more matching senses for which the most frequent senses drastically differ (*avtoritet—authority*, *artist—artist*); and words in which several senses match but others do not (*blok—block*). The authors discovered that people tend to transfer meaning structures of cognates from their own language to the other language. Thus, information about common mistakes in cognate usage and sense frequencies can be important for language learners as well as for linguists.

The question of word sense frequencies is studied as a practical application to automated word sense disambiguation tasks (Navigli, 2009). The most frequent sense detection is widely studied (Mohammad & Hirst, 2006; McCarthy et al., 2007; Loukachevitch & Chetviorkin, 2015) and is known to be an important baseline, and difficult to overcome for many word sense disambiguation systems (Agirre et al., 2007; Navigli, 2009). Furthermore, psycholinguistic experiments with homonyms and polysemes use information about sense frequency as a factor. Several studies (Klein & Murphy, 2001; Pylkkänen et al., 2006; Foraker & Murphy, 2012) showed that sense frequencies and sense dominance influence processing speed.

In this paper, we present an approach to word sense frequency estimation that is based on corpora and explanatory dictionaries. It allows us to automatically obtain sense frequency distributions from raw corpora and uses dictionary information for training. We extend previously reported works (Lopukhina et al., 2016; Lopukhina et al., in print) in a number of different directions: (1) We apply the method to verbs and

adjectives, while previous studies included only nouns. We get sense frequencies from academic and web-based corpora and compare distributions. (2) We experiment with sense granularity for nouns and evaluate our method on coarse-grained and fine-grained sense inventories. (3) We compare the *Active Dictionary of Russian* (Apresjan, 2014), that was used for sense inventory and training data, to two other dictionaries: the *Large Explanatory Dictionary of Russian* (Kuznetsov, 2014) and the *Russian Language Dictionary* (Evgenyeva, 1981–1984). Thus we aim to study whether our approach can be generalized to any explanatory dictionary. We conduct our research on the Russian language.

2. Word Sense Frequency Estimation

For the purpose of word sense frequency estimation, for each word we perform automated word sense disambiguation on contexts sampled from corpora, and then calculate relative sense frequencies in the sample. We need a word sense inventory, a source of word contexts (a corpus), and a word sense disambiguation technique. We use only existing linguistic resources, without any additional annotation except for evaluation.

2.1 Word Sense Inventories

As a source of word senses we chose an explanatory dictionary—this type of sense inventory is the most natural for our task and, besides, many languages have dictionaries, but not all possess WordNet-like resources.

For our research, we principally used the *Active Dictionary of Russian* (Apresjan, 2014). This dictionary has three major advantages: first, it is the most developed explanatory dictionary of Russian which reflects contemporary language; second, it uses a consistent and systematic approach to polysemy—each word sense is identified by a set of its unique properties and similar words are described similarly; and third, for each word sense it provides many examples and collocations. They are used by our word sense disambiguation technique for training. We have already presented the results of sense frequency estimation for 440 polysemous and homonymous nouns from the *Active Dictionary of Russian* (Lopukhina et al., 2016; Lopukhina et al., in press). Our current research is focused on verbs and adjectives from the first issue of the dictionary.

In order to answer the question of whether more coarse-grained sense distinction can boost performance (Navigli, 2006), we experimented with sense granularity of nouns, verbs and adjectives from the *Active Dictionary of Russian*. All senses that were described as components of one block and have indexes (like 1.1, 1.2, 1.3) were merged and considered as one sense. This clustering of senses inevitably leads to the loss of details: such as the loss of scope for the verb *brodit'*: 1.1 'to travel from place to place

on foot, usually without a particular direction or purpose' and 1.2 'to travel around the world with no particular purpose'; or the loss of specificity for the adjective *belyj*: 7.1 'good' (*white magic*) and 7.2 'legal' (*reported salary*). Nevertheless, coarse-grained senses are distinct, interpretable and different from other senses of a word. We aim to test whether a more coarse-grained sense inventory will provide better results in our task.

Despite its advantages, the *Active Dictionary of Russian* has one important drawback—it is an ongoing project: only 17%of the dictionary vocabulary has been described and edited (approximately 1960 words out of 11,150). Therefore, in this study, we also tested two more explanatory dictionaries: the academic *Russian Language Dictionary* (Evgenyeva, 1981–1984) and the *Large Explanatory Dictionary of Russian* (Kuznetsov, 2014). Both have electronic versions that we used in our research. These dictionaries have the most similar definitions among all the explanatory dictionaries of Russian and have similar distributions of entries by the number of senses (Kiselev et al., 2015). The major disadvantage of these dictionaries that prevented us from using them from the very beginning is the lack of collocations and illustrative sentences, which are crucial for our technique. The average number of examples and collocations for 14 nouns in the *Russian Language Dictionary* is 4.5, in the *Large Explanatory Dictionary of Russian* is 7.5, and in the *Active Dictionary of Russian* is 20. For the purpose of the current study we selected 14 polysemous nouns, extracted all the collocations and illustrative sentences in their entries from the dictionaries and compared the performance of our method for these three sense inventories.

2.2 Corpus

The corpus is a source of contexts for disambiguation. The choice of corpus influences sense frequency, because word sense distributions vary from corpus to corpus. For nouns it was found that 67 out of 440 words have different most frequent senses in the academic and in the web-based corpora (Lopukhina et al., in print). The difference was explained by the difference in content of the corpora. For purposes of the current study, we also used the contexts from the same two corpora: the Russian National Corpus (RNC, http://ruscorpora.ru/en, 230 million tokens in the main corpus), a resource created by a consortium of linguists and software developers; and the ruTenTen11 web-based corpus, the largest Russian internet corpus, consisting of 18 billion tokens integrated into the Sketch Engine system (Kilgarriff et al., 2004). Web corpora are known for having more recent data and for providing relevant and comparable linguistic evidence for lexicographic purposes (Ferraresi et al., 2010). Therefore, we expect to find differences in sense frequency distributions for verbs and adjectives in these two corpora. To estimate word sense frequency we sample 1,000 random contexts for each word in both corpora. Sample sizes yield a statistical error below 3.1%.

2.3 Word Sense Disambiguation Method

In this study we use the word sense disambiguation (WSD) method based on semantic vectors that is described in detail in Lopukhina et al. (in press). This method can achieve good disambiguation accuracy even on a small number of examples available in the dictionary, and is very robust to overfitting. The basis of the method is a vector representation of context or a dictionary example, which is obtained as a weighted sum of semantic vectors for words: this representation aims to capture the sense of a context. Context vectors for all illustrative examples, collocations, synonyms, etc. for a particular sense are averaged to form a single sense vector. Such vectors are built for all dictionary senses. When disambiguating a new context, its vector is calculated in the same way (as a weighted sum of word vectors), and the method assigns this context to the sense with the closest sense vector. In Lopukhin & Lopukhina (2016) we studied several variations of the method, and have decided to use the most simple and robust variant in this paper.

Word vectors were trained using word2vec skip-gram algorithm on a 2 billion lemmatized corpus (combined RuWaC, lib.ru and Russian Wikipedia) with vector dimension 1024, window size 5 and negative sampling. Word weights were estimated on the same corpus. Implementation of the method is available online on https://github.com/lopuhin/sensefreq.

3. Evaluation

Quantitative evaluation is comprised of three parts: evaluating WSD accuracy for different parts of speech, coarse-grained vs. fine-grained senses, and different dictionaries. In the evaluation for different parts of speech we focus on verbs and adjectives, and also include results on nouns for comparison—evaluated in more detail in Lopukhina et al. (in press). In the coarse-grained sense evaluation we compare WSD accuracy when using coarse and fine-grained senses from the *Active Dictionary of Russian* for nouns, verbs and adjectives. For the evaluation of the different dictionaries we compare WSD accuracy obtained when training on entries from the *Active Dictionary of Russian* and when training on entries from two other dictionaries.

3.1 Word Sense Disambiguation for Verbs and Adjectives

We evaluated word sense disambiguation accuracy and sense frequency estimation error of our method for words of three different parts of speech: nouns, verbs and adjectives. We used two different kinds of training data: full contexts from the corpus and entries from the *Active Dictionary of Russian* (AD). For this study at least 100 contexts were labelled for each word, and 50 random contexts were used for training, while the rest were used for evaluation in a fivefold cross-validation scheme. When training on dictionary entries, all labelled contexts from the corpus were used for

training. Frequency error was measured as maximum absolute error in sense frequency estimation averaged across all words.

Results are presented in Table 1. We provide two baselines: the first dictionary sense baseline and the MFS (most frequent sense) baseline. MFS is a powerful baseline that assigns all contexts to the most frequent sense and is often hard to beat (Navigli, 2009). The first dictionary sense baseline assigns all contexts to the first dictionary sense and is more relevant for methods trained on dictionary entries. This baseline is more powerful than a random one, because the first sense is often the most frequent.

Part of speech	**Nouns**	**Verbs**	**Adjectives**
Number of words	17	20	14
Avg. number of senses	3.82	5.00	5.93
First sense baseline	0.50	0.59	0.55
MFS baseline	0.67	0.63	0.62
Accuracy training on contexts	0.80	0.72	0.69
Accuracy training on AD entries	**0.76**	**0.69**	**0.68**
Frequency error (AD entries)	**0.10**	**0.14**	**0.14**

Table 1: WSD accuracy for nouns, verbs and adjectives

We see that training on 50 contexts from the corpus gives more accurate predictions than training on dictionary entries, although the difference for adjectives is very small. Nouns have the highest accuracy while also having the lowest number of senses, and adjectives have the lowest accuracy and the highest number of senses. Verbs have significant negative Pearson correlation between number of senses and accuracy: -0.7, while the correlation between nouns and adjectives is more moderate, at -0.3. The average number of senses given in Table 1 is for words used for evaluation, but it is similar across all polysemous words in the *Active Dictionary of Russian*: 3.33 for nouns, 5.17 for verbs and 3.79 for adjectives—only adjectives display a significant difference.

Figure 1 shows a distribution of WSD accuracy when training on AD entries. We see that verbs have a more diverse distribution, with some scoring as low as 0.2 but also many having scores above 0.9, while adjectives have few words with accuracy higher than 0.8.

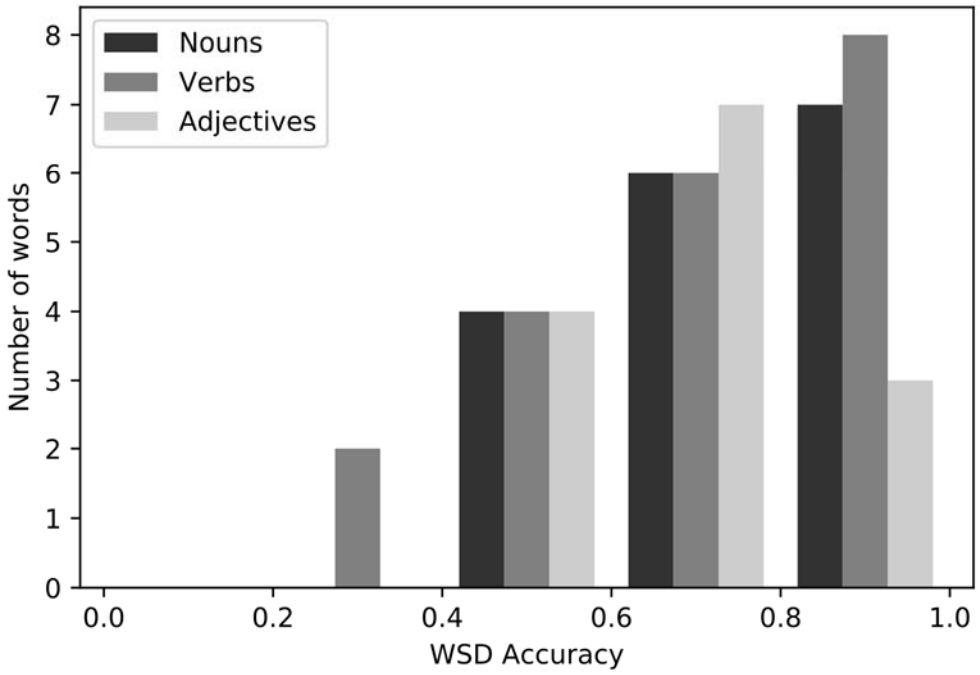


Figure 1: Distribution of WSD accuracy for nouns, verbs and adjectives

Sense frequency estimation error for verbs and adjectives is higher than for nouns, but is still low in an absolute sense, lower than 15% for all parts of speech. This means that our method gives reliable sense frequency estimation for all parts of speech.

3.2 Coarse-grained Sense Inventory

The *Active Dictionary of Russian* provides a two-level hierarchical sense inventory: senses are numbered as *x.y* (e.g. *2.1*), making it possible to evaluate word sense disambiguation on coarse-grained senses, formed by lumping together fine-grained components of one semantic block (e.g. *2.1, 2.1* $\rightarrow$ *2*). As a result, most words have fewer senses, and most senses have more training examples. Results of this evaluation are presented in Table 2. We see that all parts of speech have significantly fewer coarse-grained senses on average, and accuracy for coarse-grained senses increases for nouns and especially verbs, and is almost the same for adjectives. For verbs, the result can be explained by a general tendency to obtain a higher accuracy for fewer senses. We suppose that the lower accuracy gain for adjectives may be explained as follows: adjectives get different senses in contexts with nouns while verbs and nouns have more diverse contexts. More limited contexts for adjectives can be the reason for the results we obtained.

Part of speech		Nouns	Verbs	Adjectives
Number of senses	**Fine**	3.82	5.00	5.93
	Coarse	2.77	3.15	4.07
First sense baseline	**Fine**	0.50	0.59	0.55
	Coarse	0.56	0.66	0.60
Accuracy	**Fine**	**0.76**	**0.69**	**0.68**
	Coarse	**0.80**	**0.79**	**0.79**

Table 2: Coarse and fine sense inventories for the *Active Dictionary of Russian*

3.3 Other Dictionaries

The *Active Dictionary of Russian* is a very attractive resource for computational linguistics methods due to its very comprehensive and systematic descriptions. However, its wordlist is small compared to other dictionaries, and only the first volume has been published at the time of writing. Thus, it is interesting to check how our method works on other dictionaries with larger wordlists, namely the *Russian Language Dictionary* (Evgenyeva, 1981–1984), denoted as MAS, and the *Large Explanatory Dictionary of Russian* (Kuznetsov, 2014), denoted as BTS. Since all these dictionaries have different sense inventories, we had to perform sense mapping: each sense in MAS or BTS was mapped to one or more senses in AD. If some AD sense did not have any corresponding sense in MAS/BTS, contexts with this sense were removed from test data. Words where only one sense was left or where one AD sense corresponded to several MAS/BTS senses were discarded. Evaluation was performed only on nouns: we selected 11 nouns for MAS and 14 nouns for BTS. Results are presented in Tables 3 and 4. In order to compare the quality of training data in MAS/BTS to the *Active Dictionary of Russian*, we also measured word sense disambiguation accuracy with mapped senses but AD training data (denoted as AD* in the table).

Sense inventory	Training data	
	BTS/MAS	**AD***
MAS	**0.66**	0.75
BTS	**0.65**	0.72

Table 3: WSD accuracy for other dictionaries (MAS and BTS) compared to AD

Sense inventory	Training data	
	BTS/MAS	**AD***
MAS	**0.20**	0.13
BTS	**0.21**	0.15

Table 4: Sense frequency estimation error for MAS and BTS compared to AD

We see that both MAS and BTS perform significantly worse than AD, and that BTS performs better than MAS when compared with the *Active Dictionary of Russian*. Sense frequency estimation error for MAS and BTS is also larger but could still be useful for some tasks. In Table 5 we compare average number of examples per sense and average number of words per sense: BTS has a larger number of examples than MAS, which might explain differences in WSD performance (relative to AD) between MAS and BTS.

	MAS	**BTS**	**AD**
Number of examples per sense	4.5	7.3	20
Number of words per sense	62	49	216

Table 5: Average number of examples and words per sense in training data

4. Results and Discussion

We obtained sense frequencies for Russian verbs, adjectives (in this study) and nouns (Lopukhina et al., in press) in the academic Russian National Corpus and web-based ruTenTen11. All data are available online: http://sensefreq.ruslang.ru/. Word sense frequency distributions differ depending on the part of speech and on the corpora used. In Lopukhina et al. (in press) we reported on sense frequencies for 440 nouns. In this study, we applied our method to all homonymous and polysemous verbs and adjectives from the first issue of the *Active Dictionary of Russian* and obtained word sense frequencies for 329 Russian verbs and 256 adjectives.

First, we compared the first sense in the *Active Dictionary of Russian* with the most frequent sense in the RNC and ruTenTen11. The ratio of verbs where the first dictionary sense is the most frequent (excluding homonyms) is 50% in the RNC and 48% in ruTenTen11. For adjectives, the first dictionary sense coincides with the most frequent sense in 61% of cases in the RNC and 59% in ruTenTen11. This means that, for verbs and adjectives, the meaning described first in a dictionary differs from the most common sense of the word in contemporary language in about half of cases.

The discrepancy between the first sense of verbs in the Active Dictionary of Russian and the most frequent sense in the RNC can be observed in the following examples. The first dictionary sense of the verb *gladit'* is 'to iron', while in 83% of cases in the RNC it is used in the other sense—'to gently move your hand over skin, hair, or fur'. The first literal sense of the verb *bolet'* is 'to be ill'. In the RNC, this sense is the third most frequent (20%); the most frequent is 'to feel pain somewhere in your body' (46%) and the second most frequent, 'to be a fan, to encourage somebody's favourite sportsman or team' (31%). For several verbs, the most frequent meaning is a metaphorical one; it is normally described after a literal one in the dictionary, e.g. *vykroit'* ('to succeed in getting enough of something, especially time and money, by making a lot of effort', 87%), *vyputat's'a* ('to get yourself out of a situation that you no longer want to be involved in', 88%), *galdet'* ('to make noise (about people)', 92%), *vkluchit's'a* ('to start to take part in a particular activity that has started before', 71%), *votsarit's'a* ('something starts to happen and have an effect, and is not likely to stop for a long time', 75%), *vsplyt'* ('to appear in somebody's mind without special reason', 53%).

For adjectives, the discrepancy between the first dictionary sense and the most frequent sense in the Russian National Corpus can be illustrated by the following examples. The word *gluhoj* in the RNC is used in 30% of cases in collocations with *sound,* in the sense of 'a low sound made when one hard heavy object hits another', while its first dictionary sense is 'not able to hear anything' (12%). In some cases, a collocation may be very frequent and thus increases the frequency of an adjective. A good illustration for this observation is the word *vishn'ovyj*: its most frequent sense in the RNC is 'related with a tree that produces cherries' (57%), evidently because of the spread of the name of the Anton Chekhov play 'The Cherry Orchard', in the texts of the academic corpus. For the adjective *burnyj,* the distribution of sense frequencies is completely opposite to the ordering of senses in the dictionary: *stormy weather* (4%), *stormy wind or sea* (15%), *rapid growth* (34%) and *wild passion, stormy romance* (47%). As for verbs, for some adjectives the most frequent sense has undergone a semantic shift and is metonymical, as in the examples *bir'uzovyj* (*turquoise color*, 80%), *antikvarnyj* (*antique shop*, 59%), *belokuryj* (*fair-haired boy*, 55%), *golovnoj* (*head*, 41%).

We think that including the information about the most frequent sense and overall sense frequency distribution in explanatory dictionaries is relevant for dictionary users. Robert Lew (2013) suggested that the information about the most frequent sense would be necessary for text production (such as essay writing) but not for comprehension, as dictionary users usually do not look up a frequent sense of a word. We advocate the need for these conclusions to be tested as soon as the information about sense frequencies of words in dictionaries becomes available. Moreover, it may help to include dictionaries in natural language processing tasks like word sense disambiguation, as necessary information regarding the most frequent sense will become available in explanatory dictionaries and connected with their sense inventories.

We compared the most frequent senses for verbs and adjectives in the Russian National Corpus that contains more literary contexts, with the most frequent senses in the up-to-date web-based ruTenTen11. The corpora have a high degree of overlap: the ratio of the same most frequent sense is 80% for verbs and 82% for adjectives. The difference can be explained by the content of the corpora. The RNC provides quite literary most-frequent senses: as in the examples *close relative* for the word *blizhnij*, *boulevard bench* for *bulvarnyj* and *bitter laugh, bitter irony* for *gor'kij*, as compared with the colloquial uses *the nearest place*, *tabloid novels* and *bitter taste,* respectively. For words such as *anglijskij* and *almaznyj* the most frequent senses in ruTenTen11 are narrower and more specific than in the RNC: 'the English language'/'related to England' and 'produced using cutting diamond'/'related to a diamond' (ruTenTen11/RNC in both examples). These observations are also relevant for verbs. Moreover, we observed that for some verbs the most frequent senses in ruTenTen11 are metaphorical, while in the RNC they are literal. For example, *bazirovat's'a* 'to base a decision or idea on particular information'/'to be based somewhere', *bredit'* 'to talk nonsense'/'to be delirious' and *vooruzhit'* 'to provide yourself or other people with useful information or equipment to achieve the goal'/'to provide yourself or other people with weapons'.

Our aim was to study whether our approach to word sense frequency estimation can be generalized to any explanatory dictionary and therefore we compared the accuracy of our method for three dictionaries: the *Active Dictionary of Russian* (AD), the *Large Explanatory Dictionary of Russian* (BTS) and the *Russian Language Dictionary* (MAS). The comparison was performed on nouns, because nouns normally have more distinct senses (compared to other parts of speech), as many of them refer to objects existing in the real world (Iomdin et al., 2014). In BTS and MAS, the number of collocations and illustrative sentences is much less than in the AD. The lack of examples prevented our method from building solid sense vectors and thus the accuracy of the method trained on BTS and MAS is worse compared to that on the AD. The difference in sense inventories also influenced the results: the word *al'bom* has three senses in the AD—'a book with blank pages, used for drawing', 'a book in which you can collect things such as photographs or stamps' and 'a collection of several songs or pieces of music recorded as an MP3 file, on a CD etc'. The last is rather frequent in the Russian National Corpus (33%) and the most frequent in ruRenTen11 (73%), but is absent in both BTS and MAS. This implies that many contexts are not covered by senses described in these dictionaries. To ensure a good performance, our method requires an up-to-date sense inventory with several typical illustrative sentences and collocations for each sense used for training.

5. Conclusion

This paper continues the study of the automated word sense frequency estimation for Russian words. We applied the method based on semantic vectors and trained on collocations and illustrative sentences from the *Active Dictionary of Russian* to

ambiguous verbs and adjectives from the first issue of the dictionary. As a result, we obtained sense frequencies for 329 verbs and 256 adjectives. All the data are available on http://sensefreq.ruslang.ru. Subsequently, the word sense frequency database now contains frequency distributions for nouns, verbs and adjectives in the academic Russian National Corpus and the web-based corpus ruTenTen11 (1025 ambiguous words in total). We evaluated frequency estimation error for verbs and adjectives and found that it is slightly worse than for nouns but still below 15%.

We experimented with sense granularity in the *Active Dictionary of Russian* and found that using more coarse-grained senses improves disambiguation accuracy, and a hierarchical approach to sense description can be very helpful when fine-grained distinctions between senses are not important for the task at hand.

In order to test our approach on other dictionaries we compared word sense disambiguation accuracy obtained when training on the *Active Dictionary of Russian* to the *Large Explanatory Dictionary of Russian* and the *Russian Language Dictionary*. We found out that although the accuracy on the other two dictionaries is above the baseline, it is substantially lower than on the *Active Dictionary of Russian*. Many collocations and illustrative examples for each sense are important for achieving good disambiguation accuracy.

The information about word sense frequency may have several applications: for lexicography and language learning, for the theoretical and experimental study of polysemy, and for different NLP tasks. The method presented in this paper can be applied to any language with a sufficiently large corpus and a dictionary with contemporary vocabulary that provides several examples of each sense.

6. Acknowledgements

This research was supported by RSF (project no.16–18–02054: "Semantic, statistic, and psycholinguistic analysis of lexical polysemy as a component of a Russian linguistic worldview"). The authors would also like to thank the two anonymous reviewers for their valuable comments.

7. References

Agirre, E., Marquez, L. & Wicentowski, R. (eds.). (2007). *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*. Association for Computational Linguistics, Prague, Czech Republic.

Apresjan, J. (2000). *Systematic Lexicography*. Oxford.

Apresjan, J. (ed.). (2014). *Active Dictionary of Russian*. A-G. JSK, Moscow.

Beck, I., McKeown, M. G. & Kucan, L. (2013). *Bringing Words to Life: Robust Vocabulary Instruction*. Guilford Press.

Evgenyeva, A. (ed.). (1981–1984). *Russian Language Dictionary*. Russian language,

Moscow.

Ferraresi, A., Bernardini, S., Picci, G. & Baroni, M. (2010). Web corpora for bilingual lexicography: A pilot study of English/French collocation extraction and translation. *Using Corpora in Contrastive and Translation Studies*. Newcastle: Cambridge Scholars Publishing, pp. 337–359.

Foraker, S. & Murphy, G. L. (2012). Polysemy in sentence comprehension: Effects of meaning dominance. *Journal of memory and language* 67.4, pp. 407-425.

Iomdin, B. (2014). Polysemous words in and out of the context. *Voprosy jazykoznanija*. Vol. 4. Moscow.

Iomdin, B., Lopukhina, A. & Nosyrev, G. (2014). Towards a word sense frequency dictionary. *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2014"*. Bekasovo, Moscow, pp. 204–229.

Iomdin, B., Lopukhin, K., Lopukhina, A. & Nosyrev, G. (2016). Word sense frequency of similar polysemous words in different languages. *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2016"*. Moscow, pp. 201–211.

Kilgarriff, A., Rychly, P., Smrz, P. & Tugwell, D. (2004). The Sketch Engine. *Euralex 2004. Proceedings*. Lorient, France, pp. 105–116.

Kiselev, Y., Krizhanovsky, A., Braslavski, P., Menshikov, I., Mukhin, M. & Krizhanovskaya, N. (2015). Russian Lexicographic Landscape: a Tale of 12 Dictionaries. *Proceedings of the International Conference "Dialog 2015*", pp. 254-272.

Klein, D. & Murphy, G. L. (2001). The representation of polysemous words. *Journal of Memory and Language* 45.2, pp. 259-282.

Kruglikova, L. (2012). The big academic dictionary of Russian as a successor of Russian academic lexicography traditions. *Cuadernos de Rusistica Espanola*, 8, pp. 177-198.

Kuznetsov, S. (ed.). (2014). *Large Explanatory Dictionary of Russian*. Norint, St. Petersburg.

Lew, R. (2013). Identifying, ordering and defining senses. In H. Jackson (ed.) *The Bloomsbury Companion to Lexicography*. London: Bloomsbury Publishing, pp. 284–302.

Lopukhin, K. & Lopukhina, A. (2016). Word sense disambiguation for Russian verbs using semantic vectors and dictionary entries. *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2016"*, pp. 393-405.

Lopukhina, A., Lopukhin, K., Iomdin, B. & Nosyrev, G. (2016). The Taming of the Polysemy: Automated Word Sense Frequency Estimation for Lexicographic Purposes. *Proceedings of the XVII EURALEX International Congress. Lexicography and Linguistic Diversity* (6–10 September, 2016). Tbilisi: Ivane Javakhishvili Tbilisi State University, pp. 249-257.

Lopukhina, A., Lopukhin, K. & Nosyrev, G. (in press). Automated word sense frequency estimation for Russian nouns. In M. Kopotev, O. Lyashevskaya, A.

Mustajoki (eds), *Quantitative Approaches to the Russian Language*. Routledge.
Loukachevitch, N. & Chetviorkin, I. (2015). Determining the most frequent senses using Russian linguistic ontology RuThes. *Proceedings of the Workshop on Semantic Resources and Semantic Annotation for Natural Language Processing and the Digital Humanities at NODALIDA*, pp. 21–27
McCarthy, D., Koeling, R., Weeds, J. & Carroll, J. (2007). Unsupervised acquisition of predominant word senses. *Computational Linguistics* 33:4, pp. 553–590.
Mohammad, S. & Hirst, G. (2006). Determining word sense dominance using a thesaurus. *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006)*, pp. 121–128.
Navigli, R. (2006). Meaningful clustering of senses helps boost word sense disambiguation performance. *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of ACL*. Association for Computational Linguistics, USA, pp. 105–112.
Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)* 41:2, pp. 1–69, Article 10.
Nesi, H. & Haill, R. (2002). A study of dictionary use by international students at a British university. *International Journal of Lexicography* 15.4, pp. 277–306.
Pustejovsky, J. (1996). *Lexical semantics: The problem of polysemy*. Oxford.
Pylkkänen, L., Llinás, R. & Murphy, G. L. (2006). The representation of polysemy: MEG evidence." *Journal of cognitive neuroscience* 18.1, pp. 97-109.
ruscorpora.ru/en. Accessed at: http://ruscorpora.ru/en. (10 July 2017)
github.com/lopuhin/sensefreq. Accessed at: https://github.com/lopuhin/sensefreq. (10 July 2017)
sensefreq.ruslang.ru. Accessed at: http://sensefreq.ruslang.ru/. (10 July 2017)

LeGeDe – Towards a Corpus-based Lexical Resource of Spoken German

Christine Möhrs, Meike Meliss, Dolores Batinić

Institut für Deutsche Sprache, P.O. Box 101621, D-68016 Mannheim
E-mail: moehrs@ids-mannheim.de, meliss@ids-mannheim.de, batinic@ids-mannheim.de

Abstract

This paper gives an insight into the basic concepts for a corpus-based lexical resource of spoken German, which is being developed by the project "The Lexicon of Spoken German" (Lexik des gesprochenen Deutsch, LeGeDe) at the "Institute for the German Language" (Institut für Deutsche Sprache, IDS) in Mannheim. The focus of the paper is on initial ideas of semi-automatic and automatic resources that assist the quantitative analysis of the corpus data for the creation of dictionary content. The work is based on the "Research and Teaching Corpus of Spoken German" (Forschungs- und Lehrkorpus Gesprochenes Deutsch, FOLK).

Keywords: spoken German, corpus linguistics, internet lexicography, lexicology

1. Introduction

The purpose of the project "Lexicon of Spoken German" (Lexik des gesprochenen Deutsch, LeGeDe), which started in September 2016 at the "Institute for the German Language" (Institut für Deutsche Sprache, IDS) in Mannheim, is to build an electronic lexical resource for spoken standard German based on the empiric data of the "Research and Teaching Corpus of Spoken German" (Forschungs- und Lehrkorpus Gesprochenes Deutsch, FOLK[1]). FOLK is the largest corpus of spoken German in interactions (202h/1.95 Mio. tokens; DGD version 2.8) and is made available via the "Database for Spoken German" (Datenbank für Gesprochenes Deutsch, DGD[2]); cf. Schmidt (2014a/2014b, 2016).

LeGeDe is a third-party funded project [3] of the Leibniz Association (Leibniz Competition 2016, Funding line 1: Innovative projects[4]). For a period of three years (from 1 September 2016 to 31 August 2019) the project will be working on the creation

[1] Information about FOLK: http://agd.ids-mannheim.de/folk.shtml.

[2] URL to the DGD-Website: http://dgd.ids-mannheim.de.

[3] Applicants of the project: Annette Klosa, Arnulf Deppermann, Stefan Engelberg, Thomas Schmidt (IDS Mannheim).

[4] For more information about the competition and the funded projects, please go to: http://www.leibniz-gemeinschaft.de/en/about-us/leibniz-competition/projekte-2016/funding-line-1/.

a lexical resource of spoken German.

The project is a cooperation of two departments of the IDS in Mannheim: the Department of Pragmatics and the Department of Lexical Studies. The team consists of researchers with different research backgrounds: lexicographers (especially researchers with a special focus on electronic lexicography), corpus linguists, and researchers with a special focus on conversational analysis.

The aim of the project is twofold: (1) to develop a lexicographic resource for spoken German (language area: Germany) by benefiting from the methods of corpus-linguistics, and (2) to find an optimal solution for presenting this type of language resource by exploring and extending the possibilities offered by its digital form. The lexicographic resource of spoken German is to be designed in a dynamic (extendible) manner, and it is intended to integrate multi-modal information, such as corpus-based audio-examples and transcriptions for each entry. Hence, compiling such a resource is challenging both from the lexicographic perspective as well as from the point of view of data modelling. In the long term, the resource will be integrated into the dictionary portal OWID[5], which has been developed at the IDS in Mannheim (Online-Wortschatz-Informationssystem Deutsch; eng.: Online vocabulary system of the German language). It will cover, in an exemplary fashion, lexical units and properties typical for spoken German as it is used in conversations in private and institutional contexts.

Modern lexicographic resources of German are usually (and mainly) based on written language represented in large electronic text corpora (e.g. monolingual German dictionaries such as Duden-online, DWDS or *elexiko*). Characteristics of spoken German, especially with regard to the lexicon, are not described in great detail in these dictionaries (cf. Meliss, 2016); see the discussion in Section 5 on this aspect. LeGeDe is the first project that aims to identify the peculiarities of language in an interactional context in a systematic way (cf. Section 5). We are aware of only one similar project focusing on interjections in spoken Danish (cf. Hansen/Hansen, 2012) and another one currently being developed for Slovenian (cf. Verdonik & Sepesy Maučec, 2017).

The present paper is subdivided into six sections. The subject area of the project is presented in Section 2. In Section 3, the basis of the project's data is described. We will present aspects of the quantitative corpus analysis in Section 4 and of the data analysis in Section 5. The paper concludes in Section 6 with final remarks and comments on the project's additional objectives.

[5] URL to the OWID-Website: www.owid.de.

2. Phenomena of interest

We concentrate on those phenomena which we can characterize as "standard"—in the sense that we intend not to consider dialects (such as Bavarian), sociolects (such as adolescent language) or idiolects. Our interest is mainly directed to those phenomena of spoken German that are used more frequently, or in a different manner than in written German (e.g., regarding meaning or function in verbal interaction). A selection of phenomena that are to be dealt with in the project are listed in Table 1.

Phenomena of interest (selection)	
Verbs	*ich dachte* (tempus), *guck* (imperative), *meinste* (complementation patterns), Ich *kann* kein Deutsch (modal verbs in absolute use), *geht* (spec. semantics 3rd person) etc.
Word borrowings	German language varieties: *öko*[logisch], *wo* (as a relative pronoun) etc.; Anglicisms: *okay*, *cool*, *fuck* etc. (frequency, groups of speakers, gramm. integration, phonetic realization etc.)
Word formation	*rum-*, *rein-*, *rauf-*; *mega-, super-, sau-, ober-*; *-mäßig (randalemäßig), -i (Hirni)* etc.
Partial synonyms	*kriegen/bekommen/erhalten, gucken/ schauen/sehen; Auto/Karre/Kutsche* etc.
Conversation words	*eben, jein, hä, tss, pf, ups, hoppla* etc.; *gut, richtig, genau, sicher, einfach* etc.
Patterns	*guck mal, alles klar, einen drauf machen* etc.

Table 1: Some phenomena of interest and selected examples.

The table provides a rough guide on phenomena and specific lexical units, which should be assigned to the respective phenomena. These areas are also identified as interesting phenomena in research literature (e.g. Schwitalla, 2012; Deppermann, 2005/2007; Fiehler, 2016) and in previous studies on spoken German (Imo, 2007; Günthner, 2016; Deppermann et al. (eds.), 2017). With the help of the analysis of corpus evidence the phenomena are to be examined more closely and the candidates should be defined by means of frequency-oriented and competence-based examinations. This should make it possible to draw a clear picture of the relevant phenomena areas, following both a corpus-based and a competence-oriented methodology.

3. Corpus material

We base our research on FOLK that primarily addresses researchers from the fields of conversation analysis and corpus linguistics and comprises conversations from different interaction domains, such as institutional and private conversations, game interactions, table talk, etc. Since the data are annotated on multiple levels (meta information about speakers, interactions and word forms; cf. Westpfahl & Schmidt, 2016), FOLK provides a reliable basis for a study of interactional phenomena of spoken language, towards which our analysis is mainly directed. Schmidt (2014a) describes its aims as follows:

"[FOLK] has [...] set itself the aim of building a corpus of German conversations which:

a) covers a broad range of interaction types in private, institutional and public settings,
b) is sufficiently large and diverse and of sufficient quality to support different qualitative and quantitative research approaches,
c) is transcribed, annotated and made accessible according to current technological standards,
d) is available to the scientific community on a sound legal basis and without unnecessary restrictions of usage." (Schmidt 2014a: 383)

By today, a set of data comprising approximately 202h of recordings and close to 1.95 million transcribed tokens, has been completely processed in the FOLK corpus and has been published via the DGD.

Private interaction	**interactions**	**hours**	**tokens**
e.g. coffee table conversation, telephone conversation, conversation on a holiday trip, student everyday conversation, conversation during breakfast, conversation among friends, etc.	89	84:25	864,208
Interaction in school/university/at the workplace (non-private/non-public)			
e.g. oral exams at a university, shift change at a hospital, driving school conversation, meeting in an economic company, classroom observation, conversation during a regular meeting, etc.	117	67:53	604,121
Public interaction			
mediation talks, panel discussion	6	25:26	237,707
Other interaction domains			
maptasks, biographic interview, interview, ethnographic interview	47	24:27	246,123

Table 2: Interaction domains and examples (selection) in FOLK (status as of 17.05.2017; cf. also Schmidt, 2014a: 383).

FOLK contains transcripts as well as audio and video material on spoken German in interaction. The composition of the corpus can be observed in Table 2. Figure 1 shows the distribution of all tokens over the entire corpus with respect to major interaction domains.

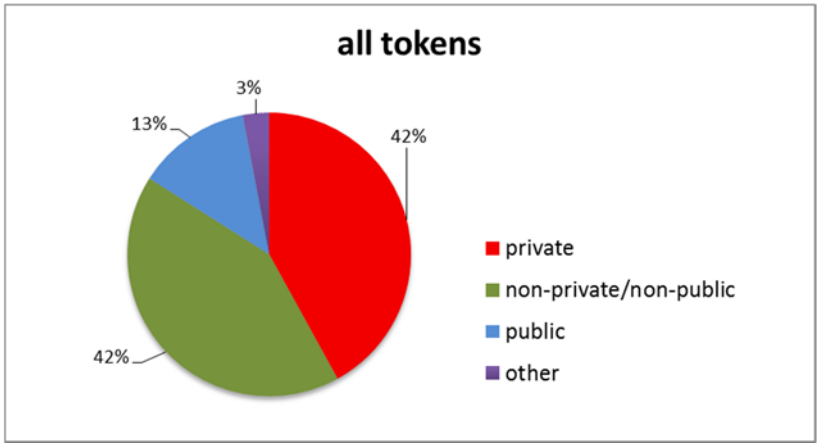


Figure 1: Major interaction domains in FOLK

The list of these different conversations (cf. Table 2) shows the broad diversity of interaction domains covered by FOLK. FOLK's special feature is to document spoken German in spontaneous interaction. This distinguishes it from most other oral corpora in the DGD (see for example the corpus "Deutsche Standardsprache: König-Korpus" which includes reading texts, in particular excerpts from the German Grundgesetz; cf. Schmidt, 2014b: 1451). After the creation of an individual account, the access to the DGD is free of charge for research and teaching purposes. This makes the data base, with which the LeGeDe project works, transparent to the scientific public. Nevertheless, one aspect with regard to FOLK is not to be neglected: Even if it is among the largest available corpora of its kind, with a total of 1.95 million transcribed tokens, it is still a relatively small corpus. Corpus-based methods, which up to now have been used in lexicography on large volumes of written German, need to be looked at in a new way.

However, FOLK is still being set up and will grow further over the project period. The coverage of different interaction domains, as well as the coverage of speakers from different regions in Germany and of additional metadata, will therefore be constantly improved and expanded over the coming years. Thus, the LeGeDe project works with the most adequate corpus for the analysis of the lexicon of spoken German on an interactional basis. Since lexicographic resources for the German language have not yet been developed for spoken language data, an important task of the LeGeDe project is to develop new approaches to the corpus-assisted analysis of interactional data. A particular challenge is to unite the methods of conversational analysis with those of lexicological and lexicographical analysis.

4. Quantitative corpus analysis

One of the challenges of the LeGeDe project is to develop automatic, semi-automatic and manual analysis methods, which serve different purposes: The results of automatic methods are used to pre-structure data sets related to different areas, e.g. information about combinatorics, formal realisation and meta linguistic data, so that they can be used for the lexicographic resource and be commented on by the lexicographers. The editorial elaboration of the dictionary entries is, of course, another important part of the project work, but this paper does not elaborate on this point.

The linguistic units to be included in the lexicographic resource should, above all, satisfy the criterion of having relevance in the spoken language. Wherever it is meaningfully possible, the aspect of distinctiveness should be taken into account in comparison to written German. In order to assist the detection of salient terms in spoken German, we work with frequency comparison between FOLK and DeReKo ("Deutsches Referenzkorpus", written German; eng.: German reference corpus[6]). DeReKo (cf. Kupietz/Keibel, 2009) is much larger: it currently comprises about 29 billion running words. Our assumption is that noticeable frequency differences may indicate to differences in meaning and use. We apply different measures for frequency comparisons, such as Log Likelihood Ratio (Dunning, 1993), Odds Ratio and frequency classes (Perkuhn et al., 2012). The comparative analyses with DeReKo, as a corpus with a wide coverage of many different types of texts, are limited to a subset of the data. For instance, we excluded the Wikipedia sources because of the conceptually spoken German used in the discussion pages. Since DeReKo and FOLK differ in corpus size (DeReKo = 29 billion text words vs. FOLK = 202 h / 1.95 million tokens) and temporal coverage of the sources (DeReKo = 1772–2015 vs. FOLK = 2003–2016) differences in metadata and text types must be judged very carefully between the two corpora. They should serve as a frequency-controlled aid to interpretation (see for example the article by Kupietz and Schmidt (2015) on written and oral corpora at IDS as the basis for empirical research).

After the frequency comparison of the two corpora, we identified different lexical units of interest, such as verbs (*gucken*, *kriegen*, *finden*, *meinen* etc.), particles in the broad sense (*mal*, *halt*, *eben*, *ah*, *oh*, *okay* etc.), adjectives (*gut*, *prima*, *schön*, *geil*, *krass* etc.), nouns (*Ding*, *Sache*, *Stress* etc.), and pronouns (*etwas*, *was*, *solch-*, *irgend-* etc.). An excerpt of the table for frequency analysis representing the particles with the highest difference in frequency classes can be observed in Table 3.

[6] Information about DeReKo: http://www.ids-mannheim.de/kl/projekte/korpora/.

Lemma	FOLK absolute frequency	DeReKo absolute frequency	FOLK frequency class	DeReKo frequency class	Difference of frequency class
okay	6477	199942	4	14	10
halt	6136	802658	4	12	8
mal	14076	8523173	2	8	6
na	3077	520673	5	12	7

Table 3: Frequency comparisons: particles (excerpt).

We also use the comparison of frequency classes for studying the distributional behaviour of pseudo-synonyms, such as between the verbs *gucken* and *schauen* (see Table 4).

Lemma	FOLK absolute frequency	DeReKo absolute frequency	FOLK frequency class	DeReKo frequency class	Difference of frequency class
gucken	2598	375327	5	13	8
schauen	570	2570951	7	10	3

Table 4: Frequency comparisons: *gucken* vs. *schauen* (excerpt).

In addition, since we categorised all the transcripts in FOLK into interaction domains such as "private", "public", "non-private/non-public" and "other" (see Section 3, Figure 1), we determine the distribution of lexical items within different categories. Such an indication can refer to a single element (example *gucken*), but it can also be considered in relation to the distribution of all lemmas in FOLK. We also use this categorisation in order to study the lexical units belonging to the same phenomenon class (example: visual perception verbs; *gucken*, *schauen*, *sehen*; cf. Figure 2).

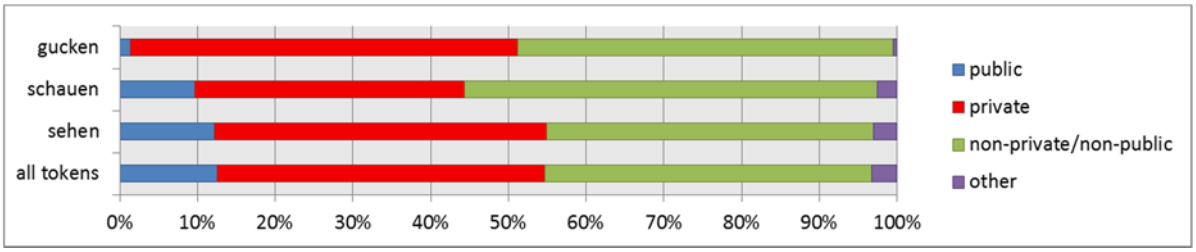


Figure 2: Distribution on different interaction domains. Comparison: visual perception verbs (*gucken*, *schauen*, *sehen)* - total amount of all tokens

The comparison in Figure 2 shows, on the one hand, that the frequency of the verb *gucken* is relatively higher in private conversations compared to the other two visual perceptual verbs (*schauen* and *sehen*); in addition, *gucken* is much less common in public conversations. On the other hand, compared to all tokens in FOLK, *gucken* rarely occurs in public conversations and with increased frequency in private contexts.

Since our first case studies focus on verbs, in order to obtain a fine-grained analysis of the verb distribution in FOLK, we perform a reconstruction of separable particle verbs in the corpus (Volk et al., 2016; Batinić & Schmidt, 2017). In that way, verbs such as *angucken* or *anschauen* can be extracted from the corpus even when they are not written together, a piece of information usually not available in the default lemmatisation of most corpora. Since FOLK contains not only transcribed words, but also their normalised and lemmatised forms, we can perform frequency measurement on each formalisation level. In order to have an overview of the word form frequencies on each level, we produce a word profile containing the frequency of transcribed word forms for each annotation level (cf. Table 5).

Lemma	**Norm**	**Transcription**
gucken	*geguckt*	*geguckt* 81, *gekuckt* 2, *geguck* 2
gucken	*gucken*	*gucken* 686, *gucke* 77, *gugge* 34, *kucken* 28, *guckn* 7, *guck* 5, *gu* 5, *kucke* 4, *kuck* 3
gucken	*guckten*	*guckten* 2
gucken	*guckte*	*guckte* 3
gucken	*gucke*	*guck* 105, *gucke* 28, *kuck* 22
gucken	*guckt*	*guckt* 111, *kuckt* 6, *guck* 3
gucken	*guckst*	*guckst* 79, *gucks* 33, *gucksch* 4, *kuckst* 3, *guckscht* 2
gucken	*guck*	*guck* 475, *gu* 82, *kuck* 13, *ku* 10, *gugg* 8, *gucke* 2, *kiek* 2

Table 5: Frequency of transcribed word forms for each annotation level (example *gucken*).

We also study word distributions by using different meta-information about region and speaker. Table 6 shows selected words that are less frequently used by men than by women.

Lemma	**Male (948,586 tokens)**	**Female (980,190 tokens)**	**Range (number of speakers)**	**Log Likelihood**	**Odds Ratio**
Gott	212	598	214	179,20	0,37
ups	17	87	48	49,27	0,20
juhu	6	47	19	34,71	0,13
boah	148	380	162	98,04	0,40

Table 6: Distribution via the parameter "gender" (excerpt).

In addition to analysing one word lemmas, we also focus on multiword expressions. We identify frequent words that co-occur with the target word as well as the most frequent bi- and tri-grams containing the target word (we work with absolute frequencies given the relatively small size of the corpus). The co-occurrence profiles are commonly used for the analysis of corpora of written language (for the creation and use of word profiles in lexicography see e.g. Adam Kilgarriff's work on Word Sketches: e.g. Kilgarriff & Kosem, 2012 or Kilgarriff, 2015). These methods have not yet been applied to data material for spoken German, especially with regard to FOLK. Missing sentence boundaries, speaker changes, uncertain word forms, and overlaps, etc. are only a few challenges in this regard. The project deals with the opportunities and limitations of such statistical procedures.

After detecting salient word combinations (e.g. *guck mal*, *müssen wir mal gucken*) we analyse them in detail in the coding part (see Section 5). An overview of some frequent co-occurrences (word combinations, patterns, etc.) of the verb *gucken* is shown in Figure 3.

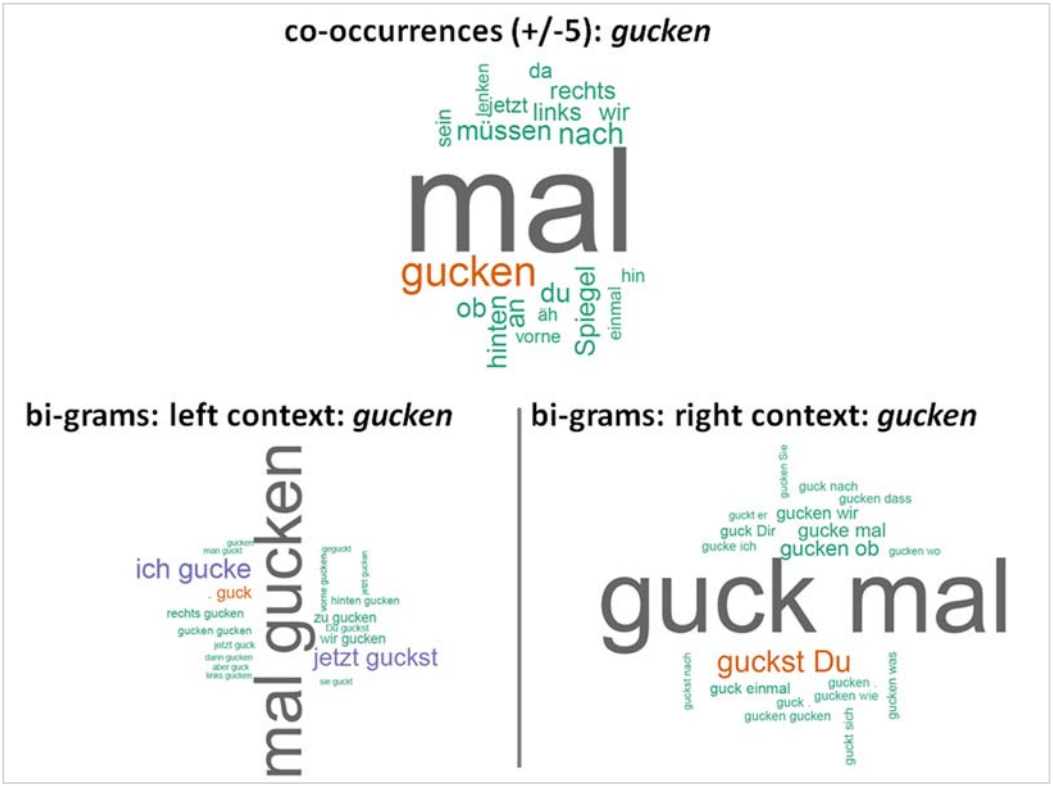


Figure 3: Co-occurrences and bi-grams with regard to the verb *gucken*

5. Data analysis

We have carried out the first in-depth analyses with verbs, which we exemplarily illustrate in this section. The first steps (sampling, creation of a coding table) involved the elaboration of a coding scheme as well as the analysis and structuring of the data – especially in connection with initial considerations about the development of a lexicographic microstructure.

In order to extract corpus samples constraining a particular lemma, we defined following preliminary steps: a) assigning all conversations to four different interaction domains ("private", "public", "non-public/non-private", "other"; see Table 2 and Figure 1), b) calculating the distribution of the lemma to the interaction domains with regard to the whole corpus and c) transferring the distribution to the proportion with regard to the sample.

Each KWIC line of our sample has a column with a link to the corresponding transcript excerpt in the database (see Figure 4; DGD, FOLK). In this way, the larger context of an occurrence and the corresponding audio recording can be inspected, both of which are essential for the various steps of the analysis (see Figure 5).

METADATEN TREFFER				
SHORT_FILE ID	LINKER KONTEXT	STICHWORT	RECHTER KONTEXT	LINKID
FOLK_E_0024	weiß ich nich wie ich	**gucke**		http://dg
FOLK_E_0002	dann	**guck**	isch heut ma was wir morgen dann noch einkaufen gehen müssen	http://dg
FOLK_E_0002	ja mir müsse mal	**gucke**	was mer für wichtige halte und was net vielleicht kannscht ja afach da irgendwie n hake dra mache	http://dg
FOLK_E_0008	ja	**guck**	mal die arbeitsumstände die werden ja immer dem land angepasst in dem du dann grade deine filiale aufmachst	http://dg
FOLK_E_0020	is immer so ich geh dann einkaufen un dann will ich ja eintlich gar nisch aber dann wenn ich immer so drinnen bin dann kann ich irgendwie dann immer so schön	**gucken**	un da guck ich immer mit	http://dg

Figure 4: Extract from an excel spreadsheet of the search results to *gucken* (eng. *to look*) (FOLK, DGD)

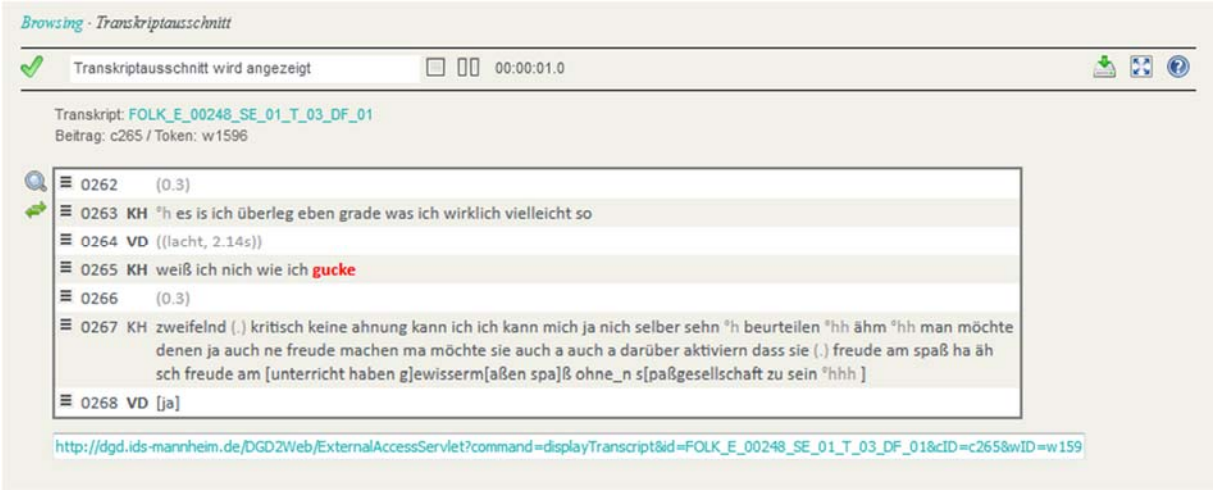


Figure 5: Corpus reference to the link from the excel sheet to the verb *gucken*, KWIC line 1

To code the data, a coding scheme has been developed for five different coding areas with different coding parameters (see Figure 6). In addition to the different automatically generated metadata regarding the hit itself (Section 1), there is automatically-generated information on meta-language data concerning the transcript (Section 5). The data are examined through a "hands-on analysis", with regard to content-functional analysis (Section 2), syntactic-formal analysis (Section 3) and grammatical information (Section 4).

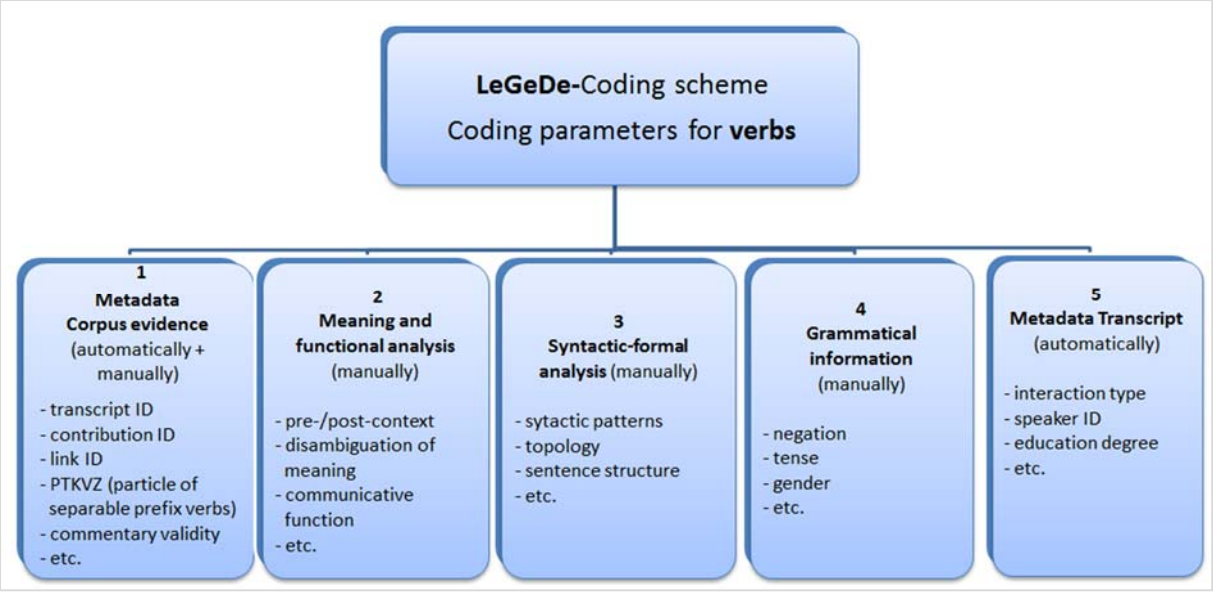


Figure 6: Coding parameters for verbs

The coding scheme is continuously refined in several encoding processes, which are carried out by several persons. Multiple encoding processes and examinations of the data by different persons are intended to increase precision in the coding and interpretation of the data, particularly in the meaning-disambiguation and the differentiation of the function of a word or a phrase in the interactional context.

As already mentioned in Section 1, the description of the peculiarities, especially in the area of the lexis of spoken German, is only inadequately documented in existing dictionaries. Figure 7 shows an extract of the dictionary article *gucken* from one of the most consulted dictionaries, the Learner dictionary for German as a foreign Language (LGWB-DaF). The extract from the dictionary article shows grammatical information (verb intransitive, sentence structure patterns, ["irgendwohin / irgendwie gucken…"]) and information on the meaning (definition, paradigmatic relations). The dictionary user also finds the very general pragmatic information that the lemma *gucken* is a lemma used in contexts of spoken German (label: "gesprochen"). Only three meanings of the lemma *gucken* are listed in this dictionary.[7]

Our analyses of the lemma *gucken* indicate that we have come to a more expanded understanding of the meanings, formal realizations, and ultimately of the function of the verb *gucken* compared to information from standard German dictionaries and, particularly, of learners' dictionaries. According to our investigations, the spectrum regarding the meanings of *gucken* is much larger. We performed the semantic disambiguation by analyzing the form ("[argument] structure pattern" in conjunction

[7] In the Pons Kompaktwörterbuch (Deutsch als Fremdsprache – German as a foreign language; 2016), two meanings are listed, the Duden 10 (Bedeutungswörterbuch - explanatory dictionary; 4th edition 2010) and the website of Duden-online show three different meanings of *gucken*.

with the corresponding "sentence structure") and content (cf. Table 7).

Figure 7: Extract from the dictionary article *gucken* from the "LGWB-DaF"

Semantic definition / meaning	**Synonyms**	**(=STM) (argument) structure pattern**	**(=SBP) sentence structure**[8]
...	...	...	...
jmd. stellt fest, dass etw. d. Fall ist	feststellen	jemand *guckt*, dass etwas der Fall ist	<Ksub, Kverb>
jmd. sieht s. etw. an	sich ansehen	jemand *guckt* etwas	<Ksub, Kakk>
jmd. beobachtet, wie etwas passiert	beobachten zuschauen	jemand *guckt*, wie etwas passiert	<Ksub, Kverb>
jmd. sucht nach etwas	suchen	jemand *guckt* nach etwas	<Ksub, $Kprp_{nach}$>
jmd. schaut sich um	umherschauen	jemand *guckt* auf eine bestimmte Art und Weise	<Ksub, Kmod>
jmd. passt auf, dass etwas (nicht) passiert	aufpassen kontrollieren	jemand *guckt* dass etwas (nicht) passiert	<Ksub, Kverb>

Table 7: Different meanings of the lemma *gucken* (excerpt)

[8] Terminology in accordance with Zifonun et al. (1997).

As FOLK constitutes our database, it is possible for us to work on interaction-specific information in particular, and to implement it for the planned lexicographic resource. The following information would be interesting and could profitably complement the offer of existing dictionaries: the interaction context or sequence context, prosody and sound realisation, large variety in functional aspects with regards to the interaction context, combination potential (cf. Figure 3 in Section 4 and the discussion about automatically generated co-occurrence profiles and the identification of combination potential), information about topology, and other aspects.

Taking into account the corresponding interaction context and the metadata, conclusions can be drawn about the respective possibilities of use and the corresponding communicative functions. With FOLK as a database, the expertise in the project on conversational analysis, as well as the expertise in the field of lexicology and lexicography, the project would like to close the gap with respect to the interaction-specific information for verbs as well as for other word classes and lexical patterns.

6. Final remarks

During the project period we want to develop corpus-based methods for analyzing and structuring spoken lexis as well as a lexicographical process that takes into acocunt the characteristics of language in interaction and the possibilities of the database. The sub-targets of the project can be described as follows: (i) determination of the peculiarities and divergences of spoken and written language usage in the lexical area at all levels (form, content/function, situation etc.), (ii) development of further corpus linguistic methods for analysing and structuring the data of spoken language, (iii) development of innovative types of lexicographical information, which refer to the function of lexical units in interaction contexts, and (iv) development of innovative description formats in a multimedia format for lexical data. The aim is to offer the user a mixture of automatically-generated data (see Section 5 in particular), as well as lexicographically-commented information (see Section 6 with regard to the analysis steps).

The lexicographically-commented information will include aspects such as peculiarities in form (form-related realization, word forms, inflection, phonetic realization, etc.), combinatorics (actants, morphosyntactic information, etc.), meaning (meaning description, conceptual reference, paradigmatic sense relations, etc.) and communicative function (combination of topology, formal aspects, interactional criteria, metadata, etc.). From the specifics of the lexicons in oral communication, new challenges arise for the macro-, micro- and medio-structure of this new type of dictionary, as well as for an electronic presentation that must combine text with multimedia forms of expressions.

Besides being used for linguistic research, the lexical resource could contribute to the acquisition of German as a foreign or second language, as well as to the development of a language-reflexive first language teaching[9].

The LeGeDe project not only contributes to a new description of contemporary German, but also to the development of lexical descriptions appropriate for the lexis of spoken German. The lexicographic resource is intended to describe the lexical competences of everyday conversation and to contribute to the better understanding of the peculiarities of the vocabulary of spoken German in interaction.

7. Acknowledgements

We would especially like to thank our colleagues Rainer Perkuhn, Cyril Belica and Marc Kupietz (program area "Corpus Linguistics" at the IDS), which support us in many corpus linguistic questions. They are of great help to the LeGeDe project in aspects relating to the corpora of spoken German and the analysis procedure for these corpora. We also thank our colleagues from the Department of Pragmatics and the Department of Lexical Studies, who have actively supported us with their advice and ideas. Our thanks go to Henrike Helmer, Julia Kaiser, Frank Michaelis, Carolin Müller-Spitzer, Nadine Proske and Arne Zeschel.

8. References

Batinić, D. & Schmidt, T. (2017). Reconstructing of Separable Particle Verbs in a Corpus of Spoken German. To appear in *Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology (GSCL 2017)*, September 13 – 15, 2017, Humboldt Universität zu Berlin, Germany.

Deppermann, A. (2005). Conversational interpretation of lexical items and conversational contrasting. In A. Hakulinen & M. Selting (eds.) *Syntax and lexis in conversation*. Amsterdam: Benjamins, pp. 289–317.

Deppermann, A. (2007). *Grammatik und Semantik aus gesprächsanalytischer Sicht*. Berlin: de Gruyter.

Deppermann, A. & Proske, N. & Zeschel, A. (eds.) (2017): *Verben im interaktiven Kontext. Bewegungsverben und mentale Verben im gesprochenen Deutsch*. Tübingen: Narr.

Dunning, T. (1993). Accurate Methods for the Statistics of Surprise and Coincide. In

[9] See e.g. keywords in "Kultusministerkonferenz" [2012: 12]: "Sprache und Sprachgebrauch reflektieren/ Reflecting language and language usage" as well as "Sich mit Texten und Medien auseinandersetzen/Dealing with texts and media".

Computational Linguistics. 19 (1), pp. 61–74.

Fiehler, R. (2016). Gesprochene Sprache. In A. Wöllstein & Dudenredaktion (eds.) *Duden – Die Grammatik.* Berlin: Dudenverlag, pp. 1181–1260.

Günthner, S. (2016). Diskursmarker in der Interaktion – Formen und Funktionen univerbierter *guck mal*- und *weißt du*-Konstruktionen. In *SpIn-Arbeitspapierreihe (Sprache und Interaktion)*, Nr. 68.

Hansen, C. & Hansen, M. H. (2012). A Dictionary of Spoken Danish. In R. V. Fjeld & J. M. Torjusen (eds.) *Proceedings of the 15th EURALEX International Congress. 7-11 August 2012*. Oslo, Norway: Department of Linguistics and Scandinavian Studies, University of Oslo, pp. 929–935.

Imo, W. (2007). *Construction Grammar und Gesprochene-Sprache-Forschung. Konstruktionen mit zehn matrixsatzfähigen Verben im gesprochenen Deutsch.* Tübingen: Niemeyer (= Germanistische Linguistik, Band 275).

Kilgarriff, A. & Kosem, I. (2012). Corpus tools for lexicographers. In S. Granger & M. Paquot (eds.) *Electronic lexicography.* Oxford: Oxford Univ. Press, pp. 31–55.

Kilgarriff, A. (2015). Using Corpora as Data Sources for Dictionaries. In H. Jackson (ed.) *The Bloomsbury Companion to Lexicography.* London/New Delhi/New York/Sydney: Bloomsbury, pp. 77–96.

Kulturministerkonferenz (2012). *Bildungsstandards im Fach Deutsch für die allgemeine Hochschulreife*. (Beschluss der Kultusministerkonferenz vom 18.10.2012).
http://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2012/2012_10_18-Bildungsstandards-Deutsch-Abi.pdf. (Accessed at: 10 July 2017).

Kupietz, M. & Keibel, H. (2009). The Mannheim German Reference Corpus (DEREKO) as a Basis for Empirical Linguistic Research. In *Working Papers in Corpus-based Linguistics and Language Education, No. 3*. Tokyo: Tokyo University of Foreign Studies (TUFS), pp. 53–59.

Kupietz, M. & Schmidt, T. (2015). Schriftliche und mündliche Korpora am IDS als Grundlage für die empirische Forschung. In L. Eichinger (ed.) *Sprachwissenschaft im Fokus*. Berlin/Boston: de Gruyter, pp. 297–322. (= Jahrbuch des Instituts für Deutsche Sprache 2015).

LGWB-DaF: *Langenscheidt Großwörterbuch Deutsch als Fremdsprache* (Neubearbeitung 2015; online access via the IDS library).

Meliss, M. (2016). Gesprochene Sprache in DaF-Lernerwörterbüchern. In B. Handwerker & R. Bäuerle, & B. Sieberg (eds.) *Gesprochene Fremdsprache Deutsch*. Baltmannsweiler: Schneider, pp. 179–199. (= Perspektiven Deutsch als Fremdsprache, Band 32).

Perkuhn, R. & Keibel, H. & Kupietz, M. (2012). *Korpuslinguistik*. (= UTB 3433). Paderborn: Fink.

Schmidt, T. (2014a). The Research and Teaching Corpus of Spoken German - FOLK. In *Proceedings of LREC'14*, Reykjavik, Iceland: ELRA, pp. 383–387.

Schmidt, T. (2014b). The Database for Spoken German - DGD2. In *Proceedings of LREC'14*, Reykjavik, Iceland: ELRA, pp. 1451–1457.

Schmidt, T. (2016). Good practices in the compilation of FOLK, the Research and Teaching Corpus of Spoken German. In J. M. Kirk & G. Andersen (eds.) *Compilation, transcription, markup and annotation of spoken corpora, Special Issue of the International Journal of Corpus Linguistics* [IJCL 21:3], pp. 396–418.

Schwitalla, J. (2012). *Gesprochenes Deutsch. Eine Einführung.* 4., neu bearbeitete und erweiterte Auflage. Berlin: Schmidt. (Grundlagen der Germanistik 33).

Verdonik, D. & Sepesy Maučec, M. (2017). A Speech Corpus as a Source of Lexical Information. In *International Journal of Lexicography.* 30 (2), pp. 143–166.

Volk, M. & Clematide, S. & Graën, J. & Ströbel, P. (2016). Bi-particle adverbs, PoS-tagging and the recognition of german separable prefix verbs. In: KONVENS 2016, Bochum, 19-21 September 2016. https://doi.org/10.5167/uzh-126372. (Accessed at: 10 July 2017).

Westpfahl, S. & Schmidt, T. (2016). FOLK-Gold – A GOLD standard for Part-of-Speech-Tagging of Spoken German. In *Proceedings of the Tenth Conference on International Language Resources and Evaluation (LREC'16), Portorož, Slovenia.* Paris: European Language Resources Association (ELRA), pp. 1493–1499.

Zifonun, G. & Hoffmann, L. & Strecker, B. (1997). *Grammatik der deutschen Sprache.* Berlin/New York: de Gruyter. 3 volumes.

Dictionaries and dictionary portals:

Duden 10 - Das Bedeutungswörterbuch (2010). Mannheim: Duden-Verlag.

Duden-online. Accessed at: http://www.duden.de. (10 July 2017).

DWDS: *Digitales Wörterbuch der Deutschen Sprache.* Accessed at: https://www.dwds.de. (10 July 2017).

elexiko. Accessed at: http://www.elexiko.de. (10 July 2017).

LGWB-DaF: Langenscheidt Großwörterbuch Deutsch als Fremdsprache (2015). München: Langenscheidt.

OWID. Accessed at: www.owid.de. (10 July 2017).

Pons Kompaktwörterbuch Deutsch als Fremdsprache (2016). Stuttgart: Pons.

Websites:

Datenbank für Gesprochenes Deutsch. Accessed at: http://dgd.ids-mannheim.de. (10 July 2017).

DeReKo. Accessed at: http://www.ids-mannheim.de/kl/projekte/korpora/. (10 July 2017).

FOLK (Information about the corpus). Accessed at: http://agd.ids-mannheim.de/folk.shtml. (10 July 2017).

Leibniz-Gemeinschaft.
http://www.leibniz-gemeinschaft.de/en/about-us/leibniz-competition/projekte-2016/funding-line-1/. (10 July 2017).

Lexik des gesprochenen Deutsch. Information about the project: http://www.ids-mannheim.de/lexik/lexik-des-gesprochenen-deutsch.html. (10 July 2017).

Building a Collaborative Workspace for Lexicography Works in Indonesia

Totok Suhardijanto[1], Arawinda Dinakaramani[2]

[1] Faculty of Humanities, University of Indonesia, Depok, East Java, Indonesia

[2] Faculty of Computer Sciences, Depok, East Java, Indonesia

E-mail: totok.suhardijanto@ui.ac.id, arawinda.dinakaramani@ui.ac.id

Abstract

This paper presents our attempt to develop a dictionary writing system for lexicographers in Indonesia. However, it does not mean that our work is only well-fitted for Indonesian languages. We developed this system from scratch to meet the basic need of lexicographers in Indonesia who are scattered in many local cities and prefer working in a team. A system which is designed and developed to meet our own demands is more easily adjusted than other existing systems. For this reason, we decided to develop this system rather than using existing ones.

In general, like other interactive lexicon viewing and editing applications, our system also provides hyperlinks for entries, category views, dictionary reversal, search engine, and export tools. However, our system is different to some extent. It is developed in a shared workspace concept to deal with lexicographers with geographic obstacles like in Indonesia. The system also comes with a corpus tool which allows users to create their own corpus. Users can store and access their language database from different locations. The corpus tool enables users to do corpus analysis and manipulation. Some major languages, such as Malay, Javanese, and Sundanese, are provided with grammatical annotation services. So, based on language corpora, users can perform lexicographic work in collaborative environments. The system also comes with a synchonization service which allows users to share and collaborate on document files, folders, and databases with other counterparts regardless of physical location. For the time being, we are developing only the web application version, but in the future, it is possible to also expand it into desktop and mobile applications.

Keywords: collaborative workspace; corpus tool; interactive lexicon viewer and editor, lexicographic application; Indonesian languages

1. Introduction

In modern lexicography, the use of computers in each step of dictionary-making process is inevitable. According to Atkins and Rundell (2008: 112), not until the beginning of 1990's did lexicographers begin to work directly on computers. Currently, it is very common to use electronic corpora in the development of lexicographic works, for example, in dictionary making. Many leading publishers have thus taken advantage of electronic corpora. The number of tokens in corpora has also increased over the years. If the corpus which was compiled for the COBUILD project consisted of eight million words, the current corpus, namely Bank of English, for the COBUILD project contains 4.5 billion words.

The use of information technology in lexicographic work is not only limited to the use of electronic corpora in dictionary making, but also extended to the use of computer tools in compiling and writing a dictionary. Atkins and Rundell (2012) mention that there are two types of software in lexicographic work. First, a corpus query system that enables us to analyze the data in a corpus in various ways and second is a software called dictionary-writing system (DWS) that enables lexicographers to compile and edit dictionary text. These lexicographic applications include all-key factors in making a dictionary, such as data collection, data analysis, and synthesis/composition. The advantage of lexicographic applications is having lexicographers to focus more on their expertise in compiling and writing a dictionary. Not only that, it has also cut and saved a lot of time and effort in the process of dictionary making.

The presence of DWSs has also given a new hope to under-resourced languages. According to Prinsloo (2012), under-resourced languages generally experience a lack of high standard dictionaries. Actually, less-resourced languages, such as many languages in Indonesia, also face deficiencies in language description and codification, including standard grammar, spelling guidelines, etc. Although in Indonesia there is a national agency for language affairs which oversees language development and conservation at a national scale, there still remains a number of languages which are under-resourced and less-described. Indonesia is the second-most lingustically diverse country in the world, with 719 languages spoken in the country. Among these languages, 386 languages have 5000 speakers or fewer. Most of them are now facing various degrees of language endangerment (Ethnologue 2015).

This paper concerns our attempts to develop a web application providing lexicographers in Indonesia with a shared workspace. This workspace is an inter-connected environment in which all the participants in dispersed locations can work and collaborate with each other in a single entity. Our DWS, called Lexcoworks, is mainly a web-based application. It means that the Lexcoworks interface can be opened with any regular browser. Furthermore, the Lexcoworks network feature makes it well-fitted for collaborative lexicographic projects, whether in a Local Area Network or on the Internet. In the case of using the Internet, people can access the project and do a lexicographic project from anywhere. It is also important to mention here that Lexcoworks is a multiuser DWS. It means that different users can log in simultaneously and work on the same project.

With regard to the nature of Lexcoworks as a web-based application, it will be a huge advantage if most users were familiar with the Internet. In Indonesia, the number of the Internet users still exhibits significant growth. However, there still remains two matters with regard to the number of Internet users in Indonesia. According to APJII, most Internet users in Indonesia regularly access the Internet through their mobile phones. Second, in Indonesia, Internet network coverage areas are still lacking compared to neighboring countries like Singapore and Malaysia. As a result, there remains many areas of Indonesia where Internet connection is not available. As a result, to anticipate

various conditions of the Internet connection in Indonesia, Lexcoworks is equipped with a file synchronization feature to ensure that computer files in two or more locations are updated in certain rules. In this way, users can work on their project regardless of the availability of an Internet connection.

As a geographically-divided country, Indonesia has a primary geographic challenge related to the distance between its myriad islands. We develop our own system from scratch to meet the basic needs of lexicographers in Indonesia who are scattered in many local cities but prefer working in a team. To help people from remote and scattered areas to become involved and engaged in a common lexicographic project, Lexcoworks is developed as a shared workspace application.

A dictionary writing system which is designed and developed to meet our own demands is more applicable and adjustable than any of the existing systems. For this reason, we decided to develop our own system. The presence of Lexcoworks will be a great help in accelerating the process of language documentation and codification. Also, this shared workspace will help Indonesian lexicographers to conduct collaborative works and to solve their geographic obstacles.

2. Dictionary Writing Systems: An Overview

In this section, we review several existing DWS software programs. In the 1990's, the big dictionary publishers in the UK had already implemented DWSs in their dictionary projects with the aim of making dictionary compiling easier. According to Atkins and Rundell (2008: 112–114), a DWS comes with a ranging version from a simple and more elaborated program. A commercial DWS program is developed to meet with the dictionary publisher's qualification and demand. This kind of software should have the ability to manage the entire process of producing a dictionary, from compiling the first entry to outputting the final product for publication in printed or electronic media. Aside from DWS, this kind of software is also referred by other terms including 'dictionary editing system' (Svensén, 2009: 422), 'dictionary compilation software', 'lexicography software', 'dictionary production software', (De Schryver & Joffe, 2006: 41; Joffe and De Schryver, 2004: 17), 'lexicographic workbench' (Ridings, 2003: 204), 'dictionary management system' or 'lexicographer's workbench' (Langemets et al., 2010: 425), 'dictionary editing tool' (Krek 2010: 928), or 'dictionary building software' (Mangeot 2006: 185).

In general, Abel (2012: 87–88) distinguishes three main characteristics of a dictionary writing system. First is the content of the dictionary. Second is related to the structure or the grammar of the dictionary. The third aspect is the data presentation which includes formatting and style (see also De Schryver & Joffee, 2006: 41). Abel suggests these three aspects to be considered individually, but specific programs are best suited to work on each of them. Although Abel considers a DWS as an independent item, it could also be regarded as a system that takes benefit from other applications.

Basically, dictionary writing comprises mainly entry-inputting and editing that can happen in many different ways. This work can also be processed by using available word-processing systems that allow dictionary text to be processed and stored linearly, in exactly the way as it should be presented in the final product (Abel, 2012: 88). In addition, to separate the works of data-entry and data-editing, a lexical database can also be implemented, where the data are structured and stored in records, as well as separated from the emerging dictionary text. According to Svensén (2009: 421), from the point of view of the dictionary producer, such a database has the advantage of generating a great variety of products based on one and the same material. For this reason, our DWS is designed to be implemented with a database management system.

Furthermore, Abel (2012: 88) also mentioned that the use of mark-up languages also offers a significant help in dictionary writing. Mark-up languages, such as the popular XML, and editing software for them allow lexicographers to manipulate and manage documents in a structured way by adding additional information to the text in the form of tags; that is, standardized labels. Such kinds of mark-up languages are very helpful in lexicographic projects, but Abel wrote that we still need an additional tool to take benefit of the tags. Although many efficient and popular programs are available, these generic tools do not necessarily meet the needs of complex dictionary projects, because they were not specifically designed for lexicographic work (Abel, 2012: 88–89). We still need more specific tools: either it is an in-house tailor-made or off-the-shelf applications because dictionary projects are complex.

Atkins and Rundell (2008: 114) mention that a typical DWS consists of three main components: a text-editing interface, a database, and set of administrative tools. With a text-editing interface, lexicographers are able to create and edit dictionary texts. A dictionary database is required to store all the emerging dictionary text. Meanwhile, administrative tools support lexicographers to manage the project and publication process. According to Abel (2012: 95) a DWS is sometimes a written in-house system, such as an XML-editor customized for one or more dictionary project, or, in other cases, an off-the-shelf dictionary writing system package.

Not surprisingly, most software in Table 1 offer three components mentioned by Atkins and Rundell (2008), that is, a text-editing interface, a database, and administrative tool packages.

SIL has produced and launched their DWSs, namely, FLEx and Lexique Pro. These two software packages are robust lexical management systems that are suited to use in fieldwork and language documentation. In addition, Lexique Pro has a variety of tools from data entry and publication. EELex is an application that is built to manage Estonian languages. Among these DWSs, only four are entirely web-based, that is, *DEB2*, *Glossword*, *Lexonomy*, and *Mātāpuna*.

Dictionary Writing System	Description
EELex	a DWS developed at The Institute for Estonian Language (Eesti Keele Instituut) (Langemets et al. 2010)
FLEx	This software is produced by SIL International (SIL) for organizing and analyzing linguistic and cultural data. It enables linguists to be highly productive when building a lexicon and interlinearizinag texts.
TschwaneLex	TLex (aka *TshwaneLex*) is a professional, feature-rich, fully internationalised, off-the-shelf software application suite for compiling dictionaries or terminology lists.
Glossword	The software is aimed at creating online multilingual dictionaries, glossaries, references. It can mix several languages in one definition and create dictionaries written in different languages, managed by a single Glossword installation.
Lexique Pro	The software is an interactive lexicon viewer and editor, with hyperlinks between entries, category views, dictionary reversal, search, and export tools.
Lexonomy	A web-based DWS developed by Michal Boleslav Měchura which has the right balance between power and ease of use. It is designed to be a tool for writing and publishing dictionaries (and other dictionary-like datasets) where users find the right balance between power (= empowering users to do what they need to do) and ease of use (= not having a steep learning curve).
Mātāpuna	It is an open-source web-based DWS developed by Dave Moskovitz of Thinktank Consulting Limited in collaboration with the Māori Language Commission of New Zealand. (Bah 2010).

Table 1: List of selected Dictionary Writing Systems

Glossword is an open source tool written in PHP and intended for creation and publishing of an online multilingual dictionary, glossary, or reference. It means that Glossword only focuses on online dictionary writing. DEB2 is a web-based application with several features such as the server running in Linux, but clients are multiplatform. However, DEB2 appearance is still so basic that users need to improve their computer skills to deal with the applications. Lexonomy is a new DWS and still an experimental prototype, which will have some (probably not all) of the new and/or improved features described in the first three sections of this document (entry editing, dictionary configuration, publishing). Finally, Mātāpuna is developed as a web-based application

that offers not only entry editing, dictionary configuration, and publishing, but also shared workspaces for collaborative work.

In terms of the availabilty of DWS basic components, our proposed DWS is quite similar to Mātāpuna. Like Mātāpuna, our system is also entirely web-based and does have functionality for collaborative work. However, due to the Internet network fluctuation, our system also implements file synchronization to anticipate it. This feature will prevent lexicographers from feeling frustrated when the Internet connection is lost, as is still the case in developing countries like Indonesia. In our system, we use character encoding UTF-8 that is capable of encoding all possible characters to accommodate any dictionary project in a different language. In the future, we are planning to extend the functionalities of our DWS by integrating a corpus query manager into our system.

3. Architecture and Functionality

In this section, we explain Lexcoworks architecture and functionality.

3.1 Design and architecture

With regard to users, this application design is divided into two main types: administrator and non-administrator (uncategorized user). Uncategorized user refers to a user who does not yet have any role in a lexicographic project.

In addition to the two main types of user, there are three additional users including chief editor, editor, and contributor. These three user types represent user roles in lexicographic works including dictionary, thesaurus, and glossary. An uncategorized user can be either a chief editor, editor, or contributor in more than one lexicographic project. Users can have different roles in different lexicographic projects. For instance, a given user can be an editor in a Javanese dictionary project and at the same time s/he plays the role of contributor in a Madurese dictionary project.

An uncategorized user automatically becomes a chief editor whenever s/he starts a lexicographic project. Once an uncategorized user has created a lexicographic project through the create feature, s/he has the authority to use all features related to the project. Meanwhile, an uncategorized user can also play a role of an editor or contributor in a given language dictionary project, when a chief editor assigns him/her to the project. An editor is considered to be part of the core members of the team in a lexicographic project, so that s/he has also an authority to use features for dictionary building, but in a more limited way. On the contrary, although a contributor is also assigned by a chief editor to a lexicographic project, s/he does not belong to the core members of the team. For this reason, a contributor can only access and use very limited features that are related to a lexicographic project. An uncategorized user can be invited and promoted by a chief editor to become a contributor. At the same time, users can also submit an application to join in a dictionary project as contributors.

3.2 Functionality

Most features of Lexcoworks are available to users who are logged in. Users who are not logged in can use the search feature and view lexicographic works that have been published online on Lexcoworks. Users can sign up to Lexcoworks to get an account. Once a user signs up, s/he can log in to Lexcoworks as an uncategorized user. Users who are logged in can start or create a lexicographic project through the Create feature, join an existing lexicographic project, and manage their lexicographic projects in addition to searching and viewing lexicographic work features. Users who are logged in to Lexcoworks have more options available to them in the search feature, and can view lexicographic works that are not published online yet, long as they have the role of a chief editor, editor, or contributor of the lexicographic work.

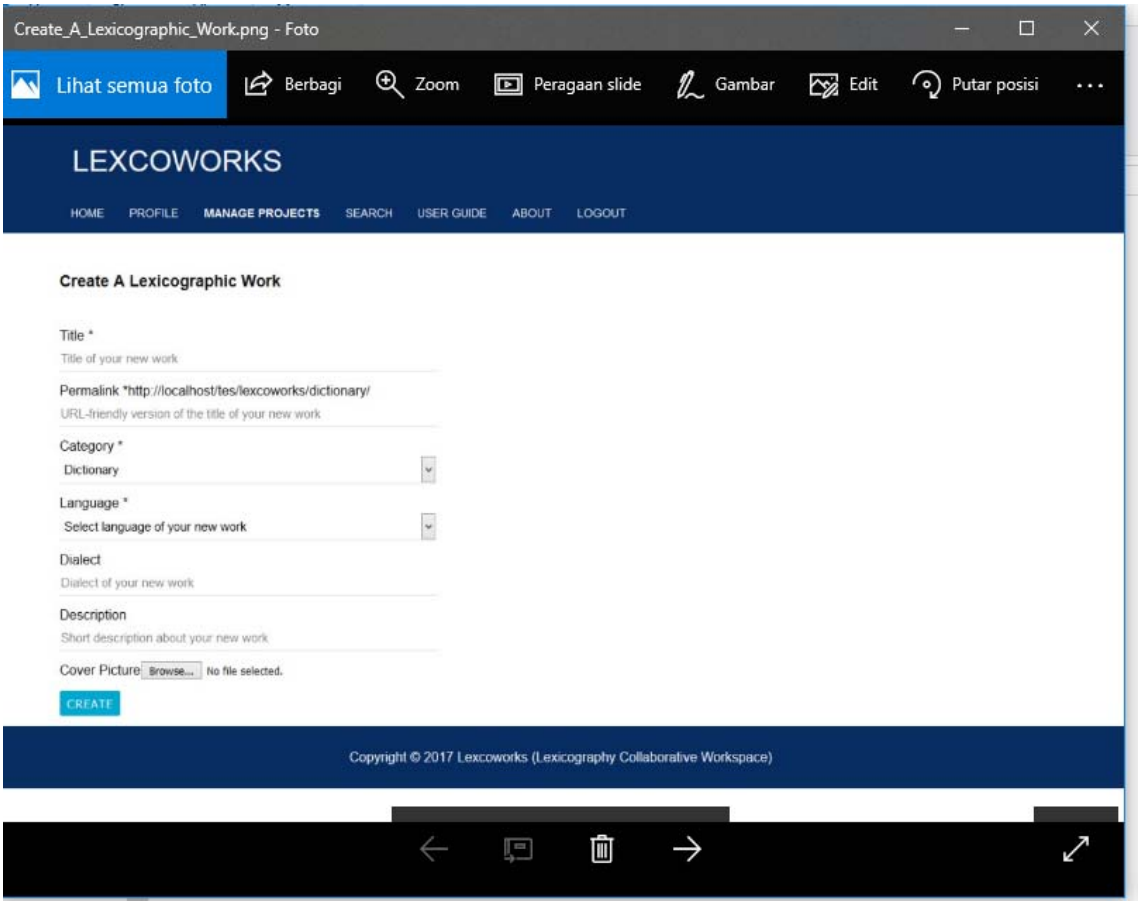


Figure 1: Screenshot of Create Menu in LEXCOWORKS

The Create feature allows users to start or create a lexicographic project. Users can provide basic information about the project including title, category (whether it is a dictionary, thesaurus, or glossary), language, and cover picture. Once a lexicographic project is created, Lexcoworks will provide users with two pages consisting of an online page and an editing page of the project. The online page displays basic information about the project, lexicographic entries, a list of the team members of the project, a link to access the editing page (only available to the chief editor, editors, and

contributors of the project), and a feature that allows users to submit a request to join the project as one of the contributors (available to uncategorized users who are logged in to Lexcoworks). If a lexicographic work has been published online, everyone can search and view its online page. If a lexicographic work has not been published online, however, only the members of the project can search and view its online page.

The editing page is a workspace for specific users to compile, edit, and add information with regard to the lexicographic project. A user can access editing features on the editing page. The availability of a feature depends on the user's role in the project. For instance, the feature to delete the project is only available to the chief editor; features to confirm entry change suggestions, publish a lexicographic work online, and print a lexicographic work are available only to the chief editor and editor; and features to edit an entry are available to all team members of the project. Lexcoworks allows users to add and edit entries online and offline. Users can add and edit entries online by filling and submitting a form directly on the editing page. For the web application version of Lexcoworks, if a user wants to add and edit entry offline or it is not possible to add and edit entry online (such as due to bad Internet connection or limited electricity access), a user can dowload a template file (tsv format) for offline editing. Users can add and edit an entry in a template file without Internet connection. Once users are connected to the Internet, they can upload the file to Lexcoworks. In addition to editing features, users can also access features to view the activity log on the editing page. The chief editor and editor can view the log of all activities in the project, such as which user edited that entry at what time, while a contributor can view only a log of his/her own activities.

Lexicographic entries of a lexicographic work in Lexcoworks could be divided into three different entities, that is entry, lemma, and sense. The entry entity represents a lexicographic entry. A lexicographic entry is assumed to consist of lemma (head word) and sense (meaning). Therefore, the entry entity includes information about identity number of the lemma entity, and identity number of the sense entity. The entry entity also includes information about the entry status, i.e. whether it is published online or not. If a lexicographic entry of a lexicographic work is published online, it is displayed on the online page of the work and can be searched by everyone. A lexicographic entry that is not published online can only be viewed on the editing page and can only be searched by members of the project.

The lemma entity represents a headword. It includes information about lemma name, lemma name with hyphenation point, pronunciation, word type, morphological structure, and homonym number. Information about word type is to indicate whether a lemma is a base or derivative. If a lemma is derivative, lemma entity also includes information about the identity number of its base. Information about morphological structure is to indicate morphological process, such as reduplication or affixation. Users can choose a morphological structure from a default set of morphological structures provided by Lexcoworks, or suggest a new morphological structure that is not included in the default set.

The sense entity represents a unit of meaning. It includes information about part of speech (such as verb, noun, and adjective), register (such as slang and formal), field (such as Chemistry and Biology), definition, example, and polysemy number. Users can choose from a default set of part of speech provided by Lexcoworks, or can suggest a new part of speech that is not included in the default set yet. Users can do the same for register and field labels; that is, choosing from default sets or suggesting a new register or field label. A sense entity is connected to a lemma entity by an entry entity. A lemma entity can be connected to more than one sense entity.

An entry entity belongs to an entity that represents a lexicographic work, namely opus entity. Therefore, the entry entity also includes information about identity number of the opus entity. The opus entity includes information about title, category (to indicate whether it is a dictionary, thesaurus, or glossary), language, cover picture, short description about the lexicographic work, identity number of the user who created the lexicographic work, and the time when the lexicographic work was created. Users can choose a language from a default set of languages provided by Lexcoworks or can suggest a new language. Languages that are included in the default set provided by Lexcoworks are referring to a list of languages from SIL.

For the time being, we have focussed on developing features for dictionary creation; therefore, the option to create a thesaurus and glossary is not available yet, and we only developed the web application version. The web application version of Lexcoworks was developed using custom PHP framework that we created, JavaScript for some functions, CSS for web design, and MySQL for database. We created custom PHP framework for Lexcoworks using model-view-controller (MVC) architectural pattern and UTF-8 character encoding. The web application version has been developed on localhost using XAMPP for Windows. The web application version will be hosted on cPanel shared web hosting. In the future, it is possible to expand Lexcoworks into a desktop application version and mobile application version.

4. Conclusion

The use of DWS is inevitable in lexicographic works. DWSs have become applications that include a range of components and modules with a great number of functions to deal with the complexity of dictionary making. Most DWSs offer three components including data-entry and editing, a database, and a set of administrative tools for publication. Our system, Lexcoworks, has all of these components including data entry interface, lexical database, and administrative tools.

We developed Lexcoworks rather than using existing systems to meet the basic needs of Indonesian lexicographers who wanted to work collaboratively from scattered remote areas. For this reason, we developed Lexcoworks as a web-based application, so it can be accessed through any web browser. The mission of Lexcoworks is to provide lexicographers with a shared collaborative workspace. The system allows users to work online and offline. If users want to add and edit entries offline or it is not possible to

add and edit entries online (e.g. due to bad Internet connection or limited electricity access), users can dowload a template file (tsv format) for offline editing. Users can add and edit entries in a template file without Internet connection. Lexcoworks is designed to support multilingual dictionary projects, thus it uses character encoding UTF-8 to accomodate all possible characters.

With all these features, Lexcoworks can be a great help for lexicographers in multilingual Indonesia. Its users can work on lexicographic tasks anytime and anywhere, online as well as offline. They can do collaborate work in a more friendly and convenient environment because our system is regularly adjusted and revised to meet the Indonesian lexicographers' needs. In the future, we are planning to extend the functionalities of our DWS by integrating a corpus query manager into our system.

5. References

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

Bah, Oumar. (2010). Matapuna Dictionary Writing System: from Thinktank Consulting Limited. *Language Documentation & Conservation*, Vol. 4 (2010), pp. 169-176. http://nflrc.hawaii.edu/ldc/, http://hdl.handle.net/10125/4477

Kilgariff, A. (2006). Word from the Chair: In G.-M. De Schryver (ed.). DWS 2006: *Proceeding of the Fourth Internasional Workshop on Dictionary Writing System 7.* Pretoria: $(SF)^2$ Press.

De Schryver, G. M., & Joffe, D. (2006). The users and uses of TshwaneLex One. In *4th International Workshop on Dictionary Writing Systems* (DWS-2006) (SF) 2 Press, pp. 41-46.

Abel, A. (2012). Dictionary writing systems and beyond. In S. Granger & M. Paquot (eds.). *Electronic Lexicography.* Oxford: Oxford University Press, pp. 83-106.

Svensén, B. (2009). A Handbook of Lexicography: The Theory and Practice of Dictionary-Making. Cambridge: Cambridge University Press.

Langemets, M., Loopman, A., & Viks, U. (2010). Dictionary management system for bilingual dictionaries. . In S. Granger & M. Paquot (eds.) *eLexicography in the 21st Century: New Challenges, New Applications.* Louvain-la-Neuve: Presses universitaires de Louvain.

Automated Identification of Domain Preferences of Collocations

Jelena Kallas[1], Vit Suchomel[2], Maria Khokhlova[3]

[1] Institute of the Estonian Language, Estonia

[2] Masaryk University, Czech Republic

[3] St. Petersburg State University, Russia

E-mail: jelena.kallas@eki.ee, xsuchom2@fi.muni.cz, m.khokhlova@spbu.ru

Abstract

This paper addresses (semi-)automatic collocations dictionary compilation in connection with the automated identification of domain preferences of collocations. The research was motivated by the process of the semi-automatic compilation of the Estonian Collocations Dictionary (ECD), where lexicographers processed a large number of terminological collocations extracted from Sketch Engine into the Dictionary Writing System EELex.

In this paper, we apply the terminology extraction module within the Corpus Query System Sketch Engine and present the results of the experiments on building military domain corpora in Russian and Estonian and extracting multiword terms. Both languages have very rich morphology and quite a large number of multiword terms, but Russian texts are well represented on the Web while Estonian ones are not. We analyze how the comparison of frequency of a collocation in a reference corpus with its frequency in a domain corpus can be used for facilitating word sketch data analysis in terms of identification of domain preference of collocations.

Keywords: collocation; multiword terms; terminological collocation; Russian; Estonian

1. Introduction

Building terminological lexicons and glossaries is a prominent task in many areas: from translators to large companies aiming to establish consistent naming in their documentation. Also for lexicographers it is quite tricky to extract terminology from texts and label it properly. As Atkins and Rundell (2008: 227) point out, domain labels play an important role in lexical databases. "A domain label indicates that *the item is used when the subject of discussion is …* (*science, hockey, plumbing, poetry etc.*)".

Traditionally, domain labels are assigned in dictionaries to word senses. However, it is also quite a common practice in collocations dictionaries. For example, the *Oxford Collocations Dictionary for Students of English* (OCDSE, 2002) presents domain specific collocations as "technical collocations" and defines them as "collocations that are used by people who specialize in a particular subject area". Altogether, eight different subject areas are distinguished (business, computing, law, mathematics, medical, military, science and sport). In addition to these labels, more specific usage restriction, such as 'in football' or 'used in journalism', are given in brackets.

As for automated collocations dictionaries, no domain labels have been provided so far. An example of an automated collocation dictionary entry is shown in Figure 1, illustrating the lexeme "operation" in the Sketch Engine for Language Learning (SkELL) system (Baisa & Suchomel, 2014).

Figure 1: An example of a word sketch for "operation" in SkELL

Among collocates, there are quite a few examples of units that belong to certain domains.[1] However, there are no labels that help learners to identify whether a particular collocation is a terminological one or not.

The same problem is significant for semi-automated compilation of collocation dictionaries. A recent survey (Tiberius et al., 2015; Gantar et al., 2016) showed that acquiring lemma lists and frequency information from corpora is a common procedure, followed by the extraction of example sentences, grammatical patterns, multiword expressions, form variations and neologisms. Less frequent are automated procedures related to semantics: word senses, lexical semantic relations, definitions and knowledge-rich contexts. Authors (Gantar et al., 2016: 211) point out that when analyzing word sketch data, lexicographers still spend a significant amount of time selecting the relevant collocates and their examples under each syntactic model.

One analytical lexicographic task that is also still performed manually is the identification of terminological collocations and making decisions about whether to exclude them from the database as not relevant or to add domain labels. This process is discussed in greater detail in Section 2. This task would be made less time-consuming with the development of new approaches within corpus tools. It should be possible to automatically identify collocations that are very frequent in particular domain corpora and provide this information to lexicographers.

This idea is not a new one and it is discussed, for example, in Rundell and Kilgarriff (2001) and Rundell (2012). "Essentially it involves comparing a word's profile in a

[1] See e.g. *military operation*, which is registered as a term in the terminology database IATE. Accessed at: www.iate.europa.eu (20 May 2017)

carefully-defined sub-corpus with its behaviour in the lexicographic corpus as a whole, in order to retrieve information about its stylistic, regional, or domain preferences" (Rundell, 2012: 28).

Figure 2 illustrates how register preference can be shown as additional information in word sketch (Kilgarriff et al., 2004) data analysis. In order to achieve it there are two subcorpora (written and spoken) compared simultaneously. The label in the upper right corner, "usually in spoken (69.9%, percentile 0.4)", indicates that this particular word is used mostly in the spoken corpus.

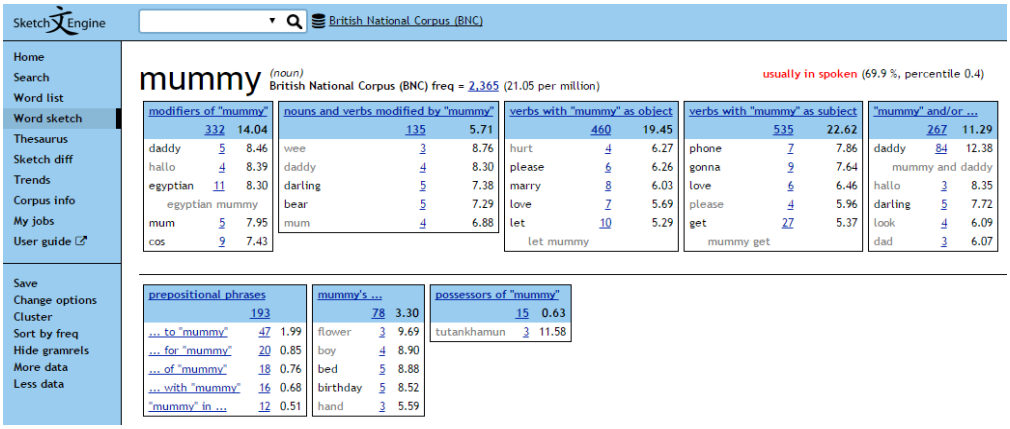


Figure 2: An example of a word sketch for "mummy" in British National Corpus, with register preference information "usually in spoken" (indicated on the right side)

Similarly, the usage of domain corpora should make it possible to apply additional filters for collocation extraction and thus to identify domain preferences of particular collocations.

In this paper, we differentiate between notions of a terminological collocation and a multiword term. For a multiword term definition, we follow the approach of Ramisch (2009). A multiword term is a term that is composed of more than one word. The unambiguous semantics of a multiword term depends on the knowledge area of the concept it describes and cannot be inferred directly from its parts (SanJuan et al., 2005; Frantzi et al., 2000). In terms of terminological collocations, we follow the conception proposed in Costa and Silva (2004). A terminological collocation can be defined as a unit consisting of a term and its collocate. For example, *баллистическая ракета* 'ballistic missile' can be viewed as a multiterm, whereas *запустить баллистическую ракету* 'to launch a ballistic missile' is a terminological collocation (however, to a certain degree the given collocation acquires the terminological status). Thus the whole item is a non-term "considering that its whole generally does not refer to a concept" (ibid). Nevertheless such terminological collocations should be presented in dictionaries with special domain labels.

2. Manual Identification of Terminological Collocations in the Estonian Collocation Dictionary Database

The Estonian Collocations Dictionary is a monolingual online scholarly dictionary aimed at learners of Estonian as a foreign or second language at the upper intermediate and advanced levels. The dictionary contains about 10,000 headwords, including single and multiword lexical items. For the automatic generation of the ECD database, the corpus query system Sketch Engine (Kilgarriff et al., 2004) functions Word List, Word Sketch and Good Dictionary Example (GDEX) were used. The main parameters used for the extraction of collocates were 1) the minimal frequency of a collocate: 10 (for the frequency I class) and five (for the frequency II class), 2) the minimal salience of a collocate: positive Dice, 3) the minimum frequency of the grammatical relation: 10, and 4) the minimum salience of the grammatical relation: positive Dice. We extracted collocates in a fixed order according to grammatical relations and ranked them by frequency (Kallas et al., 2015).

Currently, the database is being examined, edited and supplemented by lexicographers. One of the significant observations regarding editing collocations is that deleting is necessary mainly in the case of mistakes in tagging and due to insufficient disambiguation, but also in the case of specific terms that are not part of general purpose everyday Estonian. The analysis of extracted data revealed a significant number of terminological collocations that belong to different domains. The most frequent are the law, medical, mathematical, scientific, linguistic and sports domains.

Figure 3 illustrates how collocates are presented in the dictionary database. In the dictionary entry preview for the adjective *eitav* 'negative' there are three collocates that were automatically extracted and later (during the editing process) were manually identified as domain-specific collocations. These collocations are *eitav kõne* 'negative', eitav *kõneliik* 'negative' and *eitav lause* 'negative sentence'. The domain label is KEEL 'linguistics'.

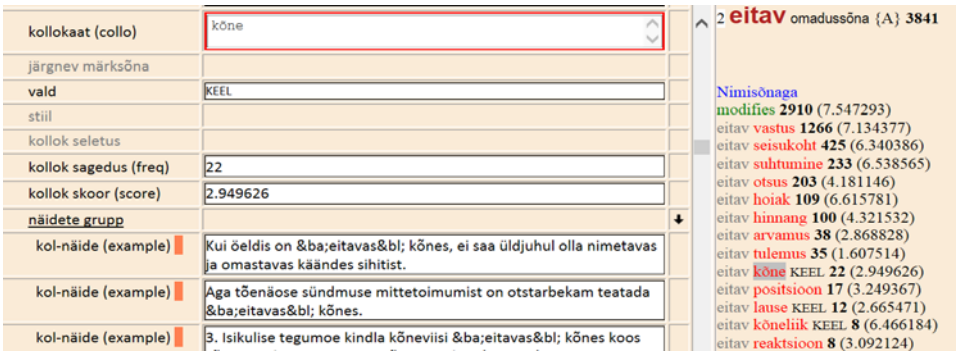


Figure 3: An example of an entry for the adjective *eitav* 'negative' in DWS EELex: the editing window in XML view (left) and the dictionary entry preview (right)

In order to identify such collocations, different approaches are used: 1) consulting terminological dictionaries and databases, 2) analyzing available domain corpora, and 3) building new domain corpora within Sketch Engine with WebBootCaT (Baroni et al., 2006) and implementing the Term Extraction function (Kilgarriff et al., 2014; Fiser et al., 2016). The latter takes a lot of effort on the part of the lexicographer.

The automation of this task would have a major impact on lexicographic word sketch data analysis and (semi-)automated collocation dictionary compilation.

3. Multiword Term Extraction within Sketch Engine: State of the Art

In this section, we present the results of our experimental study on the reliability of the data that can be identified and extracted using methods that were developed within the Sketch Engine corpus query system, particularly the tools WebBootCaT (Baroni et al., 2006) and Term Extraction (Kilgarriff et al., 2014; Fišer et al., 2016). Term Extraction is based on comparing frequencies of pre-defined units in a domain corpus and a general corpus. The resulting term candidates are sorted by the ratio of the frequencies (the keyword score).

For the experiment, Russian and Estonian were used. Russian is highly represented on the Web (estimated percentage is 6.5%) while Estonian is not (estimated percentage is 0.1%).[2]

3.1 Term Grammar and Domain Corpora

Sketch Engine implements a data-driven approach to this problem: instead of having domain experts build such a lexicon from scratch using an automatic procedure that produces a high quality lexicon from the supplied domain-specific corpus. The whole procedure is described in detail in (Kilgarriff et al., 2014). Term candidates for a language domain can be found through the following steps:

- taking a corpus for the domain, and a reference corpus for the language;
- identifying the grammatical shape of a term in the language and writing a term grammar[3];
- tokenizing, lemmatizing and POS-tagging both corpora;
- identifying (and counting) the items in each corpus which match the grammatical pattern;

[2] Accessed at: https://w3techs.com/technologies/overview/content_language/all (20 May 2017)

[3] Term Grammar: Writing term grammar. Accessed at: https://www.sketchengine.co.uk/documentation/writing-term-grammar/ (25 May 2017)

- for each item in the domain corpus, comparing its frequency with its frequency in the reference corpus.

The term identification is based on CQL—Corpus Query Language—to specify the term grammar for each language. The term grammar formalism can be defined as regular expressions over words, lemmas and morphological tags (imposing a requirement that the corpora be tagged). The format of the term grammar corresponds to the word sketch grammar and hence makes it possible to use the same indexing machinery for efficient storage and retrieval of the term candidates.

Altogether there are term definitions for 13 languages in Sketch Engine, Russian and Estonian among them. However, to the best of our knowledge, there are not many works dealing with the evaluation of these term grammars. The results of the evaluation presented in Fišer et al. (2016) were applied to the Slovene language. Adjective + noun combinations achieve 73% accuracy, whereas trigrams with prepositions have 63% accuracy.

The term grammars for Russian and Estonian were built on the assumption that terms are mostly noun phrases. This assumption is based on academic descriptions of term structures in Russian (Gerd, 1986) and Estonian (Erelt, 2007), and partly on the empirical observation of the terms structure in terminological databases (e.g., in the NATO English–Russian terminology lexicon[4], out of 300 randomly chosen terms only two were verb phrases).

The Russian term definition consists of the following lexico-grammatical patterns (Khokhlova, 2009): 1) adjective + noun, 2) adjective + adjective + noun, 3) noun + noun, 4) noun + adjective, and 5) adjective + noun + noun. For Estonian, the patterns are: 1) adjective + noun, 2) noun + noun, and 3) noun + verb. Each model involves several restrictions on the grammatical forms of words.

For Russian, the terms are built on lemmas instead of word forms so that all of the flective variants contribute to the one lemmatized item.

For Estonian, colloc-type rules were used in order to extract multiword term candidates so that one component was presented as a lemma and the other one in the particular inflectional form, e.g. *sõjaväe konvoi* (the military-SG-GEN convoy-SG-NOM) 'military convoy'.

In our experiment, as reference corpora we used large web corpora gathered using SpiderLing (Suchomel & Pomikalek, 2012). For Russian, this was Russian Web 2011 (ruTenTen11) and for Estonian Web 2013 (etTenTen13).[5]

[4] NATO database: http://www.nato.int/docu/glossary/eng/15-main.pdf (20 May 2017)

[5] Both corpora are available at https://the.sketchengine.co.uk/auth/corpora/ (20 May 2017)

Domain corpora were built by WebBootCaT (Baroni et al., 2006), a tool for gathering domain specific documents from the web. As a domain corpus, we built a military corpus due to the good quality of military lexicons that can be used both for compiling such corpora and for evaluating term extraction. For Russian we used the NATO English–Russian terminology lexicon and for Estonian the database MILITERM[6].

We used 145 monolexemic and multiword terms from the NATO list as seed words for the Russian military domain corpus. For example, *баллистическая ракета* 'ballistic missile', and *автоматическая система управления войсками* 'automated command and control system'. The resulting size of the corpus was 25 million words.

We used 1500 monolexemic and multiword terms from MILITERM as seed words to build the Estonian domain corpus. For example, *õhusõidukite liikumise miinimumala* 'minimum aircraft operating surface' and *radarihävitaja* 'wild weasel'. The resulting size of the corpus was only three million words. The reason for using a much higher count of seed terms compared to Russian was to get as many relevant texts from the web as possible. However, the resulting corpus was not big enough, as is shown in the evaluation.

To select the most relevant terms out of the term candidates set (with regard to the target domain), we compared their frequencies using the SimpleMaths method[7] and computed a score for each term.

3.2 Evaluation and Discussion

We compared the extracted terms with the original terminology database and evaluated the recall of the whole WebBootCaT and Terminology extraction method.

The full terminological database was used for the evaluation. Since the seed words were a part of the full set they naturally occurred in the result domain corpus. The benefit of creating the domain corpus is that it also contains terms which were not used as seed phrases.

The evaluation showed that the task was a precision/recall tradeoff, as can be seen in Figures 4 and 5. Taking more candidates into account, the precision dropped while the recall grew. There were enough Russian web documents in the target domain found and downloaded to cover 50% of the single word terms and 25% of the multiword terms in the top 3,000 term candidates. Thanks to the size and the satisfactory representation of the target domain, the corpus can be used by

[6] MILITERM database: http://termin.eki.ee/militerm/ (20 May 2017)

[7] https://www.sketchengine.co.uk/documentation/simple-maths/ (20 May 2017)

lexicographers to study collocations of words from the domain. The same does not hold true for the Estonian corpus: it is too small and the target domain is poorly covered.

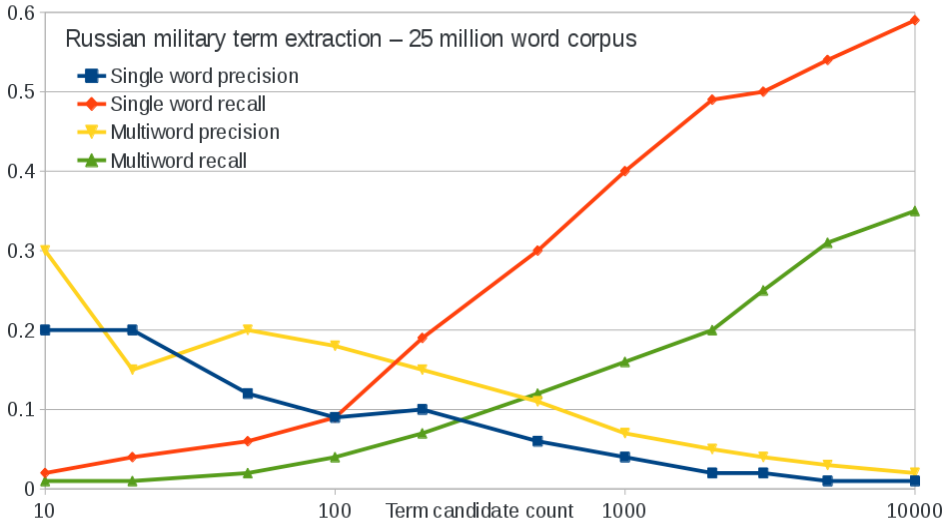


Figure 4: Evaluation of the top term candidates (with the highest keyword score) extracted from the Russian military domain corpus

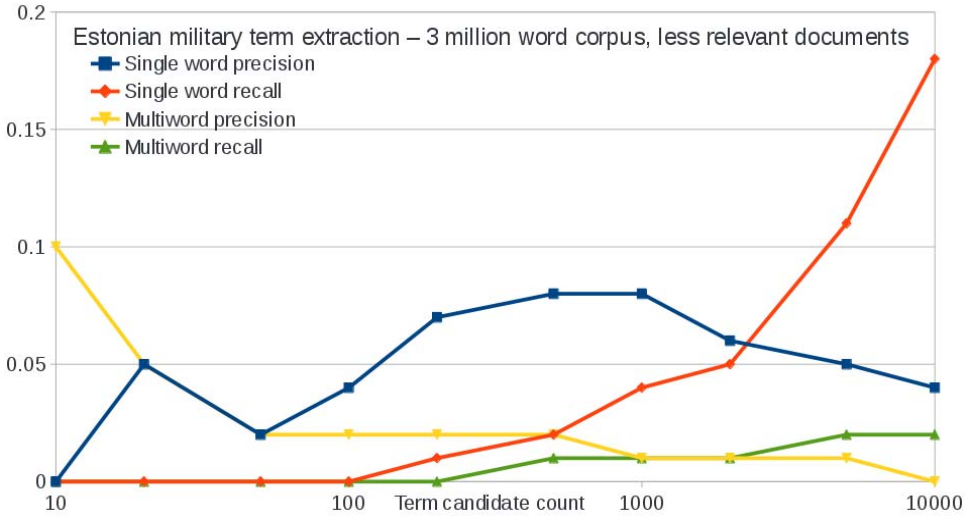


Figure 5: Evaluation of the top term candidates extracted from the Estonian military domain corpus

The most common reasons leading to a wrong classification in both languages were as follows:

- a term pattern not covered by the term grammar (e.g., more than five word terms or terms not consisting of noun phrases);
- a general noun phrase but not a term;

- a word or a phrase in the domain but not a good term;
- a part of a multiword term;
- valid terms from a different domain (e.g., politics rather than military in Estonian).

The experiment showed that this method works well only for languages that are highly represented on the Web and is insufficient for languages whose estimated percentages of the top 10 million websites is 0.1%. The result depends greatly on the size and quality of the domain corpus. The problem is that for languages with a small presence on the Web, the search engine cannot find enough documents in the domain. The minimum size for the domain corpus should be five or 10 million words.

4. Identification of Domain Preferences of Collocations in Word Sketches

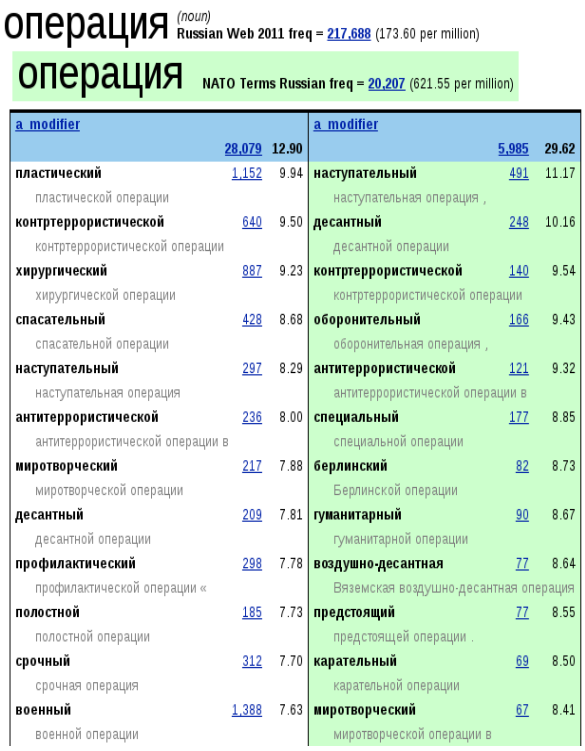
In this section, we propose two possibilities for identification of domain preferences of collocations: 1) comparing frequency in a reference and a domain corpus to identify domain preferences of a headword and its collocates, and 2) comparing word sketches of reference and domain corpora (as an example see Figure 6).

The first approach requires domain corpora to compare frequencies of collocations in a domain and the focus corpus and display domain preferences of headwords and collocations in a way similar to the indication of register preference in Figure 2. In general, any document attribute that is relevant for lexicography could be used to define a subcorpus of the focus corpus. If a collocation was mainly found in a single subcorpus based on the selected document attributes, it would be labelled by the corresponding text type in the word sketch interface. For example, taking advantage of language variety, genre and topic subcorpora, word ᶜlamerᵊ[8] could be labelled ᶜUsually American English, Internet forum, Computersᵊ which consitutes valuable information for a lexicographer.

The second approach suggests that another possible way to analyze the domain preference of collocations is to implement the procedure used in Bilingual Word Sketch function[9] (Kovář, Baisa & Jakubíček, 2016). Figure 6 illustrates the sketch for the word *операция* ᶜoperationᵊ, where adjectival collocates from a reference corpus and from a domain corpus are presented.

[8] https://en.oxforddictionaries.com/definition/us/lamer (10 July 2017)

[9] https://www.sketchengine.co.uk/user-guide/user-manual/bilingual-word-sketch/ (20 May 2017)



операция *(noun)* Russian Web 2011 freq = 217,688 (173.60 per million)

операция NATO Terms Russian freq = 20,207 (621.55 per million)

a_modifier			a_modifier		
	28,079	12.90		5,985	29.62
пластический	1,152	9.94	**наступательный**	491	11.17
пластической операции			наступательная операция ,		
контртеррористической	640	9.50	**десантный**	248	10.16
контртеррористической операции			десантной операции		
хирургический	887	9.23	**контртеррористической**	140	9.54
хирургической операции			контртеррористической операции		
спасательный	428	8.68	**оборонительный**	166	9.43
спасательной операции			оборонительная операция ,		
наступательный	297	8.29	**антитеррористической**	121	9.32
наступательная операция			антитеррористической операции в		
антитеррористической	236	8.00	**специальный**	177	8.85
антитеррористической операции в			специальной операции		
миротворческий	217	7.88	**берлинский**	82	8.73
миротворческой операции			Берлинской операции		
десантный	209	7.81	**гуманитарный**	90	8.67
десантной операции			гуманитарной операции		
профилактический	298	7.78	**воздушно-десантная**	77	8.64
профилактической операции «			Вяземская воздушно-десантная операция		
полостной	185	7.73	**предстоящий**	77	8.55
полостной операции			предстоящей операции .		
срочный	312	7.70	**карательный**	69	8.50
срочная операция			карательной операции		
военный	1,388	7.63	**миротворческий**	67	8.41
военной операции			миротворческой операции в		

Figure 6: Word sketch for the noun операция ‘operation’ with aligned grammatical relations in the Russian Web 2011 corpus and the NATO Terms Russian domain corpus

The first three collocates in the reference corpora are *пластический* ‘plastic surgery’, *контртеррористический* ‘counterterrorist (operation)’, and *хирургический* ‘surgical (operation)’. The most frequent collocates in the domain corpora are *наступательный* ‘offensive (operation)’, *десантный* ‘amphibious (operation)’, and *контртеррористический* ‘counterterrorist (operation)’. This helps to separate collocations and the word sense associated to a single topic represented by the military domain corpus.

5. Conclusion and Future Work

The results of our experiment revealed that for languages that are highly represented on the Web it is possible to create sizable domain corpora. We propose to exploit the domain corpora for automatic comparison of frequencies of collocations in a domain and a reference corpus to help lexicographers by indicating domain preferences of words and their collocates.

Our study can be implemented to improve the efficiency of word sketch data analysis and it is important to stress that the procedure itself is not language-specific, but depends on how highly a language is represented on the Web. The components required include a reference corpus, a number of different domain corpora (a minimum of 5 to 10 million words), a Sketch Grammar and a Term Grammar.

We suggest possible methodological improvements for corpus tools in order to improve automatic and semi-automatic collocations dictionary compilation by automatic indication of domain preferences. Domain preference provides useful information to users and allows to distinguish terminological collocations.

6. References

Atkins, S.B.T. & Rundell, M. (2008). *The Oxford guide to practical lexicography.* Oxford University Press.

Costa, R. & Silva, R. (2004). The Verb in the Terminological Collocations Contribution to the Development of a Morphological Analyser MorphoComp. *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004, May 26-28, 2004, Lisbon, Portugal.* European Language Resources Association.

Erelt, T. (2007) *Terminiõpetus.* Tartu: Tartu Ülikooli kirjastus.

Fišer, D., Suchomel, V., & Jakubíček, M. (2016). Terminology Extraction for Academic Slovene Using Sketch Engine. *Tenth Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2016.* Brno: Tribun EU, pp. 135–141.

Frantzi, K., Ananiadou, S., & Mima, H. (2000). Automatic recognition of multi-word terms: the C-value/NC-value method. *International Journal on Digital Libraries*, 3(2), pp. 115–130.

Gerd, A. (1986) *Osnovy naučno-texničeskoj leksikografii.* Leningrad: izd-vo LGU.

IATE: The EU's multilingual term base. Accessed at: http://iate.europa.eu. (25 May 2017)

Kallas, J., Kilgarriff, A., Koppel, K., Kudritski, E., Langemets, M., Michelfeit, J., Tuulik, M., & Viks, Ü. (2015). Automatic generation of the Estonian Collocations Dictionary database. *Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference, 11-13 August 2015, Herstmonceux Castle, United Kingdom.* Ljubljana/Brighton: Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd, pp. 1—20.

Kilgarriff, A., Jakubíček, M., Kovář, V., Rychlý, P. & Suchomel, V. (2014). Finding Terms in Corpora for Many Languages with the Sketch Engine. *Proceedings of the Demonstrations at the 14th Conference the European Chapter of the Association for Computational Linguistics.* Sweden, pp. 53–56

Kilgarriff, A., Rychly, P., Smrz, P., & Tugwell, D. (2004). The sketch engine. *Proceedings EURALEX 2004*, Lorient, France, pp. 105–116.

Khokhlova, M. (2009). Applying Word Sketches to Russian. *Proceedings of Raslan 2009. Recent Advances in Slavonic Natural Language Processing.* Brno: Masaryk University, pp. 91-–99.

Kovář, V., Baisa, V. & Jakubíček, M. (2016). Sketch Engine for Bilingual lexicography. *International Journal of Lexocography*, 29(3), pp. 339–352.

OCDSE: Oxford collocations dictionary for students of English. (2002). Oxford: Oxford University Press.

Ramisch, C. (2009). *Multi-word terminology extraction for domain-specific documents.* Master's thesis, École Nationale Supérieure d'Informatique et de Mathématiques Appliquées, Grenoble, France. Accessed at: http://www.inf.ufrgs.br/~ceramisch/download_files/publications/2009/p01.pdf (25 May 2017)

Rundell, M. (2012). The road to automated lexicography: an editor's viewpoint. In S. Granger & M. Paquot (eds) *Electronic Lexicography.* Oxford: Oxford University Press, pp. 15-30.

Rundell, M. & Kilgarriff, A. (2011). *Automating the creation of dictionaries: where will it all end?* In F. Meunier, S. De Cock, G. Gilquin & M. Paquot (eds) *A Taste for Corpora. A tribute to Professor Sylviane Granger.* Benjamins. P., pp. 257-281.

Vainik, E. (1999). Millest on tehtud õigusterminid? *Õiguskeel*, pp. 27—39.

Sanjuan, E., Dowdall, J., Ibekwe-SanJuan, F., & Rinaldi, F. (2005) A symbolic approach to automatic multiword term structuring. *Computer Speech & Language Special Issue on Multiword Expressions*, 19(4), pp. 524–542.

SkELL: Sketch Engine for Language Learning. Accessed at: http://skell.sketchengine.co.uk/run.cgi/skell. (25 May 2017)

Svensen, B. (2009). *A handbook of lexicography. The theory and practice of dictionary-making.* Cambridge: Cambridge University Press.

EcoLexiCAT: a Terminology-enhanced Translation Tool for Texts on the Environment

Pilar León-Araúz, Arianne Reimerink, Pamela Faber

Department of Translation and Interpreting, University of Granada, C/ Buensuceso, 11, 18002 Granada, Spain

E-mail: pleon@ugr.es, arianne@ugr.es, pfaber@ugr.es

Abstract

Although machine translation and computer assisted translation (CAT) are now a reality in the workflow of professional translators, terminology management is still considered complex and time-consuming and is often not seamlessly integrated into the translation process. Most terminographic resources are not designed to take into account the real search behavior of end users such as translators (Tudhope et al., 2006), and in many cases CAT tools do not provide terminological modules that go beyond a simple glossary with interlinguistic equivalents. Furthermore, corpus consultation is rarely possible in most CAT tools, despite the fact that the phraseological information extracted from a corpus is of great help for translators. To address these issues, we created a web-based tool for the terminology-enhanced translation of specialized environmental texts for the language combination English-Spanish-English. EcoLexiCAT uses the open source version of the web-based CAT tool MateCat and enriches a source text with information from: (i) EcoLexicon, a multimodal and multilingual terminological knowledge base on the environment (Faber et al., 2014; Faber et al., 2016); (ii) BabelNet, an automatically constructed multilingual encyclopedic dictionary and semantic network (Navigli & Ponzetto, 2012); (iii) and Sketch Engine, the well-known corpus query system (Kilgarriff et al., 2004).

Keywords: computer assisted translation; terminology management; specialized translation

1. Introduction

In today's world, machine translation (MT) and computer-assisted translation (CAT) are a consolidated part of the professional translation workflow. Nevertheless, terminology management is still considered complex and time-consuming and is often not seamlessly integrated into the translation process. Furthermore, most terminological tools do not take into account the real search behavior of end users such as translators (Tudhope et al., 2006; Durán Muñoz, 2012: 78) and most terminological modules in CAT tools do not go beyond a simple list of equivalences. Apart from that, access to corpora is generally not provided in most CAT tools, despite the valuable phraseological information that a corpus can provide. An exception to this is the recently added Sketch Engine plug-in (available from the SDL AppStore) in SDL Trados Studio but, generally speaking, loss of translation quality and precious time are the inevitable consequences.

An excellent example of how to improve on the current situation is the initiative

carried out by the TaaS project[1]. TaaS (Terminology as a Service) is a European project developed by a group of institutions and companies in the translation technology field who conceive 21st century terminology in a user-friendly, collaborative, cloud-based environment (Gornostay, 2014). Their aim is to create a platform for instant access to the most up-to-date terms and for user participation in the acquisition, sharing and reuse of multilingual terminological data. TaaS targets all types of language professionals, but specifically focuses on translators as end users, as it provides the following terminology services: (1) automatic extraction of term candidates; (2) automatic recognition of translation equivalents in different public and industry terminology databases; (3) automatic acquisition of translation equivalents for terms not found in term banks from parallel/comparable web data using the state-of-the-art terminology extraction and alignment methods; (4) facilities for terminology sharing and reusing within CAT tools; and (5) improvement of statistical machine translation systems through terminological data integration.

As an improvement, we developed EcoLexiCAT, a terminology-enhanced CAT tool that provides easy access to domain-specific terminological knowledge in context. This application integrates different features of the professional translation workflow in a stand-alone interface where a source text is interactively enriched with terminological information (i.e. definitions, translations, images, compound terms, corpus access, etc.) from different external resources: (i) EcoLexicon, a multimodal and multilingual terminological knowledge base on the environment (Faber et al., 2014; Faber et al., 2016); (ii) BabelNet, an automatically constructed multilingual encyclopedic dictionary and semantic network (Navigli & Ponzetto, 2012); (iii) and Sketch Engine, the well-known corpus query system (Kilgarriff et al., 2004).

The remainder of this paper is organized as follows. Section 2 explains terminology management from the perspective of the needs and expectations of professional translators. Section 3 concisely describes the web-based open source CAT Tool MateCat on which EcoLexiCAT is based as well as the external resources used for terminology enhancement. Section 4 provides a detailed explanation of EcoLexiCAT and its different modules. Finally, Section 5 presents the conclusions that can be derived from this study and outlines ideas for future research.

2. Translators' needs and expectations for terminology management

Any lexicographic or terminographic tool should take into account the needs of end users, in its structure and content as well as the way that the information is represented so that users can search and interact with the tool (Tarp, 2013). When

[1] http://taas-project.eu/

translators query a resource and do not find the information needed, they lose time and their productivity decreases (*search costs*; Nielsen, 2008). Similarly, when translators obtain too much data (*infoxication*; Cornellà, 1999), which lengthens the knowledge construction time, their *comprehension costs* (Nielsen, 2008) increase. In addition, translators do research in all phases of the translation process. This occurs during the pre-translation phase in order to understand the original text and its terminology. Research is also performed when the original message is encoded in the target text, with a view to fulfilling pragmatic requirements and searching for equivalents. Finally, in the revision phase, translators must check terminology and generally ensure the quality of their translation (Durán Muñoz, 2012: 80). Accordingly, one of the major challenges of lexicographic and terminographic resources for translators is to find the right balance between search costs and comprehension costs.

Durán Muñoz (2010, 2012) affirms that translators prefer to solve their terminological problems by consulting ready-made resources. According to her study, the most frequent resources used are (in this order): bilingual specialised dictionaries or glossaries, searches in search engines, terminological databases, monolingual specialised dictionaries, and Wikipedia (Durán Muñoz, 2012: 81). She mentions that translators do not trust the quality of multilingual resources and that searches in parallel corpora are not high on the list of preferences. However, when asked to classify the most frequent ISO fields (ISO 12620:1999) in the microstructure of terminological resources, translators considered the following to be most essential: clear and concrete definitions, equivalents, derivatives and compounds, domain specification, examples, phraseological information, definition in both languages for bilingual resources, and abbreviations and acronyms (Durán Muñoz, 2012: 82). Finally, when asked for their opinion, translators said that terminological resources should be able to do the following: (i) permit exportability and/or importability in different formats; (ii) include more pragmatic information about usage and tricky translations (old usage, false friends, specific usage in a domain or region, etc.); (iii) offer links to other resources to improve or increase the results; (iv) improve search options; and (v) provide examples taken from real texts (idem). Quite surprisingly, although the translators in this study did not show much interest in having access to corpora, they did highlight the need for more phraseological information, pragmatic information and examples taken from real texts. Even though this information can be extracted from corpora, translators were probably reticent to use them because it can take a long time if the right query methods are not provided.

Translation-oriented terminology management, or terminology-enhanced translation, should take into account all of the above. As shown in Section 4, EcoLexiCAT is a tool that includes the essential fields mentioned, links to other resources and improved search options for corpus analysis that provide the necessary pragmatic information and real text examples. All of this is available in a single-platform web-based CAT environment that has the capabilities of importing and exporting different file types and formats.

3. EcoLexiCAT sources

3.1 MateCat

MateCat, acronym of Machine Translation Enhanced Computer Assisted Translation, was originally a three-year research project led by a consortium composed of the international research center FBK (Trento, Italy), Translated SRL, the Université du Maine and the University of Edinburgh. The objective was to improve the integration of MT and human translation (Federico et al., 2014: 129). Within the project a computer-aided translation tool was developed, The MateCat Tool. This application is not only an industrial tool but also an open source platform[2]. It offers all the features of a modern CAT tool, such as a text editor that divides the text to be translated in source and target segments and saves them along with their translation in a translation memory (TM).

MateCat runs as a web server and communicates with other services through open APIs. It allows communication with pre-existing TMs, terminological databases, concordance searches within the TMs and MT engines, from which the MT provider MyMemory (a combination of Google Translate and Microsoft Translator) is freely available. The tool has been tested in professional settings and adapted for research in MT (e.g. Bertoldi et al., 2013 *apud* Federico et al., 2014: 131) and for educational purposes. The fact that it has an open-source version as well as a high level of flexibility made it a suitable option for the development of EcoLexiCAT. In addition, the features and operation of MateCat are basically the same as those found in most CAT tools used nowadays. Therefore, professional translators will not need to invest much time in learning how to use the tool and will benefit from the interoperability of CAT-related formats (TBX for glossaries, XLIFF for bilingual files, TMX for TMs, etc.). This enables them to use the resources generated during the translation process in other similar tools and reuse pre-existing resources (i.e. glossaries, bilingual files and TMs) in EcoLexiCAT.

3.2 EcoLexicon

EcoLexicon[3] is a multilingual and multimodal terminological knowledge base on environmental science (Faber et al., 2014; 2016). It is the practical application of Frame-based Terminology (Faber et al., 2011; Faber, 2012, 2015), a theory of specialized knowledge representation that uses certain aspects of Frame Semantics (Fillmore, 1982; Fillmore & Atkins, 1992) to structure specialized domains and create non-language-specific representations. Frame-based Terminology focuses on

[2] https://www.matecat.com/open-source/

[3] ecolexicon.ugr.es

conceptual organization, the multidimensional nature of specialized knowledge units, and the extraction of semantic and syntactic information through the use of multilingual corpora.

EcoLexicon is an internally coherent information system, which is organized according to conceptual and linguistic premises at the macro- as well as the micro-structural level. It currently has 3,601 concepts and 20,211 terms in Spanish, English, German, French, Modern Greek, and Russian. This terminological resource was conceived for language and domain experts as well as for the general public. It targets users such as translators, technical writers, and environmental experts who need to understand specialized environmental concepts with a view to writing and/or translating specialized and semi-specialized texts.

End users interact with EcoLexicon through a visual interface with different modules that provide conceptual, linguistic, and graphical information. Instead of viewing all information simultaneously, they can browse through the windows and select the data that is most relevant for their needs. Figure 1 shows the entry in EcoLexicon for the word FAN. When users open the application, three zones appear. The top horizontal bar gives users access to the term/concept search engine. The vertical bar on the left of the screen provides information regarding the search concept, namely its definition, term designations, associated resources, general conceptual role, and phraseology.

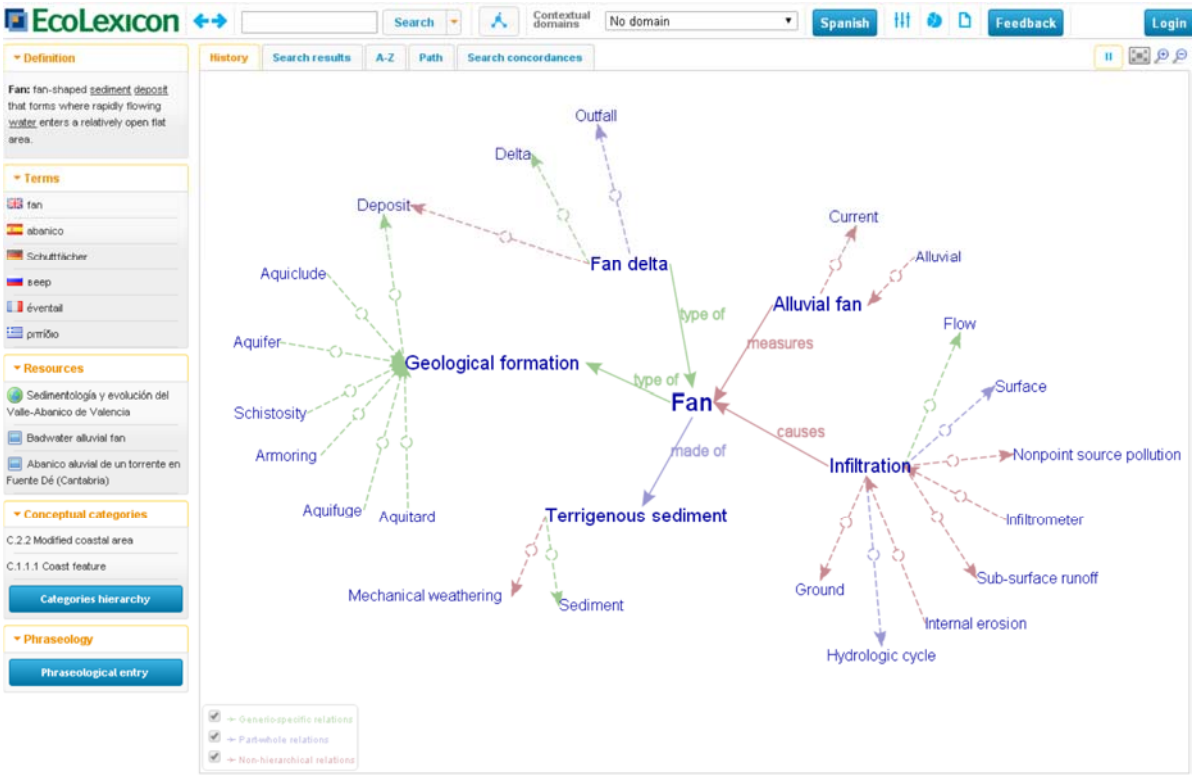


Figure 1: EcoLexicon user interface

Each definition makes category membership explicit, reflects a concept's relations with other concepts, and specifies essential attributes and features (León-Araúz, Faber & Montero-Martínez, 2012: 153-154). Accordingly, the definition is the linguistic codification of the relational structure shown in the concept map, at the center of the screen. Although users can configure the map to their needs, the standard representation mode (see Figure 1) shows a multi-level semantic network whose concepts are all linked in some way to the search concept, which is at its center.

A specialized corpus was specifically compiled for EcoLexicon in order to extract linguistic and conceptual knowledge. Currently, the corpus has over 50 million words and each of its texts has been tagged according to a set of XML-based metadata, which contain information about the language of the text, the author, date of publication, target reader, contextual domain, keywords, etc. This was done in order to provide users with a direct and flexible way of accessing the corpus. It also allows them to constrain corpus queries based on pragmatic factors, such as contextual domains or target reader. In this way, users can compare the use of the same term in different contexts. The corpus was first made available in the Search concordances tab (center area menu just above the concept map in Figure 1). However, currently, the English EcoLexicon Corpus (23 million words) is also hosted and freely available in Sketch Engine Open Corpora[4].

To fully exploit the contents and components of EcoLexicon for purposes of translation, we developed EcoLexiCAT. A terminological knowledge base (TKB) such as EcoLexicon provides a great amount of interconnected information in many different formats. However, in the professional translation workflow, especially when the source text has a high term density, searching in EcoLexicon, together with other resources, might cause high search and comprehension costs (see Section 1). EcoLexiCAT provides all this knowledge as an integral part of the translation workflow, where it is presented according to a specific context and during a specific phase of the translation process (see Section 4).

3.3 BabelNet and Babelfy

The multilingual encyclopedic dictionary and semantic network BabelNet[5] was created by linking Wikipedia to WordNet (Navigli & Ponzetto, 2012: 218). It connects concepts and named entities in a network of semantic relations, made up of about 14 million entries, called Babel synsets. Each Babel synset represents a given meaning and contains all the synonyms expressing that meaning in a range of different languages. Wikipedia and WordNet are integrated through automatic mapping and by filling in lexical gaps in resource-poor languages with MT.

[4] https://the.sketchengine.co.uk/open/

[5] babelnet.org

BabelNet is an enormous information resource that can be accessed through an open API, and was considered to be a valuable addition to EcoLexiCAT in those cases where EcoLexicon, a manually-built resource, did not include sufficient information regarding general language issues or for texts that combine environmental issues with other domains of expertise. Furthermore, the BabelNet researchers created their own algorithm, called Babelfy[6] for the disambiguation of polysemic words when found in the context of a particular text (Moro, Raganato & Navigli, 2014; Moro, Cecconi & Navigli, 2014).

Babelfy is a unified multilingual, graph-based approach to entity linking (the disambiguation of named entities) and word-sense disambiguation (the disambiguation of common nouns, verbs and adjectives). When presented with an input segment, the system extracts all the linkable fragments and lists the possible meanings of each of them according to the semantic networks of BabelNet. Evidently, this is of great help when dealing with polysemic terms. In EcoLexiCAT, the source text is disambiguated through Babelfy before matching the terms with BabelNet.

3.4 Sketch Engine

Sketch Engine (Kilgarrif et al., 2004) is an online corpus query system with a very efficient search engine and a statistical component for enhanced precision. It contains over 300 corpora in over 60 languages and allows end users to create their own corpora as well. One very interesting module is information extraction through word sketches. Word sketches are summaries of collocational information of a search term, where the term is analyzed according to the verbs, modifiers and other usual constructions that accompany it in real texts. Word sketches are created through sketch grammars that launch specific queries to a corpus. End users can create their own grammars for word sketches and therefore adapt the tool to their specific needs.

With an account, users have access to pre-loaded corpora and a corpus compiler called WebBootCaT. They can download corpora, add new documents to a corpus, extract domain keywords, view texts, and generate concordances, wordlists, frequency lists, collocations, and word sketches. Sketch Engine also hosts a set of freely available open corpora that can be queried with full Sketch Engine functionalities. This option, where the EcoLexicon corpus can be uploaded and afterwards freely accessed, made it a perfect option as a corpus query system for EcoLexiCAT.

4. EcoLexiCAT: a terminology-enhanced translation tool

When users start a new project in EcoLexiCAT, they first access the project settings interface in Figure 2, where they can do the following: (1) name the project; (2) choose

[6] babelfy.org

directionality (so far, English–Spanish or Spanish–English); (3) select a particular domain within the environment—these are in accordance with the domains according to which EcoLexicon is organized and are included in this first step as a way to classify projects and TMs for later reuse; (4) choose between general and patent segmentation rules, for the source text to be segmented accordingly; (5) optionally add an MT provider for post-editing—MyMemory is freely available, but others (e.g. Moses, DeepLingo, IP Translator) can also be added if users have an account with them; (6) optionally add users' own TMs and/or glossaries—otherwise a collective TM stored in the system will be used; and (7) upload the source text. These steps, except for (3), are default options in MateCat.

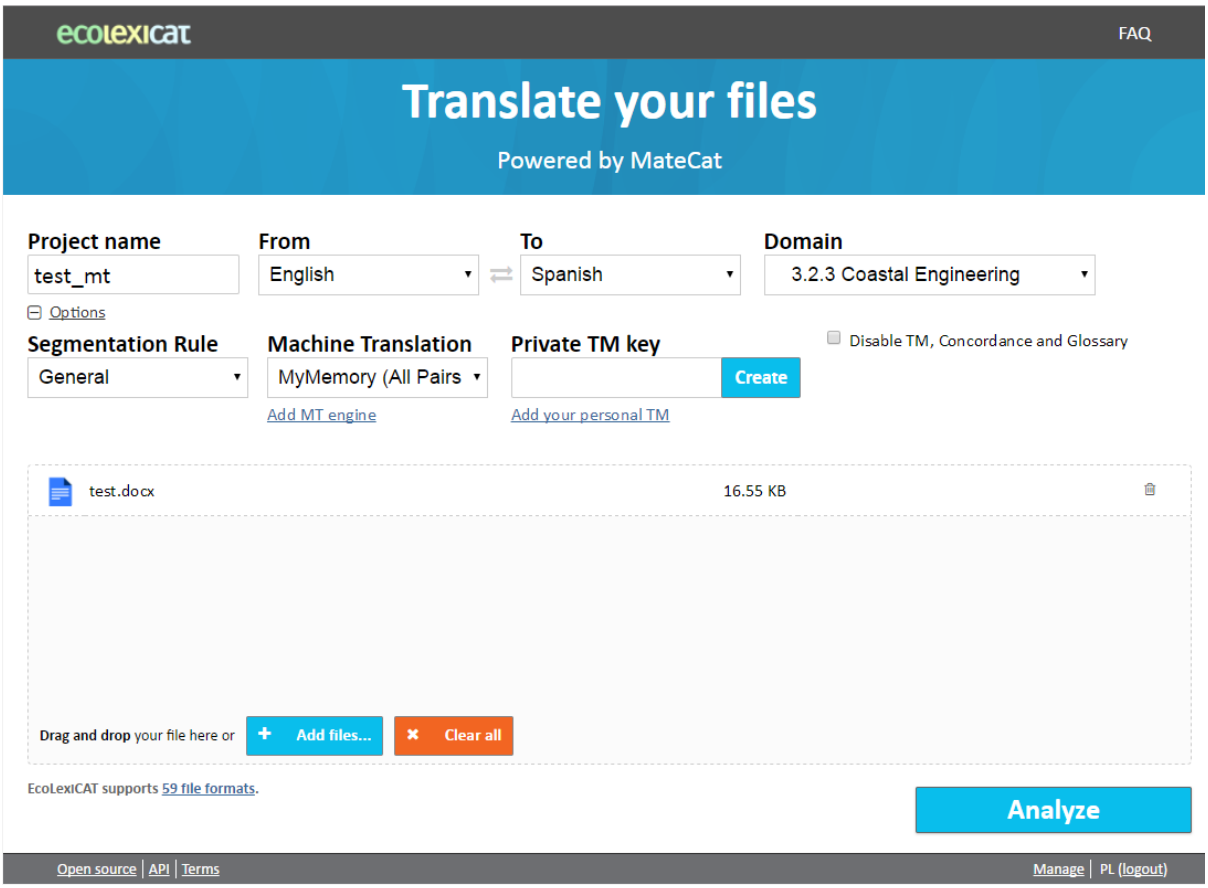


Figure 2: Project settings in EcoLexiCAT

Once the source text is processed and converted into a bilingual format (XLIFF), users can access the main interface (Figure 3), which is divided into two main sections. The left-hand section is where the three external resources (i.e. EcoLexicon, BabelNet/Babelfy and Sketch Engine) provide the terminological enhancement of the translation process. The right-hand section is where the target text is produced, an editor where the source text appears split into different segments. In the right upper part of the editor, users may download the target or the source text in their original format, and export the bilingual file in SDLXLIFF (SDL Trados Studio's native

format) or the whole project in OmegaT's native format, another desktop open source CAT tool. This, together with the possibility of downloading the TM and the glossary created during the project, ensures the interoperability of different formats across different CAT tools, an issue that professional translators must often deal with.

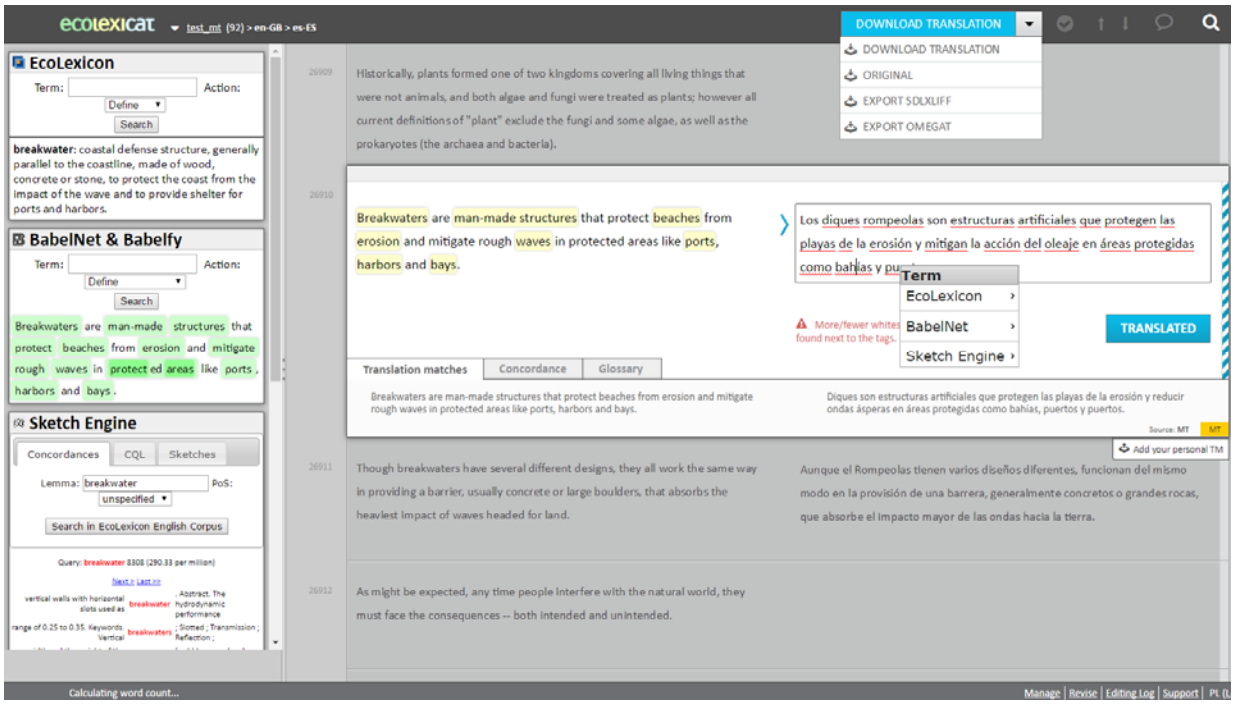


Figure 3: Main user's interface of EcoLexiCAT

Figure 4 shows a segment within the editor. This editor offers the usual editing features of any CAT tool. Users can split or merge segments, copy the source text in the target segment, benefit from a QA system that detects missing spaces or tags, create on-the-fly glossary entries, search for concordances within the TM and get suggestions from previously stored segments in the TM or, if added, from an MT engine. Once a segment is confirmed, it is stored both in the users' TM and in a collective TM from which other users can benefit. This converts the tool into a collaborative environment.

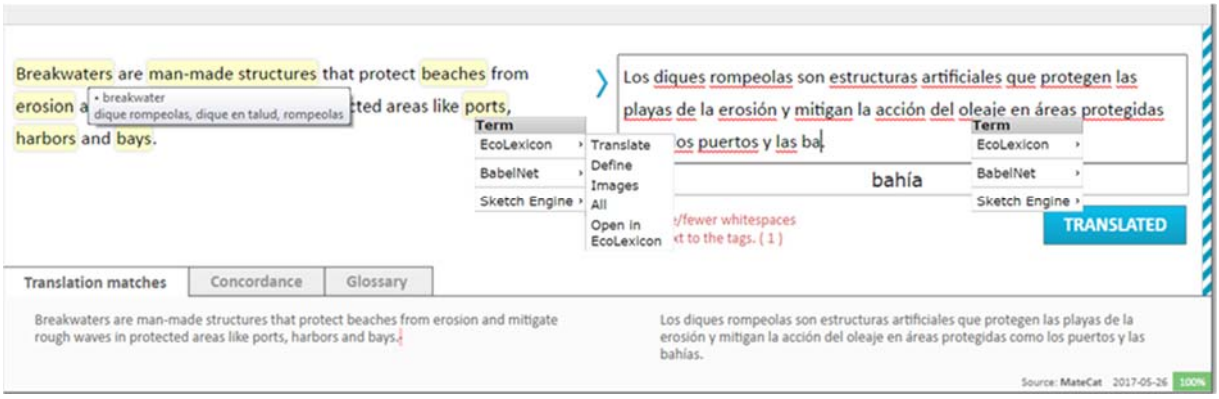


Figure 4: EcoLexiCAT editor

However, the difference between an ordinary CAT tool and EcoLexiCAT is that the EcoLexiCAT is a terminology-enhanced translation tool. This means that the editor interacts with external terminological resources that can assist the translator during the different phases of the translation workflow. First of all, the source segment is enriched with information from EcoLexicon. This is done by lemmatizing all the words in the segment and matching them against the term entries in the TKB.

All matching terms are highlighted in yellow, and users can interact with them in three ways: (1) if they hover the mouse over them, all possible translations (equivalent terms and synonyms) are displayed in an emerging box; (2) if they click on any of them, the EcoLexicon box of the left-hand side shows both the translations and the definition; and (3) if they right-click on any of them, a scroll-down menu gives access to all the different options provided by each of the resources of the left-hand section (see Figures 5–11).

For instance, in the case of EcoLexicon, these options correspond to the data categories in the TKB that usually serve for text comprehension: translations, synonyms, definitions, and images. Also from this menu, a new tab can be opened in the browser to access the EcoLexicon TKB for a more detailed analysis of the conceptual networks.

In turn, the target segment is enriched with a predictive typing feature. As soon as users start typing a word that has been matched as the translation of one of the terms in the source segment, all possible translations are shown in a drop-down list. In addition, as in the source segment, users can right-click on any term they type in the target segment and send queries to the three resources in the opposite language directionality. This is especially relevant in the case of corpus queries, since this is the resource that will usually be most useful during the text production phase.

Thus, the external resources of EcoLexiCAT interact with the segments in the editor during the different phases of the translation process, since they are terminologically enhanced for both source text comprehension and target text production tasks.

In Figures 5–11, a detailed view of the external resource boxes is provided. Figure 5 shows the EcoLexicon box as it appears when all features (i.e. translations, definition, images) are requested from any of the modules where the scroll-down menu may be activated (i.e. the EcoLexicon box itself, the BabelNet & Babelfy box, the source segment or the target segment). Users can also choose to visualize these features separately.

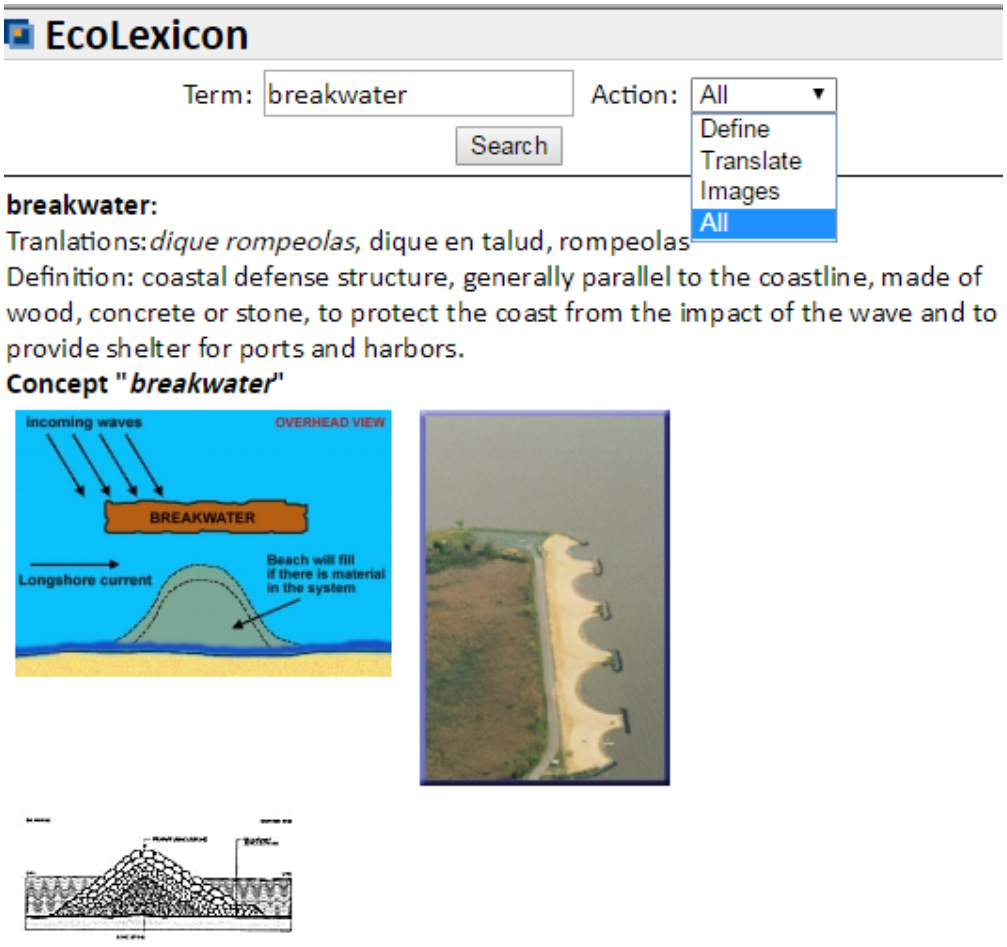


Figure 5: EcoLexicon box in EcoLexiCAT

Below the EcoLexicon box, users can find the BabelNet & Babelfy box (Figure 6), where the source text is also matched against the BabelNet network previously disambiguated by the Babelfy algorithm. This enables the system to propose statistically relevant candidate translations, which is a significant advantage taking into account that BabelNet covers any specialized or general domain and ambiguity can be frequently encountered. Furthermore, it helps the system to arrange definitions or images in the most plausible order. For instance, in Figure 6, while the first three definitions can be useful for EcoLexiCAT users, the fourth clearly belongs to a different domain and shows a different sense of the term *erosion*.

In this box, all matched terms are highlighted in green and behave in the same manner as the terms in the source segment with regard to EcoLexicon: (1) if users hover the mouse over them, all possible translations (equivalent terms and synonyms) are displayed in an emerging box; (2) if they click on any of them, the BabelNet & Babelfy box on the left-hand side shows both the translations and the definition; and (3) if they right-click on any of them, a scroll-down menu gives access to all the different options provided by each of the resources of the left-hand section. In the case of

BabelNet, these options correspond to the data categories that have been considered most interesting for translators: definitions, translations, compound words and images. Also, from the definitions option, a new tab can be opened in the browser to access the semantic networks in BabelNet.

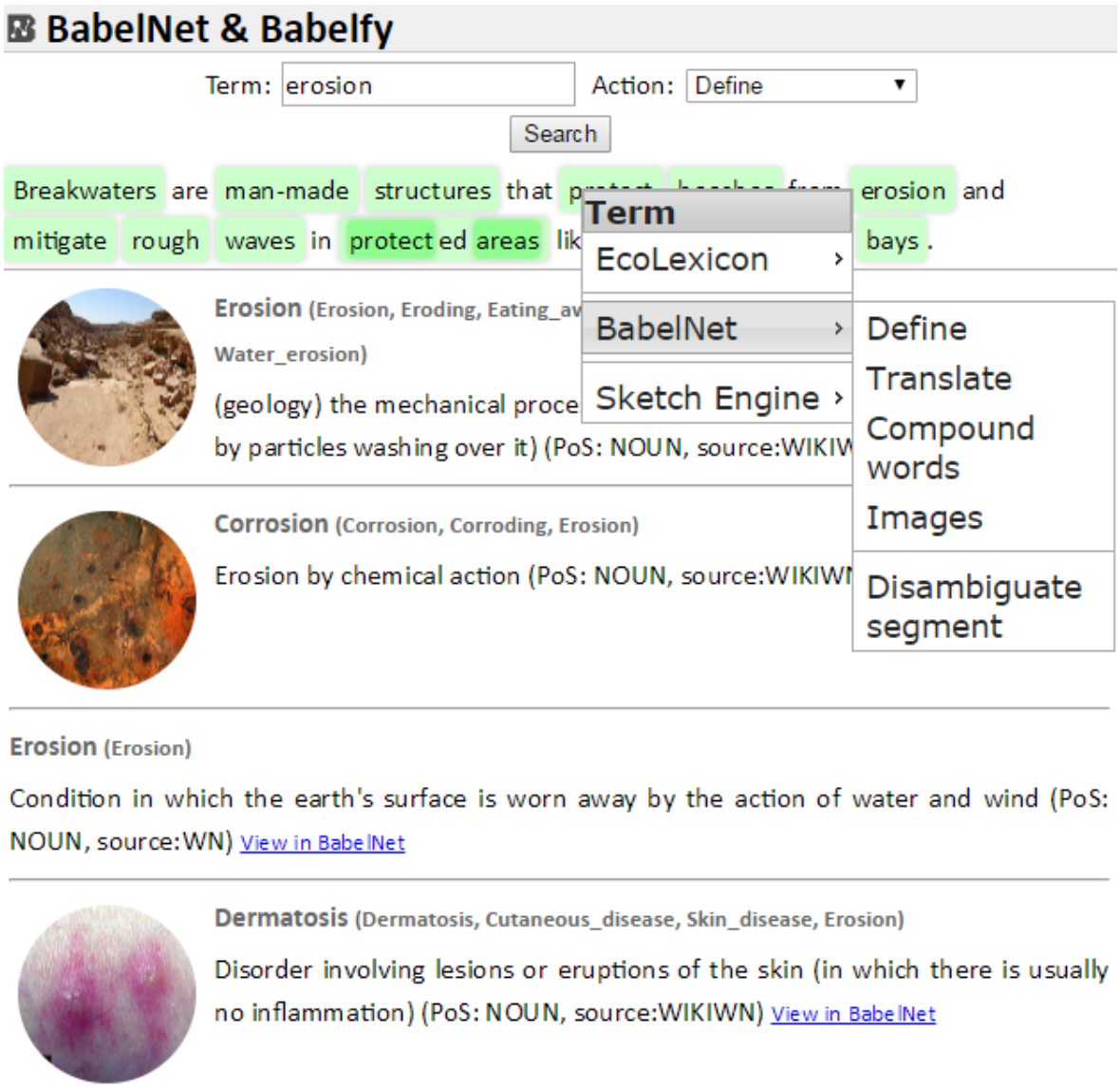


Figure 6: BabelNet and Babelfy box in EcoLexiCAT – Definitions

This box is particularly interesting for terms that are not available in EcoLexicon. This may occur when entries in EcoLexicon have not yet been included (it is a developing resource), when general language issues arise or when the source text combines environmental terms with terms from other specialized domains. Nevertheless, users should be cautious because the Babelfy algorithm may fail or produce candidate translations that do not account for domain specificity. Being an automatically built resource based on the synsets of WordNet (a general language lexical database), BabelNet often offers a set of concepts with different levels of granularity under the same entry.

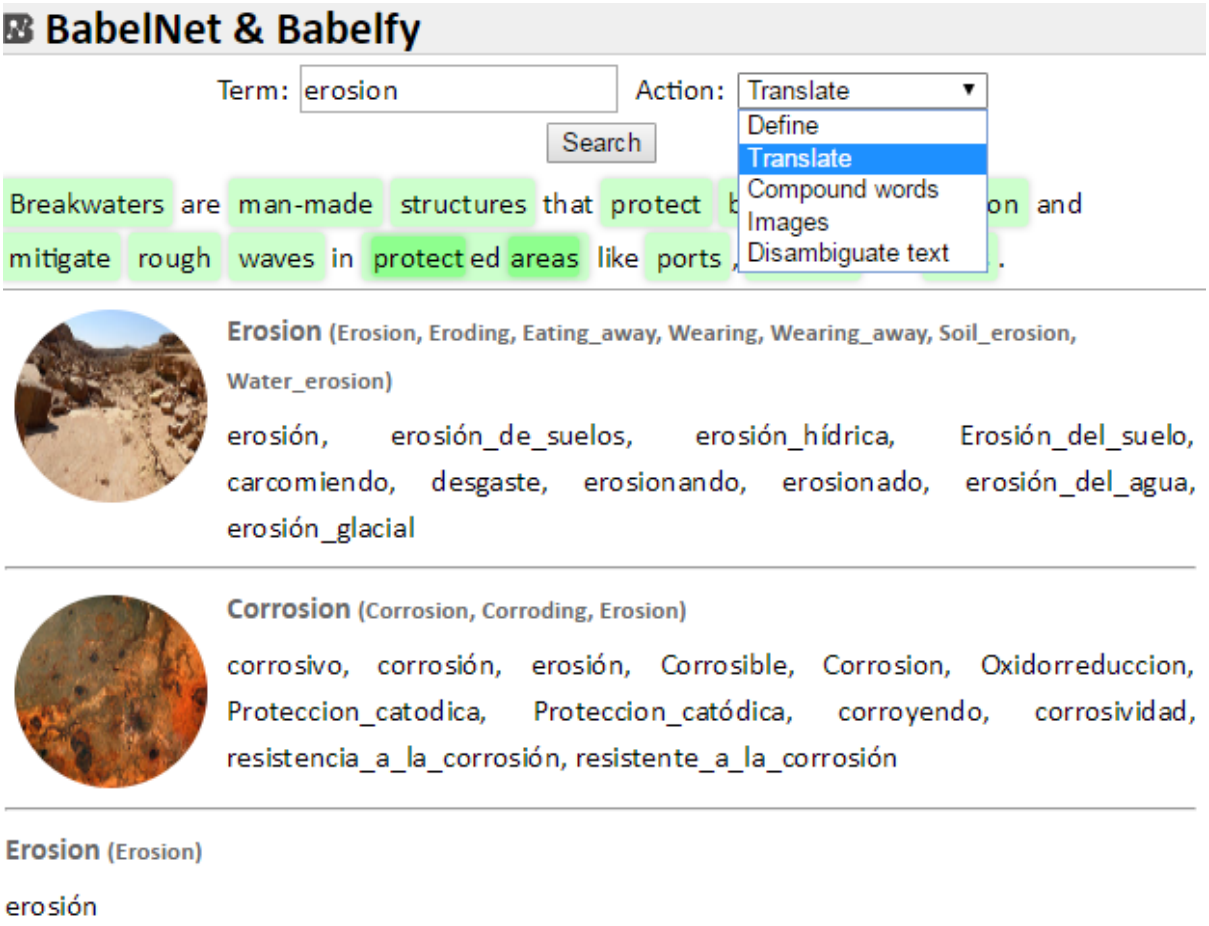


Figure 7: BabelNet & Babelfy box in EcoLexiCAT – Translations.

For instance, in Figure 7, the candidate translations go beyond equivalence, since some of the terms are hyponyms or derivatives of *erosion*. Nonetheless, when used with caution, these results can help to expand user knowledge of the semantic network of the domain

However, for this purpose, and especially for text production tasks, there is another option in this box, namely compound words. Figure 8 shows different compound terms of *erosion*, whether it acts as the head (e.g. *beach erosion*) or the modifier of the compound (e.g. *erosion control*). All of them can be clicked to access their definitions. In this way, users can browse the resource through different interconnected concepts and terms and gain a better understanding of the domain. Finally, images are the last option available from BabelNet (Figure 9). They can be very useful when understanding and translating complex concepts, such as processes or parts of entities, and can complement the images offered by EcoLexicon.

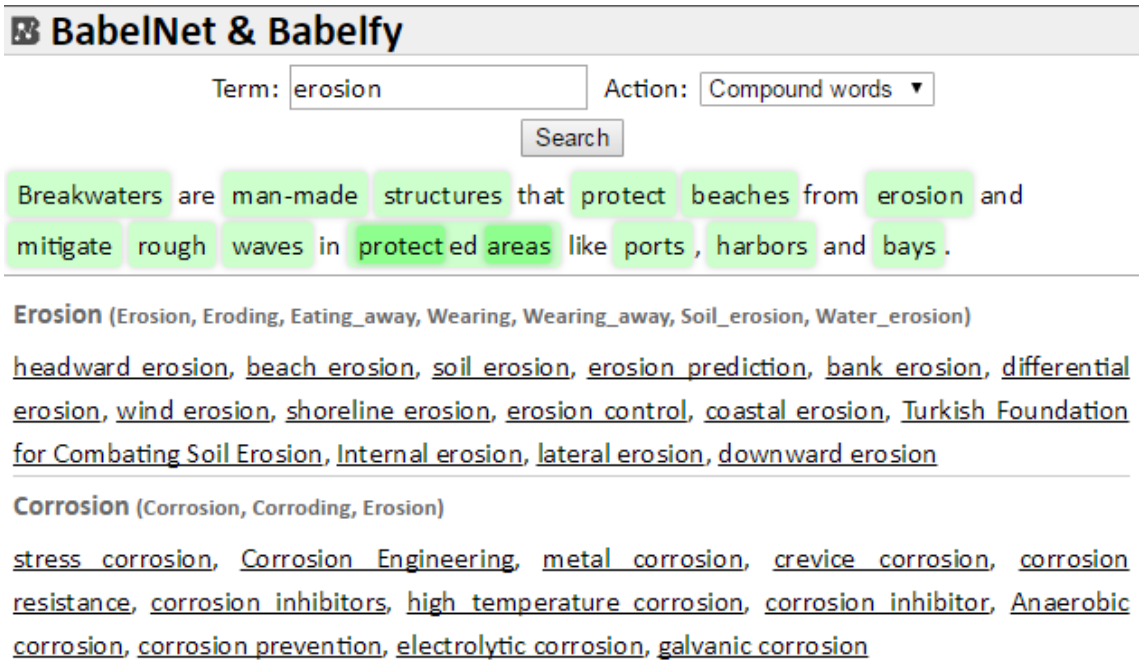


Figure 8: BabelNet & Babelfy box in EcoLexiCAT – Compound words

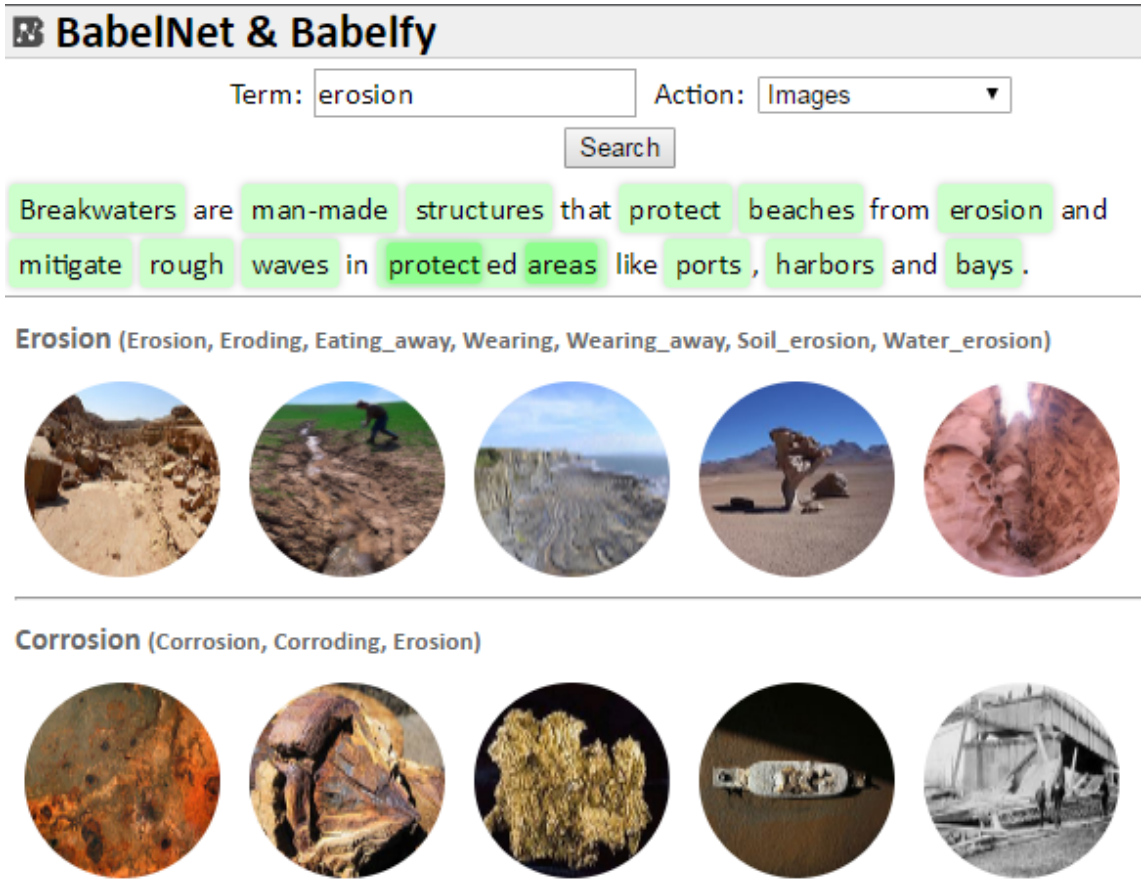


Figure 9: BabelNet & Babelfy box in EcoLexiCAT – Images

Below the BabelNet & Babelfy box, the Sketch Engine box appears (Figure 10). This box can be used to select a term from both the source and target segments and analyze its behavior in the EcoLexicon Corpus. So far, only the EcoLexicon English Corpus is hosted in Sketch Engine Open Corpora. The EcoLexicon Spanish Corpus is still in the compilation phase but will be made available in the near future.

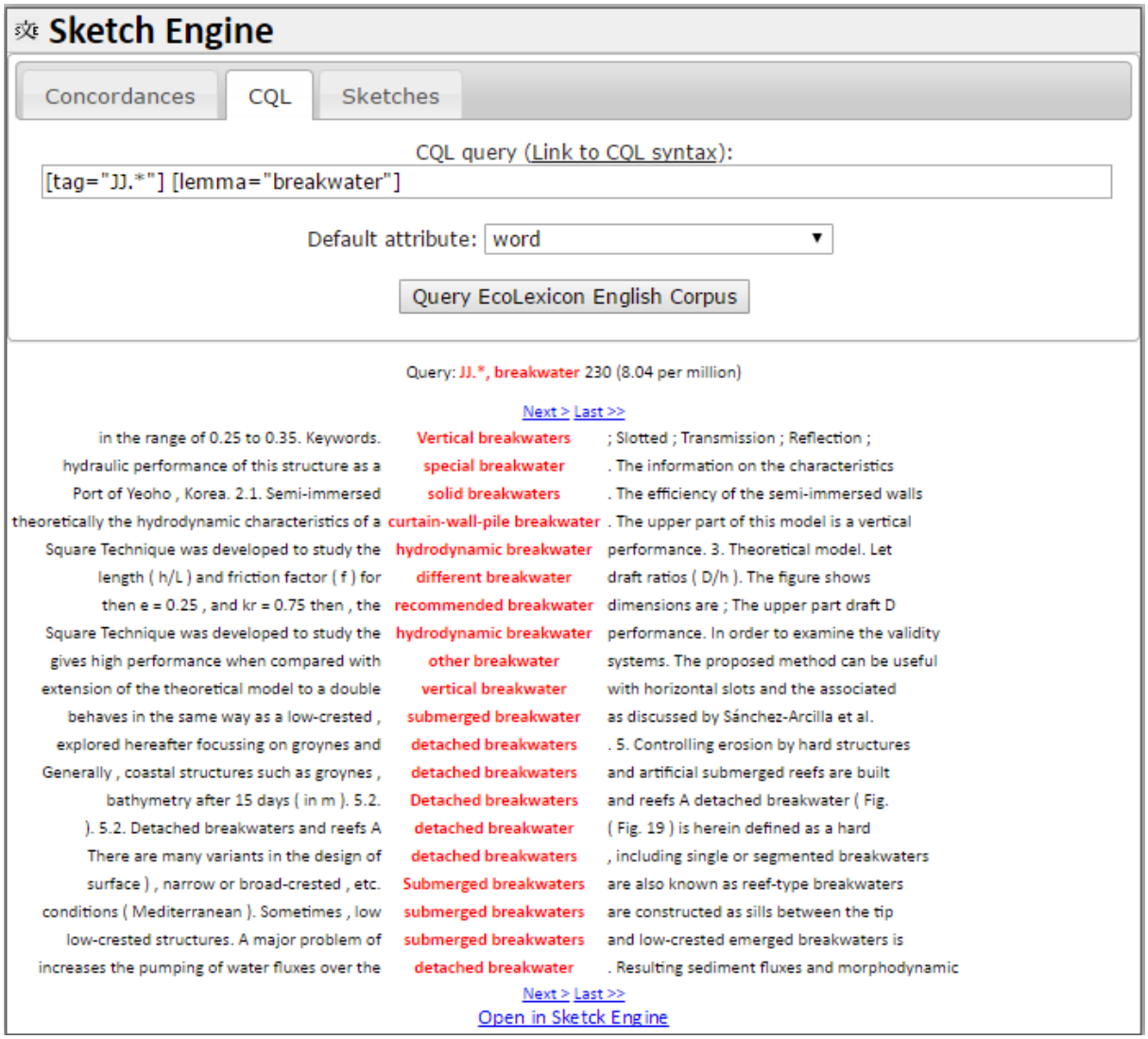


Figure 10: Sketch Engine box in EcoLexiCAT – CQL queries

The corpora can be queried through basic or CQL queries (Figure 10) as well as through word sketches (Figure 11). The output of the queries can be opened in a new tab that sends users to the website of Sketch Engine Open Corpora for a more detailed analysis. In this way, they can use all the functionalities of the tool (e.g. Context, Word list, Thesaurus, Sketch Diff, etc.) and make more specific queries filtered by the features according to which the corpus is tagged (i.e. year, genre, contextual domain, user type and linguistic variant). As previously mentioned, this information can be very useful during the text production phase (e.g. searching for modifiers or verbs that

collocate with a particular noun, looking for synonyms or frequent syntactic structures, etc.). However, corpora can also help translators to understand how concepts interrelate with each other within the domain. For this reason, corpus queries are enabled from both source and target segments.

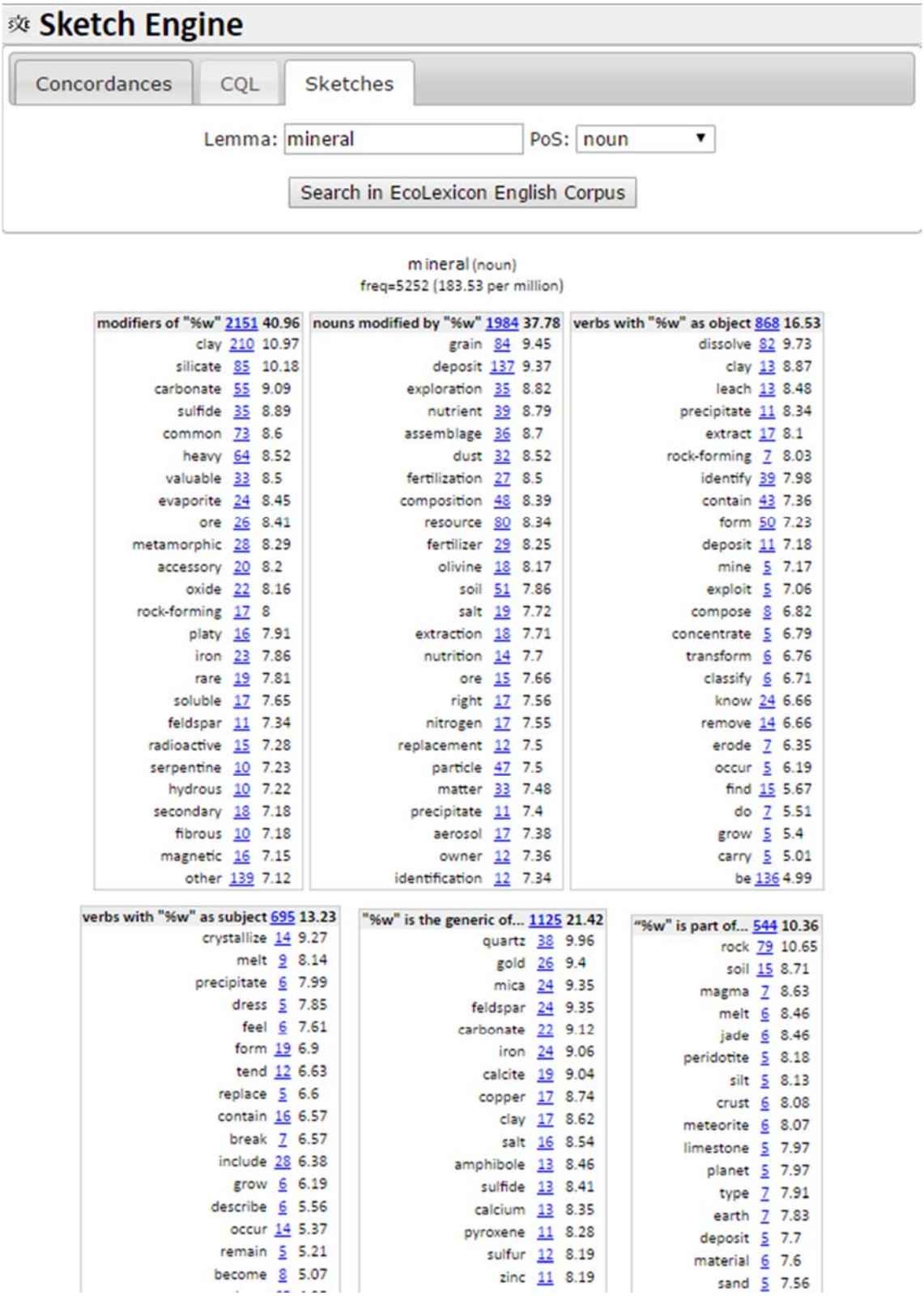


Figure 11: Sketch Engine box in EcoLexiCAT – Word Sketches

For instance, with the CQL query in Figure 10, users can not only access the adjectives that modify the term *breakwater* but also infer that breakwaters are usually classified according to position, material, function, etc. Furthermore, in Figure 11, Sketch Engine's default word sketches (e.g. modifiers and verbs) are combined with a series of customized word sketches (León-Araúz et al., 2016) especially focused on the comprehension phase, since they are based on semantic relations and thus provide knowledge rich contexts (Meyer, 2001). In Figure 11, the customized word sketches of the relations *is_the_generic_of* and *is_part_of* are shown for the term *mineral*. In this way, users can have quick access to part of the conceptual network of all concepts sufficiently represented in the corpus.

Finally, there are two other features powered by MateCat that can be of interest to professional translators, as well as to lecturers and researchers. As soon as a segment is confirmed, users can open their editing log (Figure 12) and monitor their own performance. This includes different types of information on each segment, such as: (1) the time invested in post-editing it; (2) the suggestion source, whether it comes from MT or TMs; (3) the matching percentage between the source segment and the suggestion; and (4) the post-editing effort and tracked changes of the final target segment. These data help to raise user awareness regarding their strengths and weaknesses as professional translators as well as those of the tool. For this reason, the editing log can also be exploited by Translation lecturers and researchers who are interested in assessing both the work of students and/or the performance of the tool.

Editing Details

Secs/Word	Job ID	Segment ID	Words	Suggestion source	Match percentage	Time-to-edit	PE Effort
6.4	92	26910	21.00	Machine Translation	85%	02m:13s	33%

Segment	Breakwaters are man-made structures that protect beaches from erosion and mitigate rough waves in protected areas like ports, harbors and bays.
Suggestion	Diques son estructuras artificiales que protegen las playas de la erosión y reducir ondas ásperas en áreas protegidas como bahías, puertos y puertos.
Translation	Los diques rompeolas son estructuras artificiales que protegen las playas de la erosión y mitigan la acción del oleaje en áreas protegidas como los puertos y las bahías.
Diff View	~~Diques~~ Los diques rompeolas son estructuras artificiales que protegen las playas de la erosión y ~~reducir ondas ásperas~~ mitigan la acción del oleaje en áreas protegidas como ~~bahías,~~ los puertos y ~~puertos.~~ las bahías.

Figure 12: Editing log in EcoLexiCAT

In this line, the revision panel (Figure 13) helps to perform the last phase of the translation workflow. Revisers can approve or correct all target segments. If corrected, the changes are tracked in the target cell, and revisers can use a metric for translation quality evaluation commonly used in the industry. This metric is based on different error types (i.e. tag issues, translation errors, terminology and translation consistency, language quality and style) and degrees (i.e. enhancement and error). At the end, users can generate a quality report that automatically scores the overall quality of the

translation based on the issues highlighted by the revisers. Therefore, this feature can also be used by Translation lecturers if they want to grade their students' work in a systematic way.

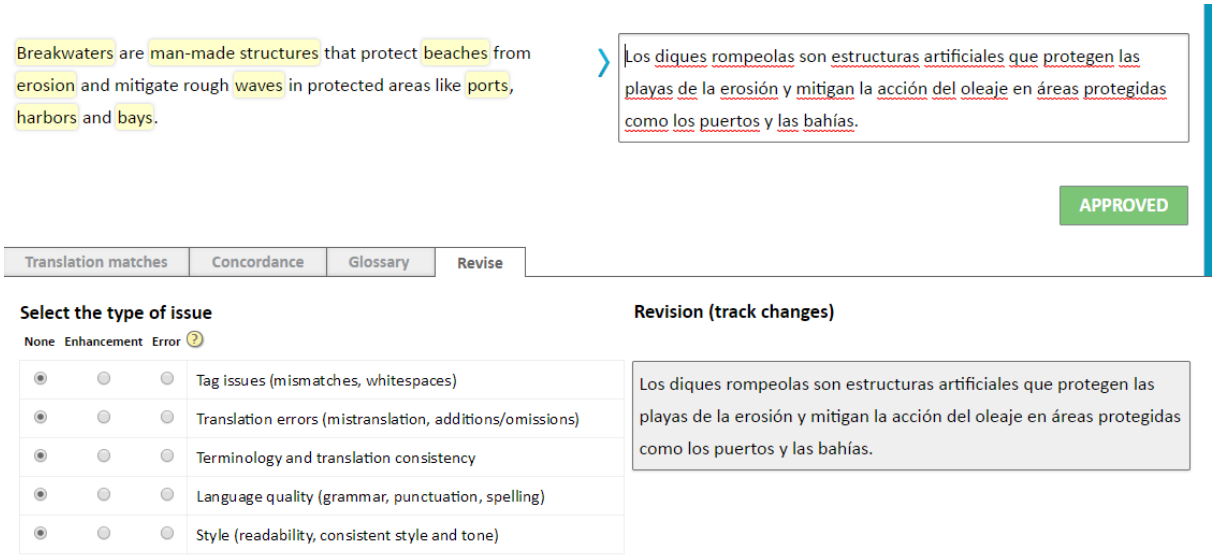


Figure 13: Revision in EcoLexiCAT

5. Conclusions and future work

In this paper we have presented the first version of EcoLexiCAT, a terminology-enhanced tool that enriches both source and target segments with terminological information from three external resources in an interactive environment. The tool has been designed to meet the expectations of professional translators regarding terminology management. However, it still needs to be evaluated by prospective users. A study comparing the performance of EcoLexiCAT users versus non EcoLexiCAT users will thus be carried out in the near future.

However, there are still other features that will be added to the tool before starting the evaluation process. For instance, EcoLexiCAT will also be enriched with other external resources. Part of the Inter-Active Terminology for Europe[7] (IATE), EU's multilingual term base, has been recently downloaded and stored in a database to interact with EcoLexiCAT as a fourth external resource. The IATE dump will cover the entries in English and Spanish belonging to environment-related domains.

Furthermore, EcoLexicon is currently being linked to other encyclopedic (i.e. DBpedia) and environmental resources (i.e. GEMET, AGROVOC) by means of Linked Data. Once the TKB is fully integrated into the Linguistic Linked Open Data, EcoLexiCAT will also benefit from reliably disambiguated encyclopedic and specialized term entries.

[7] http://iate.europa.eu/tbxPageDownload.do

In the same line, we plan to add another box enabling users to customize for each project a resource console based on the URLs of the resources that they usually consult, such as WordReference, TERMIUM Plus, MetaGlossary, Linguee, etc. This will work as the SDL Trados Studio plug-in Web Lookup or the MemoQ web search feature. Two other features from EcoLexicon will also be added once they are ready. These are the EcoLexicon Spanish Corpus and phraseological patterns from a new module that is currently under construction.

Finally, when all of these features are included in the tool, EcoLexiCAT will be made freely available for any user interested in translating English or Spanish environmental texts. Users will only need to register and indicate their educational background, translation experience and the purpose for which they will be using the tool. This will help us analyze user profiles and behaviour when interacting with the tool. Moreover, it will allow us to classify the resources generated (i.e. TMs), which can be used as a parallel corpus, thus enriching both the tool and the EcoLexicon Corpus.

6. Acknowledgements

This research was carried out as part of project FF2014-52740-P, *Cognitive and Neurological Bases for Terminology-enhanced Translation* (CONTENT), funded by the Spanish Ministry of Economy and Competitiveness.

7. References

Bertoldi, N., Cettolo, M. & Federico, M. (2013). Cache-based Online Adaptation for Machine Translation Enhanced Computer Assisted Translation. In *Proceedings of the MT Summit XIV*, Nice, France, September, pp 35–42.

Durán Muñoz, I. (2010). Specialized lexicographical resources: a survey of translators' needs. In S. Granger & M. Paquot (eds) (2010). *eLexicography in the 21st century: New Challenges, new applications. Proceedings of ELEX2009*. Cahiers du Cental. Vol. 7. Louvain-La-Neuve: Presses Universitaires de Louvain, pp. 55 – 66.

Durán Muñoz, I. (2012). Meeting translators'needs: translation-oriented terminological management and applications. *The Journal of Specialised Translation*, 18, pp. 77–92.

Faber, P., León-Araúz, P. & Reimerink, A. (2011). Knowledge representation in EcoLexicon. In N. Talaván, E. Martín Monje & F. Palazón (eds.) *Technological Innovation in the Teaching and Processing of LSPs: Proceedings of TISLID*, 10. Madrid: Universidad Nacional de Educación a Distancia, pp 367–385.

Faber, P. (ed.) (2012). *A Cognitive Linguistics View of Terminology and Specialized Language*. Berlin/New York: Mouton de Gruyter.

Faber, P. (2015) Frames as a framework for terminology. In H. J. Kockaert & F. Steurs (eds.) *Handbook of Terminology*, 1. John Benjamins Publishing Company, pp. 14–33.

Faber, P., León-Araúz, P. & Reimerink, A. (2014). Representing environmental knowledge in EcoLexicon. In *Languages for Specific Purposes in the Digital Era. Educational Linguistics*, 19. Springer, pp 267–301.

Faber, P., León-Araúz, P. & Reimerink, A. (2016) EcoLexicon: new features and challenges. In I. Kernerman, I. Kosem Trojina, S. Krek, & L. Trap-Jensen (eds.) *GLOBALEX 2016: Lexicographic Resources for Human Language Technology in conjunction with the 10th edition of the Language Resources and Evaluation Conference*, Portorož, pp. 73–80.

Federico, M. et al. (2014). The MateCat Tool. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, Dublin, Ireland, August 23-29, pp 129–132.

Fillmore, C. J. (1982). Frame Semantics. In The Linguistic Society of Korea (ed.) *Linguistics in the Morning Calm*. Seoul: Hanshin, pp. 111–137.

Fillmore, C. J. & Atkins, B. T. S. (1992). Toward a Frame-based Lexicon: The Semantics of RISK and Its Neighbors. In A. Lehrer & E. Kittay (eds.) *Frames, Fields and Contrasts: New Essays in Semantic and Lexical Organization*. Hillsdale NJ: Erlbaum, pp. 75–102.

Gornostay, T. (2014). Dreams of better terminology tools. *Multilingual Magazine* April/May, pp. 44–45.

International Organization for Standarization (ISO) (1999). *ISO 12620. Computer applications in terminology - Data categories*. Ginebra: ISO.

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 Aráuz, P., Faber, P. & Montero Martínez, S. (2012). Specialized Language Semantics. In P. Faber (ed.) *A cognitive linguistics view of terminology and specialized language.*, 20. Berlin, Boston: De Gruyter Mouton, pp. 95–175.

Meyer, I. (2001). Extracting Knowledge-Rich Contexts for Terminography: A conceptual and methodological framework. In Bourigault, Jacquemin, L'Homme (eds.), *Recent Advances in Computational Terminology*, pp. 279–302.

Moro, A, Raganato, A., Navigli, R. (2014). Entity Linking meets Word Sense Disambiguation: a Unified Approach. *Transactions of the Association for Computational Linguistics* (TACL), 2, pp. 231–244.

Moro, A., Cecconi, F., Navigli, R. (2014) Multilingual Word Sense Disambiguation and Entity Linking for Everybody (2014). *Proc. of the 13th International Semantic Web Conference, Posters and Demonstrations* (ISWC 2014), pp. 25–28, Riva del Garda, Italy, 19-23 October 2014

Navigli, R. & Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193, pp. 217–250.

Nielsen, S. (2008). The Effect of Lexicographical Information Costs on Dictionary Making and Use. *Lexikos*, 18, pp. 170–189.

Tarp, S. (2013). What should we demand from an online dictionary for specialized

translation? *Lexicographica - International Annual for Lexicography*, 29(1), pp. 146–162.

Tudhope D., Koch T. & Heery R. (2006). *Terminology Services and Technology: JISC state of the art review.* Available at: http://www.ukoln.ac.uk/terminology/JISC-review2006.html

A Corpus-assisted Approach to Paronym Categorisation

Ruth Maria Mell, Petra Storjohann

Institute for the German Language, R5, 6-13, 68161 Mannheim, Germany
E-mail: mell@ids-mannheim.de, storjohann@ids-mannheim.de

Abstract

In this paper, we will present a first attempt to classify commonly confused words in German by consulting their communicative functions in corpora. Although the use of so-called paronyms causes frequent uncertainties due to similarities in spelling, sound and semantics, up until now the phenomenon has attracted little attention either from the perspective of corpus linguistics or from cognitive linguistics. Existing investigations rely on structuralist models, which do not account for empirical evidence. Still, they have developed an elaborate model based on formal criteria, primarily on word formation (cf. Lăzărescu 1999). Looking from a corpus perspective, such classifications are incompatible with language in use and cognitive elements of misuse.
This article sketches first lexicological insights into a classification model as derived from semantic analyses of written communication. Firstly, a brief description of the project will be provided. Secondly, corpus-assisted paronym detection will be focused. Thirdly, in the main section the paper concerns the description of the datasets for paronym classification and the classification procedures. As a work in progress, new insights will continually be extended once spoken and CMC data are added to the investigations.

Keywords: paronyms; commonly confused words; e-dictionary; categorisation; semantic classification

1. Introduction

Paronyms are words that are similar in spelling, sound and / or meaning, i.e. *formell / formal / förmlich* (*formal*), *Technik / Technologie* (*technology*), *elektrisch / elektronisch* (*electric / electrical / electronic*), *Methode / Methodik / Methodologie* (*method / methodology*)[1] etc.[2] In this sense, paronyms are easily confused words which regularly cause problems for language learners and native speakers. Generally, such pairs of paronyms are not regarded synonymous although corpus analyses suggest that some items undergo meaning change due to the rivalry of two or more paronyms:

> "Sometimes, [paronyms] can develop synonymous notions and simply become lexical alternatives (cf. Storjohann, 2015). In other cases, they remain similar in meaning but show subtle differences and restrictions in usage. Inevitably,

[1] The first group are all essentially *formal*; the second are essentially *technology*; English has the same problem with *electric / electrical / electronic* and *method / methodology*.

[2] For more examples see Schnörch (2015).

situations of confusion arise when speakers' intuitions contradict information in existing reference works." (Storjohann, forthcoming)

So far, paronyms have been looked at only from a structuralist point of view and mainly from a language learners' perspective (cf. Lăzărescu, 1999). Up until now, the phenomenon has attracted little attention from the perspectives of corpus linguistics and cognitive linguistics. With the availability of diverse corpora, particularly spoken data and the development of new semantic approaches, only recently has paronymy become the focus of a new project ("Paronyme – Dynamisch im Kontrast"). The project lexicographically documents paronyms in a new corpus-based e-dictionary. Furthermore, it focusses on research on paronymy as a lexical-conceptual phenomenon and aims to develop an empirically-driven classification of paronyms using diverse genres of corpora including written and spoken texts as well as CMC data. In the past, investigations have relied on models accounting for language as a formal and logic system and not requiring empirical evidence in real communicative situations.

While a detailed description of the e-dictionary with respect to structure, content, navigation and visualisation is provided by Storjohann (in this volume), the central aim of this paper is to attempt to classify commonly confused words in German by consulting their communicative functions and semantic manifestations in written corpora.

2. The Paronym Dictionary ("Paronymwörterbuch")

The new German online dictionary "Paronymwörterbuch" (*Paronym Dictionary*) (Storjohann, 2014; Storjohann and Schnörch, 2017), which is currently being developed at the Institute for the German Language (IDS, Mannheim), breaks new ground by adopting a more conceptual and encyclopaedic approach to meaning by incorporating cognitive features. It will be published in 2017 and is publically accessible free of charge.[3] It is the very first corpus-assisted reference guide to the contemporary use of paronyms with regard to German.[4] The online dictionary strives to exploit the possibilities of using the electronic medium more effectively and in order to create an innovative, flexible and user-friendly instrument instead of listing traditional, linear and static entries. In doing so, this dictionary represents a step towards a dynamic, multi-functional cognitive-oriented online reference work with adaptive navigation (for details see Storjohann in this volume).

3. Corpus-assisted paronym detection and paronym analysis

Language data used for compiling dictionaries is often outdated, or lexicographic

[3] It will be published in the dictionary portal OWID (www.owid.de) in 2017.

[4] To our knowledge there is no corpus-guided, electronic reference guide of easily confused pairs in any other language.

practice is rather conventional and does not take advantage of corpus-assisted approaches to semantic analysis. The objective of the "Paronymwörterbuch" is to compile a new kind of dictionary with contrastive entries which will be a useful reference tool in situations of language doubt. At the same time, it aims to sensitise users to context dependency and language change.

> "As the subject of paronyms has not been revisited with empirical, data-driven methods either in terms of semantic theory or in terms of practical lexicography suitable corpus methods for contrastive investigation needed to be tested. Currently, complementary software-driven resources facilitating the search for similarity and difference are being exploited, each of which is based on the analysis and interpretation of contextual profiles, collocations and colligations, corresponding semantic roles and syntactic functions." (Storjohann, forthcoming)

To create the new online dictionary "Paronymwörterbuch", innovative approaches to empirical lexicographic work that pave the way for a new data-driven, descriptive reference work of confusable German terms have been adopted. An index (lemma list) is an essential pillar of every type of dictionary. For this reason, the concept, corpus extraction and compiling of a lemma list is a key task in the initial phase of every lexicographic project (cf. Schnörch, 2015: 16).

The first step in the paronym dictionary project was to find potential candidates for a paronym index. Consulting traditional print dictionaries such as Pollmann & Wolk (2010), Duden 9, and Müller (1973) provided us with typical pairs and their morphological features. We were then able to establish groups of candidates based on a variety of formal patterns (Schnörch, 2015), e.g.:

-al/-istisch (*natural*/*naturalistisch*)

-end/-lich (*dringend*/*dringlich*)

-ig/-lich (*fremdsprachig*/*fremdsprachlich*)

-sam/-lich (*betriebsam*/*betrieblich*)

Approximately 154 such formal categories were detected through the study of texts and dictionaries.

With the help of large corpora, all pairs which differed with respect to such patterns (often regular suffixes) but were identical in their root were automatically extracted using the 'string comparison' method. As a database, we used DeReWo (version derewo-v-ww-bll-320000g-2012-12-31-1.0). DeReWo consists of frequency-based rankings of lemmata and word forms on the basis of virtual corpora. These lists of

lemmata and word forms in use in the German language (for example the lemma candidate list with 350,000 entries for *elexiko*[5], the online dictionary of contemporary German) are generated by applying the methods for creating frequency-based rankings of lemmata and word forms on DeReKo – the German Reference Corpus (cf. http://www1.ids-mannheim.de/direktion/kl/projekte/methoden/derewo.html?L=1).[6]

In the next step, all automatically retrieved pairs were analysed manually. Overall about 9000 cases were scrutinised, 2000 were considered potential candidates. They were then categorised according to frequency (Storjohann & Schnörch, 2017). Two years ago, semantic analyses and lexicographic descriptions of the most frequent pairs started using different analysing tools and methods. An examination of the paronym list reveals a remarkable attribute of all these words. The candidates of the index are not an arbitrary jumble of words; by segmenting the character strings, morphological patterns and regular occurrences can be found. Among them is the study of significant collocations as identified by the corpus tool COSMAS II[7] – the Corpus Search and Management Analysis System. A further effective procedure is the use of the contrasting-near-synonym-method (CNS). This is profitably employed for contrastive analyses.

4. Datasets / Corpora for paronym classification

In this chapter, we will describe the corpora we are currently using for the analysis of paronyms. We will also present further options using different corpora for a future comprehensive classification of paronyms, paying particular attention to our base corpus "Paronymkorpus" (which is the basis for detailed paronym analysis). These different data resources will hopefully enable us to define a wider spectrum of variational properties and specific communicative idiosyncrasies otherwise not detected through the sole use of newspaper texts.

4.1 Paronymkorpus

As all analyses are guided by large corpora, for our initial investigations we have compiled a special, publically accessible corpus (the so-called Paronymkorpus) that contains written texts from between 1990 and 2015, comprising around 2.3 billion tokens. We have built a corpus based on DeReKo (the German Reference Corpus Collection, hosted by the Institute for the German Language (IDS) in Mannheim). DeReKo includes vast amounts of texts from genres as diverse as newspapers, fiction, parliamentary debates, and specialised text with different terminologies from more technical language use (cf. Kupietz & Lüngen, 2014).

---

[5] *elexiko*: http://www.owid.de/wb/elexiko/start.html.

[6] COSMAS II: https://cosmas2.ids-mannheim.de/cosmas2-web/.

[7] For details on analysing methods see Storjohann and Schnörch (2017).

With respect to German, the Paronymkorpus is the first lexicographic data resource that is completely open to the public. As it contains texts without restrictions of copyright it allows lexicological investigations and lexicographic documentation to be completely transparent. Concerning the regional distribution of the newspaper data (Figures 1 and 2), the corpus can be defined as relatively well-balanced (Paronyme – Dynamisch im Kontrast: project description, http://www1.ids-mannheim.de/lexik/paronymwoerterbuch/dasparonymkorpus.html) compared to others, e.g. elexiko http://www.owid.de/wb/elexiko/glossar/elexiko-Korpus.html).

Figure 1: Regional distribution of newspapers in the Paronymkorpus

Currently, the main focus of the project is on the analysis and description of the most frequent paronyms in written language data, especially in newspapers. Besides dialectal diversity of smaller regional newspapers and standardised nation-wide reception of larger journals, one major advantage of this text type is its variety of authors and subjects and genre (e.g. weather forecasts, adverts, political and scientific reports, readers' letters etc.). The underlying paronym corpus consists of the following texts in more detail (see Figure 2):

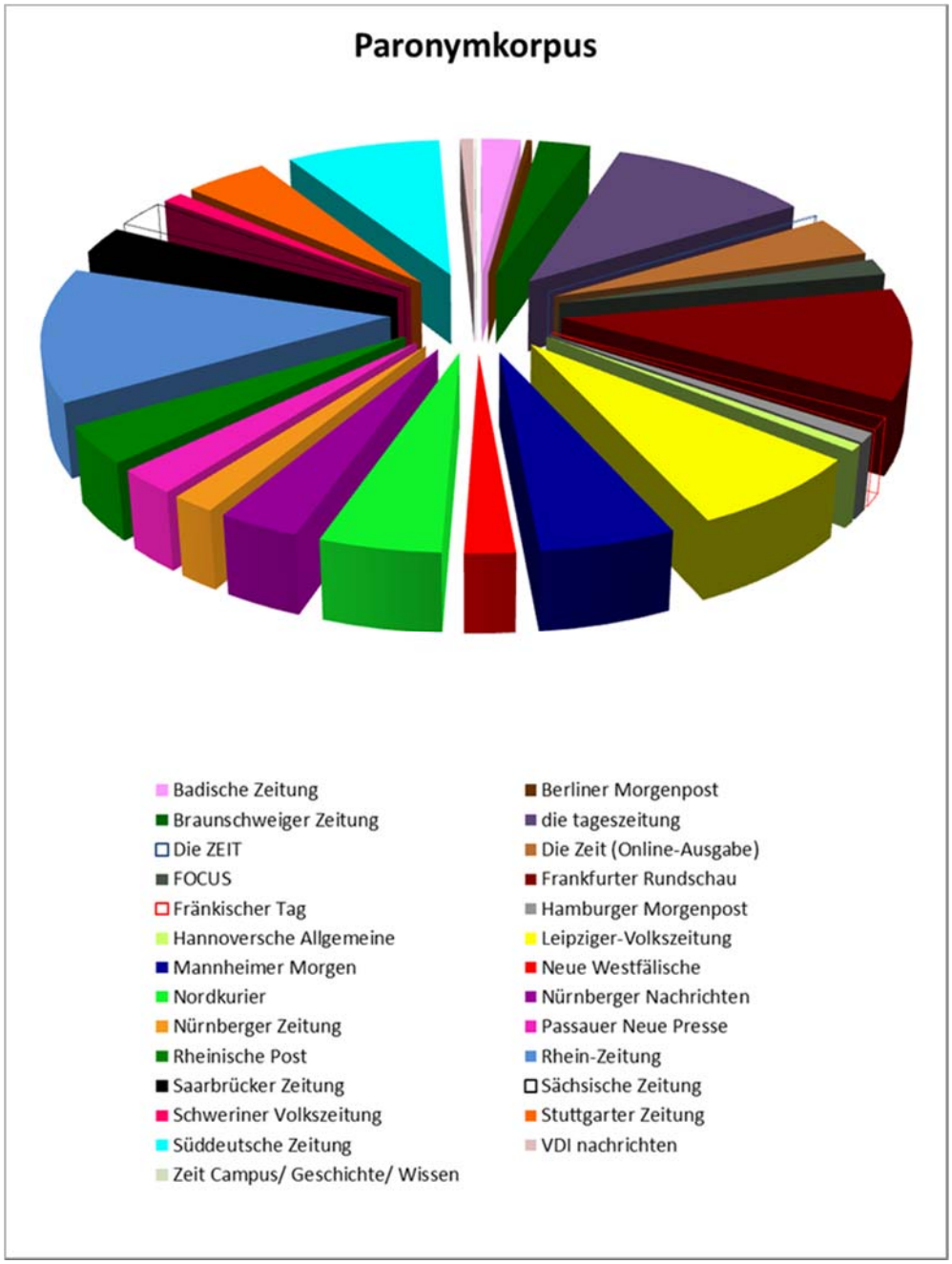


Figure 2: Percentage of newspapers in the Paronymkorpus

4.2 FOLK – the Research and Teaching Corpus of Spoken German

In a further step, we will look at technical terms and easily confused pairs in spoken data as a lexical database for expert communication and the Datenbank für Gesprochenes Deutsch (DGD-IDS 2012-2017) (Database of Spoken German) as a resource for spoken communication. Specifically, FOLK, the Research and Teaching Corpus of Spoken German, which is part of the DGD will be used for our linguistic

research and lexicographic investigations (cf. Stift & Schmidt, 2014; Schmidt, 2016: 398). FOLK is a large corpus of spontaneous verbal interactions in German (Schmidt, 2016: 396–397), containing a growing number of TV-interviews and conversations. As Schmidt (2016: 397) points out, FOLK

> "i. covers a broad range of interaction types in private, institutional and public settings;
>
> ii. is sufficiently large and diverse and of sufficient quality to support different qualitative and quantitative research approaches;
>
> iii. is transcribed, annotated and made accessible according to current technological standards;
>
> iv. is available to the scientific community on a sound legal basis and without unnecessary restrictions of usage." (Schmidt, 2016: 397)

Another reason for using FOLK is that it is a balanced corpus. Schmidt writes: "FOLK also attempts to control for some secondary variables, like regional variation, sex and age of speakers, in order to achieve a balanced corpus" (Schmidt, 2016: 398). FOLK currently contains data from 259 different conversations. This makes 202 recorded hours and 1.95 million tokens. (DGD: New version of DGD, http://www1.ids-mannheim.de/prag/artikelansicht/article/neue-version-der-dgd-3.html). Unfortunately, so far spoken and written data cannot be analysed using one and the same corpus tool since they are incorporated into different systems. As a consequence, results have to be individually interpreted and their underlying data need to be explicitly mentioned in order to relate findings to their source of information. Hopefully, the corpus systems of the next generation will be able to process both types of data.

4.3 Wikipedia Corpus

In a final step, we will additionally use the German Wikipedia Corpus[8] (hosted at the Institute for the German Language) for analysing the use of paronyms in computer-mediated communication (CMC). It is through the research of paronyms in a third textual variety that our findings can cover a larger spectrum of the German language than would be possible by looking at written corpora only. Margaretha & Lüngen (2014) describe Wikipedia as a large and rich online encyclopaedia that covers an unbelievably wide range of subjects including history, sport, arts and culture in articles and talk pages (discussions). As a language repository, Wikipedia provides a wealth of multilingual natural language data, also useful for the analysis of knowledge concepts and ontological categories. Since the content of Wikipedia has not been

[8] Available under http://www.ids-mannheim.de/cosmas2/.

written by a single author, but collaboratively by many users, it is particularly interesting for the study of computer-mediated communication (CMC) (Margaretha & Lüngen, 2014: 59), as aspects of dialog and mediation need to be considered. Of particular importance, might be the Version Control System (VCS) for documenting the various versions of an entry, including editorial comments and remarks.

Analytical relevance is given, as this kind of corpus data gives us the opportunity to analyse CMC language data spontaneously and dialogically. The Wikipedia corpora are also available as a virtual corpus in the COSMAS II corpus search and analysis system. Currently, only research of written texts is being conducted; this will be followed by further investigations of spoken data and analyses of Internet texts in the following years. The findings will be documented as part of the dictionary in different sections.

5. Paronym Categorisation and Classification Procedures

At the moment, our paronym classification is solely based on written corpora and it only relates to analyses of roughly a hundred paronym pairs. Needless to say, it cannot lead to a sufficient classification model but has already provided us with valuable insights into functions in thematic domain, discourse and style, text types, and degrees of semantic similarity or contrast of easily confused words. It is expected that in the future we might be able to come up with a detailed terminology covering paronyms from different angles.

A closer look at the different communicative and discursive functions of paronyms has so far suggested the following cases:

i. general (non-technical) paronyms with some conceptual overlap but individual constructional preferences, e.g. *praktisch / praktikabel* (*practical*), *nötig / notwendig / notwendigerweise* (*necessary / necessarily*),

ii. discourse-identifying word pairs, i.e. paronyms strictly determined through specific (critical) discourse, e.g. *national / nationalistisch* (*national / nationalistic*) in political discourse; *unehelich / nichtehelich* (*illegitimate / out of wedlock*) in official language discourse. The wrong choice between them can lead to politically incorrect use,

iii. pairs with different connotations with the tendency to be misused more frequently in spoken conversations, e.g. *bäuerlich / bäurisch* (*rural / peasant*), *weiblich / weibisch* (*feminine / effeminate*); one item has a neutral connotation while the other is of negative pragmatic value,

iv. opposites denoting similar concepts but with concrete contrary specifications, e.g. *konkav / konvex* (*concave / convex*), *Stalagmit / Stalaktit* (*stalagmite /* stalactite); users are usually aware of a distinction but lack factual knowledge

in specific situations,

v. paronyms with strong similarities in spelling but no semantic closeness, e.g. *ethisch / ethnisch* (*ethnic / ethical*); There is no overlap on the designated concept and confusion leads to clear mistakes,

vi. pairs with different syntactic functions, e.g. *fraglich / fragwürdig* (*questionable / dubious*); there are restrictions of grammatical usage for one member of a pair, such as adverbial, attribute or predicative role of adjectives,

vii. synonyms which specifically occur in different thematic domains, e.g. *sportlich / sportiv* (*athletic / sporty*); these are identical in meaning but are preferably used in different subjects,

viii. pairs with a very different distribution and frequency pattern, e.g. *Adaption / Adaptation* (*adaption / adaptation*), *herzlich / herzig* (*warm, lovingly / cute, heart-shaped*).

Taking the class of thematically related synonyms (vii) as an example, the differences between the adjectives *sportlich / sportiv* can be summarised as follows: Generally, both denote a person as physically fit, healthy and athletic. Hence, they can be used synonymously. Still, they differ with respect to their collocates.

Collocates of *sportlich* are, for example, *Figur* (*figure*), *Fitness* (*fitness*), *Statur* (*stature*), *Mann* (*man*), *Täter* (*culprit*), *Pensionär* (*pensioner*) (all of which refer to people and their appearances). Contexts in which *sportlich* occurs together with these collocates are predominantly found in police reports, illustrating the thematic domain of descriptions of criminal offenders (see examples 1, 2 and 3) [9]:

1. *Ein **Täter** soll 18 bis 20 Jahre alt und 1,65 Meter groß sein. Er soll eine **sportliche** muskulöse **Figur** und kurze schwarze leicht gelockte Haare haben. Bekleidet war er mit weißem T-Shirt, dunklen Jeans und weißen Schuhen.* (Frankfurter Rundschau, 29.05.2007, S. 36)

2. *Nach übereinstimmenden Aussagen mehrerer Zeugen ist er 20 bis 22 Jahre alt, 1,80 Meter groß, hat kurze Haare und eine **sportlich**, kräftige **Statur**. Bekleidet war er mit schwarz-weiß karierten Bermudashorts, dunkelblauem T-Shirt und Basecap.* (Leipziger-Volkszeitung, 31.05.2014, S. 19)

3. *Freitagvormittag sah Schiefer zufällig, wie ein Einbrecher in das Haus seines Sohnes auf der anderen Seite der Gustav-Mahler-Straße einstieg. Seine Schwiegertochter mit ihrer kleinen Tochter war glücklicherweise nicht mehr im Haus, stellte er nach einer Schrecksekunde mit Blick auf den Parkplatz*

[9] The examples are taken from the Paronymkorpus.

fest. Der ***sportliche Pensionär*** *alarmierte die Polizei über Handy, bewaffnete sich mit einem Golfschläger und filmte das Haus von der anderen Straßenseite aus.*(Rheinische Post, 16.11.2006, Diebe bei Einbruch gefilmt)

Collocates of *sportiv* are, for example, *Typ* (*type*), *Menschen* (*people*), *Erscheinung* (*appearance*), *Biker* (*biker*), *Models* (*models*), *Damen* (*ladies*), all of which refer to general denotations of humans. Frequently, these can be found in contexts of sports and health issues (see citations 4 and 5):

4. "*Fit for Life*" *lautet das Motto zweier Grundlagenseminare, die "rz sporty" am Mittwoch, den 7. bzw. 14. Februar, zwischen 18 und 21 Uhr im RZ-Haus in Koblenz veranstaltet. Sportmediziner Prof. Dr. Peter Billigmann und die Diplom-Ernährungsberaterin Birgit Binninger-Heid vermitteln dabei Ernährungstipps für* ***sportive Menschen****. Folgende Themenkomplexe werden behandelt: Weg mit dem Winterspeck - wie nehme ich gesund ab; Fitnesssport und Ernährung - zehn Prinzipien für Essen und Trinken im Sport; Herzkraft und sportliche Leistung; Träge im Winter, topfit im Sommer - das wichtigste über das Immunsystem.* (Rhein-Zeitung, 25.01.2001, Die letzten Reste werden gesucht.)

5. *Petras ist nicht nur äußerlich, als notorischer Baseballkappenträger, der* ***sportive, kämpferische Typ****, er ist es auch in seinem Verständnis vom Theatermachen.* (Die Zeit, 12.10.2006, S. 53, Im Hagel der Stücke.)

As emphasised before, in a second step, spoken data and CMC data will be investigated in terms of paronym behavior. We have indicative evidence that specific aspects occur in different genres, styles and registers only or preferably. For instance:

i. There are paronyms that are more typically confused in spoken communication, e.g. *anvisieren / avisieren* (*to target / to notify*). In such situations, mistakes occur more frequently as "side effects" of spontaneous, unreflected speech. These are particularly revealing in terms of cognitive processing.

ii. There is a class of technical terms, i.e. paronyms originally from expert communication, mostly in written language, but also in spoken language, e.g. *Parodontose / Parodontitis* (*periodontosis / parodontitis*), *Arthrose / Arthritis* (*arthrosis / arthritis*). Confusion occurs in everyday language but not in technical terms. In public discourse, such terms are treated differently from medical contexts.

The list is neither complete nor homogeneous, but it accounts for some formal and linguistic elements. Without doubt, these distinctions and classes listed above are only a first sketch approaching the phenomenon of paronymy from a usage-based perspective. These first findings do not constitute a uniform classification but suggest that different linguistic aspects need to be taken into account and any adequate

approach to grouping paronyms requires a multi-layered, cross-classification. Hopefully, on the one hand, the features mentioned above enable us to find usage-based definitions and restrictions of paronyms, and on the other hand, they inform us about guiding principles of semantic change in authentic language in use.[10] In order to be able to identify classes and to be able to provide an adequate and comprehensive description of the various kinds of paronyms, it is, however, necessary to use different data sets for a more refined classification model: As results vary according to which corpus we use for our analysis, we distinguish between paronyms that are most frequent in written, spoken and CMC-language data.

Overall, the findings concerning the classification of paronyms are not only theoretically relevant. They help us to find criteria which reflect usage behavior, context-dependent functions and cognitive principles rather than formal, logical distinctive aspects isolated from contexts. As a result, information on their features as described here are implemented in the dictionary entries (or will be in the future) in different ways, e.g. through specific sections, guidewords, explicit reference in the paraphrase.

6. Conclusion

The focus of this paper was to present a first attempt to classify commonly confused words (so-called paronyms) in German by studying their communicative and discursive functions in written corpora. These unveil different categories compared to traditional models and principles. Paronyms have not been studied empirically in language use so far. Sound corpus-guided studies of paronyms show different meanings from traditional dictionaries; a contextual usage-based approach leads to different categories of classifications than structural accounts. Our categories of paronyms exemplify text-functional aspects with regard to contextual relations as illustrated by collocation constructions. These uncover complex semantic structures and relational networks and we are able see how paronyms behave differently in contextual patterns and discourse.

At the moment, the bases of our investigations are very large written corpora. In the future, additional text types and genres of written as well as spoken language (see section 4) will play a vital role in defining paronyms and in embedding the phenomenon into a larger semantic framework. This necessarily has to imply approaches to real language in use and a variety of registers for a more objective view on communication and language in general.

[10] Another interesting aspect of research implies the rivalry of paronyms and their mutual contextual as well as cognitive influence on each other.

7. References

Duden 9. Richtiges und gutes Deutsch (2007). *Wörterbuch der sprachlichen Zweifelsfälle*. Edited by P. Eisenberg, F. Münzberg & K. Kunkel-Razum (Dudenredaktion). 6. ed. Mannheim et al.: Dudenverlag.

Kupietz, M. & Lüngen, H. (2014). Recent Developments in DeReKo. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis (eds.) *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (LREC'14), Reykjavik, Iceland. European Language Resources Association (ELRA).

Lăzărescu, I. (1999). Die Paronymie als lexikalisches Phänomen und die Paronomasie als Stilfigur im Deutschen. Bukarest: Anima Verlag.

Margaretha, E. & Lüngen, H. (2014). Building Linguistic Corpora from Wikipedia Articles and Discussions. Available at: http://www.jlcl.org/2014_Heft2/3MargarethaLuengen.pdf.

Müller, W. (1973). *Leicht verwechselbare Wörter*. Duden-Taschenwörterbücher, Bd. 17. Mannheim: Bibliographisches Institut.

Pollmann, C. & Wolk, U. (2010). *Wörterbuch der verwechselten Wörter*. Stuttgart: Pons.

Schmidt, T. (2016). Good practices in the compilation of FOLK, the Research and Teaching Corpus of Spoken Language. In *International Journal of Corpus Linguistics* 21: 3 (2016), pp. 396–418.

Stift, U.-M., & Schmidt, T. (2014). Mündliche Korpora am IDS: Vom Deutschen Spracharchiv zur Datenbank für Gesprochenes Deutsch. In Institut für Deutsche Sprache (ed.) *Ansichten und Einsichten. 50 Jahre Institut für Deutsche Sprache*. Mannheim: Institut für Deutsche Sprache, pp. 360–375.

Schnörch, U. (2015). Wie viele Paronympaare gibt es eigentlich? Das Zusammenspiel aus korpuslinguistischen und redaktionellen Verfahren zur Ermittlung einer Paronymstichwortliste. *Sprachreport*, 4/2015, pp. 16–26.

Storjohann, P. (forthcoming): Cognitive descriptions in a corpus-based dictionary of German paronyms. In Anatol Stefanowitsch & Stefan Hartmann (eds.) *GSCL Yearbook 2017*. Berlin: de Gruyter.

Storjohann, P. & Schnörch, U. (2017). Sprachlicher Wandel der Gegenwart und seine Dokumentation in einem Wörterbuch. In M. Vachková, M. Šemelík & V. Kloudová (eds.) *Themenheft Lexikographie, Germanistica Pragensia, AUC Philologica*, 4/2016, pp. 133–172.

Websites:

Cosmas II. Accessed at: https://cosmas2.ids-mannheim.de/cosmas2-web/ (18 May 2017)

DGD-IDS (2012-2017). DGD. Datenbank für Gesprochenes Deutsch. Accessed at: http://dgd.ids-mannheim.de/dgd/pragdb.dgd_extern.welcome?v_session_id (17 May 2017)

DGD: New version of DGD. Available at: http://www1.ids-mannheim.de/prag/artikelansicht/article/neue-version-der-dgd-3.html (17 May 2017)

DEREKO: Deutsches Referenzkorpus. Available at: www1.ids-mannheim.de/kl/projekte/korpora/ (10 May 2017)

DeReWo: version derewo-v-ww-bll-320000g-2012-12-31-1.0. Download: http://www1.ids-mannheim.de/direktion/kl/projekte/methoden/derewo.html?L=1 (5 May 2017)

elexiko. Accessed at: http://www.owid.de/wb/elexiko/start.html (30 April 2017)

FOLK: Forschungs- und Lehrkorpus gesprochenes Deutsch. Available at: http://dgd.ids-mannheim.de (26 April 2017)

Paronyme – Dynamisch im Kontrast. Project description. Accessed at: http://www1.ids-mannheim.de/lexik/paronymwoerterbuch/dasparonymkorpus.html (1 May 2017)

PARONYM-Korpus. Available at: https://cosmas2.ids-mannheim.de/cosmas2-web/faces/investigation/corpus.xhtml.

WIKIPEDIA-Korpus (2013/2015). Available at: https://cosmas2.ids-mannheim.de/cosmas2-web/faces/investigation/archive.xhtml.

A Limburgish Corpus Dictionary: Digital Solutions for the Lexicography of a Non-standardized Regional Language

Yuri Michielsen-Tallman[1], Ligeia Lugli[2], Michael Schuler[3]

[1]Maastricht University, FASoS, Grote Gracht 90, 6211 SZ Maastricht.

[2]King's College London, Virginia Woolf Building, 22 Kingsway, London, WC2B 6LE.

[3]Unaffiliated, now at Google Inc.

E-mail: j.michielsen@maastrichtuniversity.nl, ligeia.lugli@kcl.ac.uk, masmas@google.com

Abstract

This paper presents the Limburgish Corpus Dictionary (LCD), a newly-started project at Maastricht University that aims to create an online corpus and dictionary of Limburgish from scratch.

Limburgish comprises a set of West Germanic dialects spoken in the Dutch and Belgian provinces of Limburg. Due to a variety of factors, including its history and geographic spread, Limburgish exhibits an extremely high degree of spelling variation. In conformity with current policies, our dictionary strives to give equal visibility to all local dialects and variant spellings, with a view to enabling users to search for and retrieve lexical entries using their preferred spelling of a lemma.

After a brief outline of the Limburgish language, the history of writing in Limburgish, and Limburgish lexicography, this paper presents the dynamic and multi-layered entry structure that we have devised to represent information about spelling variation. Subsequently, it discusses how our lexicographic model impacts the way we prepare our corpus for analysis. It concludes with a description of our tentative corpus-processing pipeline and the results of some initial NLP software testing.

Keywords: minority language; Limburgish; spelling variation; normalization; lemmatization

1. Introduction

This paper introduces the Limburgish Corpus Dictionary (LCD), a project recently started at Maastricht University in cooperation with the *Meertens Instituut* and other partners. Much befitting the eLex theme of this year, this project starts completely from scratch. Despite a long history of Limburgish lexicography, the LCD will be the first lexicographic resource of its kind. It is the first dictionary to be derived from a digitized corpus of texts written in Limburgish and the first to include all spelling variations found in varieties of Limburgish. This requires unprecedented efforts and raises new challenges. In this paper, we focus only on those efforts and challenges that stem from the lack of an agreed upon standard written variety and the consequent abundance of co-existing spelling variants for every lemma.

The paper comprises four parts. First, it opens with a brief overview of Limburgish, its writing and spelling practices, and lexicography history. It proceeds to describe a model to represent different dialectal varieties in a single online dictionary.

Subsequently, it outlines how spelling variation complicates corpus processing and describes a set of heuristics and computational tools available to address these issues. Finally, it delves into future lines of development, especially regarding a possible NLP software pipeline.

2. Limburgish

2.1 Limburgish language

Limburgish refers to a language variety that is part of a continuum of West Germanic dialects, traditionally referred to as East Low Franconian in Dutch and Flemish dialectology and South Low Franconian in German dialectology (Belemans, 2009: 29). Limburgish consists of several dialects that share fundamental common characteristics (Schutter & Hermans, 2013), are mutually intelligible (Leerssen et al., 1996), and exhibit linguistic variety (Draye, 2007: 15). Its demarcation is subject to debate, but in many definitions Limburgish refers to most, though not all, of the dialects spoken in the Dutch and Belgian provinces of Limburg and some adjacent areas in the German Rhineland region, delimited by the Ürdinger isogloss (*ik-ich*) and the Benrather isogloss (*maken-machen*) (Belemans, 2009: 14; Notten, 1988: 71). For the purposes of the Limburgish Corpus Dictionary (LCD) we will adhere to the demarcation of Limburgish as used by the *Woordenboek van de Limburgse Dialecten* (Dictionary of the Limburgish dialects)[1] and illustrated below in Figure 1.

Limburgish developed separately from other Low Franconian varieties. It has a different phonetic system, grammar, and vocabulary. Unlike other Low Franconian varieties it only marginally contributed to the development of standard Dutch (Opgenort, 2012; Leerssen et al., 1996). According to some measures, the dialects of Limburgish are further removed from standard Dutch than any dialect or other regional language in the Netherlands and the Dutch-speaking part of Belgium (Hoppenbrouwers & Hoppenbrouwers, 2001; van Hout & Münstermann, 1981). Moreover, strikingly different from Dutch, as part of a continuum of Low and Central Franconian tonal dialects, most Limburgish dialects exhibit binary tone contrast on long vowels and diphthongs (Boersma, 2013; Gussenhoven & Peters, 2008; Fournier et al., 2004).

In the Netherlands, since 1997, Limburgish has enjoyed some official recognition as a regional language according to Part II European Charter for Regional or Minority Languages (Swanenberg, 2013). This legal recognition applies to all dialects spoken in the province of Dutch Limburg. This includes the small Kleverland and Ripuarian dialect regions that under some definitions are viewed as part of respectively Brabantian-Dutch and High German dialects (see below Figure 1) (Belemans et al., 1998; Daan & Blok, 1969). As part of this recognition, at the regional level, the

[1] See Belemans et al. (1998) and Weijnen et al. (1983: 7-11, 22).

Dutch province of Limburg has established an advisory body *Raod veur 't Limburgs* (Council for Limburgish) to tend to Limburgish. However, this is not the case in Belgium and Germany, where Limburgish has no official status.

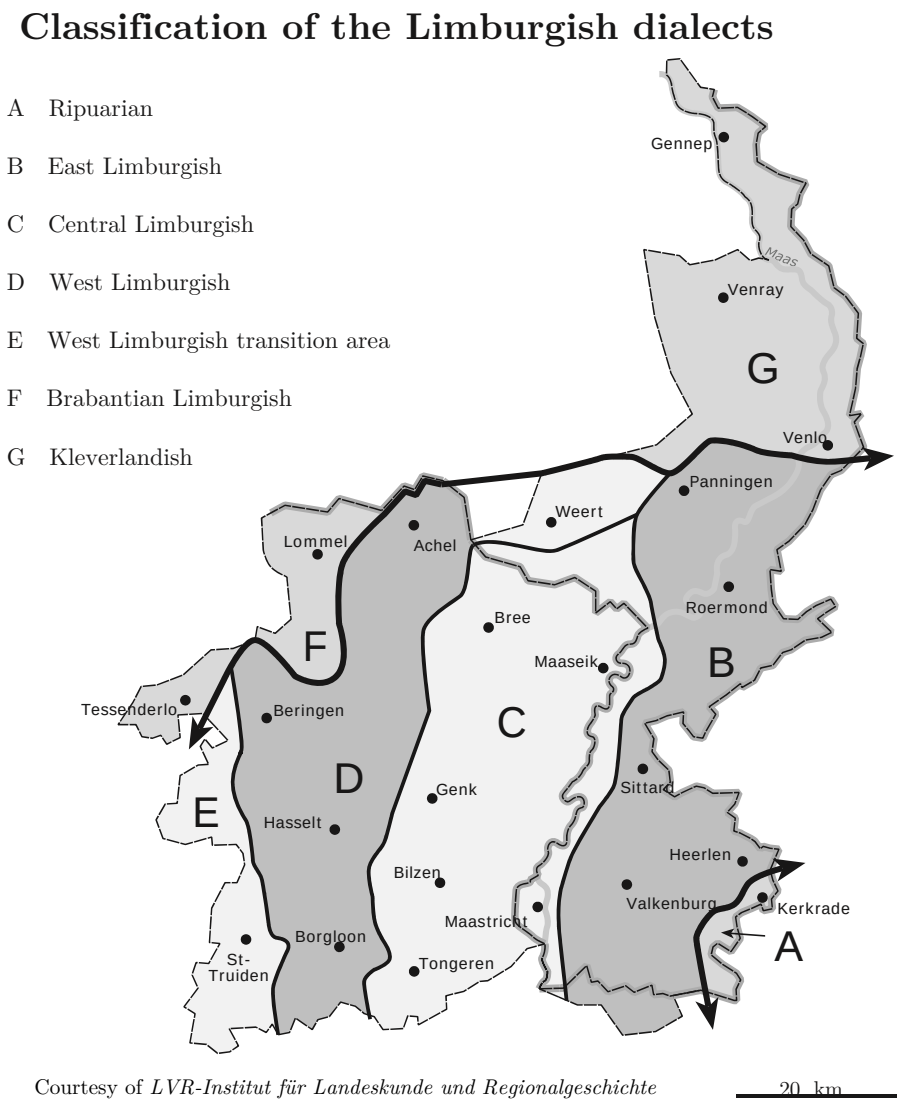


Figure 1: Map of the main Limburgish dialect areas

2.2 Written Limburgish

Since the LCD is a based on a diachronic written corpus (see 2.4 below), a brief history of writing in Limburgish against the backdrop of Limburg's history might be useful. Writing in Limburgish has a long history. The Wachtendonck Codex of around 900 CE contains the oldest known Limburgish fragment (Jongen, 2016: 25; Robinson, 1992: 205). During the Middle Ages, Limburgish was an important literary language (Tervooren, 2006) and was used as a language of government and administration (Willemyns, 2003; Moors, 1952). Wars fought in the territories of present-day Limburg during the 16th and 17th centuries led to increasing political

fragmentation, due to which either French, German or Dutch replaced Limburgish as a language of government (limburgs.org; Otten, 1977). As a result of economic and cultural decline, literary production stagnated (van Horen & van Horen-Verhoosel, 2016: 67). In 1795 the fragmented Limburgish territories were unified and incorporated by France as a *département*. Subsequently, in 1815, they were placed under Dutch control by the Congress of Vienna. During the Belgian uprising in 1830, Limburg seceded to become part of Belgium. In 1839, the east of Limburg was returned to the Netherlands, splitting the region into a Dutch and a Belgian province. For reasons that are unclear, at the end of the 18th century, writing in Limburgish slowly revived (Spronck, 1962: 436). From 1840 onwards, literary production started gathering pace (Spronck, 2016; Nissen, 1986), especially in literary societies in the urban centers of Dutch Limburg. In 1926 with the foundation of Veldeke, a Limburg-wide organization to promote the use of Limburgish, writing in Limburgish became more common practice (Spronck, 2016).

2.3 Spelling variation

Spelling variation is very much part of Limburgish writing. Possibly as a result of its past political fragmentation, Limburgish speakers strongly identify with their native locality and its dialect. Virtually all published (or online) texts are accompanied by an indication of the dialect that is used. This practice both testifies to and likely reinforces such identification. An attempt to unify the written standard faltered in the Limburgish parliament in 2000 (limburgs.org).

The official policy of the Council for Limburgish is to treat all dialects of Limburgish equally (Weusten et al., 2013; van Hout, 2007) and to support the current variation in spelling practices. To this end, in 2003, the Council for Limburgish created a normative orthography, which links graphemes and phonemes and can be used for writing in the different Limburgish dialects (Opgenort, 2012; Bakkes et al., 2003). This orthography is based on a succession of previous spelling guidelines created by Veldeke, the main regional language organization, since 1934 (Wolters, 2016), which in turn was influenced by the orthographic tradition that developed in the wake of the literary revival of the 19th century. Much, though not all, of the writing since 1934 is based on the Veldeke guidelines (limburgs.org). Yet, this does not ensure spelling homogeneity, and the result is a phonological and sometimes idiosyncratic spelling that reflects each writer's own dialectal pronunciation and spelling practices. An example of some of the regional spelling variation, based on local dictionary forms, is given in Table 1 and illustrated in Figure 2.

Hasselt	Tongeren	**Maastricht**	Weert	Maasbree	Thorn	Elsloo	Echt
stoan	stún	**stoon**	staon	staòn	staon	staon	staon

Venlo	Sittard	Roermond	Posterholt	Valkenburg	Simpelveld	Heerlen	Kerkrade
staon	Sjtaon	sjtaon	sjtaon	sjtaon sjtoon	sjtoa	sjtoa	sjtoa

Table 1: Representation of spelling variation of some Limburgish dialect-specific lemmas associated with the Maastricht lemma <stoon> [stʊ · ²n] 'to stand' taken from local dialect dictionaries of Belgian and Dutch Limburg.

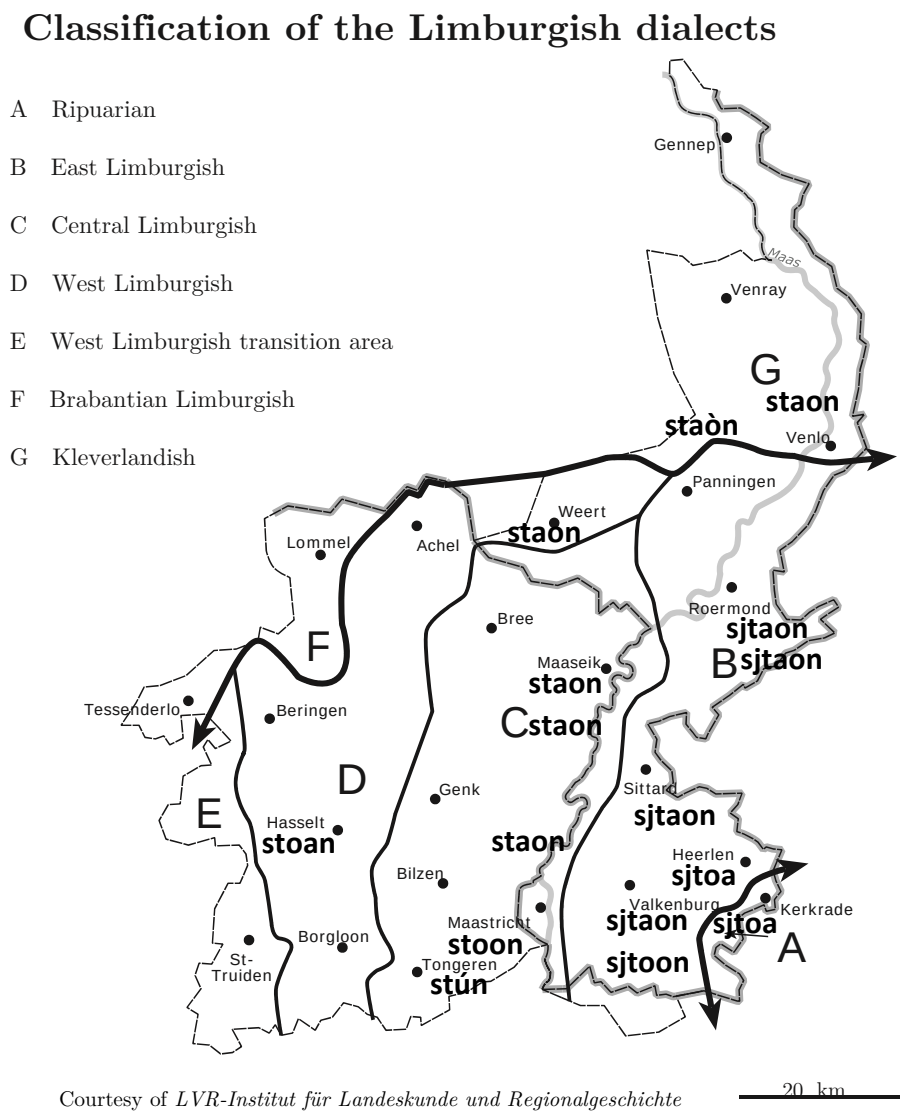


Figure 2: Illustration of spelling variation of some Limburgish dialect-specific lemmas associated with the Maastricht lemma <stoon> [stʊ · ²n] 'to stand' taken from local dialect dictionaries of Belgian and Dutch Limburg (lemma forms added to original table)

2.4 Limburgish lexicography

Glossaries of Limburgish dialects exist from the Middle Ages (Jongen, 2016: 25). Since the end of the 19th century around 80 dictionaries and glossaries of local

dialects of Limburgish have been created. These vary in size and the methodology used, but virtually all are bilingual to or from Dutch. For the Limburgish content, most adhere to the spelling guidelines mentioned above, applied to the local variant. A few are online[2].

So far, only three lexicographic projects have covered all dialects in Limburg; the *Woordenboek van de Limburgse Dialecten,* the *Taal van de Maas*, and the Limburgish Academy dictionaries. The *Woordenboek van de Limburgse Dialecten* (Dictionary of the Limburgish dialects), completed in 2008, is a thematically-organized dictionary created by the universities of Nijmegen and Leuven. Sources for the dictionary were questionnaires, dictionaries of local dialects and other sources that included research focused on the lexicon. The spelling of the Limburgish lexicon is adapted to standard Dutch, whereby the original Limburgish is spelled according to Dutch phonology and orthography. An online version is available[3]. In the 1990s, the *Werkgroup Algemeen Geschreven Limburgs* (working group General Written Limburgish) created the *Taal van de Maas* (Language of the Meuse), a Dutch–Limburgish dictionary (Prikken, 1994). Its sources and the selection criteria for the Limburgish lexicon are unclear. A spelling system was developed that differed from traditional Limburgish spelling in that it was not based on phonology. An online version gives access to Dutch–Limburgish and Limburgish–Dutch word lists[4]. Finally, on the basis of written and online sources, the Limburgish Academy Foundation created two online dictionaries: a Limburgish–Dutch and a Limburgish–English dictionary. The spelling of Limburgish words is mostly based on the 2003 normative orthography of the Council for Limburgish applied to phonology of the Maastricht dialect. These dictionaries are only available online[5].

The LCD will be the first corpus-driven dictionary of Limburgish. It is based on ideally every extant sample of written, transcribed from spoken, internet, and social media text in every dialect from both provinces of Limburg and the Limburgish territories that preceded their existence[6]. The corpus will be diachronic, encompassing texts from about 1775 until the present, though most texts date from 1926 until the present. The LCD will be a free online dictionary. In line with Limburger writing practices and the official position of the Council for Limburgish, the LCD will strive to give equal representation to all dialectal varieties in Dutch Limburg, as well as Belgian Limburg, and the resultant spelling variation. This has some important lexicographical implications.

[2] See for Gronsveld woordenboek.gronsveld.com, Maastricht mestreechtertaol.nl, and Thorn limburgsewoordenboeken.nl.

[3] See e-wld.nl.

[4] See limburghuis.nl.

[5] See limburgs.org.

[6] For a complete demarcation of Limburgish we use the definition of the Dictionary of the Limburgish dialects (see 2.1 above).

3. Requirements for a Limburgish Corpus Dictionary

Spelling variation, Limburger writing practices, as well as language policy, all impact our project on the level of the designs of both corpus and dictionary. The lexicographer needs to be able to retrieve all instances of a lemma in the corpus, determine how they are distributed, and identify whether the variation is purely formal or somehow correlates with semantic variation. This calls for processing our corpus in a way that clusters all spelling variation under a single lemma form. The users of our dictionary need to retrieve an entry for a word, regardless of which local spelling they enter in the search box. This would necessitate the possibility of displaying headwords in all the local spelling variations to allow users to see 'their' preferred spelling in the online dictionary.

The LCD is aimed at a range of audiences spanning from general Limburgish-speaking users to linguists. Its primary focus is on non-specialist Limburgish users who will be interested in referencing only limited information in each entry. To facilitate perusal of the dictionary on the part of such non-specialist users, search results will only display the lemma in the user's preferred spelling. In addition to that spelling, the dictionary entry will also display the most frequent spelling of that lemma in the corpus (for problems related to calculating the relative frequency of different spellings of a lemma see Section 4.3 below) to inform the user of a more general spelling of the term throughout Limburgish (see Figure 3).

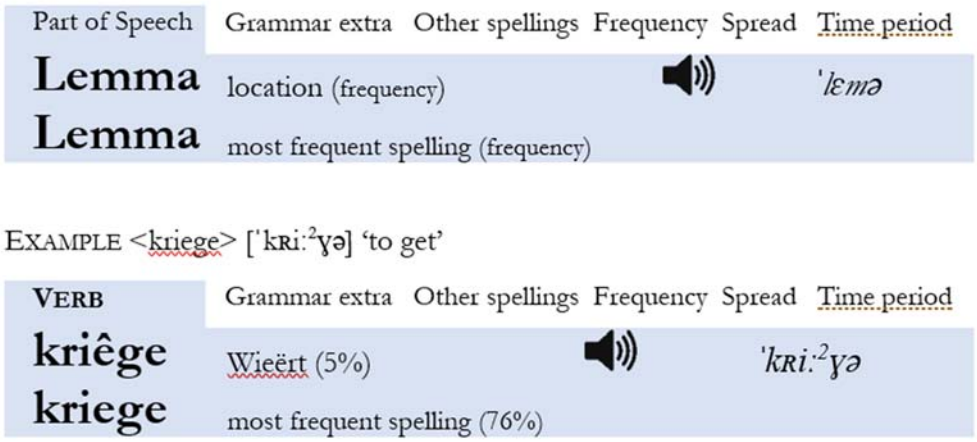


Figure 3: Representation of the display of a lemma in the online dictionary of the user's spelling and the most frequent spelling of that lemma

Users interested in accessing more information about a lemma will be able, by clicking on a tab, to access all spelling variations of a Limburgish lemma as attested by the corpus, including the location[7] where this variant is found and its frequency in the corpus (see Figure 4).

[7] Based on authors' practice in indicating the dialect of a written text, we assign a location with a Kloeke code, a location code commonly used in Netherlandic dialectology: meertens.knaw.nl/kloeke.

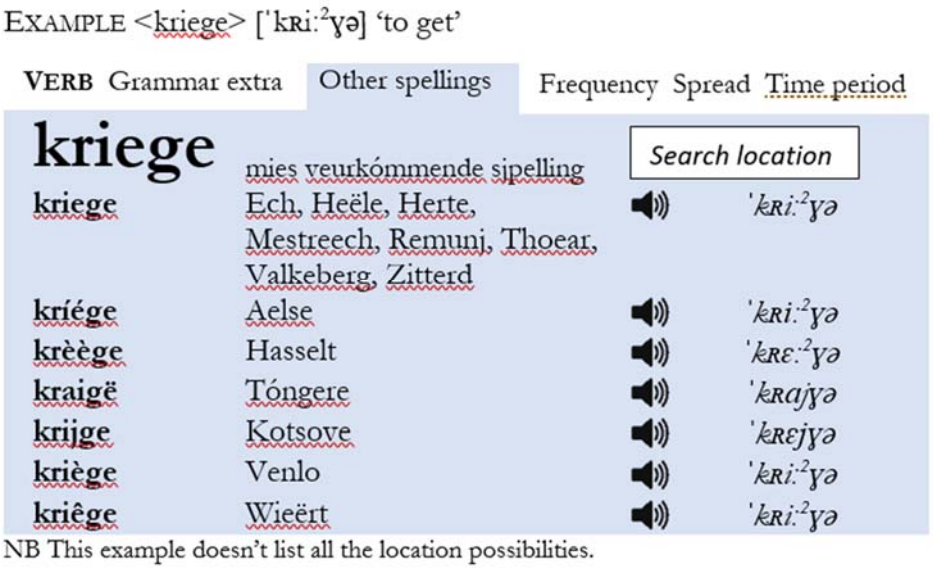


Figure 4: Representation of the display of spelling variety of a lemma in the online dictionary

Two further viewing modalities will be available to access information about the geographic spread of a lemma throughout Limburg as attested in the corpus (see Figure 5) and a diachronic table indicating the time period of a lemma (see Figure 6).

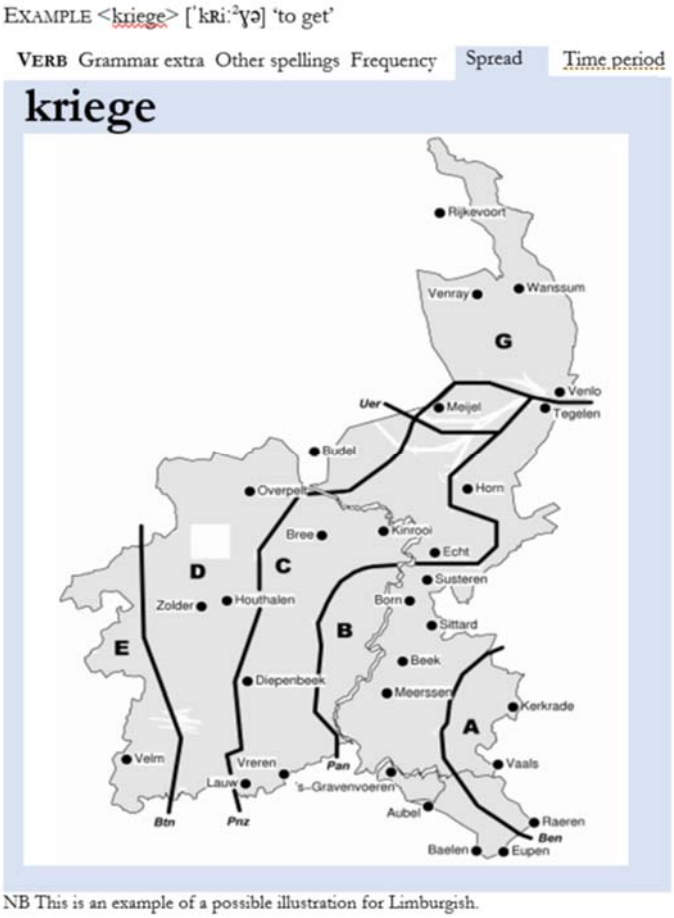


Figure 5: Representation of the geographic spread of a lemma in the online dictionary

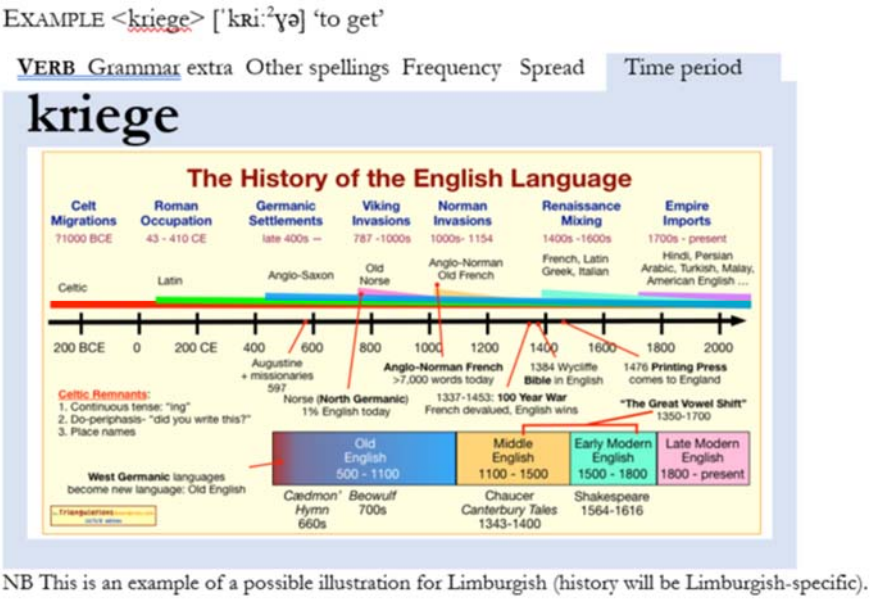


Figure 6: Representation of a possible display of the time period of a lemma in the online dictionary

Finally, the online dictionary will provide a 'concordance feature' where lexicographers and linguists, after log-in, will have direct access to the corpus (see Figure 7). This feature will be required to portray Limburgish texts in their original spellings.

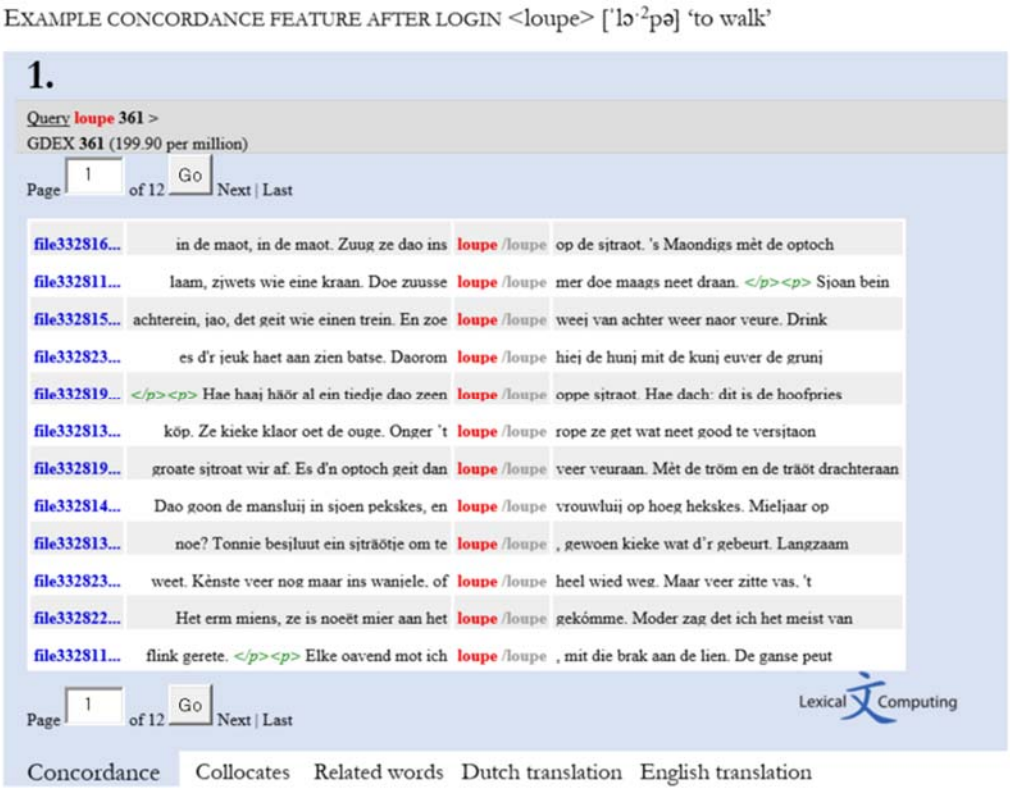


Figure 7: Representation of the 'concordance feature' for the lemma <loupe> [ˈlɔˑˀpə] 'to walk' in the online dictionary

To enable this entry structure in the dictionary and to allow lexicographers to retrieve and analyze all the relevant information about spelling variation, we outline the following considerations to ensure that our corpus is adequately processed.

4. NLP tools and Limburgish spelling variation

4.1 NLP tools and spelling variation

NLP tools have mostly been developed to process standardized languages and are not designed to deal with languages rich in spelling variation. Several NLP tools have been developed to process spelling variation, especially for historical corpora (see, e.g., van Halteren & Rem, 2013). The main pathway has been to apply a preprocessing tool before lemmatizers or Part of Speech (PoS)-taggers to normalize all orthographic variants of a token to a single spelling (Barteld et al., 2016). This normalization leads to more accurate processing in subsequent NLP tools, (Hendrickx & Marquilha, 2011). This practice presumes the existence of a standardized language that can be used for normalization. For standardized languages, unary normalization of diachronic corpora is possible, but has also proven problematic (Archer et al., 2015). For a non-standardized contemporary language like Limburgish, the issues are more complex. We will first outline some general issues pertaining to corpus normalization and lemmatization that have arisen in our project, and we will then describe a tentative processing pipeline and the result of some initial software testing.

4.2 Normalization for spelling variation in Limburgish

Our corpus exhibits both diachronic and synchronic spelling variation. Its diachronic and multi-dialectal nature, combined with idiosyncratic spellings and the lack of an agreed upon written standard, lead to an extremely high degree of spelling variation in a Limburgish corpus. This problem is by no means unique to this project. It has indeed already been treated effectively within several other projects, mostly of a historical nature, where the texts were normalized to a single standardized variety of the language, typically the contemporary form of the language[8].

In our project, however, the policy of treating all dialectal varieties equally adds a layer of complexity to the task of corpus normalization. The rationale for text-normalization is that in other cases it facilitates information retrieval because the language to which the text is normalized is more standardized and more widely accessible than the original. In the case of Limburgish, however, we face a multitude of similarly non-standardized varieties, none of which is more universally accessible than the others.

[8] For a survey of technical approaches used for normalizing historical texts see e.g. Barteldet al. (2016); Archer et al. (2015); Piotrowsky (2012: 74ff); Pilz et al. (2008).

To bypass this difficulty, we initially considered normalizing the Limburgish corpus to Dutch. *Prima facie*, this would seem like a good solution. Dutch is a standardized language and it is known to all Limburgish speakers in the Netherlands and Belgium. It would be relatively easy to find Limburgish staff able to supervise the semi-automatic normalization process from any Limburgish variety into Dutch. Despite these undeniable advantages, we discarded this solution, as introducing Dutch in a Limburgish corpus would have two major drawbacks. First, it would effectively amount to translating the corpus into another language and possibly obfuscate features peculiar to Limburgish. Second, it would rely on an assumption of extreme lexical similarity between Dutch and Limburgish, which a study of the corpus may or may not confirm.

To avoid embedding such assumptions in the design of our corpus, we opted for an alternative strategy. We decided to pick one of the Limburgish varieties as a target for normalization. This was done with the understanding that this would not affect the way other varieties will be represented in the dictionary, but would only facilitate information retrieval in the corpus, mostly for the use of researchers and lexicographers working on the dictionary. Since the largest single-dialect database available to us is the dictionary of the Limburgish Academy Foundation[9], which is easily rendered into contemporary Maastricht-Limburgish, we decided to normalize to the contemporary spelling of the Maastricht dialect. In those cases, where no corresponding Maastricht form exists, a pseudo-Maastricht form will be created on the basis of regular inter-dialectal phonological transformation[10]. To distinguish it from the Maastricht forms attested in the corpus, such pseudo-Maastricht renderings of other dialects will be preceded by an asterisk (*) (see below Table 2).

Elsloo	Roermond	Sittard	Thorn	Valkenburg	Venlo	Weert	Maastricht
spóéze	sjpoeze	sjpoeze	spoeze	sjpoeze	spoeze	spoeze	*spoeze

Table 2: Example of normalization to a pseudo-Maastricht form.

These normalized Maastricht forms will then be added alongside the original dialectal forms, including cases in which the dialectal form is in an idiosyncratic or historical spelling. In the case of an idiosyncratic or historical Maastricht spelling, the form will be paired with a normalized form based on contemporary Maastricht spelling.

[9] See limburgs.org.

[10] Cf. the creation of pseudo-modern forms for historical forms that do not exist anymore in modern languages (e.g. for historical Dutch see Brugman et al., 2016; van Halteren & Rem, 2013).

4.3 Lemmatization and dialect-specific lemma forms

Following our normalization strategy, we will lemmatize the corpus to Maastricht-Limburgish and then tag it for part of speech (PoS-tag) on the basis of grammatical information derived from a Maastricht-Limburgish dictionary. It is important to note that the original tokens will be retained alongside the normalized forms, so that the PoS-tags will be associated with both the Maastricht and the original form (see Table 3). This will allow lexicographers and researchers to analyze the different spellings associated with each lemma and derive dialect-specific lemma forms (see above Table 1 for an example of dialect-specific lemma forms). These dialect-specific forms will eventually feature as headwords in the LCD and enable users to search for and retrieve their preferred spelling of any Limburgish word included in the dictionary (see above Table 1). They will also serve as an indicator of the frequency and distribution of different spelling of a word across Limburg.

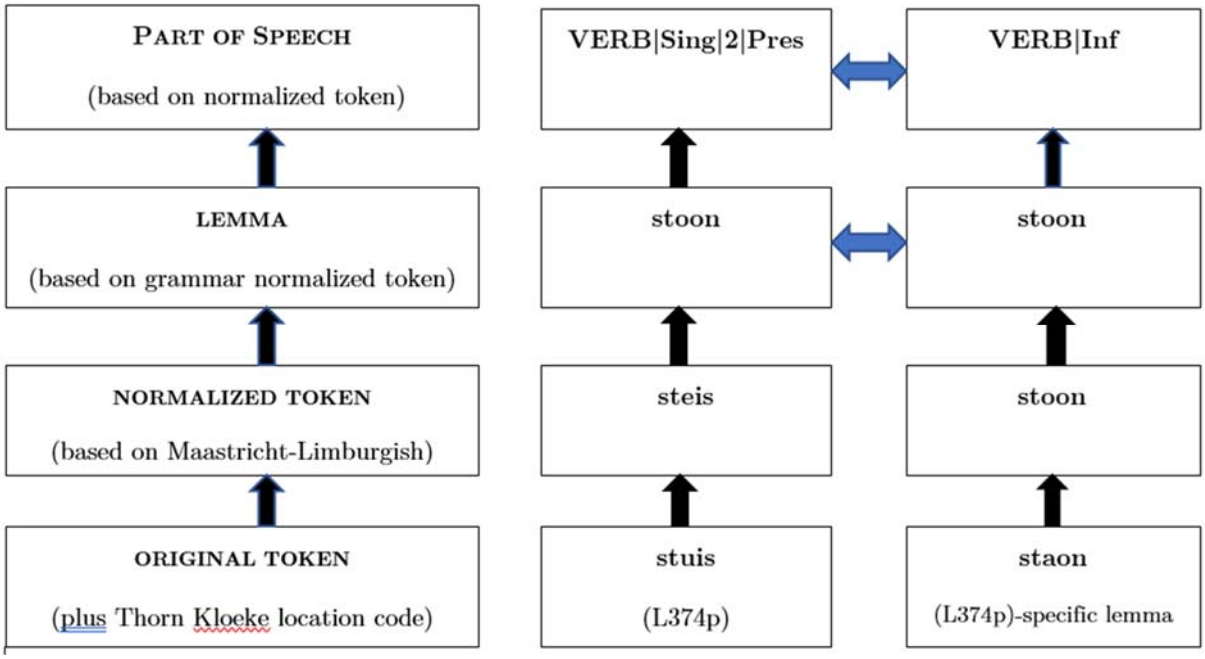


Table 3: General form (left column) and example of normalization, lemmatization, and PoS tagging of a conjugated form (middle) found in a specific dialect and the connection pathway to its dialect-specific lemma (right).

Given the importance of dialect-specific lemma-forms in this project, we initially intended to perform a double lemmatization and pair each token with both its dialect-specific lemma and the corresponding lemma in Maastricht-Limburgish. After much consideration we discarded this approach. In the rest of this section we outline the options we had initially favored and the rationale for choosing a different strategy. We hope that our experience may benefit other projects dealing with the lexicographic representation of regional spelling variation in a corpus.

Initially, we considered relying on existent lexicography and location metadata to pair each token with the corresponding lemma form recorded in dictionaries of the relevant token. This approach presupposes that all words associated with a certain location are amenable to the same lemmatized form, thus not allowing for variant spellings within the dialect. We discarded this idea in favor of a corpus-driven approach which would allow us to derive lemma-forms directly from the corpus and thus account for intra-dialect variation. To this end, we initially aimed to pair each token with a corpus-derived lemma-form that would match the regional spelling of the token. We soon realized that this model, too, was not viable, because it assumes a morphological correspondence between a token and its lemma form. Unfortunately, several Limburgish verbs violate this assumption. For example, in the dialect of Valkenburg the indicative second-person singular of the verb 'to stand' is <sjteis>. At the current state of research, there is no morphological transformation rule that determines whether this form should be matched to <sjtaon> or <sjtoon>, both of which are possible spellings of the infinitive of this verb in this location (see Tables 2 and 3 above). It is possible that predictable transformation patterns for these verbs will emerge from a study of our corpus and make automated lemmatization to a dialect-specific lemma form possible. In the meantime, we will have to dispense with dialect-specific lemmatization and derive lexicographic information on dialect-specific lemma forms only from tokens morphologically identical to the lemma[11]. Thus, the frequency of the Valkenburg lemma form <sjtaon> as opposed to the Maastricht form <stoon> will be calculated on the basis of tokens spelled <sjtaon> only (i.e. the infinitive and indicative first and third person plural), and will not be derived from other conjugated forms. It remains to be determined whether the tokens morphologically identical to the lemma form will constitute a sufficient and reliable indicator of the overall frequency and distribution of a spelling variant. Information about the frequency and distribution of the spelling of other selected conjugated forms (e.g. indicative second person singular or sample past tense forms) may be added to provide a more complete representation of spelling variation across Limburg.

[11] The full verbal paradigm for this verb in Valkenburg-Limburgish based on the local dictionary is the following.

Present tense:		Past tense:	
1s <sjtaon>	1p <sjtoon> / <sjtaon>	1s <sjtóng> / <sjting>	1p <sjtónge> / <sjtinge>
2s <sjteis>	2p <sjtaot>	2s <sjtóngs>	2p <sjtóngt>
3s <sjteit>	3p <sjtoon> / <sjtaon>	3s <sjtóng>	3p <sjtónge> / <sjtinge>

Pp <gesjtange>. Imperative s <sjtank>, p <sjtaot>.

4.4 Considering NLP tools for Limburgish spelling variation

Several software options have been identified for a tentative pipeline. VARD[12] is being considered as a spelling normalizer, Frog[13] for tokenization, lemmatization and PoS-tagging, and Sketch Engine[14] for corpus analysis. Dictionary writing software, such as TshwaneLex[15] and DPS from IDM[16], are also being considered, but will not be further discussed in this article.

At the time of writing this article, testing is still in a very preliminary stage. Only some general comments about the usefulness of VARD and Frog to our project can be made, whereby the focus will be on Limburgish spelling variation.

4.4.1 VARD

VARD was initially built to deal with spelling variation in Early Modern English (Baron & Rayson, 2009), but can potentially be re-trained for other languages[17]. VARD normalizes spelling by inserting a normalized lemma in the place of the spelling variant and retains the original form in an XML tag. VARD can be used in two ways: to manually standardize texts or to automatically standardize a set of texts or corpora (Baron & Rayson, 2009). VARD is a well-known tool and we will not elaborate on it further, except insofar as evaluating it as a potential option for our project.

Since we are still in the process of collecting our corpus, and Limburgish writing exhibits such a high degree of spelling variation, we do not yet know all the variants we will encounter. To gain some preliminary understanding of how much spelling variation we can expect to encounter in our project, we used VARD 2.5.4 for an initial assessment of variation. We first tested diachronic texts from the Maastricht dialect and subsequently synchronic texts in different spellings from the main Limburgish dialect areas for token recognition based solely on a curated word list *before training* VARD. Employing contemporary Maastricht-Limburgish spelling, we created a curated word list for VARD. It contains all parts of speech with inflected forms and consists of 85,731 unique words out of a total of 126,755 words, whereby duplicates existed for separate entries for polysemous words, verbal inflections of the past tense, homonyms and tonal opposites.#

[12] ucrel.lancs.ac.uk/vard/about/.

[13] languagemachines.github.io/frog/.

[14] sketchengine.co.uk.

[15] tshwanedje.com/tshwanelex/.

[16] idm.fr.

[17] For example for historical Dutch (Tjong Kim Sang, 2015), historical Portuguese (Reynaert et al., 2012), and historical German (Pilz et al., 2008).

We tested nine diachronic Maastricht-Limburgish text samples, including literary and Wikipedia texts, in their original spellings spanning the period of ca. 1775–2017. All texts were about 4500 tokens each, except three of the older texts which only have about half as many tokens each. As expected, the percentage of tokens recognized is well over 90% for texts written after 2010. The Wikipedia text samples, although from 2017, registered a recognition percentage of 78.5%. The lower token recognition is at least in part due to more idiosyncratic spellings, unknown proper nouns, foreign script, foreign tokens, more specialized compounds, and typos. For 20th-century texts, token recognition was 75–85%. Surprisingly, for 19th-century and older texts 45–60% of tokens are still recognized.

For the second test on the same Maastricht-Limburgish texts, replacement rules were added to VARD for spelling phenomena that affected most texts. Baron and Rayson (2009) indicate that VARD's user-defined list of letter replacement rules to compute alternative forms results in a significant increase in performance when automatically normalizing the corpus. These replacement rules for Maastricht-Limburgish included replacements for spelling changes made in 2004 and some 19th-century spelling peculiarities. Some of these rules will also benefit token recognition for many East Limburgish spellings, as these were closer to the Maastricht spelling before the 2004 spelling change. The results of the second test enhanced token recognition on average by about four percentage points, whereby texts from the 21st century gained 2.1%, 20th-century texts 5.7% and pre-1900 texts 4.6%.

Subsequently, we tested nine synchronic text samples from Wikipedia of about 2000 tokens each from all main Limburgish dialect areas[18]. We first used the same approach as mentioned above for the first VARD test. Token recognition for non-Maastricht spellings had a mean of 45%. The range was between 37% for the spelling of the Kerkrade Ripuarian dialect and 56% for the spelling of the Valkenburg East Limburgish dialect, which is geographically close to Maastricht. The results for the second test, with the replacement rules indicated above, resulted in a mean recognition of 51%. There was a range of about 40% for the spelling of the Kerkrade dialect to 62% for the spelling of the Valkenburg dialect.

The results from the diachronic Maastricht texts and the synchronic texts from all main dialect areas can be interpreted as indicators of the different levels of spelling

[18] These included the following dialects: Alken* (West Limburgish), Geleen (East Limburgish), Heerlen (East Limburgish Ripuarian transition area), Kerkrade* (Ripuarian), Montfort (East Limburgish), Ool (East Limburgish), Roermond (East Limburgish), Valkenburg (East Limburgish), Venlo (Mich Quarter transition area). Those with an asterisk (*) only had about half of the tokens.

variation in Limburgish. Considering the fact that this is a pre-trained version of VARD, these results are encouraging. We are contemplating to test and train VARD on a large corpus, which, according to Baron and Rayson (2009: 9), should allow it to better find and rank candidate equivalents for variants found in the remainder of the corpus.

We are considering the following steps regarding VARD. We shall start by training VARD and creating a Maastricht-Limburgish word list that is as extensive as possible. This is crucial, since we normalize to the contemporary spelling of this dialect. We will start with contemporary texts in the Maastricht spelling and subsequently process all Maastricht texts diachronically. Thereafter we intend to process texts in spellings from other dialects. On the basis of a mapping of Limburgish spelling variation we are examining whether to first process texts from dialects with spellings closest to Maastricht-Limburgish, followed by texts that in terms of spelling are progressively farther removed. For each dialect we will first normalize contemporary texts followed by increasingly older texts. Finally, on the basis of a mapping of Limburgish spelling variation, we will also determine whether to create a more extensive list of replacement rules. Some replacement rules to normalize to the Maastricht spelling are common to all dialectal spellings. For the spelling of some (groups of) dialects we might have to create a separate set of replacement rules. Depending on how extensive these separate replacement rules are for different (groups of) dialects we are contemplating training separate VARD applications.

One last issue we need to resolve is how to disambiguate homographs with different meanings in different dialects. Since Limburgish spelling is phonological and the normative spelling tags a grapheme with a particular phoneme, in some instances a word spelled according to the phonology of one dialect exists in another dialect, but with a different meaning. For example, <eur>, a possessive pronoun in Maastricht dialect meaning ‘your’ (singular polite form and plural), is the possessive pronoun for ‘her’ in the Venlo dialect. The Maastricht form for ‘her’, to which it has to be normalized, is <häör>. In a Venlo text, the Maastricht-trained VARD will recognize the token, but will not recognize that it is a variant spelling. Further experimentation will be required to optimize for the tool’s maximum effectiveness in normalizing the spelling variation in Limburgish texts. That optimization might include forking the normalization rules for individual dialects, or dialect areas, to force certain normalizations that are specific to only that dialect.

4.4.2 Frog

Frog is a Natural Language Processing suite originally developed for standard Dutch (Van den Bosch et al., 2007). It integrates a series of modules including a tokenizer, lemmatizer, morphological segmenter, and a PoS tagger. It also includes a named entity recognizer, phrase chunker, and dependency parser, but it is still to be

determined whether these tools are useful for our project. Frog is originally intended to work on modern standard Dutch, but has been used amongst others by the Nederlab project for historical Dutch (Brugman et al., 2016: 1279). It also contains Froggen, a trainer module part of Toad[19], that allows Frog to be trained for another language. However, Frog relies on a standard spelling to perform its analysis and is not equipped to deal with rich spelling variation. Normalizing Limburgish spelling variation by a pre-processing tool like VARD is therefore a prerequisite.

When evaluating Frog's usefulness for our pipeline, we first need to consider the output content and format from VARD. VARD can create two output formats. One is a version of the text with fully normalized spelling. Another version is the normalized text with XML tags, each encapsulating the original token along with the details of the normalization. As Frog cannot parse the VARD XML output out-of-the-box, we have a few pathways to experiment with to determine which is most compatible and without data loss.

Since Frog does not natively deal with spelling variation we have had to investigate options how to preserve the data of both the text in original spelling and the normalized version. One option is to configure Frog so that it can accept the pseudo-XML that VARD produces. This seems feasible as one of Frog's native formats is an XML format, namely FoLiA XML[20]. We are investigating whether it is possible to adapt Frog's parser to read VARD's pseudo-XML format. This, potentially, would allow us to preserve the connection between original token and normalized token through Frog's processing. Another, possibly simpler, option would be to insert an original token column to Frog's tab-delimited output of processed normalized tokens. This means that only normalized tokens are present in Frog's processing, but the connection to the original text is re-established in a secondarily, post-Frog processed output file.

We will not consider here in depth the steps in the pipeline after this point. However, one possible Frog output option is a tab-delimited text file, which Sketch Engine can process. The content of the Frog output will certainly include token, lemma, and PoS columns. Additional output from Frog may be included, depending on Sketch Engine's ability to parse the information and include it in its word sketches. Finally, header information, including items like location code, date, and author information, will be appended to the file to be read into Sketch Engine. At this point we will have attempted to preserve all the data from the source texts including all the tagging, and the corpus would be ready for analysis in Sketch Engine.

[19] github.com/LanguageMachines/toad/releases/tag/v0.3.

[20] For FoLiA XML see van Gompel and Reynaert (2013).

5. Conclusion

In this paper, we introduced a new project at Maastricht University for the creation of a Limburgish Corpus Dictionary (LCD). Limburgish spelling variation, diachronic spelling data, writing practices and language policy present us with the possibility to look for novel ways to process and display this non-standardized regional language. We first presented a model of how to display the spelling variation in Limburgish in an online dictionary, based on how Limburgers use their language and the policy to treat all dialects and spelling variation equally. For NLP processing purposes we then discussed the reasons to use the Maastricht-Limburgish variety as a normalizing standard. We also developed a set of heuristics to retrieve dialect-specific forms that will eventually feature as headwords in the LCD. This will enable users to search for and retrieve 'their' preferred spelling of any Limburgish headword included in the dictionary from the myriad of spellings that a Limburgish lemma can have. The dialect-specific forms will also serve as an indicator of the frequency and distribution of different spellings of a lemma across Limburg. We then discussed possible software options for a tentative pipeline and the steps we consider taking to further investigate their usefulness for our project. Our focus will now be on determining how the available NLP software options will allow us to execute our project in conformity with the lexicographic model we have developed for Limburgish. This will enable us to present Limburgish-speakers with a free online dictionary that represents their real language usage.

6. References

Archer, D., Kytö, M., Baron, A. & Rayson, P. (2015). Guidelines for normalising Early Modern English corpora: Decisions and justifications, *ICAME Journal*, 39, pp. 5-25.

Bakkes, P., Crompvoets, H., Notten, J. & Walraven, F. (2003). *Spelling 2003 voor de Limburgse dialecten.* Available at: http://www.limburgsedialecten.nl/download/spelling2003.pdf.

Baron, A. & Rayson, P. (2009). Automatic standardization of texts containing spelling variation. How much training data do you need? In M. Mahlberg, V. González-Díaz and C. Smith (eds.) *Proceedings of the Corpus Linguistics Conference*, CL2009, University of Liverpool, UK, 20-23 July 2009.

Barteld, F., Schröder, I. & Zinsmeister, H. (2016). Dealing with word-internal modification and spelling variation in data-driven lemmatization, *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pp. 52–62.

Belemans, R. (2009). *Van (Limburgse) dialecten naar Europees erkende streektaal en/of immaterieel cultureel erfgoed? De invloed van nationale taalpolitiek en van internationaal erfgoedbeleid op de perceptie van en op de overheidszorg voor endogene taalvariatie in Vlaanderen.* Doctoral dissertation Universiteit Leuven.

Belemans, R., Kruijsen, J. & van Keymeulen, J. (1998). Gebiedsindeling van de zuidelijk-Nederlandse dialecten. *Taal en Tongval*, 50, pp. 25-42.

Boersma, P. (2013). *The history of the Franconian tone contrast*. Available at: http://www.fon.hum.uva.nl/paul/papers/FranconianToneHistory68.pdf.

Brugman, H., Reynaert, M., van der Sijs, N., van Stipriaan, R., Tjong Kim Sang, E. & Van den Bosch, A. (2016). *Proceedings of LREC*, Portoroz: ELRA, pp. 1277-1281.

Daan, J. & Blok, D. (1969). Van Randstad tot Landrand. In *Bijdragen en Mededelingen der Dialectcommissie van de KNAW XXXVI*. Amsterdam: Noord-Hollandsche Uitgevers Maatschappij.

Draye, L. (2007). Enkele klank- en vormkenmerken van de Limburgse dialecten. In R. Keulen, T. van de Wijngaard, H. Crompvoets & F. Walraven (eds.). *Riek van Klank; Inleiding in de Limburgse dialecten*. Sittard: Veldeke Limburg, pp. 24-44.

Fournier, R., Verhoeven, J., Swerts, M. & Gussenhoven, C. (2004). Prosodic and segmental cues to the perception of grammatical number in two Limburgian dialects of Dutch. *Proceedings of the Speech Prosody 2004 Conference*, pp. 713-716.

Gussenhoven, C. (2007). De Limburgse tonen. In L. Heijenrath & S. Kroon (eds.), *Jaarboek 2006*. Roermond: Veldeke Limburg, pp. 21-32.

Gussenhoven, C. & Peters, J. (2008). De tonen van het Limburgs. In *Nederlandse Taalkunde*, 13, pp. 87-114.

Hendrickx I. & Marquilhas, R. (2011). From Old Texts to Modern Spellings: An Experiment in Automatic Normalisation, *Journal for Language Technology and Computational Linguistics*, 26(2), pp. 65-76.

Hoppenbrouwers C. & Hoppenbrouwers G. (2001). *De indeling van de Nederlandse streektalen, Dialecten van 156 steden en dorpen geklasseerd volgens de FFM*. Assen: van Gorcum.

Jongen, L. (2016). Van het begin tot 1500. Van geschreven naar gedrukte letters. In L. Spronck, B. van Melick & W. Kusters (eds.) (2016). *Geschiedenis van de literatuur in Limburg*. Nijmegen: Uitgeverij Vantilt, pp. 23-65.

Kestemont, M., Daelemans, W. & De Pauw, G. (2010). Weigh your words - memory-based lemmatization for Middle Dutch, *Literary and Linguistic Computing*, 25(3), pp. 287-301.

Leerssen, J.Th., Crompvoets, H., Walraven, F., Segers, J., Belemans, R., Bakkes, P. & Gillessen, L. (1996). *Verslag werkgroep erkenning Limburgs als streektaal*. Available at: http://jonckbloet.hum.uva.nl/leerssen/images/limburgs/adindex.html.

Opgenort, J. R. (2012). *Limburgse taal*. Available at: http://www.opgenort.nl/limburgse_taal.

Pilz, Th., Ernst-Gerlach, A., Kempken, S., Rayson, P. & Archer, D. (2008). The identification of spelling variants in English and German historical texts: manual or automatic?, *Literary and Linguistic Computing*, 23(1), pp. 65-72.

Priotrowski, M. (2012). *Natural Language Processing for Historical Texts*. San

Rafael: Morgan & Claypool.

Reynaert, M. (2014). TICCLops: Text-Induced Corpus Clean-up as online processing system, *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, Dublin, Ireland, August 23-29, pp. 52-56.

Reynaert, M., Hendrickx, I. & Marquilhas, R. (2012). Historical spelling normalization. A comparison of two statistical methods: TICCL and VARD2, *Proceedings of the Second Workshop on Annotation of Corpora for Research in the Humanities (ACRH-2)*, Lisbon, pp. 87-98.

Robinson, O.W. (1992). *Old English and its closest relatives; A survey of the earliest Germanic languages*. Stanford: University Press.

Schutter, de, G. & Hermans, B. (2013). The Limburg dialects: Grammatical properties. In F. Hinskens & J. Taeldeman (eds.) *Language and Space. An International Handbook of Linguistic Variation*, 3. Berlin/Boston: De Gruyter Mouton, pp. 356-377.

Souvay, G. & Pierrel, J.-M. (2009). LGeRM Lemmatisation des mots en Moyen Français. Traitement Automatique des Langues, *ATALA*, 50(2), pp. 149-172.

Spronck, L. (2018). *Van Sermoen tot Percessie. Het Maastrichts rond 1800* (to be published).

Spronck, L. (2016). 1793-1893. Verandering van het blikveld: verlies en winst. In L. Spronck, B. van Melick, & W. Kusters (eds.) (2016). *Geschiedenis van de literatuur in Limburg*. Nijmegen: Uitgeverij Vantilt, pp. 203-309.

Spronck, L. (1962). De Maastrichtse dialektliteratuur voor 1840. In Miscellanea Trajectensia; *Bijdragen tot de geschiedenis van Maastricht*, Werken LGOG nr. 4, Maastricht: LGOG, pp. 435-495.

Spronck, L., Salemans, B. & Schrijnemakers, S. (2007). Maastricht: Het Maastrichts anno 1807: boers? de gelijkenis in het Maastrichts besproken. In F. Bakker & J. Kruijsen (eds.), *Het Limburgs onder Napoleon. Achttien Limburgse en Rijnlandse dialectvertalingen van 'De verloren zoon' uit 1806-1807*. Utrecht: Gopher, pp. 177-215.

Swanenberg, J. (2013). All dialects are equal, but some dialects are more equal than others, In *Tilburg Papers in Culture Studies*, Paper 43.

Tervooren, H. (2005). *Van der Masen tot op den Rijn; ein Handbuch zur Geschichte der mittelalterlichen volkssprachlichen Literatur im Raum von Rhein und Maas*. Band 105. Geldern: Historisches Verein für Geldern und Umgegend.

Tjong Kim Sang, E. (2015). *Converting seventeenth century Dutch to modern Dutch*. Presented at the Workshop Morphosyntactic Enrichment of Historical Texts, Utrecht, The Netherlands.

Van den Bosch, A., Busser, G., Canisius, S. & Daelemans, W. (2007). An efficient memory-based morpho-syntactic tagger and parser for Dutch. In P. Dirix et al. (eds.), *Computational Linguistics in the Netherlands: Selected Papers from the Seventeenth CLIN Meeting*, Leuven, Belgium, pp. 99-114.

van Gompel, M. & Reynaert, M. (2013). FoLiA: A practical XML format for

linguistic annotation: A descriptive and comparative study. *Computational Linguistics in the Netherlands Journal*, 3, pp. 63-81.

van Halteren, H. & Rem, M. (2013). Dealing with orthographic variation in a tagger-lemmatizer for fourteenth century Dutch charters, *Language Resources and Evaluation*, 47, pp. 1233-1259.

van Horen, H. & van Horen-Hoosel, H. (2016). 1500-1793. Duistere eeuwen? In L. Spronck, B. van Melick & W. Kusters (eds.), *Geschiedenis van de literatuur in Limburg*. Nijmegen: Uitgeverij Vantilt, pp. 66-199.

van Hout, R. (2007). Het Europese Handvest en het Limburgs: het politieke en taalkundige discours. In H. Bloemhoff & P. Hemminga (eds.), *Streektaal en duurzaamheid. Lezingen van de internationale streektaalconferentie in Noordwolde*, 25 mei 2007, Berkoop/Oldeberkoop: Stichting Stellingwarver Schrieversronte, pp. 33-47.

van Hout, R., & Münstermann, H. (1981). Linguistische afstand, dialect en attitude. *Gramma: Nijmeegs Tijdschrift voor Taalkunde*, 5(2), pp. 101-123.

Weijnen, A., Goossens, J. & Goossens, P. (1983). *Woordenboek van de Limburgse dialecten. Inleiding & I. Agrarische terminologie*. Aflevering 1. Assen: Van Gorcum.

Weusten, S., Grondelaers S. & Van Hout, R. (2013). De herkenning en waardering van zes Limburgse 'stadse' dialecten. Leve het Maastrichts? In P. Bakkes (ed.), *Jaarboek Veldeke Limburg*, 20. Roermond: Vereniging Veldeke Limburg, pp. 61-75.

Willemyns, R. (2003). *Het verhaal van het Vlaams; De geschiedenis van het Nederlands in de Zuidelijke Nederlanden*. Antwerpen: Standaard Uitgeverij.

Wolters, L. (2016). *Veldeke Limburg 1926-2016*. Roermond: Veldeke Limburg.

Websites:

e-wld.nl. Accessed at: e-wld.nl. (20 April 2017)

github.com/LanguageMachines/toad/releases/tag/v0.3. Accessed at: https://github.com/LanguageMachines/toad/releases/tag/v0.3. (20 March 2017)

idm.fr. Accessed at: http://www.idm.fr/. (27 March 2017)

languagemachines.github.io/frog/. Accessed at: https://languagemachines.github.io/frog/ (27 March 2017)

limburghuis.nl. Accessed at: https://limburghuis.nl/. (20 April 2017)

limburgs.org. Accessed at: http://www.limburgs.org/en/limburgish. (26 March 2017)

limburgsewoordenboeken.nl. Accessed at: www.limburgsewoordenboeken.nl. (20 April 2017)

meertens.knaw.nl/kloeke. Accessed at: https://www.meertens.knaw.nl/kloeke/. (22 April 2017)

mestreechtertaol.nl. Accessed at: http://www.mestreechtertaol.nl/dictionair/mst. (20 April 2017)

sketchengine.co.uk. Accessed at: https://www.sketchengine.co.uk/. (27 March 2017)
tshwanedje.com/tshwanelex. Accessed at: http://tshwanedje.com/tshwanelex/. (27 March 2017)
ucrel.lancs.ac.uk/vard/about. Accessed at: http://ucrel.lancs.ac.uk/vard/about/. (22 March 2017).
woordenboek.gronsveld.com. Accessed at: woordenboek.gronsveld.com. (20 April 2017)

Language Policy in Slovenia: Language Users' Needs with a Special Focus on Lexicography and Translation Tools

Mojca Šorli[1], Nina Ledinek[2]

[1] Faculty of Arts, University of Ljubljana, Aškerčeva 2

[2] Research Centre of the Slovenian Academy of Sciences and Arts, Novi trg 2

E-mail: mojca.sorli@guest.arnes.si, NLedinek@zrc-sazu.si

Abstract

In the following contribution we present the design and the sociolinguistic background of the government-funded **Slovenian Language Policy and User Needs CRP 2016** project conducted between October 2016 and September 2017 under the leadership of the Research Centre of the Slovenian Academy of Sciences and Arts. Broadly speaking, the survey, which constitutes the core of the project, focuses on the language needs of four main categories: speakers of Slovene as their mother tongue; Slovenian minorities living across the border in Italy, Austria, Hungary and Croatia, with their specific linguistic and cultural background (bilingualism); users/learners of foreign languages; and users with special needs. All of these are investigated from the perspective of the legal framework regulating language use in individual fields, communicative practices, empirical evaluation of users' habits and attitudes; and, of particular importance for the present contribution, the current state-of-the-art in language infrastructure, including language technologies and digitisation. "Language description and language infrastructure in Slovenia" is a topic covered by the ZRC SAZU CRP 2016 project that will be treated in this paper in more detail, with special attention given to the questions asked about the use of the existing monolingual and bilingual (multilingual) language resources, in particular, dictionaries and other lexical resources. An in-depth survey will cover different groups of language professionals who use Slovene/foreign languages on a regular basis in the production of written and spoken texts for public use, such as journalists, publicists, fiction writers, bloggers, researchers, copywriters, PR professionals, legal document compilers, business and public administrators, as well as proofreaders and language editors and, last but not least, translators and interpreters.

Keywords: language infrastructure; interlingual resources; translation tools; user needs; online survey

1. Introduction

The **Slovenian Language Policy and User Needs CRP 2016** project conducted by the Research Centre of the Slovenian Academy of Sciences and Arts (henceforth: ZRC SAZU)—specifically its subtopic "**Language description and language infrastructure in Slovenia**", which is treated in more detail in the present paper—includes a study of the use of monolingual as well as bilingual and multilingual

(henceforth: interlingual) language resources, in particular, **dictionaries, corpora and other lexical resources**. The term "infrastructure" is used to incorporate language resources, i.e., sources and tools, as well as language technologies. Due attention is given to the supporting government documents, notably the *Slovenian 2014–2018 Action Plan* (*for Interlingual Resources*) (henceforth: SAPIR) and the legal framework necessary for the implementation of a language policy. A significant lack of data about the needs and expectations outside the formal education system, especially those of expert user groups, and increasingly intertwined private and public interests in the development of language resources, calls for an in-depth, comprehensive and, as far as possible, unbiased study of the actual habits and attitudes of the various user groups. Such a study could form the basis for future action plans and other binding language planning documents at the national level. Given that, in view of a general lack of surveys, the role of language resources in formal education has been reasonably well investigated, in the part of the survey presented here we focus on the different groups of **language experts who use Slovenian and foreign languages on a regular basis** in the production of written and spoken texts for public use, such as journalists, publicists, fiction writers, bloggers, researchers, copywriters, PR professionals, legal document compilers, business and public administrators, as well as proofreaders and language editors and, last but not least, translators and interpreters. In so doing, we indirectly address the question of contexts of the use of dictionaries and other, primarily lexical, resources. Furthermore, the survey questions elicit information on how specific user needs/aims are related to the use of specific sources and translation tools. In view of the targeted user groups, it was mandatory to survey the use of both interlingual and monolingual resources for both the language of origin (Slovene) and the target language (various foreign, mainly European, languages).

2. (Socio)linguistic background

The Slovenian language community is somewhat specific due to the small number of speakers, resulting in an imbalance between the number of users of the language of origin and any target language. At the same time, it is universal and conditioned by numerous radical changes in attitudes towards the use and planning of language resources, as well as access to them.

In light of the (linguistic, social and sociolinguistic) issues raised in the following paragraphs, our aim here is to relate this (socio)linguistic reality of Slovenian speakers to their actual needs and expectations, thus providing language policy makers and language infrastructure developers with some hard evidence on what they should prioritise. As the final results of the related online study will be available by September 2017, in the present paper we have chosen to showcase the vast range of issues addressed, directly or indirectly, by the survey.

While monolingual dictionaries and other reference books have been published by the Fran Ramovš Institute of the Slovenian Language (henceforth: FRISL) of ZRC SAZU,

an institution that regulates the standardised use of written (and spoken) language, the production of bilingual resources has not been subject to any systematic planning and control at the national level. Since the collapse of most commercially driven bilingual dictionary publishing about a decade ago, the development of interlingual resources has simply been determined by the demands of the free market. However, considering the (small) number of Slovenian speakers and the importance of cross-cultural exchange for Slovenia, providing resources for Slovene is as vital as ensuring the ongoing production of high quality interlingual resources (and research), especially for the leading European languages, including English as the lingua franca. The production of high quality interlingual resources is vital due to the significance of foreign language use and instruction **for Slovenian speakers, as well as foreign users/learners of Slovene**.

Here we will focus on the attitudes of (Slovenian) foreign language users with respect to both the available language resources and those that are lacking, particularly dictionaries and other lexical resources, as well as translation tools.

2.1 The changed role of translators (and dictionaries) in the (semi)automated translation business

In this section, we highlight two (sociolinguistically) relevant factors demonstrating the impact that technological advances, especially in automated translation, have had on the way we now perceive the professional field of linguistic mediation and the translator's role in it. Rapid technological advances have enabled numerous (semi)automated processes whereby human translators are declared (semi)redundant, perhaps one of the most common being the widespread practice of automated website translation. If on the one hand, a general leniency towards clearly inadequate but increasingly widespread fully automated translations of web content can even be detected in academic settings, which by definition (would be expected to) deal with both the theory and practice of translation, notably translation studies, it should not be overlooked that minor language speakers, in particular, are expected to accept linguistic degradation as part of the presumably necessary collateral damage of technical progress. The consequences for the development and status of minor languages are yet to be fully understood. Technological (individual) initiatives along the lines of the multilingual, partly crowdsourced web dictionary **Glosbe** (www.glosbe.com) enable users to access multiple international multilingual databases, which is essentially a positive development, as they aim to improve the level of (human) translation. Paradoxically, however, the low quality of automated website content demonstrates a surprisingly high level of tolerance for linguistic inadequacy.

A third factor, related to the two factors mentioned above, should really be addressed here in order to give a more complex and therefore more adequate picture of the

translation "market": the intertwining of academic interests and the increasingly commercialised framework to which translation and interpreting as a professional field has been assigned. For reasons of space, however, we have to leave this issue aside.

2.2 The motivation and rationale behind the research of user needs and habits

An overview of research to date reveals at least three different settings in which studies of user habits take place, each with its own set of objectives and motivation. Research on monolingual dictionaries and the role of dictionaries in the teaching process seems to be carried out by: 1) faculty members and doctoral or postdoctoral students at the Department of Slovene Studies, within special projects and in collaboration with other research/academic partners; individual minor scale studies are undertaken occasionally by graduate or master's students (such as Čebulj 2013 for general monolingual resources, etc.); 2) the central cultural institution traditionally in charge of language resources (the aforementioned Fran Ramovš Institute of the Slovenian Language of ZRC SAZU), which partly operates on continuous financial support at the national level; and 3) recently formed and relatively exclusive initiatives with an explicit interest in the research and production of digital language resources involving language technologies and crowdsourcing, combining public and private initiatives. Whereas the second group is composed mainly of linguists and lexicographers, in the third group the presence of "practicing lexicographers" is notably modest. Nevertheless, the third group have largely been driven by aspirations to compile an already envisaged new corpus-based dictionary of Slovene.

3. Survey and Analysis of Studies to Date and the Relevant Literature

There has been no comprehensive study to date in Slovenia covering both monolingual and interlingual resources, and including more than one or two specific user groups.

3.1 Monolingual studies

Most studies regarding general **monolingual resources for Slovene** have been carried out in the context of formal education amongst primary and secondary school students. Mostly, the role of the dictionary as a basic tool in the teaching/learning process has been examined, focusing on the comprehensiveness and accessibility of dictionary data. Specifically, the use of the Dictionary of Standard Slovene has been examined as well as its inclusion in teaching Slovenian language at school (Stabej et al., 2008; Rozman et al., 2010; Čebulj 2013). A more detailed overview of the research into monolingual dictionary use in teaching can be found in Rozman et al. (2015). As

found in another attempt at a (monolingual) dictionary survey (Arhar Holdt et al., 2015), the specifics of "professional dictionary use" have not been sufficiently examined compared to dictionary use for pedagogical purposes. There is also a lack of research in the field of Slovene as a second or foreign language (Rozman et al., 2015). Despite its limited data on actual dictionary use, the survey on Slovene language teaching (Rozman et al., 2010) has identified/highlighted the growing use of ICT amongst students, thus suggesting that a similar trend could be expected for foreign language resources for Slovene.

3.2 Foreign language studies

The use of foreign language resources has been investigated mainly in the rather narrow and specific field of formal education amongst university students of translation. In fact, the first of the two most recent studies was conducted on trends in the use of language resources (sources and tools) amongst trainee translators (Hirci 2013), while the second focused on translation queries performed by users of the "Translators, help!" Internet forum (Čibej et al., 2015).

A few earlier studies carried out at the Translation Department, Faculty of Arts, University of Ljubljana, need to be pointed out; namely, Hirci (2007; 2009), Mikolič Južnič (2009), Pisanski Peterlin (2003) and Vintar (e.g., 1999). All of these studies were conducted in the context of university translator training with a focus on the use of text corpora. However, there have been no studies involving various groups of more general users, including language experts of various backgrounds, which would provide a more objective picture of the current state of affairs. In view of the above, the ZRC SAZU survey—more precisely, the section on interlingual infrastructure—foresees a systematic analysis of the actual needs and attitudes of the various professional groups and actors in translation and interpreting, especially as resulting from and conditioned by their professional affiliations and status. In any case, the analysis of interlingual resources was designed to minimise biased interpretations of users' needs relying on specific groups, such as students of particular subjects, with a maximum dispersion of target groups in terms of age, professional background, education, etc.

Below we highlight two pieces of research (by faculty members) that shed some light on translation practice (and translation practicalities) in Slovenia in the past two decades. The first focuses on the development and use of new translation resources amongst translation students, while the second summarises changes in the translation market that have radically reshaped the profession of translating and interpreting.

3.3 On the application of translation resources

The results from the questionnaires of 2005 and 2012 show changing trends in the application of translation tools (Hirci, 2013: 154–158). While the results show a stable

use of bilingual and general monolingual dictionaries as resources in first and second places respectively, a change is evident in that, in 2012, electronic dictionaries were used almost exclusively, unlike in 2005, when paper and electronic dictionaries were used on a much more equal basis. The vast majority of the respondents in 2012 thus reported using only those resources that can be accessed electronically. Furthermore, the proportion of those respondents who regularly consult text corpora and parallel texts found on Google had considerably increased by 2012 (Hirci, 2013: 155). Another research question showed that the consultation of dictionaries, glossaries, encyclopaedias remains stable, albeit now almost exclusively in digital form (in fact, the proportion of those who consult bilingual dictionaries increases to virtually 100%), as does the use of the Internet (parallel texts). The use of corpora (monolingual and bilingual) is on the increase (from 12 out of 20 in 2005 to 18 in 2012). Perhaps the most striking difference is seen in the decrease in the use of mobile phones as a platform for accessing linguistic information, which in the 2012 survey is not reported at all.[1] On the other hand, there are more users of CAT software (4 out of 20 in 2012 as opposed to 2 in 2005). It can be concluded that the structure of, and familiarity with, the resources used in the examined period is largely unchanged, but the resources themselves are increasingly electronic, i.e., digital. The results of both questionnaires are also highly consistent on the issue of the usefulness of translation resources: roughly 30% of respondents in 2005 and 2012 indicated dictionaries, glossaries and encyclopaedias as the most useful resources, followed by the Internet (parallel texts), with 22% in 2005 and 26% in 2012. An increase is seen in the benefits ascribed to various computer corpora (15% in 2005 vs. 25% in 2012). Very similar results are yielded by reporting on the frequency of use of the listed categories of resources. A change is identified in the use of various monolingual and bilingual corpora (with an increase from 14% in 2005 to 26% in 2012), while a serious drop is also detected in the use of e-mail and translation forums for seeking advice from friends/experts (from 14% in 2005 to 6% in 2012). There is a considerable increase in the use of CAT systems (Hirci, 2013: 156–158). Whereas, in 2005 only 11 (out of 20) respondents believed that their translation work was considerably influenced by the use of electronic tools, in 2012 19 out of 20 believed that to be the case (ibid.: 158–159).

3.4 Changes in the "translation market"

The second study is more recent and addresses the problem of the radical reshaping of the translation market, which bears considerably on translator training programmes and, in particular, on the status of professional translators in Slovenia. In addition to the latest developments in lexicology, lexicography and translation studies, radical shifts in translation practice and the use of language resources have been caused by the

[1] Caution is needed when researching the use of mobile devices as sources of linguistic information. At least two very distinct scenarios are at work here: contacting people/experts as sources of information or using mobile applications, such as dictionaries.

drastically changed translation market. This has resulted in the deteriorating status of professional translators in Slovenia and worldwide. The number of translation agencies and companies has gone up since 2004 when Slovenia joined the EU, as has the need for translations. Until about that time, however, in a steadily growing Slovenian (and European) translation market, accompanied by an ever greater accessibility of contemporary translation tools, the relationships and roles of all the stakeholders remained basically unchanged. Kocijančič-Pokorn recently made a repeat of a survey carried out in 2007 by Fišer of the situation in the translation market, concluding that the translation market is still on the rise and the trends established in 2006 still valid: only a few translation companies/agencies seemed to be reaping the fruits of this growth, despite the fact that since 2004 the number of (small) businesses and (self-employed) individuals engaged in translation activity has grown considerably (Kocijančič-Pokorn, 2016: 5; data are based on business entities stating translation as their key activity in the Business Registry of the Agency for Public Legal Records and Related Services, AJPES, for 2014). As implied in the cited article, the growing market and the increased automation of (some aspects) of translation work have caused a disconcerting degradation and led to the increasingly precarious status of the profession.

Technological advances resulting in translation memories, applications for editing terminology databases and automated translation project management have further increased the individual translator's dependence on larger teams and translation agencies. This is corroborated by the fact that even seriously underpaid literary translators, who are only marginally, if at all, replaceable by machine translation software, must often seek additional financial means (from European or national funds) in order to ensure fair payment for their work, thus sharing the fate of their Western European counterparts (see Kocijančič-Pokorn, 2016). The new situation is characterised by demands for virtually instant translations, often into more than one language, which means that complex or larger translation jobs are only manageable by large translation teams. Individual translators are unable to meet the demands of such clients (ibid.: 13). Despite ever greater demands on translators in terms of the speed of their services and the quantity of texts, human translating has become increasingly undervalued, with translators increasingly hired for the so-called (full/partial) 'post-editing' of large portions of machine translated texts.

3.5 General studies on user attitudes towards language resources and language policy

In May 2017, a European survey on dictionary use was launched in 29 countries with the support of the European network for e-lexicography (http://www.elexicography.eu/events/european-survey-on-dictionary-use/), which was partly aimed at users regardless of their country of origin and partly

country/language-specific. The survey "aims to explore the attitude of language users towards general monolingual dictionaries of their native language" (Corpora list, 9 May 2017). This survey appears to be the first international survey of native language user needs and attitudes of its kind that—in addition to the anticipated similarities in attitudes related to the technology-driven changes in the use of language resources—might provide an insight into potential culture-specific differences in the attitudes of the respondents. The initiative has resulted from ENeL activities aimed at unifying and standardising cross-linguistic lexicographic tools and infrastructures across Europe.

In October 2016, a comprehensive and systematic (sociolinguistic) national study was launched by the FRISL called **Slovenian Language Policy and User Needs** (hence: ZRC SAZU CRP 2016 Study, http://isjfr.zrc-sazu.si/sl/programi-in-projekti/jezikovna-politika-republike-slovenije-in-potrebe-uporabnikov#v). A part of this study deals with language resources, and, within that, **multilingual resources**[2] addressing primarily, but not exclusively, groups of language experts—translators, interpreters and other language professionals using at least one foreign language, including language teachers—with their established daily working routines and strategies for dealing with professional challenges. Naturally, an insight into the use of monolingual resources is of crucial importance and therefore was not excluded from the section on interlingual resources.

More generally speaking, the interdisciplinary research project, which involves many experts, such as legal experts, educationalists, etc.,[3] focuses on the language needs of four main categories: speakers of Slovene as their mother tongue; Slovenian minorities living across the border in Italy, Austria, Hungary and Croatia, with their specific linguistic and cultural backgrounds (bilingualism); users/learners of foreign languages; and users with special needs. All these categories are investigated from the perspective of the legal framework regulating language use in individual fields, communicative practices, empirical evaluation of user needs and attitudes and, of particular importance for the present contribution, the current state-of-the-art of language infrastructure, including language technologies and digitisation. The results of the survey will provide an overview of the sociolinguistic situation in Slovenia as well as a description of user needs to help create a platform for the new national language policy agenda. In the following chapters, special prominence is given to some aspects of the interlingual resources survey.

[2] In the actual study, the term "multilingual (society)" is used to mean the ability of a group of speakers to communicate in more than one language, but the term "interlingual" is used instead in the present paper in the context of language resources to denote the type of both bilingual and multilingual resources.

[3] To name just a few participating partners: Academy for Theatre, Radio, Film and Television, two Law faculties, Faculty of Arts, Institute for Ethnic Studies, Centre for Slovenian as a foreign/second language, Pedagogical Institute, etc.

4. The ZRC SAZU CRP 2016 Study - Interlingual Resources: Content and Method

The **online survey** on interlingual resources has been designed with a view to exploring the **actual needs, practices and attitudes** of language users that can be aggregated to help identify the potential need to amend legislation regulating language use and speaker rights. In this paper, we present the design of the section focusing on key general and specialised resources for foreign languages. Overall, the aim of the section on interlingual resources is to give an illustrative insight into how users themselves reflect on their use of language resources, particularly with regard to the various categories of these resources.

4.1 Target groups

As stated above, and in view of research carried out to date, such as Hirci (2013) focusing on students of translation and Čibej et al. (2015) examining the habits of professional translators, we have sought to design from scratch first and foremost a survey of the use of language resources on the part of: **a)** professional translators, interpreters; **b)** other language experts using foreign languages professionally on a regular basis in the production of written and spoken texts for public use; and **c)** general users. While the section on interlingual resources is very much focused on professional use, both in opposition to private use and non-expert use, the part of the ZRC SAZU CRP 2016 Study investigating monolingual resources for Slovene seeks to investigate language issues from the perspective of field experts as well as non-experts. The differences between expert and non-expert users are, in fact, in themselves an interesting research topic, and we expect questions to arise in the process.

As stated above, the survey on interlingual resources investigates the habits and needs of translators and interpreters, proofreaders and language editors, journalists, publicists, legal document compilers, business and public administrators, etc. Also invited to participate in the survey, specifically through their professional associations, are the Slovenian Scientific and Technical Translators, Slovenian Association for Permanent Court Interpreters and Translators, Slovenian Association of Literary Translators, Slovenian Association of Conference Interpreters, Association of Slovenian Film and Television Translators, Slovenian Proofreaders' Association, Slovene Association of Journalists, Translation Services with the Government Secretariat-General, etc. In addition, potential respondents will be invited to participate in the survey through the appropriate mailing lists, forums and language-related websites, as the aim is to include as many general users as possible.

4.2 A note on Slovenian language policy drafting: the need for an adequate language technology taxonomy

The popular discourse on the need to develop language technologies for Slovene, particularly vociferous on the part of (technology-oriented) stakeholders/project partners, is often very general, disregarding the specifics of individual infrastructural units and the actual needs and attitudes of their potential users. From this point, and for the purposes of more efficient language policy drafting, it would be necessary to adopt a functional taxonomy of language technologies as well as setting priorities according to a clear set of criteria.

In the narrow sense, language technologies (henceforth: LT) are generally believed to include all forms of language processing and pre-processing (tokenisation, named-entity recognition, etc.), tagging, parsing, semantic analysis, (morphological and phonetic) lexicons, etc., while speech technologies include speech recognition and synthesis and other speech-related technological products. While these listings/facts are relevant for field experts, from a general user's point of view, LT can fundamentally (and more intuitively) be divided into: 1) tools serving the compilation of digital dictionaries, corpora, other manuals, etc. (mainly tools designed for researchers and experts in the field); and 2) tools designed to solve language-related problems encountered by general users (machine translation, data summarisation, speech recognition and synthesis, etc.). Probably the most widely used LT applications at this stage are grammar and spellcheckers (Krek, 2012: 14). On the other hand, LT include speech technologies, technologies for users with special needs, specialised technologies, such as those for translators, interpreters, etc. Another related distinction that should be made when talking about language resources is between so-called "applied" resources and LT applications, a distinction that is also important in terms of financial transparency. For one thing, lumping "applied" bilingual (general and specialised) dictionaries, parallel and monolingual corpora with taggers, parsers, machine-readable syntactic or semantic lexicons, etc., is manipulative in view of the fact that the two groups have different end users and do not serve the same purpose. In other words, it should be clear which resources are designed primarily for NLP and which are primarily for human users.

Even though it makes sense to conceive of the various tools as part of a broader category of LT, we would like to emphasise that, from the perspective of Slovenian language policy, each tool requires individual treatment: it is assigned a place in the priority list with regards to its design and objectives, as well as in relation to language policy as a whole. Although perhaps not so important from a purely technological perspective, this analysis is crucial from the point of view of meeting the needs of the various user groups. To give an example: unless it is to be an end in itself, listing a machine translation system as a priority should be supported with empirical evidence

on the types of text such a system should be developed for: general language or, perhaps, specialised texts with a high proportion of terminologies, such as MT@EC,[4] the online machine translation service provided by the European Commission. These tools have been specifically developed and used in combination with specialised translation memories. Currently free of charge, the conditions of use will eventually become "part of the sustainability plan for the new EU Automated Translation platform (eTranslation), which is funded through the Connecting Europe Facility programme".

Moreover, the rapidly developing relationship between public-private initiatives and the technology-driven economy, on the one hand, and the (digital) humanities and LT, on the other, is accompanied by a potential lack of transparency in determining the goals and priorities of language policy. In view of this, there is a need to draw a clear line between the requirements and expectations of the LT community as a professional field—which has its own (commercial) interests, whether in public research institutions or private institutions—and that which represents the common interests of the community at large, which will contribute to the development of LT indirectly, in the form of taxes through the state budget.

4.3 The significance of the production and development of interlingual resources for Slovene

According to the Eurobarometer 386 of 2012, as many as 92% of Slovenians (aged 25–64) speak at least one foreign language, which places Slovenia in fifth place among EU countries (the average for EU is 54%) (Krek, 2012: 1): "37.2% of these 92% can use two and 34.1% even three or more languages. In the 50 plus population 27.8% speak English, the proportion rising to 50% in the 35 to 49 age group and to as high as 75.5% in the lowest age group (25 to 34)" (ibid.: 14). On the other hand, knowledge of German, French and Italian is more constant, with the first at around 30% and the last at around 10%. We can safely predict that English is by far the leading foreign language and perhaps soon to become a second language (L2) in Slovenia, as is already the case in the Netherlands, for example, which for a while has prioritised English as L2. However, if these data seem encouraging we cannot say the same for language resources supporting multilingualism (SAPIR: 12). As rightly established in *The Slovene Language in the Digital Age study* (Krek, 2012; http://www.meta-net.eu/whitepapers/e-book/slovene.pdf), translators and

[4] See MT@EC: https://ec.europa.eu/info/resources-partners/machine-translation-public-administrations-mtec_en (Accessed 24 May 2017)

interpreters use dictionaries, corpora and translation memories, which need to meet the desired standards in terms of quality and scope (ibid.). According to the above study, Slovene is rated very low on the scale of MT development, particularly with regard to resource and technology enablers, which include general and specialised bilingual dictionaries and lexicons, parallel and comparable corpora, taggers, syntactic and semantic parsers, etc.

4.4 Type of inquiry in the ZRC SAZU CRP 2016 Study

There are four main categories of questions, of which not exclusively but mainly category d), in particular the part on language infrastructure, is addressed in this paper:

a) attitudes related to language use;
b) communicative practices and usages;
c) language users' needs;
d) language description and language resources.

The aim is to verify some of the generally accepted truths and assumptions regarding the state of needs in the area of interlingual resources against the responses acquired in the survey. These data should assist in determining the possible priorities for a 5–10 year action plan at the national level.

4.4.1 Question type profiling

As a number of drawbacks can be identified in studies based solely on data analysis (e.g., log-files) or social media (e.g., translation forums), we have opted at this point to conduct an online questionnaire survey, which includes mainly closed-ended but also open-ended questions (as a free comment). On the whole, according to Müller-Spitzer (2012: 5), these are expected to elicit more informative responses and are actually more appropriate "in web surveys than in paper surveys, especially when the response field is large". The idea is for the results to be later complemented with data/query analysis, such as that found in translation forums, e.g., Facebook groups, mailing lists, online chat rooms, etc.

One of the open-ended questions is dedicated to language pairs that users prioritise in the scheme of planning publicly (co-)funded revisions or new dictionary editions. More importantly, the questions also investigate attitudes, albeit indirectly, towards actions that were drafted in the SAPIR 2014–2018; for instance, respondents' views with regards to investing public money in the digitisation of out-of-date dictionaries with the aim of integrating them into multilingual portals, as drafted in the SAPIR. Moreover, we aspire in this study to establish the current trends and the relationship between the role of traditional (electronic) resources and those of CAT and MT

systems. This includes obtaining information on the type of source or tool, such as a CD-ROM, a web dictionary, a CAT or MT system.

Furthermore, drawing on extensive lexicographic practice, the authors of this paper can safely conclude that language policy in general and language planning documents in particular focus almost exclusively on the infrastructure for teaching foreign languages; therefore, the needs of other types of users have been prioritised in the survey in order to bridge the gap.

4.4.2 The questionnaire

The questions targeting language experts as compiled in the first version of the online survey include[5]:

1) the type of texts they most frequently translate:

   a. *Literary*
   b. *Journalism/media*
   c. *Commercial*
   d. *Sworn court translations*
   e. *Expert and academic/scholarly*
   f. *Technical*
   g. *Other*

2) the language resources and translation tools they most frequently use:

   a. *Bi- or multilingual dictionaries*
   b. *Machine translation*
   c. *Monolingual dictionaries of the target language(s)*
   d. *General lexical and terminological databases on the Web*
   e. *Translation memories*
   f. *Bi- or multilingual text corpora*
   g. *Monolingual text corpora*
   h. *Translation forums and other social media*
   i. *Speech repositories (e.g., of the EU Directorate-General for Interpreting)*
   j. *Other*

3) the role of social media:

[5] The questions have been translated from Slovene into English by the author for the purposes of this article.

Is using a translation forum or other social network your first choice or do you consult the forum only when you cannot find the answer in a standard lexical resource, such as a dictionary or lexical database?

a. *A translators' forum or other social network is my first choice followed by standard language resources*
b. *Standard language resources are my first choice followed by a translators' forum or other social network*

4) the role of translation technologies (memories, MT, etc.):

What percentage of your work is completed via a translation desktop, i.e., a translation memory and a machine translation system (MT)?

a. *Translation memory: less than 20%/between 20% and 40%/between 50% and 70%/over 70%*
b. *MT: less than 20%/between 20% and 40%/between 50% and 70%/over 70%*

5) the use of native language (i.e., Slovene) dictionaries, corpora, etc.

6) attitudes and opinions showing user priorities in the development of language infrastructure and its financing:

Which of the below resources in the field of interlingual resources for Slovenian users should, in your opinion, become priority in the next 5–10 years in terms of upgrade or development funded with public money (you can choose up to 2 resources)?

a. *Bilingual text corpora, i.e., parallel or comparable text/translation corpora*
b. *Machine translation*
c. *Bilingual dictionaries and bilingual lexical databases for the prioritised language pairs*
d. *Slovenian Wikidictionary, Slovenian Wikipedia, Wikisource and other collaborative interlingual resources*
e. *Terminological databases and a terminological portal*
f. *Multilingual information portal (with links to the existing sources and tools)*
g. *Other*

Comments:

7) users' favourite IT platforms for accessing multilingual information, and similar:

Which information-communicative platforms do you normally use to access information in/on foreign languages? (Mark with 1 to 7 whereby the most frequently used platform is 1 and the least frequently used one is 7.)

a. *Google*
b. *Websites*
c. *Mobile applications*

d. Social media (Facebook, Instagram, Pinterest, Twitter and similar)
e. Electronic resources (e.g., CD-ROM)
f. Paper resources
g. Other

Comments:

8) what source and/or tool do you miss the most when you are working in a multilingual context within the field of your expertise? (Please, list your language combination(s))

9) you perform your language services:

a. As a self-employed language expert with privately owned language resources
b. Employed in a private company/agency with access to language and translation technologies
c. *Employed as a civil servant or public administrator with access to language and translation technologies*
d. *Other*

5. Conclusion

The questions are designed so that they enable the assessment of the efficiency and actual benefits of some of the already financed public projects for the development of language resources. Any future action plan needs to take into account the limited financial means allocated to the development of language infrastructure and the fact that Slovenian speakers are yet to see the compilation and publication of some of the most basic corpus-based (monolingual and interlingual) resources, such as a comprehensive monolingual dictionary of contemporary Slovene, a pedagogical monolingual dictionary, an SFL (Slovene as a Foreign Language) dictionary, a Slovenian–English dictionary, etc., calling for a sensible judgement on which of all the possible language resources, including language technologies, are truly urgently needed in the most imminent future. The missing resources are, in fact, as has often been pointed out, critical for the development of language technologies for Slovene.

While in a systematic analysis we study the actual needs and habits of members of all of the major language-related professional associations, we also ask questions in the online survey that will give an insight into the daily (social) reality of individual language experts: forms of employment, working conditions, the degree of professional autonomy, social status, etc. This will fill the gap in data regarding the needs and expectations beyond the system of formal education, particularly of professional groups who are the most actively involved in language mediation and in the production of texts for public use in foreign languages and in Slovene. On the basis of the final results (projected for September 2017), it will be possible to plan the development of language resources and LT, whereby adequate, more intuitive

functional distinctions within the field of LT, as suggested in the present paper, will serve the purposes of greater transparency in language planning.

The online survey on interlingual resources is part of a broader interdisciplinary research project, the aim of which is to provide data on the sociolinguistic situation and user needs in Slovenia for the compilers of the key Slovenian language policy documents (resolutions and action plans). Sociolinguistic as well as legal aspects will be examined due to the fact that any language policy as a public policy in the interest of all the speakers must necessarily be adequately legislated. The comprehensive on-line survey analyses Slovenian speaker attitudes, communicative practices and language infrastructure, bringing all of these into a meaningful relationship with the need to develop language technologies. Ultimately, this project is to produce a comprehensive and empirically based study of key sociolinguistic issues, including the attitudes of language users towards the existing language infrastructure and that which is lacking, for the new *National Language Policy Programme* and a future *Slovenian Action Plan* (*for Interlingual Resources*).

6. Acknowledgements

This contribution is based on research conducted within the *Slovenian Language Policy and User Needs* national project (V6-1647) and partly financed by the Faculty of Arts, University of Ljubljana.

The authors would also like to thank the anonymous reviewers for their useful and much appreciated comments.

7. References

Arhar Holdt, Š. (2015). Uporabniške raziskave za potrebe slovenskega slovaropisja: prvi koraki. In V. Gorjanc, P. Gantar, I. Kosem, S. Krek (eds.) *Slovar sodobne slovenščine: problemi in rešitve.* Ljubljana: Znanstvena založba Filozofske fakultete UL, pp. 136–149.

Arhar Holdt, Š., Čibej, J. & Zwitter Vitez, A. (2016). Value of language-related questions and comments in digital media for lexicographical user research. *International Journal of Lexicography.* doi:10.1093/ijl/ecw017.

Arhar Holdt, Š., Čibej, J., Zwitter Vitez, A. (2015). S pomočjo uporabniških jezikovnih vprašanj in mnenj do boljšega slovarja. In V. Gorjanc, P. Gantar, I. Kosem & S. Krek (eds.) *Slovar sodobne slovenščine: problemi in rešitve.* Ljubljana: Znanstvena založba Filozofske fakultete UL, pp. 196–214.

Arhar Holdt, Š., Kosem, I. & Gantar, P. (2016). The Dictionary Typology: the Case of Slovene. In T. Margalitadze & G. Meladze (eds.) *Lexicography and linguistic diversity: proceedings of the XVII EURALEX International Congress, 6-10 September, 2016.* Tbilisi: Ivane Javakhishvili Tbilisi State University, pp. 179–187.

Čebulj, M. (2013). *Raba slovarja v 1. in 2. triletju osnovne šole*. University graduation thesis. University of Ljubljana: Faculty of Education.

Čibej, J., Gorjanc, V. & Popič, D. (2015). Vloga jezikovnih vprašanj prevajalcev pri načrtovanju novega enojezičnega slovarja. In V. Gorjanc, P. Gantar, I. Kosem & S. Krek (eds.) *Slovar sodobne slovenščine: problemi in rešitve*. Ljubljana: Znanstvena založba Filozofske fakultete UL, pp. 168–181.

Fišer, D. (2008). Recent trends in the translation industry in Slovenia. *The Journal of Specialised Translation*, Issue 10. http://www.jotrans.org/issue10/art fiser.pdf (18 February 2016).

Hirci, N. (2003). Prevajanje danes in jutri: delo s sodobnimi prevajalskimi viri in orodji. *Jezik in slovstvo* 3-4/48, pp. 89–102.

Hirci, N. (2007). *Učinkovitost uporabe sodobnih prevajalskih virov pri prevajanju v nematerni jezik*. PhD Dissertation. University of Ljubljana. Filozofska fakulteta.

Hirci, N. (2009). Uporaba sodobnih prevajalskih virov pri izobraževanju prevajalcev. In V. Mikolič (ed.) *Jezikovni korpusi v medkulturni komunikaciji*. ZRS Koper: Založba Annales, pp. 57–87.

Hirci, N. (2013). Changing trends in the use of translation resources: the case of trainee translators in Slovenia. *ELOPE* 10, pp. 149–165.

Gorjanc, V., Gantar, P., Kosem, I. & Krek, S. (eds.) (2015). *Slovar sodobne slovenščine: problemi in rešitve.* Ljubljana: Znanstvena založba Filozofske fakultete UL.

Kocijančič-Pokorn, N. (2016). Nič več obljubljena dežela : dinamični premiki na slovenskem prevajalskem trgu in področju izobraževanja prevajalcev. *Vestnik za tuje jezike*, 2016, 8 (1), pp. 9–21 http://revije.ff.uni-lj.si/Vestnik/article/view/7174/6878. (31 May 2017).

Krek, S. (ed.) (2012). The Slovene Language in the Digital Age/Slovenski jezik v digitalni dobi. *METANET White Paper Series*. Springer.

Lew, R. & de Schryver, G.-M. (2014). Dictionary Users in the Digital Revolution. *International Journal of Lexicography 27.4*, pp. 341–359. doi:10.1093/ijl/ecu011.

Mikolič, V. (2015). Slovarski uporabniki – ustvarjalci: ustvarjati v jeziku in z jezikom. In V. Gorjanc, P. Gantar, I. Kosem & S. Krek (eds.) *Slovar sodobne slovenščine: problemi in rešitve*. Ljubljana: Znanstvena založba Filozofske fakultete UL, pp. 182–195.

Mikolič Južnič, T. (2009). Vzporedni korpus - prevajalsko orodje in orodje za jezikoslovne analize. In V. Mikolič (ed.) *Jezikovni korpusi v medkulturni komunikaciji*. ZRS Koper: Založba Annales, pp. 117–132.

Mikolič, V. (2015). Slovarski uporabniki – ustvarjalci: ustvarjati v jeziku in z jezikom. In V. Gorjanc, P. Gantar, I. Kosem & S. Krek (eds.) *Slovar sodobne slovenščine: problemi in rešitve*. Ljubljana: Znanstvena založba Filozofske fakultete UL, pp. 182–195.

Müller-Spitzer, C., Koplenig, A. & Töpel, A. (2012). Online dictionary use: Key findings from an empirical research project. In S. Granger & M. Paquot (eds.) *Electronic Lexicography*. Oxford: Oxford University Press, pp. 425–457.

Pisanski Peterlin, A. (2003). Uporaba novih tehnologij pri jezikovnem pouku. *Jezik in slovstvo* 3-4/48, pp. 103–112.

Rozman, T., Krapš Vodopivec, I. Stritar, M. & Kosem, I. (2010). *Nova didaktika poučevanja slovenskega jezika - projekt »Sporazumevanje v slovenskem jeziku«* - Kazalnik 15 http://www.slovenscina.eu/Vsebine/Sl/Kazalniki/K15.aspx (12 June 2015).

Rozman, T., Kosem, I., Pirih Svetina, N. & Ferbežar, I. (2015) Slovarji in učenje slovenščine. In V. Gorjanc, P. Gantar, I. Kosem & S. Krek (eds.) *Slovar sodobne slovenščine: problemi in rešitve.* Ljubljana: Znanstvena založba Filozofske fakultete UL, pp. 150–167.

Stabej, M., Rozman, T., Pirih Svetina, N., Modrijan, N. & Bajec, B. (2008). Jezikovni viri pri jezikovnem pouku v osnovni in srednji šoli: končno poročilo z rezultati dela. Ljubljana: Pedagoški inštitut. http://www.trojina.si/p/jezikovni-viri-pri-jezikovnem-pouku-v-osnovni-in-srednji-soli/ (22 June 2015).

Vintar, Š. (1999). Računalniške tehnologije za prevajanje. *Uporabna informatika* VIII/1, pp. 17–24.

Vintar, Š. (2016). *Prevajalske tehnologije.* University of Ljubljana. Filozofska fakulteta.

Websites:

European network for e-lexicography (European survey). Accessed at: http://www.elexicography.eu/events/european-survey-on-dictionary-use/. (20 May 2017)

Glosbe. Accessed at: www.glosbe.com. (20 May 2017)

Meta-Net. Accessed at: http://www.meta-net.eu/whitepapers/e-book/slovene.pdf. (20 May 2017)

MT@EC. Accessed at: https://ec.europa.eu/info/resources-partners/machine-translation-public-administrations-mtec_en. (24 May 2017)

ZRC SAZU CRP 2016 Project: Accessed at: http://isjfr.zrc-sazu.si/sl/programi-in-projekti/jezikovna-politika-republike-slovenije-in-potrebe-uporabnikov#v (20 May 2017)

Lexicography: What is the Business Model?

Henrik Køhler Simonsen

Copenhagen Business School, Dalgas Have 15, 2000 Frederiksberg
E-mail: hks.msc@cbs.dk

Abstract

Lexicography is a four thousand year old discipline and dictionaries have been an integral part of commerce and human cultural history for centuries. But lexicography is also a business activity undertaken by individuals or companies with a view to generating profits or creating value. And any discipline, movement, organization or company needs a plan of how it intends to create, deliver and capture cultural or monetary value.

The analysis and discussion of the lexicographic business model uses a kaleidoscopic approach where the concept business model is seen and analyzed by means of the five lenses: strategy, core competencies, innovation, business understanding and organizational inertia. By means of these lenses, the paper explores the business model of lexicography in Denmark, and it analyzes and discusses whether the Danish lexicographic industry understands the concept business model at all, and if so, to what extent it applies business model thinking. Furthermore, this paper discusses different categories of lexicographic business models, potential elements of a new lexicographic business model and finally it formulates six theses on a new, more viable lexicographic business model.

Keywords: Business model; strategy, core competencies, innovation; organizational inertia

1. Introduction and Research Questions

Conventional dictionaries seem to have over-exploited their current business model and seem for too long to have had a disproportionate "relation between the exploration of new possibilities and the exploitation of old certainties" (March, 1991).

On the basis of the empirical data presented in this article, it is in fact argued that dictionaries as we know them seem to have been disrupted (Christensen, 1997), and a large number of dictionary publishers have in fact closed down, merged or changed focus over the past 10-15 years.

In other words, lexicography seems to be in need of a new business model, which is viable and geared for the future, and it is argued that we must start to define new ways of creating value. So perhaps lexicography needs to start from scratch. This disruptive transformation, the current status quo of dictionaries, and the business model of lexicography in Denmark were some of the topics discussed in my MBA Thesis (Simonsen, 2016), which was built on approx. 25 years of experience with and research within lexicography.

Over the past two decades I have been witnessing the deteriorating performance,

viability and relevance of lexicographic products, so I decided to research and analyze whether the lexicographic business model has in fact already disappeared, whether and how the lexicographic industry understands and uses the concept business model (Osterwalder & Pigneur, 2010) and what might constitute elements of a new, viable business model based on the Value Proposition Canvas theory proposed by Osterwalder et al. (2014).

This paper is based on the empirical data collected and discussed in Simonsen (2016 and Simonsen (2017), but focusses on three new research questions with a clear business modelling focus. The objectives of this article are to analyze and discuss lexicographic business models in Denmark via the following research questions:

1. What characterizes the current understanding and application of the concept business model in the Danish lexicographic industry?

2. What characterizes the different lexicographic business models?

3. What characterizes potential new elements of new and more viable lexicographic business models?

The structural approach of this article is kaleidoscopic, meaning that the concept business model will be analyzed and discussed through five theoretical lenses.

First, the delimitations, research methods and empirical basis of this article will be outlined; then, the five theoretical lenses consisting of a number of relevant theories and models will be outlined and discussed. Third, this article offers an in-depth discussion of the understanding and application of the concept business model based on interview data. Fourth, building on the insights from the analysis and discussion of the interview data and the kaleidoscopic lenses, this article discusses different types of lexicographic business models and potential elements of a new and more viable lexicographic business model. Finally, based on the analysis and discussion, this article discusses six theses on a new and more viable lexicographic business model.

2. Delimitations, Research Methods and Empirical Basis

To analyze and discuss the above research questions the following delimitations and methodological and empirical considerations must be discussed.

First, the term “business model” is used to refer to the creation of value, i.e. both monetary and cultural value, which means that the term can be used about both private companies and public organizations. Second, the term “lexicographic industry” covers both private and public companies and organizations, which design, compile and market dictionaries, reference works and lexicographic data. This means that the term is used in its widest possible sense and covers conventional dictionary publishers, educational publishers, information suppliers, data distributors, etc.

Third, the collection of the empirical data was delimited geographically to Denmark and temporally to October-November 2015. To ensure validity, reliability and relevance, and to add an international perspective, two interviews with international CEOs were conducted. However, the overall focus of this paper is the Danish dictionary market.

To increase the validity, reliability and relevance of the data in relation to the research questions, the 15 interview subjects were carefully selected based on five selection criteria. First, it was important to recruit interview subjects, who were very experienced in publishing conventional and online dictionaries, i.e. both small and large publishers and both private and public organizations. Second, it was important to recruit interview subjects from educational publishers and from large public dictionary associations. Third, I selected interview subjects from industry specific publishing companies including interview subjects from the information and data industry. Finally, I interviewed two international CEOs to obtain an international perspective. This means that 15 different interview subjects from ten different types of organizations were interviewed, and it is argued that this particular approach enhances validity, reliability and relevance.

Validity deals with questions like whether or not data can be trusted, whether or not the subsequent findings address the research questions and whether or not the author has been able to process the data in an unbiased and critical way (see Saunders et al., 2009: 157 for a detailed discussion of validity). On the basis of the analysis of the interviews it is argued that the data are valid and relevant for the discussion; however, it is also important to reflect upon the personal bias of one's own work, which, I believe has been very much the case.

Reliability deals with questions like whether or not the research method used, in this case semi-structured research interviews, is used in a consistent and structured way to ensure that what is measured is measured consistently (Saunders et al., 2009: 156) for a detailed discussion of reliability. Again it is argued that the method was consistent and systematic. Prior to each interview the interview person received an interview guide with 15 questions and the interviews were conducted consistently and systematically by means of open, semi-structured interviews (Simonsen, 2016). For the purpose of this article all statements were translated into English.

The overall philosophy of science used can be described as social-constructivist, and the research method is the interview method (see Kvale, 2007 for a description of both collection and analysis of data on the basis of qualitative research interviews). The approach used was based on Kvale's seven stages of an interview investigation, and the interviews can be described as open, semi-structured research interviews.

Each interview lasted approx. 45 minutes and was recorded for subsequent meaning extraction and analysis. During the analysis and the meaning extraction, six overall themes were identified, see also Figure 1.

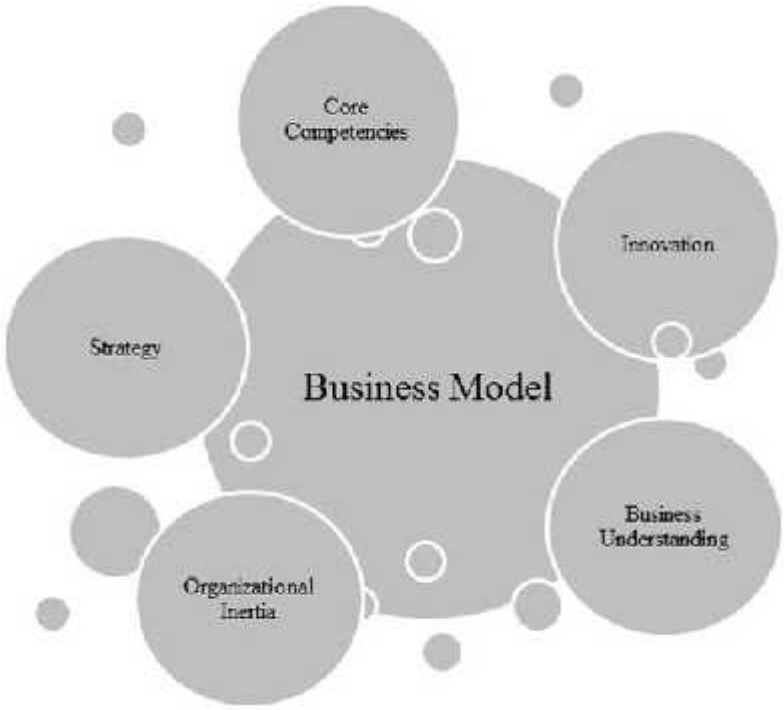


Figure 1: Themes Identified in Empirical Data

The article is thus based on interviews with 15 CEOs and/or senior managers from 10 different types of organizations totaling 15 x 45 minutes of interview data. Relevant statements from the interview subjects are used in the analysis and discussion below.

The next part of this paper focuses on the five theoretical kaleidoscopic lenses through which the concept "business model" is analyzed.

3. Theory and Models

The research object of this article is business models, and it might be argued that all discussion of business models in fact starts with Drucker (1994), who made a strong argument for what he referred to as "a theory of the business".

Drucker describes the theory of the business as follows: *"These are the assumptions that shape any organization's behaviour, dictate its decisions about what to do and what not to do, and define what the organization considers meaningful results. These assumptions are about markets. They are about identifying customers and competitors, their values and behaviour. They are about technology and its dynamics, about a company's strengths and weaknesses. These assumptions are about what a company gets paid for"* (Drucker, 1994: 95) (my underlining). Drucker does not use the term business model, but it is argued that this particular theoretical contribution started the entire business model discipline.

Later, Osterwalder & Pigneur (2010), building on Drucker's theory, introduced a comprehensive and graphically appealing approach to business model generation called the Business Model Canvas (BMC). Osterwalder & Pigneur define a business model as *"the rationale of how an organization creates, delivers and captures value"* (my underlining) (Osterwalder & Pigneur, 2010: 18) and this is in fact the definition upon which this article is based. Osterwalder & Pigneur's definition and understanding of business model is very useful, because it covers the creation of all types of value, including monetary, cultural and experiential value. The BMC is shown in Figure 2.

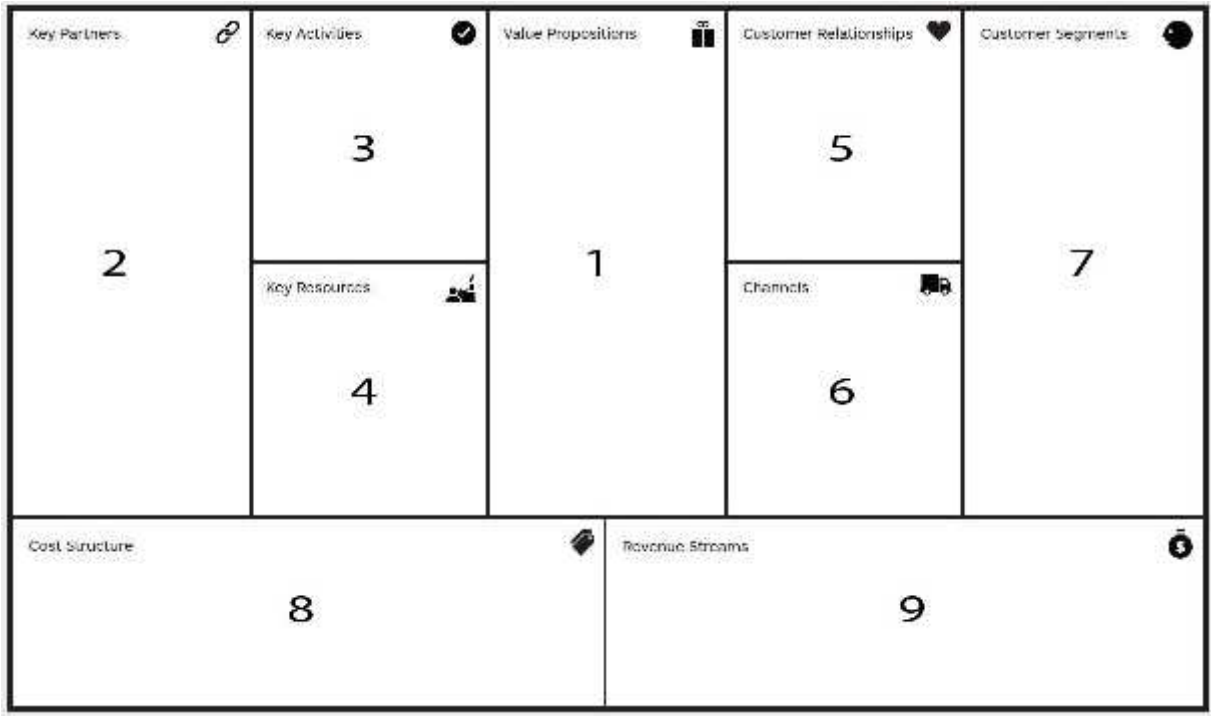


Figure 2: Business Model Canvas

The BMC consists of nine fields, which for the purpose of this discussion have been numbered 1-9 (my insertion), as these numbers will be referred to later on. The nine fields in numerical order are "Value Proposition", "Key Partners", "Key Activities", "Key Resources", "Customer Relationships", "Channels", "Customer Segments", "Cost Structure" and "Revenue Streams". The starting point of the BMC is the Value Proposition field, which in fact is the most important field of the BMC (Osterwalder & Pigneur, 2010: 16–19). The nine fields describe four overall business areas: Field 1 is the offer, fields 2-3-4 are the infrastructure, fields 5-6-7 are the customers and fields 8-9 are the financial visibility. This article will primarily focus on the business areas offer and its customers, and secondarily on infrastructure and financial visibility. The Value Proposition field is particularly relevant and may be described as the line of products or services, which create the required value for the customer segment in question (Osterwalder & Pigneur, 2010: 16–19).

As will be shown below in the Value Proposition Canvas (VPC), it is crucial that an organization ensures that there is a "fit" between the value proposition, or what the company offers, and what the customer segment in question demands. This particular line of thinking is not alien to lexicography, which is why this model is so useful when analyzing and discussing a new lexicographic business model. A company can have different value propositions to different customer segments, but the important aspect is to ensure that they are aligned and that there is a "fit" between what is offered and what is needed. According to Osterwalder & Pigneur (2010: 23–25) value propositions may create value for the customer segment through elements like "*newness, performance, customization, getting the job done, design, brand/status, price, cost reduction, risk reduction, accessibility and convenience/availability".*

To help define and describe value propositions (Osterwalder et al., 2014:10) designed the Value Proposition Canvas (VPC), see also Figure 3.

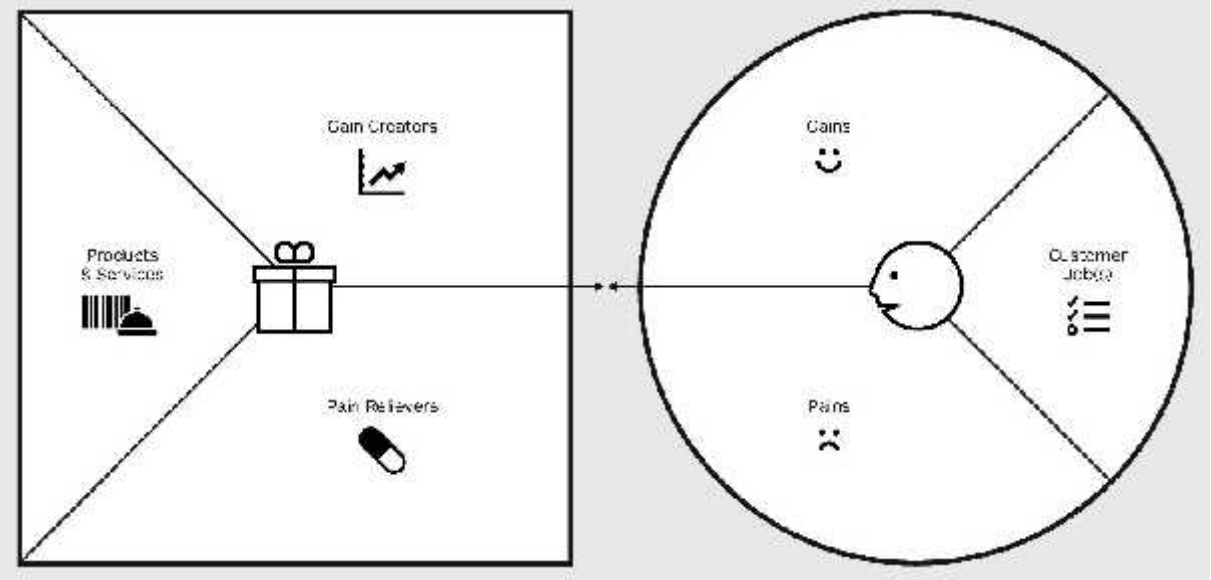


Figure 3: Value Proposition Canvas

The VPC consists of two building blocks. The circle on the right is termed the "Customer Profile" (what the customer needs) and the square on the left is termed the "Value Map" (what the company offers).

The "Customer Profile" has three fields named "Customer Jobs", "Pains" and "Gains" and according to Osterwalder et al. (2014: 10–11) a "Customer Profile" can be used to describe the customer's job functions and his pains and gains. This outward-inward description of the customer enables the company to get a detailed understanding of what the customer actually needs. The "Value Map" on the left also has three fields named "Products & Services", "Pain Relievers" and "Gain Creators". The "Value Map" is used to describe the company's products and services and how they may relieve the customer's pain and/or create even more gain for him.

The value propositions are then created on the basis of the Pain and Gain approach and by ensuring that there is a "fit" between the "Customer Profile" on the one hand and the "Value Map" on the other. This "fit" is obtained by analyzing and aligning "Customer Jobs vs. Products and Services", "Customer Gains vs. Customer Pains" and finally "Gain Creators vs. Pain Relievers".

As argued above, the analysis of the lexicographic business model is conducted by means of five theoretical lenses, which are strategy, core competencies, strategic innovation, business understanding and organizational inertia.

The first theoretical lens, through which business model generation is analyzed, is "strategy". For the purpose of this paper, "strategy" as a theoretical and practical concept is defined as *"the long-term direction of an organisation"* (Johnson et al., 2012: 3) and it deals with the ability and performance of an organization to plan ahead, but also the ability of an organization to plan on an ad hoc basis and in accordance with market developments (Mintzberg & Lampel, 1994).

The second theoretical lens is "core competencies", which is shown in Figure 4 below. Hamel & Prahalad's (1994: 227) understanding of core competencies is quite relevant for the understanding and management of competencies in the lexicographic industry. According to Hamel & Prahalad, core competencies are the human and technological core competencies required for a company to be successful. Hamel & Prahalad's core competency matrix, see also Figure 4 below, is particularly useful, because it can be used to understand how to build, retain and divest competencies in accordance with the type of market in question (see also Hamel & Prahalad 1994: 227).

Core Competences			
	New	*Premier plus 10* What new core competencies will we need to build to protect and extend our franchise in current markets?	*Mega-opportunities* What new core competencies would we need to build to participate in the most exciting markets of the future?
	Existing	*Fill in the blanks* What is the opportunity to improve our position in existing markets by better leveraging our existing core competencies?	*White spaces* What new products or services could we create by creatively redeploying or recombining our current core competencies?
		Existing	**New**
		Market	

Figure 4: Core Competence Matrix

The core competence matrix is a conventional 2 × 2 table with four quadrants. On the X axis we have the market dimension, i.e. either of an existing market or a new market, and on the Y axis we have core competencies, i.e. either existing core competencies or new core competencies. If, for example, a company decides to focus on a new market or new product it will probably need to build or add new core competencies, i.e. New-New and thus Mega-opportunities (the upper, right-hand quadrant) (see Hamel & Prahalad, 1994: 227 for a detailed discussion of core competencies).

The third theoretical lens, through which business model generation is analyzed, is "innovation or strategic innovation". Strategic innovation is about making the right strategic choices in innovation, i.e. innovating what is strategically in focus. According to Afuah (2009: 1), strategic innovation may be defined thus: *"It often entails changing the rules of the game"*, which in fact is what seems to have happened in the lexicographic industry. Innovation is also about not "exploiting", but "exploring" (March, 1991) and even inventing new services and products and disrupting markets or inventing new markets in line with the Blue Ocean Strategy, as proposed by Kim & Mauborgne (2004). For the purpose of this discussion, (Christensen, 1997), is highly relevant. According to Christensen, innovation and disruptive technologies are *"Disruptive technologies bring to a market a very different value proposition than had been available previously. Generally, disruptive technologies underperform established products in mainstream markets. But they have other features that a few fringe (and generally new) customers value"* (Christensen, 1997: XV) (my underlining). The deliberate underperformance on certain parameters is relevant, as will be argued later in the discussion and analysis.

The fourth theoretical lens is a relatively broad concept referred to as "business understanding". Business understanding is here defined as an organization's ability to analyze, interpret and act on the fluctuating market conditions, competitor strength, etc. Business understanding is also the ability of an organization to interpret and act on shifting demands and its ability to understand the value chain in which it plays a role. In this connection, Adner (2012), in particular, offers a useful theoretical contribution, because Adner discusses the ability of an organization to apply a wide lens approach and to design the value and adoption chain. Furthermore, Beckmann et al. (2016) expand the discussion of business model generation with the concept "Value Creation Architectures" (VCA), which in many ways resembles the wide lens approach proposed by Adner (2012). The VCA approach discussed by Beckmann et al. (2016) is particularly relevant when discussing how to convert old technology-based business models into new value-creating business models.

The fifth theoretical lens is "organizational inertia". Organizational inertia is discussed by, for example, Gavetti (2005) and Sull (1999), and is about a company's ability and tendency to accept and embrace change or perhaps even more relevant its ability and tendency not to accept and embrace change. Organizational inertia is in fact closely

related to core competencies, organizational culture and of course the concepts "exploitation" vs. "exploration" (March, 1991) and related to the study of what happens when successful companies suddenly go bad, which in fact seems to be what has partly happened in the Danish lexicographic industry.

4. Analysis of the Existing Lexicographic Business Model

This part of the article discusses the lexicographic business model on the basis of the interview data and the theoretical models above and discusses the first research question. In the discussion, where the author is aware of both interviewee and interviewer bias, it is argued that the data acquired are valid, and that the conclusions and insights are reliable.

On the basis of the 15 interviews, the Danish lexicographic industry seems neither to understand nor use the concept business model. In fact, only two of the interviewed CEOs indicate that they actively use business model generation. A number of statements from the interview subjects substantiate this argument as one CEO says *"we neither have a strategy nor a business model – it is a gut feeling and we live by it".* A similar approach can be seen in a statement from another CEO, who says *"We have never worked with a business model. We have just followed the path".*

In contrast to this somewhat reactive approach are two statements from the two CEOs whose companies use a business model. The first CEO says *"In fact it was quite easy to steal the market. The existing players were just not competent enough"* and the other CEO says *"How can you do business without knowing your market, customers and competitors and without having considered how to make money".* Finally, a third CEO explains, when talking about the market for reference works and dictionaries in Denmark, *"What we have witnessed is a shift in quality parameters. To be frank – company X launched inferior lexicographic content, but had a superior business model and distribution platform. Company Y had superior lexicographic content, but a very poor distribution platform".* The last statement resembles Christensen's (1997: XV) view on innovation and disruption where you deliberately underperform on some parameters.

Denmark has a population of 5.75 million and Danish is a very small language in terms of number of speakers in comparison to, for example, Chinese, English, Spanish, German or French. The Danish dictionary market is also relatively unique because of its limited size, its English-competent users and the limited number of dictionary publishers. This obviously frames the discussion, and it may be argued that the findings and conclusions presented in this paper would perhaps have been different had the analysis been made in, for example, the United Kingdom.

The kaleidoscopic approach, where business model generation in Denmark has been analysed by means of five selected theoretical lenses, reveals additional relevant

insights in how to design a new and more viable lexicographic business model.

When looking at business models through the lens of strategy, it seems as if there is a lack of strategic planning and execution in the vast majority of the surveyed case organizations. One example from a company, which in fact has had a clear strategic approach to innovation processes and core competencies, illustrates that a clear plan seems to be working. When asked about its strategy, the CEO said *"We decided on a clear digital strategy and started to contact our customers directly. And it worked".*

When applying the lens of core competencies it also seems as if the large majority of the 15 companies have had no or little strategic direction in their treatment and management of their core competencies. When asked about core competencies one CEO said *"We have had a very large employee turnover. What we did was to outsource a number of functions but in line with our digital strategy we bought an entire software company and added 12 new software specialists".* This is both an example of a company with a very clear strategy of what it wants to focus on and also a company with a clear approach to capabilities in the form of core competencies. The company has decided to enter a new market (for digital learning materials) and a clear consequence of that is to add new core competencies in this field, see also the core competence matrix in Figure 4 above (Hamel & Prahalad, 1994: 227). In stark contrast to an active and strategic approach to one's core competencies is the statement from a CEO, who says *"We do not have an active approach. We keep having technical challenges, so no, we do not prioritise that".*

Closely connected to the previous two lenses is the lens of strategic innovation. When applying the lens of strategic innovation on the interview data it also seems as if there is a very limited active approach to innovation in the majority of the 15 case companies. One CEO from one of the very successful companies said, when asked about strategic innovation, that *"We closely analysed what the other companies did. And then we disrupted everything by doing something entirely different",* which in fact is in line with Christensen's (1997) recommendations on disruptive technologies and Kim & Mauborgne's (2004) recommendations on creating a new Blue Ocean with no competitors. When asked about innovation in the Danish lexicographic industry, another CEO succinctly said *"The business model of the established dictionary companies has indeed been challenged. There has been too little innovation and too little focus on the customers. I think we should have acted quicker".*

The fourth kaleidoscopic lens is business understanding and again it must be concluded, based on the empirical data, that the vast majority of the surveyed case companies seem to have limited business understanding. It may sound unfair, but many of the interviews with CEOs seem to indicate that most companies are neither competent enough in conducting market, competitor and customer analyses nor analysing and acting on the data and the changing market conditions. When asked about the ability to understand the value chain and the market, one CEO in fact said:

"We contacted the decision makers. And it really hit off when we managed to convince the Ministry of Education to allow pupils and students to use electronic dictionaries during exams. And from that point we went from 0 to all but two municipalities. This is an example of a company, which has been able to analyse and interpret the value chain and to focus its sales organization on the decision makers, which resembles the adoption chain approach (Adner, 2012). The point is that the value chain has changed dramatically in the lexicographic industry (Hall, 2013), which discusses the business of digital publishing.

Finally, the fifth theoretical lens of organizational inertia also reveals a number of interesting insights. Again, it is argued that the majority of the surveyed case companies seem to have suffered from negative organizational inertia (Sull, 1999 and Gavetti, 2005). This argument can in fact best be supported by a statement from a CEO, who reflects on the lexicographic industry's ability to innovate and develop. He said *"As a whole, I think the industry as such has had a very closed mind-set. We have isolated ourselves, we have not developed and we have placed ourselves on a pedestal – you know – something with public funding and the literary element etc. And in that process we have been disrupted because we thought things would not change".*

If the above theoretical lenses had been applied on international dictionary markets, a somewhat different picture would probably have appeared. One example is when MacMillan decided to go 100% online almost 10 years ago, which spurred a heated debate in lexicographic circles. Another example is the host of partnerships and contributions on the future of dictionary-making described in Kernerman Dictionary News in the years 2008-2009. So on the international dictionary markets, dictionary publishers have no doubt already been working with new business models for a decade or so.

However, in conclusion the analysis and discussion of the empirical data from the Danish market by means of the five theoretical lenses have nevertheless resulted in a number of insights on how we might develop a new and more viable lexicographic business model. The focus of the next part of this article is thus to analyse and discuss the last two research questions.

5. Elements of a New Lexicographic Business Model

Milton Friedman allegedly said in 1970 that *"the business of business is business",* i.e. that businesses should only engage in activities with a view to create profits. This is of course a somewhat bold statement in this context, but it is argued that this approach is perhaps what we need in lexicography. The question is: have we focussed too much on developing the quality and amount of linguistic data and too little on developing new distribution platforms and new business models?

5.1 The Business of Lexicography is Business

On the basis of the insights gained from the analysis of the empirical data, it is in fact argued that the business of lexicography is business. Commercial lexicography is business. Publicly-funded lexicography is business. And business is about making strategic decisions about focus and innovation, etc. And this seems to have been one of the challenges of the Danish lexicographic industry: i.e. that there has been too little focus on making business decisions.

It all starts with strategic decisions and leadership and about having a clear strategic focus and not being "*stuck in the middle*" (Porter, 1985: 11–15). This not only includes decisions on differentiation and cost, but also important decisions on strategic innovation and investments. And as the empirical data indicate, this seems to have been one of the biggest challenges of the companies surveyed.

So it all boils down to strategic decisions not having been made in time, and instead almost the entire Danish lexicographic industry has been suffering from what is sometimes referred to as the "*sailing ship effect*" (Gilfillan, 1935: 156). The "sailing ship effect" is the typical reaction of companies when they face new disruptive technologies. They simply continue to invest in old technologies to retain their competitive position in that market. The "sailing ship effect" refers to the situation whereby sailing ships were heavily improved the moment the steam ship emerged during the 19th century. This reaction also resembles what March calls "exploitation of old certainties" (March, 1991).

It is always easy to be "Captain Hindsight", but obviously an entire industry has been suffering from the "sailing ship effect" for too long. Instead, the Danish lexicographic industry could have made a number of strategic decisions. Investing in old certainties and continuing to make small, incremental improvements of the lexicographic data is a natural reaction, but it is only a good idea if the decision is made strategically. Because arguably lexicographic companies can, in line with Ansoff's growth strategy matrix (Ansoff & McDonnell, 1988: 109), make four fundamental types of decisions:

- To improve the performance and characteristics of old lexicographic technology, i.e. existing market and existing product (Penetration).
- To develop the performance and characteristics of old lexicographic technology into new lexicographic technology, i.e. existing market and new product (Product Development).
- To introduce old lexicographic technology into new markets, i.e. new market and existing product (Market Development).
- To diversify and develop old lexicographic technology into new lexicographic technology and establish a new market, i.e. new market and new product (Diversification).

It seems as if the lexicographic industry has focussed too much on the two options *Stay* or *Go*. However, it is of course not as simple as that. When a lexicographic company faces huge technological changes it has to consider investments already made, the cost of future investments and the cost of its existing production systems. Again, everything should be based on rational business decisions. And rational business decisions also sometimes mean the need to discontinue business activities to limit losses. Especially when there is the risk of disruption and drastically changed market conditions (Christensen, 2007).

In principle, the decision is synonymous to the hit song "*Should I Stay or Should I Go*" by The Clash. I would argue that there are four types of decisions:

- Go and exit. Leave the market in the long term, but try to harvest as much value as possible in the short term with a view to leaving the market (exit strategy).
- Go and relocate. Leave the market but relocate in new adjacent markets and industries to apply core competencies and technologies (disruption strategy).
- Stay and contract. Retrench and try to sustain a competitive position in a niche market with a view to contracting and surviving (technological retreat).
- Stay and expand. Retrench and invest in new technologies and new platforms with a view to create new value (strategic innovation).

Having a clear and rational business mind-set is a precondition for making sound business decisions. And it is argued that the Danish lexicographic industry neither seem to have had a sufficiently focussed business mind-set nor to have had enough focus on business core competencies.

To sum up, it is argued that the business of lexicography is business. And with all due respect, calls like "bridging the gap between the general public and scholarly dictionaries" (cf. www.elexicography.eu) are not business. Calls like that do not solve the underlying problem: that the demand for lexicographic products has plummeted, because the business model of many existing lexicographic products has disappeared. At least that seems to be the case in the Danish market, where online dictionaries are playing an increasingly smaller role, for e.g. professional translators (Bundgaard, 2017). Lexicographic data are neither sufficiently integrated in our job-related tools nor sufficiently integrated with or related to the tasks that we solve. And I would argue that that is one of the biggest challenges of the existing business model.

Instead we should focus our efforts on either staying or leaving. Sometimes we have to leave a market in time to avoid becoming the next in a long line of disrupted companies like Kodak or Blockbuster, etc. And, according to a recent survey among more than 2,000 C-level executives, the media industry is expected to be the most disrupted industry in the next 12 months (Grossman, 2016). I argue that the lexicographic industry is similar to the media industry.

If we decide to stay in the business we need to form new and value-creating partnerships with, for example, robot or A.I. companies. We could also develop new lexicographic products which, to a much higher extent than the existing products, create value for the customer and would be in demand by the general public, for example by focussing on the distribution platforms and task relevance.

When diversifying into adjacent markets we need to ask ourselves how the assets and core competencies of our company can be used in an adjacent market with potentially millions of new customers; how our value system is performing and moving us upwards in the value chain (Adner, 2012); and finally we must find where customers are underserved and decide where we could solve their problems.

5.2 Different Lexicographic Business Models for Different Services & Markets

First of all it is important to realize that we cannot develop a one-size-fits-all type of business model. A lexicographic business model and its underlying unique value propositions are naturally dependent on the Value Map of the company (what the company offers), the Customer Profile of the company (what the customer wants) and of course the Market in which the company operates (market conditions).

On the basis of the empirical data it is argued that there are at least five different types of lexicographic business models.

A. Commercial, private dictionary publisher

The typical mission of this type of company is to create monetary revenue. The focus of the activity is on the delivery of linguistic data. Example: ordbog.gyldendal.dk.

B. Commercial, private, educational dictionary publisher

The typical mission of this type of company is also to create monetary revenue. The focus of the activity is on the delivery of linguistic data with a learner focus. Example: ordbog.gyldendal.dk or ordbogen.com.

C. Non-commercial, public dictionary publisher

The typical mission of this type of company is to create cultural value or national language value. The focus of this type of activity is on the delivery of linguistic data. Example: dsn.dk or ordnet.dk.

D. Commercial, private, industry-specific lexicographic activity

The typical mission of this type of activity is to create monetary revenue and to create

branding value for the industry in question. The focus of this type of activity is on the delivery of industry-specific lexicographic data. Example: Medicin.dk.

E. Commercial, private, company-specific lexicographic activity

The typical mission of this type of activity is to create monetary revenue and to create branding value for the company in question. The focus of this type of activity is on the delivery of company-specific lexicographic data. Example: TeleLex, ZooLex, COWILex (Simonsen, 2002 and 2007).

5.3 Considerations on a Lexicographic Value Proposition and Business Model

In addition to the above theoretical considerations on different lexicographic business models, it is now time to discuss proposals for new and more sustainable lexicographic business models.

The discussion starts with customer value, defined by Drucker (1999: 57) as *"What the customer buys and considers of value is never a product. It is always a utility – that is – what a product does for him"* and it must be argued that conventional lexicographic products do not do anything – or at least not enough - for the customer. According to Drucker (ibid) customer value can be defined as:

$$\textit{Customer Value} = \frac{F \quad b \quad +e \quad b}{M \quad c \quad +Ti \quad c \quad +E \quad c \quad +P \quad hli \; c}$$

So we need to do something about both the functional benefits and the emotional benefits. And we need to learn so much more about not only the user, but also his job tasks, his functional benefits and his emotional benefits. And we can do that by means of value stream analyses, whereby the value stream while completing different job tasks is analyzed and measured (Martin & Osterling, 2014: 9–20). Having established what customer value is, we can now venture into the analysis and discussion of the value proposition of "new lexicography".

As it was already pointed out at the beginning of this paper, the biggest problem of lexicography is that lexicographic products are no longer perceived as relevant for the vast majority of people. Most people in fact do not use dictionaries, and if they need to find help when communicating or when looking for data, they simply use the Internet instead.

So dictionaries are in fact not being used as much as we want them to be. The most important question is: why do not people use online or mobile dictionaries? Obviously, there are a number of reasons, but I would argue that the most important reason is

that most lexicographic resources are not tool-integrated and not specifically related to the user's job tasks.

In order to get an overview of what really is needed by the user, a Customer Profile (Osterwalder et al., 2014) should be used. According to Osterwalder et al. (2014: 53) we first need to categorize the arena in which our value proposition should have effect and four different arenas are described: *Financial, Digital, Physical/Tangible and Intangible*. For the purpose of this discussion, the lexicographic industry operates in the *Digital, Physical/Tangible and Intangible* arenas.

As explained above, a Customer Profile includes an analysis and identification of *Customer Jobs* (the tasks that the customers are trying to complete)*; Customer Pains* (the obstacles, hassles and risks that occur when the customers are trying to complete the job); and finally *Customer Gains* (the outcomes and benefits that the customers are harvesting when completing the job). What the customers do can be characterized as functional, emotional, personal and supporting jobs (Osterwalder et al., 2014).

If we use the Value Proposition Canvas shown in Figure 3 above and we start listing typical customer jobs of the particular type of customer in mind in the *Customer Job* field, it soon becomes clear that many of the jobs listed and the associated benefits seem to be mainly functional and here lies perhaps one of the fundamental reasons why people do not use lexicographic products as much as we would like to. In the Social Age, functional benefits are not enough and we need to consider how to give users more emotional and personal benefits.

The final steps in building a Customer Profile are the Customer Pains field, which lists the customer pains connected to solving customer jobs, and the Customer Gains field, which covers the benefits of the jobs and the positive outcomes associated with completing the jobs. When listing these pains and gains, it soon becomes clear that customers would probably list pains like time-consuming, task un-related etc. and that they look for gains like convenience, task-related and learning gratification.

Once considerations on the Customer Profile are done we can take another look at the Value Map of the Value Proposition Canvas in Figure 3 above. Depending on the type of customer, market, product or service in mind we can now start listing some of the most important Pain Relievers vis-à-vis the pains listed in the Customer Profile. We can then move on and list some of the most apparent Gain Creators, which in this example probably would be tool and task integration. So, having established what the customer does and what he likes and dislikes, etc., it is now possible to describe the Products & Services and design what we are going to offer and thus create the fit between what customers need and what we offer.

On the basis of these considerations and taking into account that this is a general and non-exhaustive example, it is now possible to design a new lexicographic business model canvas based on Osterwalder and Pigneur (2010).

Figure 5 shows what might be described as a general proposal for a lexicographic business model, and as it will appear I have listed in blue a number of ideas in each of the nine fields. Obviously, this is a general example and it is based on the decision Stay and expand. It is argued that the most important effort is to focus more on integrating the lexicographic data with the tools that we use and to make lexicographic data task-related and thus integrate them into the customer's value chain.

Figure 5: General Example of a Lexicographic Business Model

All these considerations can be summarized as six theses on a new and more viable lexicographic business model. The six theses are:

⟩ Thesis 1: From lexicographic products to lexicographic services

We need to move upstream in the value chain and offer lexicographic services instead of just lexicographic products. So this thesis takes its starting point in field 1 in Figure 2. By moving upwards in the value chain, lexicographic data can be integrated vertically into the customer's value chain and thus become an important, indispensable and value-creating element for the customer. Customers need solutions and advice on real-world problems and lexicographic data could be value-adding by offering communications consultancy services.

⟩ Thesis 2: From lexicographic data to lexicographic platform and distribution

The empirical data also clearly indicate that we need to focus on lexicographic platforms and new ways of distribution instead of making small incremental improvements of the linguistic data. So this thesis takes its starting point in field 6 in Figure 2. The channels are extremely important, especially in the Social Age, and customers need easy access to, not simply more, linguistic data.

⟩ Thesis 3: From lexicographic data competencies to platform competencies

The analysis also clearly showed that we need to focus more on platform competencies than data competencies (Hamel & Prahalad, 1994). This thesis starts in field 4 in Figure 2 and concerns replacing old core competencies with new core competencies, which focus on developing innovative platforms and access methods.

⟩ Thesis 4: From lexicographic data to lexicographic user and user job

The analysis and the discussion also showed that we need to focus more on the user jobs to make the lexicographic data in question related to a real-world task. We need to focus more on understanding the value stream of the customer and making the appropriate quality adjustments. This thesis is primarily based in fields 5 and 7 in Figure 2, as it deals with customer relationships and the customer segments.

⟩ Thesis 5: From dictionary to lexicographic data in software, artificial intelligence and augmented reality

On the basis of the data and the analysis above, thesis 5 argues that we need to focus on the lexicographic data used in adjacent industries, in co-creation initiatives, in partnerships and in artificial intelligence or augmented reality industries. One challenge of these industries is to facilitate interaction, and lexicographic data would be a huge asset. This thesis focuses on using lexicographic data in new and adjacent industries and takes its starting point in fields 1 and 3 in Figure 2.

⟩ Thesis 6: From dictionary to experience and sales-based services

This thesis also argues that lexicographic data can be used in adjacent industries and in alternative setups. The closer the integration with the customer's value chain, the better, and especially in operative functions, in experience-based functions and in sales functions, cf. www.altomhus.dk, which is an example of lexicographic data being used in a double-loop sales channel. This thesis takes its starting point in fields 1 and 9 in Figure 2.

6. Conclusions

In this paper, the lexicographic business model on the Danish market was analyzed and discussed. The analysis of 15 interviews with senior executives and CEOs from the Danish lexicographic industry shows that it is time to start lexicography from scratch and to design new and more sustainable lexicographic business models. The analysis indicates that the value chain has shifted from lexicographic content to lexicographic platform. The paper addressed three research questions.

The first research question was to analyze and discuss what characterizes the current understanding and application of the concept business model in the Danish lexicographic industry. The answer to this question was that an overwhelming majority of the interviewed senior executives and CEOs neither knew the term business model nor had a business model. The analysis of the interviews also indicated that the reason why they did not have a working business model was that the lexicographic industry had had too little focus on strategic innovation and on core competencies and that the industry for too long has suffered from negative organizational inertia.

The second research question was to discuss and describe the different types of lexicographic business models. On the basis of the interview data and the theoretical models and considerations discussed, it was first argued that the business of lexicography is business, and four overall strategic avenues for lexicographic companies were proposed (*Go and exit, Go and relocate, Stay and contract, Stay and expand*). With the insights from the analysis, it was also possible to develop and describe at least five different lexicographic business models based on parameters such as value creation type (Value Proposition) and focus of activity (Value Map).

The third research question was to discuss and develop potential new elements of a new and more viable lexicographic business models. On the basis of the analysis of the empirical data and the lexicographic considerations made, six theoretical theses on lexicographic business models were developed and discussed. The validity of the theses were supported by means of the lexicographic Business Model Canvas (Osterwalder & Pigneur, 2010).

In conclusion, it is important to remember that there are different lexicographic business models for different services and markets and that a business model should be company-specific. Moreover, even though this analysis focused on the Danish dictionary market, it is argued that selected insights and conclusions from this research can be generalized and applied to a number of other dictionary markets.

Further research in lexicographic business modelling is needed and time will show whether the lexicographic industry is up to the challenge and able to reinvent itself and start from scratch.

7. References

Adner, R. (2012). *The Wide Lens*. Penguin Putnam.

Afuah, A. (2009). *Strategic Innovation: New Game Strategies for Competitive Advantage.* New York. Routledge.

Ansoff, I.H., & McDonell, E.J. (1988). *The new corporate strategy*. New York. Wiley.

Beckmann, O.C., Royer, S. & Schiavone, F. (2016). Old but sexy: Value Creation of old technology-based business models, *Journal of Business Models*, Vol. 4, No. 2 pp. 1-21

Bundgaard, K. (2017). (Post-)editing – A Workplace Study of Translator-Computer Interaction at Textminded Danmark A/S. Ph.d.-afhandling. Aarhus. Aarhus BSS, Aarhus University, Department of Management.

Christensen, C. (1997). *The innovator's dilemma. When new technologies cause great firms to fail*. Boston: Harvard Business School Press.

Drucker, P. F. (1994). The Theory of the Business, *Harvard Business Review*, Sept-Oct, 95-104.

Drucker, P. F. (1999). *Management: Tasks, Responsibilities, Practices*. (Oxford, United Kingdom: Butterworth-Heinemann, 1999, p. 57.

Gavetti, G. (2005). Strategy Formulation and Inertia. In. *Harvard Business Review.*

Grossman, R. (2016). The Industries That Are Being Disrupted the Most by Digital. In: *Harvard Business Review*. March 2016.

Hall, F. (2013). *The Business of Digital Publishing*. London. Routledge.

Hamel, G.H. & Prahalad, C.K. (1994). *Competing for the Future*. Boston, MA: Harvard Business School Press.

Johnson, G., Scholes, K., Whittington, R. (2012). *Fundamentals of Strategy*, 2/E. London. Financial Times Press.

Kim, W. Chan & Mauborgne, Renee (2004). Blue Ocean Strategy. In: *Harvard Business Review*. pp. 76-85.

Kvale, S. (2007). *Doing Interviews*. London. SAGE Publications Ltd.

March, J.G. (1991). Exploration and exploitation in organizational learning. *Organization Science*, vol. 2, 71–87.

Martin, K. & Osterling, M. (2014). *Value Stream Mapping*. McGraw-Hill.

Mintzberg H. & J. Lampel (1998). Reflecting on the Strategy Process, *Sloan Management Review*, Spring 1999, pp. 21-30.

Osterwalder, A., & Pigneur, Y. (2010). *Business Model Generation: A Handbook For Visionaries, Game Changers, and Challengers*. Hoboken. John Wiley & Sons, Inc.

Osterwalder, A., Pigneur, Y., Bernada, G., & Smith, A. (2014). *Value Proposition Canvas: How to create products and services customers want*. Hoboken. John Wiley & Sons, Inc.

Porter, M.E, (1985). *Competitive advantage: Creating and Sustaining Superior Performance*, The Free Press, New York.

Saunders, M., Lewis, P., & Thornhill, A. (2009). *Research methods for business*

students. Harlow: Pearson Education.

Simonsen, H. K. (2002). *TeleLex - Theoretical Considerations on Corporate LSP Intranet Lexicography: Design and Development of TeleLex - an Intranet-based Lexicographic Knowledge and Communications Management System.* Ph.d.-afhandling, 436 sider. Handelshøjskolen i Århus.

Simonsen, H. K. (2007): Virksomhedsleksikografien viser tænder: leksikografiske løsninger i København Zoo og Fagerberg A/S. In: Nordiske studier i leksikografi 10. Rapport fra Konference om leksikografi i Norden. Island 22.-26. maj 2007. Akureyri: Nordisk forening for leksikografi 2007.

Simonsen, H. K. (2016). *Hvor er forretningsmodellen? En analyse af de forretningsmæssige udfordringer i forlags- og informationsindustrien med særlig fokus på opslagsværker.* MBA-afhandling. Institut for Økonomi og Ledelse. Aalborg Universitet.

Simonsen, H. K. (2017). Hvor er forretningsmodellen? In: *LEDA-nyt nummer 63 - april 2017*, pp. 5-17.

Sull, D. N. (1999). Why Good Companies Go Bad. In. *Harvard Business Review*, July-August, 2-11.

Open Access to Frisian Language Material

Eduard Drenth, Pieter Duijff, Hindrik Sijens

Fryske Akademy, Box 54, 8900 AB Leeuwarden (NL)

E-mail: edrenth@fryske-akademy.nl; pduijff@fryske-akademy.nl; hsijens@fryske-akademy.nl

Abstract

The Fryske Akademy has a long history—since 1938—of developing printed Frisian dictionaries and word lists, usually with Frisian and Dutch or Dutch and Frisian as the source and target languages, respectively. In the 1990s, the Akademy also began working on digital language resources for Frisian: a language database, various digitized dictionaries, a digital preferred vocabulary for Frisian and an Online Dutch–Frisian translation dictionary.
This paper briefly describes the available digital language resources and how access to them can be improved by means of a yet-to-be-developed application programming interface (API). The Fryske Akademy has three primary user groups in mind: language users, linguists and developers. A list of superlemmas will be compiled to link the information in the different systems.
Several examples are used to illustrate the requirements demanded of the API. Underpinning all this are the questions that might be asked by the three user groups of the language resources. Sections 5 and 6 describe the work and projects that are required to implement the API. The final section outlines a roadmap for potential future developments.

Keywords: linguistics; API; service; corpora; dictionaries

1. Introduction

The Fryske Akademy (FA) has a number of digital language resources, but these are largely independent of one another. In addition, some cannot be accessed by the public from outside the FA, despite the Akademy's aim to make its products available through open access wherever possible. Taking the needs of its target groups as the starting point, the FA plans to use an application programming interface (API) to provide access to data in the language resources. This paper aims to show how the FA will serve its target groups via the API. The API will not be discussed in detail here; instead, we will use examples to demonstrate how the API can be used to retrieve information from different data sources in a coherent way. Key principles for the API are standardization of the interface, and ease of access and service provision for users.

Before discussing the technical provisions and requirements that the API must satisfy, we first describe the language functionalities at the FA that will underpin the development of the API. We then identify the target groups we need to serve: language users, linguists and developers. We describe how these groups are currently utilizing our resources and the options we will offer in the future for making digital language material accessible for language users, researchers and developers.

2. Current language resources

2.1 Preferred vocabulary

Frisian, the second national language of the Netherlands, is a minority language with a limited written tradition, even within the province of Friesland where it is the native language. Frisian spelling was officially established for the first time in 1879 and it was not until after the Second World War that these spelling rules were officially adopted—in a slightly modified form—by the province of Friesland. Standard Frisian did not develop until the latter half of the nineteenth century; much later than, for example, Dutch. The standard language has been recorded in dictionaries and teaching resources during the past 120 years. A preferred vocabulary, which is essentially a list of standard forms (Taalweb.frl), has been made available online by the FA since 2015. For a detailed description of the preferred vocabulary, see Duijff (2016).

Since the development phase of Frisian, it has been common practice in written Frisian to accept different dialect variants alongside one another. Even though increasingly fewer variants are to be found in Standard Frisian dictionaries and vocabularies, standard Frisian continues to display greater variation than Dutch (Breuker, 2001; Duijff, 2008; 2016; Duijff & Van der Kuip; 2017). This variation also applies both to dialect forms and spelling variants in the preferred vocabulary. Because Frisian has acquired a growing role within education and as a written language, this has sparked a need for a list of standard or preferred forms, which the provincial government subsequently commissioned. In 2014 the FA created a database of preferred forms, in which the different variants are linked to the respective preferred forms (see Figure 1).

The database underpinning this vocabulary currently contains 96,146 lemmas, whose sources are the lists of lemmas for various Frisian dictionaries, supplemented by recent material from a range of sources. Of the 96,146 lemmas in the database, 85,730 can be labelled as standard forms (89.2%) and 10,146 (10.8%) as variants of these forms. In addition to lemma forms, the database provides word information in the form of word type, paradigm information and hyphenation. This database of standard forms is already being used in an application, namely a spelling checker (see Sijens & Dykstra, 2013: 96-99).

The preferred vocabulary is stored in an access database, which comprises several tables that are linked via IDs. The main tables are ‘lemma’ and ‘paradigm’. In addition to a column with the lemma form, the lemma table has columns with part-of-speech information and preferred form marking, etc. The paradigm table contains, in addition to a column with the paradigm forms, a column with hyphenation forms and a column where the form can be marked as the preferred form.

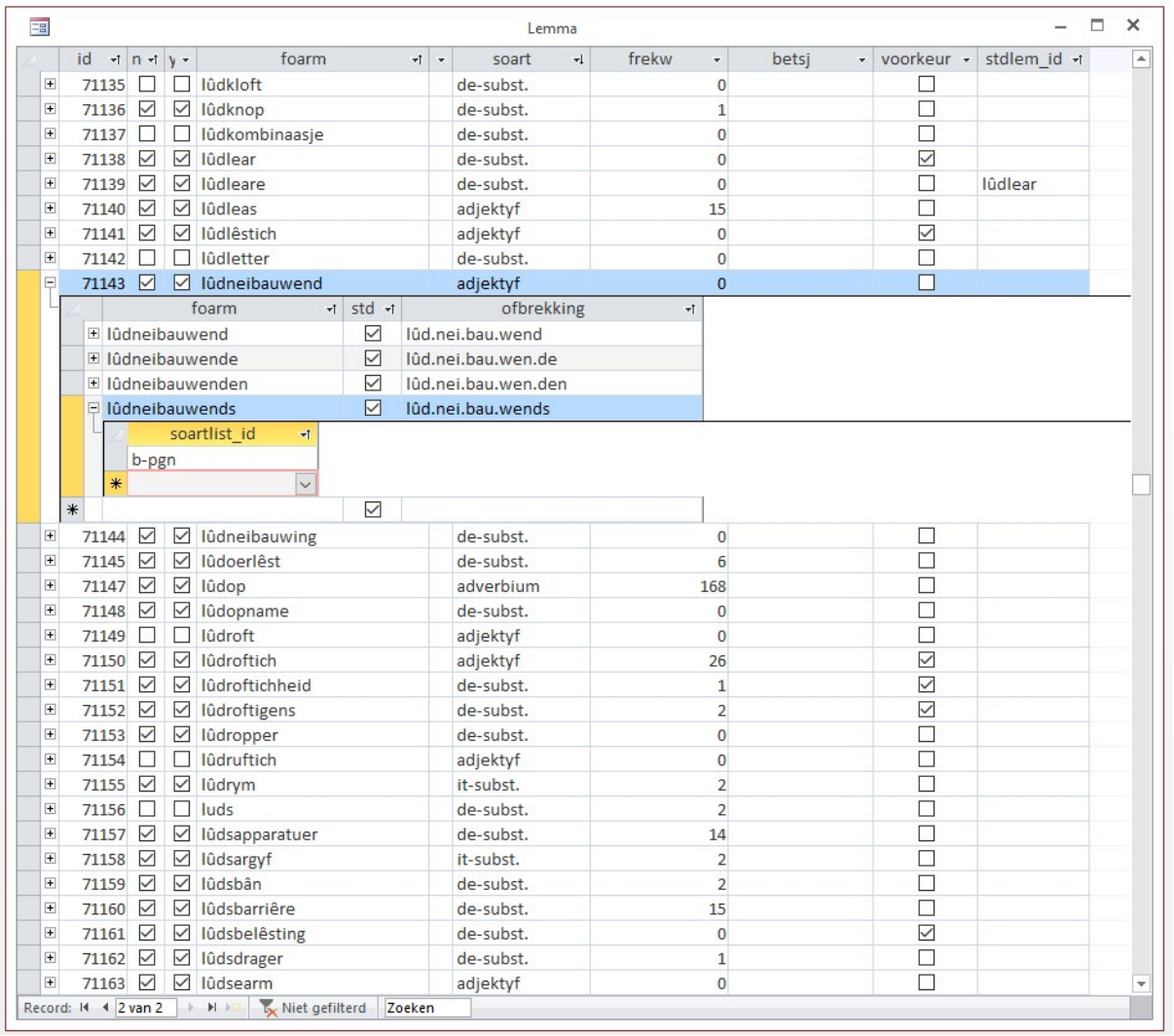


Figure 1: Screen of the preferred vocabulary database

2.2 Digital dictionaries

Since its establishment, the FA has compiled several dictionaries of Frisian. They are almost all bilingual, with mostly Frisian and Dutch alternating as the source and target languages. The most frequently used and most comprehensive translation dictionaries are still Zantema (1984), with 55,000 lemmas, and Visser (1985), with 45,000 lemmas. The historical/academic dictionary WFT (1984-2011) is also a bilingual dictionary, in the sense that Dutch is used to describe the Frisian language material. The most recent comprehensive desk dictionary with 70,000 lemmas is the monolingual Frysk Hânwurdboek/FHW (2008). Together with other dictionaries, these desk dictionaries can be consulted online at *Taalweb.frl*. The WFT can be consulted and searched online at Gtb.inl.nl (Depuydt et al., 2017). All these dictionaries were first developed as paper dictionaries and were only later made available online to language users.

2.3 Online Dutch–Frisian Dictionary

To meet the need for a modern, contemporary Dutch translation dictionary, the FA has begun compiling the Online Nederlands–Fries Woordenboek ('Online Dutch–Frisian Dictionary'/ONFW). The ONFW is an online production dictionary that takes modern standard Dutch as its source language and the standard Frisian equivalent as its target language. The dictionary will present not only the meaning and use of words and phrases, but also grammatical information. The dictionary will appear in parts from 2018 to 2022, after which it will continue to be updated and expanded (Duijff & Van der Kuip, 2017). For the source language, the ONFW will draw on the language corpus of the Algemeen Nederlands Woordenboek (ANW), an online dictionary of contemporary standard Dutch in the Netherlands and Flanders that describes Dutch vocabulary since 1970 (Schoonheim & Tempelaars, 2010: 718).

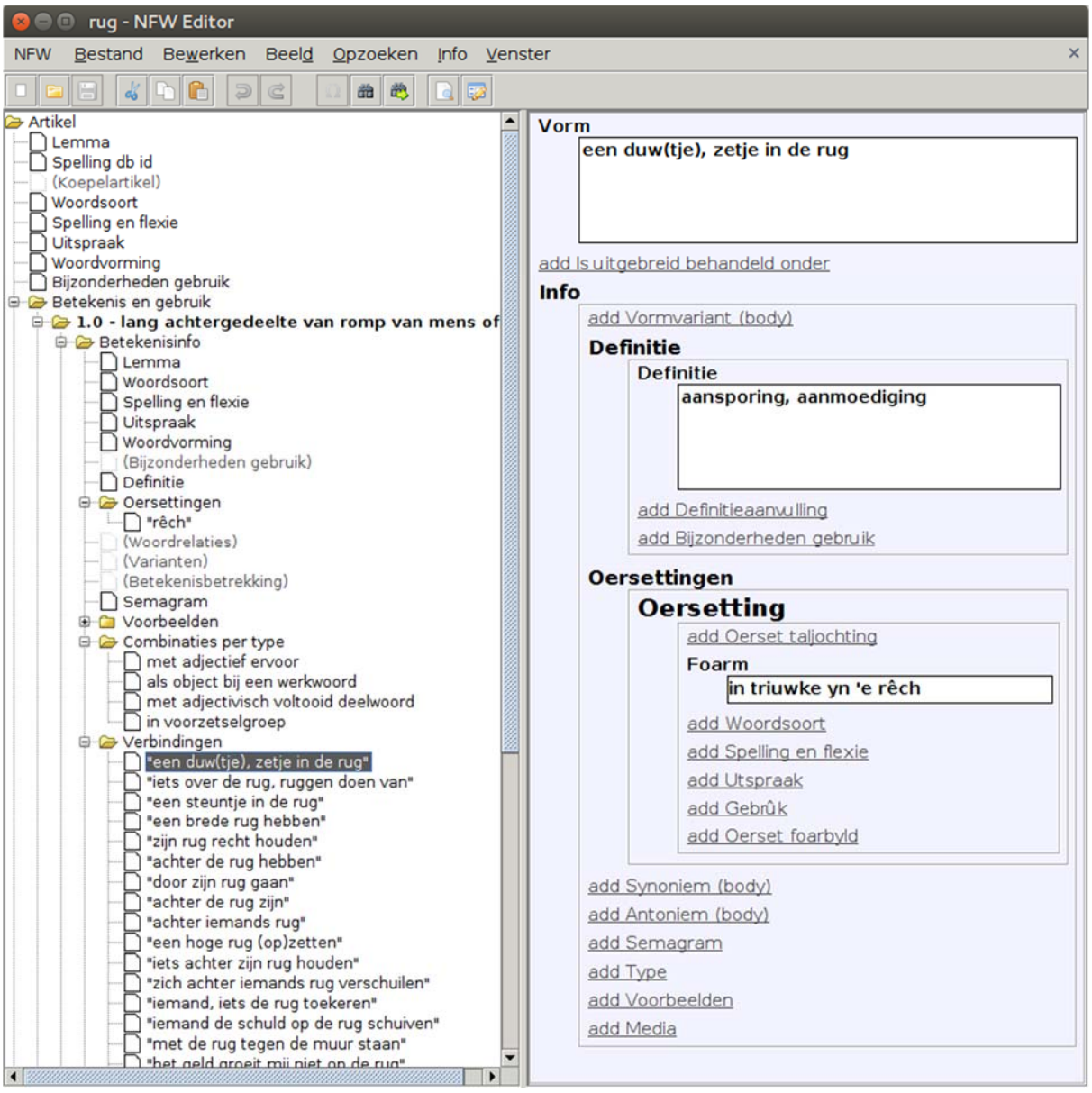


Figure 2: DWS screenshot of one page

The ANW, which is still being compiled, can be accessed at anw.inl.nl. The dictionary writing system (DWS) for the monolingual ANW has been modified so that it can be used for the bilingual ONFW. Figure 2 gives an idea of the DWS for the ONFW.

Using DWS enables editing of XML to conform a schema; below a snippet of the schema is shown.

```
<?xml version="1.0" encoding="UTF-8"?><xs:schema
xmlns:xs="http://www.w3.org/2001/XMLSchema" version="1.0">
    <!-- xmlns:vc="http://www.w3.org/2007/XMLSchema-versioning" vc:minVersion="1.0"
vc:maxVersion="1.0" -->

    <xs:element name="Oersettingen">
        <xs:complexType>
            <xs:sequence>
                <xs:element name="Taljochting" type="xs:string" minOccurs="0"/>
                <xs:element ref="Oersetting" maxOccurs="unbounded"/>
            </xs:sequence>
        </xs:complexType>
    </xs:element>

    <xs:element name="Oersetting">
        <xs:complexType>
            <xs:sequence>
                <xs:element name="OersetTaljochting" type="xs:string" minOccurs="0"/>
                <xs:element name="Foarm" type="xs:string"/>
                <xs:element ref="Woordsoort" minOccurs="0"/>
                <xs:element ref="SpellingEnFlexie" minOccurs="0"/>
                <xs:element ref="Utspraak" minOccurs="0"/>
                <xs:element ref="Gebrûk" minOccurs="0"/>
                <xs:element name="OersetFoarbyld" type="xs:string" minOccurs="0"/>
            </xs:sequence>
        </xs:complexType>
    </xs:element>
```

The bilingual DWS can also be used for other translation dictionaries that have Dutch as the source language. With some modifications, the system has already been made suitable for a yet-to-be-developed bilingual online dictionary of Dutch–Stellingwerfs. Stellingwerfs is a Saxon language variety spoken in the southeast of the Dutch province of Friesland and northwest of the Dutch province of Overijssel.

The ONFW will consist of a MySQL database and a Java application. The key feature of the database is a field with an XML that complies with an XML Schema. The database also contains user information, logging, status information for articles and workflow information. The XML can be edited using the Java application and the work status can be updated. The XML contains detailed information about Dutch entries and their Frisian translation, including spelling and inflection, word forms, pronunciation, combinations and fixed expressions.

2.4 Language databases

The FA has various Frisian text corpora containing data from the period 500–2017. The main ones are the corpus of Old Frisian (500–c. 1550), Middle Frisian (c. 1550–1800) and Modern Frisian (1800–2017). In addition to these three databases of exclusively written material, there is also a corpus of spoken Frisian, compiled from the period 2002–2006. Additional contemporary spoken material is currently being collected and an old spoken corpus and will be made available once more. In this contribution, we confine ourselves to written Frisian.

2.4.1 Corpora

The Old Frisian corpus comprises texts from the entire Old Frisian period until about 1550. It is a closed corpus with about 323,000 tokens. The material is for the most part linguistically annotated (lemmatized and tagged with part of speech). The Middle Frisian corpus contains all surviving Frisian texts from the period 1550–1800. This closed corpus is linguistically annotated and contains 488,000 tokens and 19,000 lemmas.

The corpus of Modern Frisian comprises a selection of written texts from the nineteenth and twentieth centuries and contains approximately 25 million tokens. The nineteenth-century part contains a small selection of prose written at that time. Efforts have been made to ensure that the twentieth-century part of the corpus is as representative as possible to provide maximum coverage of the Frisian vocabulary (Dykstra & Reitsma, 1995: 63). The corpus is not linguistically annotated.

2.4.2 Web interfaces

There are various interfaces that provide access to the corpora. Three of them can be used via the internet, and one can only be used internally within the FA for copyright reasons. The interfaces were all developed at different times, using different techniques and with different aims. The oldest interface gives the option of searching the various corpora by word form, possibly with the help of wildcards. Figure 3 gives an impression of a search with results in the oldest corpus. The results are presented in a concordance that can be ordered alphabetically by the word occurring to the left or right of the keyword.

Searches can also be made in sub corpora. A distinction is made between the three different language phases for Frisian: Old, Middle and Modern Frisian. Frisian in the period 1900 can in turn be broken down into different periods distinguished by clearly identifiable spellings. This interface was developed in 1998, primarily for the lexicographical projects that the FA was, and continues to, work on.

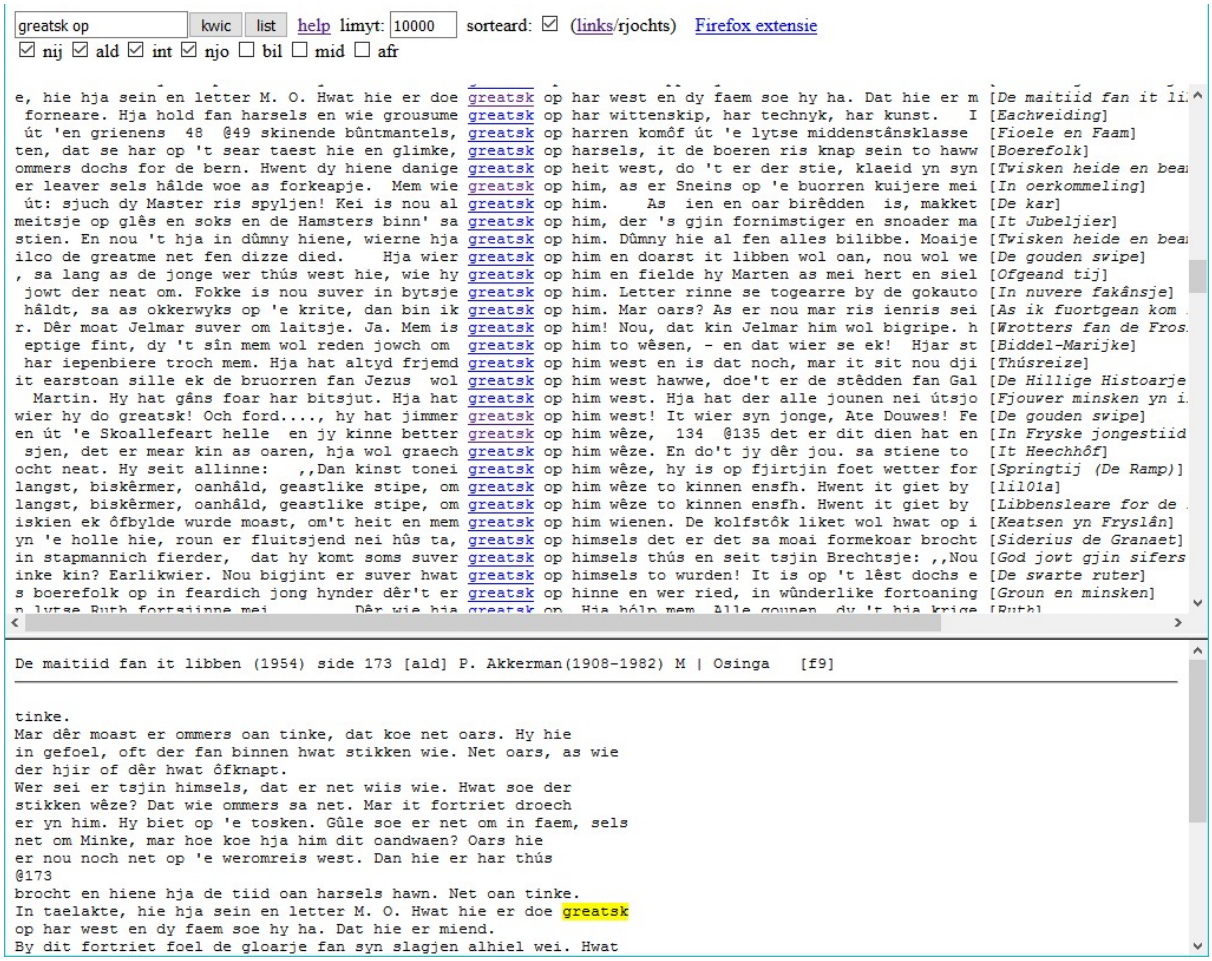


Figure 3: Search and results in the oldest corpus

The second interface provides access to the linguistically annotated corpus of Middle Frisian from the period c. 1550–1800, combined with corpus material from the earlier and later periods. Users can search by lemmas and word forms, possibly with the help of wildcards. They can opt to have the results presented in a concordance or in a list of word forms. This corpus is linked to a bibliography of secondary literature. Another special feature is that geographical information that is linked to lemmas can be downloaded. With a designated account, the database is freely accessible via http://pc245.fa.knaw.nl:8020/tdbport/. It was developed in 2002 to give easy access to Middle Frisian material, and with the option of adding more corpora. This interface continues to fulfil a need, namely searching by word form, or by morphological, diachronic and paradigm information.

A third interface, developed in 2009, makes the Old, Middle and Modern Frisian material accessible to a wider audience. With this interface, users can access sources directly or can search by lemmas. The results can be shown in KWIC (keyword in context) view, with an option to show the sources. Zantema (1984) is also integrated and his bibliographies are linked to this interface, which is freely accessible at

http://tdb.fryske-akademy.eu/tdb/. Unique in the language database are the integrated scans of medieval manuscripts, the integrated Old Frisian dictionary (Hofmann & Popkema, 2008), clickable words in the corpora and linked secondary literature. In the interface, it is not possible to search on word form or with wildcards; only lemmas are accessible. The offered language information in the results is restricted to word type and period.

None of the existing language databases can be searched by linguistic information or by other meta-information that is present. All versions are interactive and there is no interface to conduct searches from other applications.

3. Target groups

We have identified three different user groups for FA's digital language resources: professional and non-professional language users, linguists and developers.

3.1 Language users

The preferred vocabulary is used in education—in schools and within adult education—and serves as a foundation for the creation of teaching materials. Journalists, authors and publishers use it as a reference work when editing and correcting publications. Officials, lawyers and staff in public sector institutions use it to assist in document writing. In all these instances, the vocabulary serves as a lexicographical resource to enable users to write Standard Frisian. Users can check which variants are acceptable. The vocabulary also provides information about the basic inflection and hyphenation of lemmas. The preferred vocabulary is the basis of a spelling checker for spelling errors and typos, and to check for standard forms and Dutchisms. The ONFW is a lexicographical Dutch–Frisian translation resource for a target group made up of language learners and native speakers of Frisian. Language learners are primarily interested in finding translations and grammatical information, while native speakers also use the dictionary for text production, such as searching for the right word forms, collocations and idioms. Users will consult the language databases to find contexts for a particular word form, information about Old Frisian manuscripts, etc. Interested individuals can look at facsimiles of manuscripts.

3.2 Linguists

Linguists utilize the digital language resources of the FA, although the three resources offer differing possibilities.

The preferred vocabulary gives researchers only a limited range of options. It presents a preferred form for Standard Frisian. Researchers who want to find out how standardization has developed can view the vocabulary as the modern-day final stage

in this process. For example, they can investigate whether and to what extent the vocabulary differs from the one in current Frisian dictionaries. They can also explore which Frisian dialects have contributed preferred forms to the standard language, or they can use the vocabulary to check which articles go with nouns, since each noun is accompanied by the correct article.

The database underlying the preferred vocabulary provides researchers with more options. The grammatical information included with each entry, for example, is an invaluable source of information. Researchers studying inflection variation in spoken Frisian can check which inflections verbs take compared to the standard. This variation is on the increase, mainly as a result of the dominance of Standard Dutch, particularly among younger generations. The database also contains many grammatically correct variants of the standardized inflection.

The language database is used for a wide range of linguistic research. Firstly, the Old, Middle and Modern Frisian texts in the database can be used to compile lexicographical resources. To date there is no lexicographical access to the Old and Middle Frisian language material. The language database can be used to describe word forms and the grammatical and semantic properties of lemmas. The link between KWIC and manuscripts or text editions means that it is easy to illustrate the lemma descriptions with text fragments linked to the source. The language database offers almost unlimited opportunities for the study of Frisian grammar. Because images of the Old Frisian text sources are linked to the texts in the language database, philologists can work on text editions. Thanks to the availability of texts from all three stages of the Frisian language, the database can be used to conduct detailed research on Frisian language change over the centuries. An example of one such study is Versloot (2008), which describes vowel reduction in fifteenth-century West Old Frisian, on the basis of material in the language database.

Like the language database, the ONFW can be used for grammatical research. The inclusion of grammatical information with the Frisian translations is a feature of the ONFW. This information is generated from the preferred vocabulary database. The bilingual Dutch–Frisian dictionary will enable researchers to make lexicological and semantic comparisons of the two languages. Because the dictionary includes many examples of idiomatic usage, it is an ideal tool for studying the use of idioms in Frisian and the differences with Dutch.

3.3 Developers

In the future, the idea is that stakeholders within education or culture, for example, will be given opportunities to develop applications on the basis of the API, such as massive open online courses (MOOC) or apps for mobile devices. Examples are apps with lexicographical applications (translating or looking for definitions), apps that check and assess texts for style or grammar, or apps with spelling exercises and

language games (puzzles, Scrabble-type games). Other applications involving the API include serious games or applications in healthcare (care robots that understand and speak Frisian).

4. API

4.1 CLARIN

CLARIN offers solutions and technology services for deploying, connecting, analyzing and sustaining digital language data and tools. With the API, we hope to achieve at FA level what CLARIN is seeking to achieve at supra-organizational level: standardized access to digital language material for teaching, research and other purposes. The API will also serve as a springboard for the development of services within the CLARIN infrastructure.

4.2 Design

Figure 4 below shows what the API will look like, with the main data sources, the target groups and the subdivision into editing and production environments.

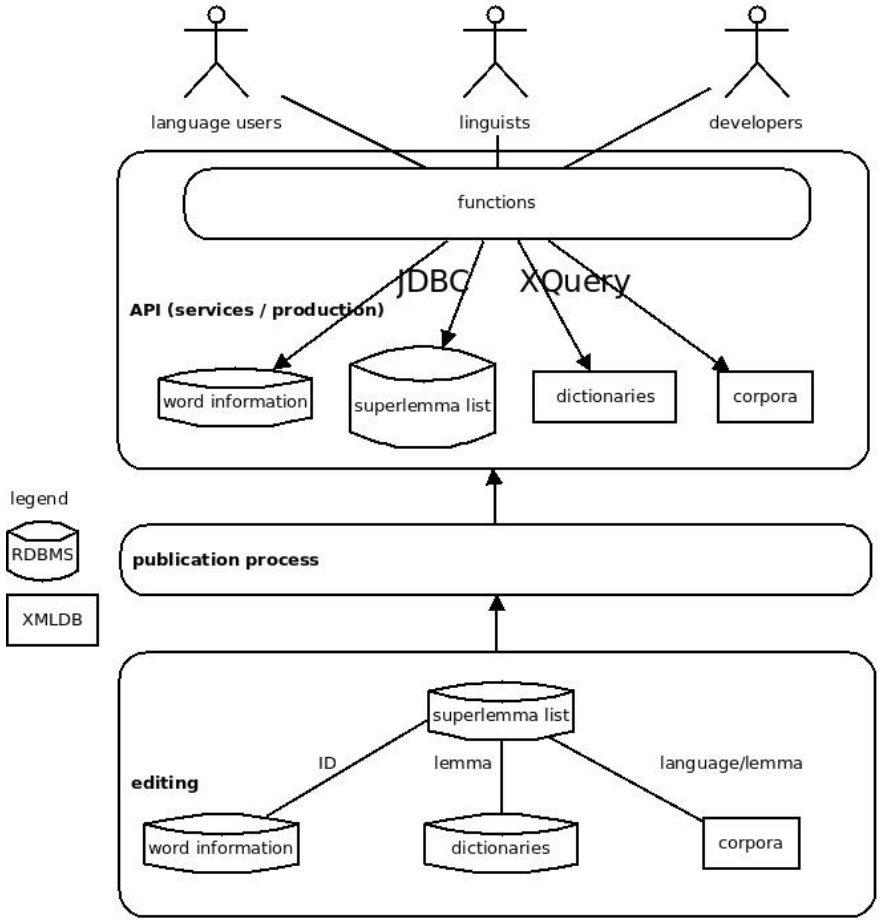


Figure 4: Diagram of the API

The principles are: information in one place, the separation of editing and production, and good support for work on the data.

Word information: Information at word level, including paradigm, morphology, preferred forms, word type.

Dictionaries: translation dictionaries from Dutch, based on the ANW.

Corpora: TEI-encoded texts with numerous possibilities for linguistic coding

Superlemma list: List of superlemmas with corresponding lemmas in a language category (Old, Middle and Modern Frisian, etc.)

4.2.1 Superlemma list

The superlemma list will play a key role. 'Superlemma' refers to an abstract lemma form in modern Frisian spelling to which the Old, Middle and Modern Frisian forms/lemma forms are linked. For example, the lemmas *sjitte* (Modern Frisian), *sjiette* (Middle Frisian) and *skiâta* (Old Frisian) are linked to the superlemma *sjitte* ('to shoot').

The aim is to arrive at, via a superlemma from the systems, information in another system. Lemmas in a particular language category are included under each superlemma. Superlemmas can be searched via a modern Frisian lemma or a lemma in another language category, together with that category. Each superlemma is also assigned an ID so that it can be selected directly. The available language categories will make up a list (for example: runen, old_frisian, bildts) to be published in the API.

Superlemmas for extinct words from older language phases will be reconstructions based on etymological patterns.

4.2.2 Links

The systems will be managed separately. The links between the systems are the information they contain that also appears in another system. See Table 1.

A possible consequence of these separate links is that systems could become 'out of sync'. This is particularly true of the superlemma IDs. Checks will be built into the management environments to help prevent links from no longer being valid.

The various data sources will be managed separately in the management environment. There will be a minimal relationship between the systems, just enough to gather information in the production environment. 'Word information' and 'superlemma list' will be linked via the ID that uniquely identifies each superlemma. 'Corpora' and

'superlemma list' will be linked via language category and lemma, while 'dictionaries' and 'superlemma list' will be linked via the new Frisian lemma.

word information	superlemma list	Each superlemma has an ID that can be included with an entry in 'word information'
corpora	superlemma list	Lemma annotation together with an annotation for language category can be found under a superlemma
dictionaries	superlemma list	A Frisian lemma in a dictionary can be found in a superlemma
dictionaries	word information	A Frisian lemma can be found in a dictionary with an entry in 'word information'

Table 1: Overview of links between systems.

4.2.3 Publication

Information will enter the production environment through a publication process, whereby data in the systems—with the exception of the dictionaries—will in principle be transferred one to one. Because of optimizations, users may choose to save certain data in the production environment twice. In the case of the dictionaries, the XML of the articles is removed from the database field in question and put into eXist-db. The publication process is also the place where transformations to standardized formats, etc. will be made.

4.2.4 Service

In production, the service offers functionality for the development of applications. See Section 4.3 for some detailed examples of functions within the service. The service uses standard technical links—JDBC and XQuery—to access information from the underlying systems. These technical interfaces provide access to all the information in the data sources and offer expressive query options. The technical interfaces can be used directly, but this requires extensive knowledge of the interface and the underlying data. The service offers a more user-friendly portal to information in the data sources. Filtering, sorting, pagination and other important functionalities that users require when querying data sources are built into the API.

4.3 Functions

In the use cases below, we will demonstrate how questions from target groups will be answered by means of a set of functions in the API. The functions will be defined in such a way that they can be used in different use cases. To maintain the focus on functionality, we have not included filtering, sorting, pagination and other general functionalities such as error handling in the examples.

4.3.1 Language users

Use Case: translation

For a Dutch word and its Frisian translation, a user also wants to find the inflection and pronunciation for that translation, as well as examples of contexts in which the Frisian translation is used. For this, the API offers the following functions.

Firstly, it must be possible to translate text from a language (in this case Dutch) into Frisian via a function. Input characters, possibly with wildcards, are used to search for matching Frisian lemmas. The result is a list of found Frisian lemmas, in which each found lemma is accompanied by the ID of the associated superlemma, the word type and the description. First, a search is made in dictionaries (ONFW in this case) for the Frisian translation of a text. This translation is then used to search in 'word information':

Signature: FrisianLemma* translate(text, language category)
Input: Text with wildcard support * and ? and a language category (Dutch in this case)
Output: 0 or more FrisianLemmas, with the superlemma ID, word type and description
Data used: 'word information' and dictionaries (ONFW)

Second, a function for retrieving inflection information on the basis of the superlemma ID:

Signature: Inflection getInflection(ID)
Input: superlemma ID
Output: Inflection
Data used: 'word information'

Third, a function for retrieving pronunciation information on the basis of the superlemma ID:

Signature: Pronunciation getPronunciation(ID)
Input: superlemma ID
Output: Pronunciation

Data used: 'word information'

Finally, a function is needed to show context information (KWIC) on the basis of a Frisian lemma. Searching for context information can be confined to a particular language category:

Signature: KWIC* getKWIC(lemma, language category)
Input: Frisian lemma, language category
Output: 0 or more KWIC showing text before the searched lemma, the lemma itself (or word forms of that lemma) and subsequent text.
Data used: corpora

Use Case: corpora

While searching the language database (TDB), a user finds a word form in the Old Frisian corpus that he cannot place. He therefore wishes to find a Dutch lemma for the word form. For this, the API offers the following functions.

Firstly, a function is needed to find superlemmas on the basis of a lemma in a particular language category (Old Frisian in this case). The principle here is that the word form in the corpus is annotated with the associated Old Frisian lemma. In the superlemma list, superlemmas are searched on the basis of the Old Frisian lemma (lemma + language category Old Frisian):

Signature: superLemma* findSuperLemma(lemma, language category)
Input: lemma and language category
Output: 0 or more SuperLemma, with ID and associated lemmas
Data used: superlemma list

The superlemma now has to be searched in dictionaries (ONFW) to find Dutch translations:

Signature: DutchLemma* translate(FrisianLemma)
Input: FrisianLemma, the Frisian lemma used to search for Dutch lemmas
Output: 0 or more DutchLemma, with meaning
Data used: dictionaries (ONFW)

The superlemma can be used to retrieve the new Frisian inflection in 'word information', for example, to compare it with the Old Frisian (and possibly Middle Frisian) inflection.

4.3.2 Linguists

Usage

The integration of the databases offers extensive opportunities for comparative research across time and space. With the help of the superlemma, information about lemmas can be selected in 'word information'. Dictionaries can be searched for Dutch translations of the lemmas, and in the TDB searches can be made in the corpora, for example by word form and their dialect distribution, or by linguistic information.

Researchers can also investigate, for example, the differences between separable verbs in Dutch and Frisian, in modern Frisian and Dutch and in older phases of these languages. A query in 'word information' will give a list of all separable and inseparable verbs with their paradigm, plus morphological information. The 'translation' field can be used to establish a link between this information and Dutch verbs in the ONFW. Paradigm and morphological information can also be retrieved from that database. Finally, the TDB can be searched for corpus evidence.

Functions in the API that support research are presented below. A function for finding all separable/non-separable verbs:

Signature: FrisianLemma* getVerbs(separable)
Input: Boolean separable
Output: 0 or more FrisianLemma, with the superlemma ID, word type and description
Data used: 'word information'

Next, a function for finding words with particular linguistic annotations. This function also supports separable verbs. The result contains the superlemma for the found words; this can be used to retrieve information in 'word information' and dictionaries. The linguistic annotations that are available for searches are published and updated in the API https://bitbucket.org/teibestpractices/linguistic-customization.

Signature: Result* find(text, linguistics*)
Input: Text with wildcard support * and ?; combinations of linguistic properties that are searched by
Output: 0 or more Results, showing found words in context, the superlemma for found words and metainformation on the corpus
Data used: corpora and superlemma list

As well as this function for retrieving results, there is also a function simply for counting:

Signature: CountResult count(text, linguistics*)
Input: Text with wildcard support * and ?; combinations of linguistic properties that are searched by

Output: The number of results and metainformation on the corpora
Data used: corpora and superlemma list

Researchers can also search corpora on the basis of information in 'word information'. An example is searching for neologisms, whereby 'word information' is searched for entries labelled 'neologism', possibly restricted to certain lemmas. The associated superlemmas are then retrieved. Under the superlemma are lemmas with a language category that can be used to search the corpora.

Signature: Result* findNeologisms(text)
Input: Text with wildcard support * and ?
Output: 0 or more Results, showing found words in context, the superlemma for found words and metainformation on the corpus
Data used: 'word information', corpora and superlemma list

5. Further development of data sources

5.1 Word system

The current access database for the preferred vocabulary will be transformed into a server database, such as MySQL. The database will be redesigned, bearing in mind the merging of information from the preferred vocabulary with information from other systems such as a morphological database. A management application will then be designed and built and a conversion will be written for converting data. In this conversion, linguistic terms will be converted into terms from linguistic-customization.

5.2 Online Dutch–Frisian Dictionary

The XML from the online dictionaries will be published to an XML database (eXist-db). This database will become the source in which searches will be made from the API via XQuery and/or REST. A website will also be generated for the ONFW so that people can engage interactively with the dictionary.

5.3 Language database

A new version of the language database is being developed. It is based on TEI XML, with a linguistic expansion based on universaldependencies.org (see linguistic-customization). We are thus opting for reputable, internationally supported open standards which enable digital publication with minimum effort and which offer a foundation for research. Tei-c.org makes this possible by choosing customization as a base. This occurs via One Document Does all (ODD), which will manage validation, support/editing support and presentation.

The XML contains information about the manuscript, such as author, repository, location, the manuscript text, linguistic annotations at word level and a reference to the superlemma list.

The Oxygen XML Editor is used for editing and offers support for TEI and the linguistic expansion.

eXist-db is used for storing and accessing the material. eXist-db offers the option of querying the manuscripts using the standard XQuery language. There are no restrictions here; all information present can be queried.

There is a need to generate a website with TEI Publisher for the corpora, where manuscripts, including scans, can be viewed, where the material can be searched by text, with KWIC results, and where manuscripts can be downloaded as PDF files.

The material in the language database is not always free of copyright. This will be taken into consideration, including technically via the availability element in TEI.

6. Implementation

Implementation mainly involves upgrading the current systems, setting up a management and production environment and publication processes, designing and building links (via superlemma) and designing and building the API. The steps in this implementation process will be set up as projects that will be assessed, prioritized and scheduled in relation to one another. At the very least, the building projects will involve versioning, dependency management and issue management. Ideally, we will also work with continuous build, with a test environment and with other solutions that are customary in a development process, for example Docker.

6.1 Projects

Table 2 provides an overview of the work required to implement the API. It does not include scope, prioritization, phasing, etc.

6.2 Service and support

An online help desk will be set up for language users, researchers and application developers. There will also be built-in options for reporting problems and suggestions and for monitoring their status. System monitoring will also be set up to maintain automated monitoring of system use.

Component	**Work**
Word information	Design and build data model in management screens
Word information	Migration and conversion of existing material from the preferred vocabulary and morphological database, etc.
Word information	Design and build technical interface
Dictionaries	Import information from the preferred vocabulary
Dictionaries	Design and build publication to eXist-db
Dictionaries	Design and build technical interface
Corpora	Design and build functions
Corpora	Design and build technical interface
Superlemma list	Design and build data model and management screens
Superlemma list	Design and build technical interface
API	Design and build functions, with input, output and error handling functionality
API	Build the implementation of the API
General	Set up publication processes, including automation

Table 2: Work to realize the API.

7. Conclusion

The information about functionality in this paper is based on information already contained in the databases. Meanwhile, the FA has a number of databases and language resources. Lexicographical tools are digitally available since the end of the twentieth century. In our paper the available databases of the Frisian language are described. The reader has been able to conclude that these databases can certainly be improved. The aim of the linguistic department of the FA is to expand the databases and to create, add and link more databases. We see it as in important task of the FA to optimize the databases and their use. Finally, we summarize this task in the following roadmap, divided into three important items: (a) expanding the databases, (b) including spoken language material, and (c) linking data.

- Expanding
    - The Modern Frisian corpus will be expanded to include texts from domains and genres that are currently underrepresented, supplemented by texts from social media, weblogs. Distribution over time is also out of balance: nineteenth-century Frisian, in particular, is underrepresented, and the period after 1990 also requires attention.
    - The WFT can be linked to the Modern Frisian corpus to make more corpus evidence available.
    - There is a long-held wish to create a WordNet-type lexical semantic database of Frisian.
    - On the basis of the Old and Middle Frisian corpora, lexicographical resources should be made for each of these two language phases.
- Speech
    - Spoken corpora must have a place in the landscape described here, and possibly in the API as well.
    - The preferred vocabulary will be supplemented in future by pronunciation information in the International Phonetic Alphabet (IPA) and by morphological information about the keywords.
- Linked data
    - Geographical and diachronic information are already present, but are not yet clear and unequivocal. In the new system, solutions will be sought to give both data types a good home, in data sources and the API.
    - Digitized dialect-geographical material could be linked to the various databases.
    - The corpora, in particular, contain information that can be made available as Linked Open Data (LOD). This could involve metadata information, such as author, location, year of publication and publisher, as well as information in the text, such as geonames and named entities. Through LOD, links can be made to other data sources, such as HISGIS (Historisch Geografisch Informatiesysteem / Historical Geographical Information System), which also works with LOD.

8. References

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

Breuker, P. (2001). The Development of Standard West Frisian. In H.H. Munske (ed.) Handbuch des Friesischen / Handbook of Frisian Studies. Tübingen: Max Niemeyer Verlag, pp. 711-721.

Depuydt, K., De Does, J., Duijff, P. & Sijens, H. (2017). Making the Dictionary of the Frisian Language available in the Dutch historical dictionary portal. In J. Odijk & A. van Hessen (eds.) CLARIN in the Low Countries. London: Ubiquity Press, chapter 13.

Duijff, P. (2008). Towards Standard Frisian in the Friesch Woordenboek. In M. Mooijaart & M. van der Wal (eds.) Yesterday's Words: Contemporary, Current and Future Lexicography. Newcastle: Cambridge Scholars Publishing, pp. 53-66.

Duijff, P. (2016). Towards Modern Standard Frisian. In I. Tieken-Boon van Ostade & C. Percy (eds.). Prescription and Tradition in Language. Establishing Standards across Time and Space. Bristol / Blue Ridge Summit: Multilingual Matters, pp. 532-549.

Duijff, P. & Van der Kuip, F. (2017). Lexicography in a minority language: a polyfunctional online Dutch-Frisian dictionary (in progress).

Dykstra, A. & Reitsma, J. 1993, De struktuer en de ynhâld fan 'e Taaldatabank fan it Frysk. It Beaken, 55, pp. 55-82.

Schoonheim, T. & Tempelaars, R. (2010). Dutch Lexicography in Progress: the Algemeen Nederlands Woordenboek (ANW). In A. Dykstra and T. Schoonheim (eds.), Proceedings of the XIV Euralex International Congress. Ljouwert: Fryske Akademy / Afûk, pp. 718-725.

Sijens, H. & Dykstra, A. (2013). Language Web for Frisian. In I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets & M. Tuulik (eds.) Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference. Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut, pp. 93-105.

Versloot, A.P. (2008). Mechanisms of Language Change, Vowel Reduction in 15th Century West Frisian. Utrecht: LOT.

Zantema, J.W. (1984). Frysk Wurdboek. Frysk-Nederlânsk. Leeuwarden: A.J. Osinga Uitgeverij.

Dictionaries and websites:

OCDSE: *Oxford Collocations Dictionary for Students of English.* (2009). 2nd edition. Oxford: Oxford University Press.

Anw.inl.nl. Accessed at: http://anw.inl.nl. (9 May 2017).

FHW: Frysk Hânwurdboek. (2008). Ljouwert: Fryske Akademy / Afûk.

Gtb.inl.nl. Accessed at: http://gtb.inl.nl/ (9 May 2017).

HISGIS. Accessed at: www.hisgis.nl/ (26 May 2017).
Hofmann, D. & Popkema, A.T. (2008). Altfriesisches Handwörterbuch. Heidelberg: Universitätsverlag Winter.
Linguistic-customization. Accessed at: https://bitbucket.org/teibestpractices/linguistic-customization (12 April 2017)
ONFW: Online Nederlands-Fries woordenboek. In progress.
Taalweb.frl. Accessed at: www.taalweb.frl https://taalweb.frl/. (28 April 2017)
Tei-c.org. Accessed at: http://www.tei-c.org/index.xml/. (19 July 2016)
Universaldependencies.org. Accessed at: http://universaldependencies.org/. (15 May 2017)
Visser, W. (1985). Frysk Wurdboek. Nederlânsk-Frysk. Leeuwarden: A.J. Osinga Uitgeverij.
WFT: Wurdboek fan de Fryske taal / Woordenboek der Friese taal. (1984-2011): Ljouwert / Leeuwarden: Fryske Akademy.

Designing a Learner's Dictionary Based on Sinclair's Lexical Units by Means of Corpus Pattern Analysis and the Sketch Engine

Paolo Vito DiMuccio-Failla, Laura Giacomini[1]

[1] Heidelberg University/ Hildesheim University

E-mail: paolodimuccio@gmail.com, laura.giacomini@iued.uni-heidelberg.de

Abstract

This paper is part of a study for the design of an advanced learner's dictionary (in Italian) aimed at implementing Sinclair's vision of 'the ultimate dictionary' (see Sinclair et al., 2004: xxiv) and based on his conception of lexical units. Our present goal is to exhaustively portray the meaning profile of verbs, systematically distinguishing their meanings by their normal patterns of usage. To achieve this, we apply Hanks's Corpus Pattern Analysis by means of Kilgarriff and Rychlý's Sketch Engine.

The first chapter presents and discusses the theoretical background to our work. The second gives a description of our methodology, which is then exemplified by a case study of the Italian verb *seguire*. The final part of the paper draws a few conclusions on the feasibility and usefulness of Sinclair's 'ultimate dictionary' and hints at future steps of the dictionary-making process.

The dictionary project is in its design stage and is intended to be a platform for cooperation between the Italian publisher Zanichelli and a network of international universities and research institutes.

Keywords: learner's dictionary; Italian learner's dictionary; lexical units; Sinclair's thesis; Sinclair patterns

1. Methodological background

1.1 COBUILD's scientific revolution

At the start of the 1980s, lexicography, and linguistics in general, were undergoing a far-reaching paradigm shift thanks to the new availability of huge quantities of machine-readable text made possible by advances in computer technology. According to John Sinclair, a then leading linguist at the University of Birmingham, the situation was similar to that of the physical sciences in the first half of the 17th century, when they started to rely on empirical observation (Sinclair, 1991: 1). If the intuition of a single individual had been, up to that moment, the key to all linguistic investigation, lexicography finally had the possibility to utilize "objective evidence" (ibid.).

Given these premises, Sinclair (with funding from Collins publishers) founded the COBUILD (Collins Birmingham University International Language Database) project with the aim of producing innovative language reference works (Sinclair, 1991: 2). Together with his collaborators, he started building a large and representative

electronic corpus of contemporary English (Sinclair, 1987: 1), based on which, in the following years, "a completely new set of techniques for language observation, analysis, and recording" was developed (Sinclair, 1991: 2). Many consider this the very first study in corpus-driven lexicography (Tognini-Bonelli, 2001: 85), initiating the now thriving tradition of empirical lexical analysis (Hanks, 2008a: 222).

The main result was the compilation of the COBUILD English Dictionary (COBUILD 1987), the first dictionary based on evidence of actual contemporary usage, and the first to give a central role to the "spectacular" regularities of language patterning which had been displayed by corpus analysis (Sinclair, 1991: 4) and had lead Sinclair to conclude that "by far the majority of text is made of the occurrence of common words in common patterns, or in slight variants of those common patterns" (Sinclair, 1991: 108). This phenomenon goes far beyond that which the pioneer lexicographers like Palmer and Hornby had shown, since different senses of the same word present, in general, different characteristic patterns, as we will explain in the following subsection.

In order to display the "typical features" of the characteristic co-texts of words (Sinclair, 2004b: 5; see also Hanks, 1987), Sinclair systematically utilized full-sentence definitions, which he considered theoretically sounder and easier to understand (Hanks, 2008a: 221) than traditional ones (for a balanced discussion see Rundell, 2006). Furthermore, in the COBUILD dictionary, every observation about language was accompanied by at least one example, and all examples were taken from the corpus in order to obtain "genuine instances of language in use" (Sinclair, 1991: 4–5). All this was thought to help students to speak and write naturally and idiomatically (see Hanks, 2008a: 219).

An important point should be made about Sinclair's empirical corpus analysis. On the one hand, it proceeds along the standard scientific method of inspecting the data, discerning regularities, formulating hypotheses, and testing the hypotheses on the data (Sinclair, 2004a: 10 ff.). On the other hand, "intuition and introspection still play an important role, since perceiving meaning is a subjective experience, and descriptions in dictionaries need to satisfy intuition" (Moon, 1987). The role of introspection is to evaluate evidence rather than to create it (Sinclair, 1991: 39), whereas intuition is crucial exactly when introspection "is not in accordance with the newly observed facts of usage" (Sinclair, 1991: 4). Therefore, intuition, introspection, and data analysis must work together (Sinclair, 2004a: 115).

This is why Sinclair does not, in principle, discard traditional kinds of evidence, obtained for example by consulting other dictionaries or by testing native speakers (Sinclair, 1991: 38–39). Most of all, "ultimately... the lexicographic decisions will be personal evaluations by the lexicographer, giving due consideration to all evidence that he or she has amassed" (ibid.: 39). For these reasons, Sinclair takes a balanced stance in the debate between descriptive and prescriptive studies, stating that "a

purely objective description of text will not contain adequate generalization" (ibid.: 60) and that "prescriptive studies fall into disrepute only when they ignore or become detached from evidence" (ibid.: 61).

A second important point to be made for our purposes is that Sinclair distinguishes between typical language patterns on one side and extended, displaced, and distorted usages on the other side (Sinclair, 1991: 61). A synchronic dictionary of usage should be filled with norms (ibid.: 61), not with unusual language events, and should warn against specialized use (ibid.: 38).

1.2 Sinclair's thesis about lexical units

When lexical information began to be extracted from multi-million word corpora in the early 1980s, several long-accepted conventions in lexicography were called into question, for example the idea that a polysemous word could inherently, by itself, have several distinct meanings (Sinclair, 1998; Sinclair, 2004a: 132), and that any occurrence of such a word could signal any of those meanings (Sinclair, 1986: 60). Sinclair recognized that if this were actually the case, ambiguity would make communication virtually impossible (see Sinclair, 1998), because the meanings of polysemous words, though related, can be very diverse (this later became known as the *polysemy paradox* - see Falkum, 2011: 13 ff.). On the contrary, in continuous discourse, whether written or spoken, ambiguity is rare, except when intended (see Moon, 1987).

In the course of the survey leading up to the publication of the COBUILD dictionary, evidence gradually accumulated for an alternative hypothesis which, at first, had been ridiculed (Sinclair, 1991: 10): that of a general correspondence between observable patterns of words and distinctions of meaning. In fact, Sinclair came to the conclusion that not single, isolated words, but rather *words in their contextual patterns* are the true bearers of meaning, and that every such pattern has only one meaning (not considering sub-meanings given by trivial generalization or specification - cf. for example Sinclair, 1991: 55–56). This claim can be stated in a more rigorous fashion, which we might call 'Sinclair's thesis':

> In general, each (major) (normal) sense of a word can be associated with a distinctive pattern of usage (see Moon, 1987: 89 ff.; Sinclair, 1991: 6 ff.; Sinclair, 2004b: 5; Sinclair, 2004c: 281; Sinclair et al., 2004: xxiv) determined by the following features (see also Sinclair 1996; Sinclair, 1998; Sinclair, 2003: 145 ff.; Sinclair, 2004a: 39 and 141):

1) collocation, i.e., the co-occurrence of particular words (with the given word);
2) colligation, i.e., the co-occurrence of particular grammatical patterns;
3) semantic preference, i.e., the co-occurrence of words with particular meanings;
4) semantic prosody, i.e., a co-text implying a particular connotation of the described

state of affairs or a particular attitude of the speaker[1].

Take for instance the word *put*. It can be part of a phrasal verb, in which case its meanings are co-determined by other parts of speech, or it can be a non-phrasal verb, in which case its senses mostly correspond to the (choices of the) semantic types of the referents associated with its arguments (i.e., its selectional preferences). As an illustration of this correspondence, we look at the first three senses of *put* in the corresponding entry of the latest edition (2014) of the COBUILD dictionary (see also Moon, 1987: 91):

1. "When you **put** something in a particular place or position, you move it into that place or position"
2. "If you **put** someone ... [in a particular place or position], you cause them to go there and to stay there for a period of time"
3. "To **put** someone or something in a particular state or situation means to cause them to be in that state or situation".

Sinclair's analysis even allows the finding of hidden senses of words. Consider for example the word *feeling*. No corpus analysis is needed to know that it frequently co-occurs with the adjective *true* in the phrase *true feelings*. Such a collocation would not be considered idiomatic and hardly given any special treatment in a traditional dictionary (Sinclair, 1996: 89). An accurate pattern analysis (cf. Sinclair, 2003) will in fact show statistical restrictions on the choice of its co-text. *True feelings* is usually preceded by a possessive adjective, which is in turn preceded by a verb synonymous with *express*, *show*, or *hide*. This constitutes a syntactic tendency, a colligation, but also a semantic preference for verbs of expression. In the case of semantically 'positive' expression, there is usually an even broader context, i.e. a semantic prosody, hinting at a reluctance or difficulty in expressing those true feelings. Hence the actual lexical unit here can be presented by

"to hide one's true feelings or show them with/after some reluctance/difficulty".

Thus, Sinclair arrived at the conclusion that the true units of meaning of a language are largely phrasal and that, as a consequence, phraseology is due to become central in the description of language (cf. Sinclair, 2004a). Sinclair used the term '(extended) canonical form' to refer to the most explicit, full and unambiguous presentation of a lexical unit (Sinclair, 2004c: 298), like the one we just proposed for *true feelings*. The shortest unambiguous presentation of the lexical unit (in our case, simply *true feelings*) he called 'short canonical form' (Sinclair et al., 2004: xxiv). In the final years of his career, he was convinced that a new kind of dictionary based on the canonical forms of lexical units "would be the ultimate dictionary" and would allow students to truly master a language (ibid.).

[1] The notion of semantic prosody was implicitly introduced by Sinclair (1987: 155; 1991: 75) and first defined by Louw (1993). It is actually rather controversial (see Whitsitt, 2005; Stewart, 2010) and hard to work with.

1.3 Hanks's analysis of corpus patterns

Patrick Hanks, one of the main collaborators of John Sinclair at COBUILD, has since been a committed supporter of the corpus-driven approach to lexicography and of Sinclair's thesis about lexical items (Hanks, 2004a: 87; Hanks & Pustejovsky, 2005; cf. also Krishnamurthy, 2008: 239). Hanks's focus on NLP has lead him to develop and standardize a technique, which he dubbed 'Corpus Pattern Analysis' (CPA), to analyze large corpora and find the "normal patterns of usage" associated with each word, with the aim "to link word use and word meaning in a machine tractable way" (Hanks & Pustejovsky, 2005: 64). The main result will be a dictionary for use in NLP (ibid.) and in language teaching (cf. PDEV website). In the Pattern Dictionary of English Verbs (PDEV), a pilot study currently in development under the supervision of Hanks, many verbs are being analysed, having priority over nouns (cf. Hanks, 2008b; Hanks, 2004a: 92).

Hanks' "semantically motivated syntagmatic patterns" (Hanks, 2004a: 88) are simplified and strictly formalized versions of Sinclair's word patterns. In the case of a verb, they consist of an argument structure, assigned together with the most general semantic types (and possibly semantic roles[2]) which the arguments of the verb *normally* refer to (ibid.: 87–88). The last bit is a tricky one: identifying the right semantic types as selectional preferences, in particular not leaving out normal usage on one side and not generalizing into abnormal usage on the other side, requires linguistic and ontological expertise: "Among the most difficult of all lexicographic decisions is the selection of an appropriate level of generalization on the basis of which senses are to be distinguished" (cf. Hanks, 2004a and PDEV website). In general, "the identification of a syntagmatic pattern is not an automatic procedure: it calls for a great deal of lexicographic art" (Hanks, 2004a: 88).

In CPA, one starts with concordance lines and groups them into patterns, whereas "associating a 'meaning' with each pattern is a secondary step, carried out in close coordination with the assignment of concordance lines to patterns" (ibid.). "The 'meaning' of a pattern is expressed as a set of basic implicatures" (ibid.). Let us look for example at the syntagmatic patterns of the verb *lead* according to the PDEV (cf. PDEV website)[3]:

1. Pattern: [Eventuality]$_1$ **leads** TO [Eventuality]$_2$
Implicature: [Eventuality]$_1$ is the cause of [Eventuality]$_2$

[2] Roles are not considered types by Hanks (cf. Hanks et al., 2007: 5). We will discuss the use of semantic roles in word patterns in the following section.

[3] According to CPA conventions (cf. Hanks, 2004a: 93), double square brackets indicate semantic types and curly brackets (braces) indicate sets of specific lexical items. The keyword is written in bold letters. For readability reasons, we have slightly modified the convention regarding types by using simple square brackets.

2. Pattern: [Eventuality]$_1$ **leads** UP TO [Eventuality]$_2$
Implicature: [Eventuality]$_1$ precedes but may not be the cause of [Eventuality]$_2$

3. Pattern: [Eventuality] **leads** [Human]/[Institution] TO-INFINITIVE
Implicature: [Eventuality] causes, enables, or encourages [Human]/[Institution] TO...

4. Pattern: [Human]/[Institution]$_1$ **leads** [Human group]/[Institution]$_2$
Implicature: [Human]/[Institution]$_1$ organizes, directs, or provides a model for activity of [Human group]/[Institution]$_2$

The choice of appropriate selectional preferences can be hard not only because of the inherent difficulty in building a coherent ontology compatible with everyday language, but also because it is not always immediately clear what semantic types normal usage can possibly refer to. Take for instance the English verb *toast* in the sense of "cook food by exposure to a grill or fire" (as in Hanks, 2004a: 91 and Jezek & Hanks, 2010). A quick look at the word sketch of *toast* on the Sketch Engine (see Kilgarriff et al., 2004) shows that the most frequent direct objects of *toast* are *bread*, *almonds*, *marshmallows*, *buns*, *walnuts*, *pecans*, *coconut*, *bagels*, *nuts*, *hazelnuts*, *sandwiches*, *baguettes*, *brioche*, *muffin*... In such cases, Hanks proposes to either use a general semantic type (as in the PDEV), like

[Human] **toasts** [Food],

or, when possible, to insert directly into the pattern (see Hanks, 2004a: 91) the paradigmatic lexical set of the most frequent collocates. In our case, this results in

[Human] **toasts** {*bread*, *almonds*, *marshmallows*, *buns*, *walnuts*, *pecans*, *coconut*, *bagels*, *nuts*, *hazelnuts*, *sandwiches*, *baguettes*, *brioche*, *muffin*}.

However, in the first case the type [Food] can be seen as too general and uninformative, whereas in the second case the list was truncated at *muffin* for no statistical reason: the actual progression of collocates slowly fades into statistical insignificance without any apparent discontinuity. This raises a semantic issue (see for example Jezek & Hanks, 2010), which we will try to resolve in the following section.

As an ontology for CPA, Hanks uses a shallow hierarchy of types selected for their prevalence in the manual identification of patterns (Pustejovsky et al., 2004). The number of types is kept to a minimum, as perfect ontological coherence is required. "New types are added occasionally, but only when all possibilities of using existing types prove inadequate" (Pustejovsky et al., 2004). Currently, there are 253 types in the PDEV.

Corpus Pattern Analysis hinges on the Theory of Norms and Exploitations (see Hanks, 2013), which makes a strict (conceptual) distinction between normal and

abnormal usage of language (Hanks, 2013: 3; Hanks 2004a: 89), be it because of anomalous syntactic structures, anomalous semantic arguments, or figurative uses (El Maarouf, 2013: 125; see also Hanks, 2004a: 92). When abnormal usage is intentional, it is called an 'exploitation' of a norm (Hanks, 2013: 8). This theory led Hanks to conclude that "attempts to account for all possible meanings [of words] are misguided. Projects with this aim tend to produce impractical results, because normal usage becomes buried in a welter of remote possibilities" (Hanks & Pustejovsky, 2005: 64). On the contrary, "the number of normal combinations is remarkably small and computationally manageable" (Hanks & Pustejovsky, 2004: 15).

2. Our purpose and method

We are convinced that Sinclair's concept of fundamental lexical units is the right one. We know this is still a controversial issue: many linguists do not even agree on the existence of objective criteria for correctly lumping/splitting the senses of polysemous words (see for example Kilgarriff, 1997: 100). However, by comparing the results of our present research with ItalWordNet, the Italian wordnet (see Roventini et al., 2003) created in the framework of the EuroWordNet project (see Vossen, 2002), we discovered a stunning overlap of the meanings of Sinclair's lexical units with the single senses of words implicit in the synsets of ItalWordNet. Such senses result from a completely different approach and it is hard to see how this could be a coincidence. We will explain our findings in detail for the Italian verb *seguire* in the following section.

We also share Sinclair's opinion that a dictionary extensively describing the canonical forms of each lexical unit would be 'the ultimate dictionary', because it would potentially contain all semantic information about word usage. This is why we started investigating the feasibility and actual utility of implementing Sinclair's vision. The two main problems we encountered are the following:

1) Extracting lexical units from a corpus and accurately studying their canonical forms can be difficult and time consuming.

2) It is not easy to present the extended canonical form of a lexical unit without overloading its entry with information of various degrees of importance. This is exactly the problem we have mentioned about the full-sentence definitions found in the COUBILD dictionary.

We will explain how we coped with both problems in a series of papers. For now, we will concentrate on the first one, showing in particular how we adapted Hanks's CPA and applied it by means of Kilgarriff and Rychlý's Sketch Engine (see Kilgarriff et al., 2004) to find all senses of the Italian verb *seguire*.

2.1 Building an ontology

We need to build an ontology not only because Sinclair's word patterns refer to semantic types, but also because ontologies facilitate homogenous definitions and a clean overview of any lexical domain. Our approach to the upper part of the hierarchy is similar to that of EuroWordNet (see Vossen et al., 1998), which distinguishes, along the lines of Lyons (1977), the category of concrete objects and substances (first-order entities) from that of properties, relations, situations, and events (second-order entities). We will discuss the details in a future paper. Concrete entities can be further classified into types according to the four independent criteria advocated by Pustejovski (1995): origin, form, composition, and function. Second-order entities can be classified into types according to more sophisticated criteria, which will also be examined in a future paper.

Fortunately, for the purpose of monolingual learner lexicography, hierarchies of types only have to be as systematic and coherent as normal language usage. Hence, in principle, we accept the possibility that semantic types assigned in different word patterns might not be perfectly compatible. Furthermore, it is natural to add to the ontology not only any lexicalized semantic role, like [Patient] or [Monarch], but also a distinct type of entity for every nominal lexical unit (of the language in question), like [Means of public transport] and [Job creation scheme].

Not having to use a limited, perfectly coherent ontology can make things a lot easier. Consider the example of the verb *toast* in the sense of "cook food by exposure to a grill or fire", which we have discussed in the previous section. Allowing for relatively uncommon concepts like [Breadstuff], a word pattern can be assigned which is easier for a human to read and understand:

[Human] **toasts** [Breadstuff]/[Marshmallow]/[Nut]/[Seed]

Differentiating between prototypical, common, and possible usage of words is also an option:

[Human] **toasts** *prototypically* [Bread]/[Sandwich]
usually [Breadstuff]/[Marshmallow]/[Nut]/[Seed]
possibly [Food]

2.2 Identifying lexical units

In general, we employ a bottom-up, empirical strategy to identify the semantic types selected by a word for its argument slots, following the clues provided by the word sketches of the Sketch Engine. Notice, for example, that the paradigmatic lexical set of collocates found in a particular argument slot of a particular word sense can be partially ordered, in a mathematical sense, according to the hyponym-hypernym relation. If it presents a maximum, i.e. a hypernym of all other words in the set, such

a hypernym denotes the needed semantic type. For instance, a paradigmatic lexical set of nouns associated, as subjects, with the Italian verb *fermare* (to stop) is

> {*bus, autobus, corriera* (coach)*, tram, treno* (train)*, metro, metropolitana, mezzo pubblico* (means of public transport)}.

Clearly, *mezzo* (*di trasporto*) *pubblico* is a hypernym of all other words in the set, and thus identifies the most appropriate semantic type for the subject slot of the corresponding lexical unit.

It must be stressed that Sinclair's patterns do not come out of a corpus by themselves: they must be properly looked for by means of the scientific method, as we mentioned in the previous section. Consider for instance the Italian word *braccio* (arm). We analysed the word sketches of *braccio* by taking into account all its possible syntactic constructions. Two of them turned out to be particularly informative: (N + Adj) and (N + *di* + N). After finding many lexical units, like *braccio di un essere umano*, *braccio di un carcere*, *braccio di terra/mare/fiume*, and others, we were left with what we thought to be a paradigmatic lexical set of a single remaining unit:

> {*mobile*, *meccanico* (mechanical), *flessibile* (flexible), *regolabile* (adjustable), *articolato* (jointed), *snodabile* (hinged), *estensibile* (extendable)}

Since *braccio mobile* (mobile arm) is a hypernym of *braccio meccanico*, *braccio flessibile*, and so on, we selected it as a candidate. However, in trying to confirm the hypothesis, we indeed falsified it when we found out that *braccio fisso* (fixed arm) also exists and that it refers to the same kind of objects: supporting arms of devices. Most of the adjectives in the set were confirmed to be, in fact, collocates of the lexical unit *braccio di sostegno di uno strumento* (supporting arm of a device). The remaining adjective, *mechanical*, must hence build a separate lexical unit: *braccio meccanico* (mechanical arm).

2.3 Formalizing word patterns

Our final objective is to compile an Italian learner's dictionary. Hence, an accurate adherence of word senses to actual normal usage is of paramount importance. However, since we are convinced that the best way to achieve this goal is by means of Sinclair's patterns of word usage, we do not want to exclude *a priori* an application of the dictionary for NLP, like the PDEV. Therefore, we will adopt a semi-formal approach: our patterns will have in general a formal part adapted from CPA and an informal expansion for human readers.

Now consider the first sense of the verb *follow* in the COBUILD dictionary (2014):

> "If you follow someone, who is going somewhere, you…"

The phrase "who is going somewhere" predicates a necessary stage-dependent (cf. Kratzer, 1995) condition for the action of following to take place. Such semantic prerequisites are often not needed for the disambiguation of a polysemous word because it is constant in all of its senses. However, they will inevitably be part of the semantic preference of any given word, and therefore we will always make them explicit, as they are in the COBUILD dictionary:

> "If you repair something that... is not working..."
> "When you unzip something which is fastened by a zip..."
> "If you find something that you need or want…"

In some cases, semantic conditions are essential for disambiguation. Suppose, for example, that you were just told to follow a man who is standing. If he is talking, you were probably told to listen to him. The prerequisite for the literal sense of the verb *follow* to be activated is that the person to be followed must be going somewhere. Its formalized canonical form could be

> [Human]$_1$ **follows** [Human]$_2$ *SUCH THAT* ([Human]$_2$ *IS A* [Goer]).

Notice that [Goer] is a rather unusual semantic role. To avoid cluttering our ontology with unnatural concepts, we prefer a different approach to the formalization of Sinclair patterns, allowing formulas to refer to meanings of predicates defined in the dictionary itself, as long as this does not result in a circular definition. The previous pattern can thus become more readable:

> [Human]$_1$ **follows** [Human]$_2$ *SUCH THAT* ([Human]$_2$ goes TO *SOME* [Place]).

Let us confront this pattern with the first sense found in the PDEV (we are ignoring the presence of the type [Animal] for the sake of clarity):

> [Human]$_1$/[Vehicle]$_1$ **follows** [Human]$_2$/[Vehicle]$_2$

This sense is not disambiguated from the second one in the same entry:

> [Human]$_1$ **follows** [Human]$_2$

Furthermore, the type [Vehicle] was needed because it was not possible to rely on the general regular alternation substituting people moving in vehicles for the vehicles themselves (when describing their motion). Incidentally, we conjecture that Hanks's question as to why semantic types do not seem to match well with paradigmatic lexical sets (see Hanks et al., 2007; Hanks & Jezek, 2008; Jezek & Hanks, 2010) can be at least partially answered by taking into consideration stage-dependent semantic conditions, which are not always easy to identify.

As a final remark, we will conform to the standard lexicographic practice of using in general (with a few natural exceptions) the type [Person] instead of both [Human] and [Animal], as this distinction is rarely needed for word sense disambiguation, and action verbs are principally thought to apply to any real or imaginary person. Similarly, if the type selected by a verb sense for the subject slot is [Person] we will omit it.

3. Case study: the Italian verb *seguire*

On the Sketch Engine, we selected the 2010 itTenTen corpus (see Jakubíček et al., 2013) and set out to identify and study the Sinclair patterns of the Italian verb *seguire* (to follow).

3.1 Patterns

We analysed the first 500 concordances of *seguire* chosen as "good examples" by the Sketch Engine (cf. Kilgarriff et al., 2008). It quickly became clear that the main distinction to be made was between transitive and intransitive patterns. The intransitive patterns could then be distinguished according to their argument structure, and the transitive ones according to their semantic preference. We progressively classified the instances of *seguire* according to those criteria and also, subordinately, depending on whether we deemed them to be normal or abnormal. Regular alternations as described by Pustejovsky in his Generative Lexicon Theory (see Pustejovsky, 1995) were classified as normal usage, whereas *ad hoc* metaphors, metonyms and other figures of speech were considered exploitations.

One by one, we identified the following lexical units, here arranged in an order which facilitates an overview:

T1) Seguire qu. presente che sta andando da qualche parte (to follow sb. present who is going somewhere)

This is the most basic pattern of *seguire*, used as a transitive verb with the literal meaning of "andare dietro a qu." (to move along behind sb.). As already mentioned, in our approach, we attempt to identify semantic types by finding the most general semantic restriction which disambiguates the present sense from the other senses. In this case, however, the only such restriction is that, normally (excluding occasional extensions to small objects), in Italian you follow persons (possibly alternating with animals, as in the case of many other verbs of motion). As previously discussed, the disambiguating information for this sense is actually a stage-level semantic condition, i.e., the fact that the followed person is (present and) going somewhere.

T2) Seguire un certo tragitto o una certa descrizione di un tragitto (to follow a particular route or a particular description of a route)

This pattern has the meaning of "andare lungo un certo tragitto" (to move along a particular route). It displays a metonymical alternation between routes and descriptions of routes (*indicazioni*). By means of the word sketches provided by the Sketch Engine, we found a large number of collocates in the direct object position which refer to types of routes: *percorso, corso, traccia, sentiero, strada, itinerario, pista, via, cammino, tracciato, rotta, traiettoria*. Since *tragitto* (route) is a hypernym of all members of the lexical set in question, we selected it as the name of the associated type. We did not choose *percorso* (path), because its most common meaning is concrete, whereas, as confirmed by standard dictionaries (e.g., TRECCANI and DE MAURO), the basic meaning of *tragitto* is abstract. Definition no. 8 of *follow* in the COBUILD (2014) dictionary perfectly matches our pattern:

"If you follow a path, route, or set of signs, you go somewhere using the path, route, or signs to direct you."

T3) Seguire qu. presente che sta svolgendo una sequenza di azioni (to follow sb. present who is performing a sequence of actions)

This pattern has the meaning of "fare ciò che si vede/ sente fare a qu., imitare qu." (to do what you see/ hear sb. do, to imitate sb.). Definition no. 13 of *follow* in the COBUILD dictionary loosely corresponds to our pattern:

"If you follow what someone else has done, you do it too because you think it is a good thing or because you want to copy them."

T4) Seguire una certa linea di condotta o una certa descrizione di una linea di condotta (to follow a particular course of action or a particular description (of a course of action))

This pattern has the meaning "agire secondo una certa linea di condotta" (to act according to a particular course of action). Typical collocates we found are *dieta, esempio, moda, metodo, modello, tendenza, trend*. Definitions no. 17 and 12 in the COBUILD dictionary loosely correspond to our pattern:

"If you follow a particular religion or political belief, you have that religion or belief."
"If you follow advice, an instruction, or a recipe, you act or do something in the way that it indicates."

T5) Seguire con lo sguardo qu. che si sta spostando (to follow with your eyes sb. who is moving)

This pattern has the meaning "mantenere lo sguardo su qu. che si sta spostando" (to keep one's eyes on sb. who is moving). It is disambiguated by the prepositional phrase "con lo sguardo", which is an idiomatic argument of *seguire*. Definition no. 10 in the COBUILD dictionary corresponds to our pattern:

"If you follow something with your eyes, or if your eyes follow it, you watch it as it moves or you look along its route or course."

T6) Seguire una certa scena in corso (to follow a particular scene in progress)

This pattern has the meaning of "fare attenzione e percepire/ capire il progredire di una certa scena in corso" (to pay attention and perceive/ understand the progression of a particular scene). Typical collocates are *partita*, *concerto*, *trasmissione*, *discussione*, which may refer to actual shows or, more in general, to collective activities progressing with time (jumping in place would not qualify as one) and in which the perceiver does not take part. As a spectator, she or he may witness the activity in person or via a medium, for instance the TV.

T7) Seguire una certa attività remota/ regolare in corso (to follow a particular remote/ regular activity in progress)

The meaning is "tenersi aggiornati sul procedere di una certa attività remota/ regolare" (to keep up to date on the progress of a particular remote/ regular activity). Typical collocates are *sport*, *calcio*, *vicenda*, *movimenti di qu.* Definition no. 16 of the COBUILD dictionary corresponds to our pattern:

"If you follow something, you take an interest in it and keep informed about what happens."

T8) Seguire qu. che sta narrando, spiegando o argomentando (to follow sb. who is telling a story, explaining, or making an argument)

The meaning is "fare attenzione e capire lo svolgimento della narrazione, della spiegazione o dell'argomentazione di qu." (to pay attention and understand the progression of sb.'s story, explanation, argument). Notice that here the activity in progress is not only perceived, but must be interpreted.

T9) Seguire una certa narrazione, spiegazione o argomentazione (to follow a particular story, explanation, or argument)

We found several typical collocates for the direct object, such as *lezione*, *logica*, *filo*, *ragionamento*, *argomentazione*, *racconto*, *spiegazione*. In the COBUILD, definition no. 15 corresponds to our pattern:

"If you are able to follow something such as an explanation or the story of a film, you understand it as it continues and develops."

I1) A un primo periodo/ situazione/ evento SEGUE un secondo periodo/ situazione/ evento (a second period/ situation/ event follows a first period/ situation/ event)

This pattern has the meaning "un secondo periodo/... viene immediatamente dopo un primo periodo/... in ordine temporale" (a second event/... comes immediately after a first event/... in time order). Collocates appearing as arguments were quite easy to identify and extremely heterogeneous: *caduta*, *dissoluzione*, *rielezione*, *proclamazione*, *bocciatura*, *sconfitta*, *dichiarazione*, *crollo*, *scoppio*, *terremoto*, *tracollo*, *sisma*, *ritrovamento* and many others. Definition no. 4 in the COBUILD dictionary corresponds to our pattern:

"An event, activity, or period of time that follows a particular thing happens or comes after that thing, at a later time."

I2) A una prima persona/ oggetto SEGUE una seconda persona/ oggetto (a second person/ object follows a first person/ object)

This pattern has the meaning "una seconda persona/ oggetto viene immediatamente dopo una prima persona/ oggetto in un ordine spaziale/ convenzionale" (a second person/ object comes immediately after a first person/ object in a spatial/ conventional order). The word sketches revealed no typical collocates. Hence we chose very general semantic types by introspection. Definition no. 7 in the COBUILD dictionary corresponds to our pattern:

"If you refer to the words that follow or followed, you are referring to the words that come next or came next in a piece of writing or speech."

I3) Un evento SEGUE DA un altro evento (an event follows from another event)

This pattern has the meaning "un evento è effetto di un altro evento" (an event is the effect of another event). Word sketches have not been particularly useful in this case. The only typical (idiomatic) collocation we could identify is "ne seguì" + [Evento] (an event followed from that), which hints at the fact that, in this pattern, *seguire* is just an abbreviation of *conseguire*, with precisely this meaning.

I4) Un'affermazione SEGUE DA un'altra affermazione (a statement follows from another statement)

This pattern has the meaning "un'affermazione è vera se è vera un'altra affermazione" (a statement is true if another statement is true). A statement is here the logic consequence of another. Also in this case, *seguire* seems to be an abbreviation of *conseguire* with the same meaning. Definition no. 6 in the COBUILD dictionary loosely corresponds to our pattern:

"If it follows that a particular thing is the case, that thing is a logical result of something else being true or being the case."

I5) Un testo SEGUE IN una parte di supporto testuale diversa dalla presente (a text follows in a different part of a textual carrier)

This pattern has the meaning "proseguire in un'altra parte di supporto testuale" (to continue in a different part of a textual carrier). The only typical collocate in the locative slot that emerges from the word sketches is *pagina*, indicating a 'textual place'. For the subject role we have chosen the semantic type [Testo], which covers all typical lexical items. In this case, *seguire* seems to be an abbreviation of *proseguire*.

3.2 Idiomatic sub-patterns and notable exploitations

We assigned the idiomatic expressions *seguire la corrente* and *seguire i passi/ le orme di qualcuno* to pattern T4. We did the same with a limited but significant number of similar figurative expressions, like *seguire il cuore/ l'istinto/ le inclinazioni/ gli impulsi* (to follow one's heart, instinct, inclinations, impulses)

In the concordance list found in Figure 1, we encounter the expression *seguire la voce di qu.* (to follow sb.'s voice). This is an exploitative alternation: *voce* (voice) → *narrazione* (story). In the same list, we also see an interesting example of a multiple exploitation. In the clause *lo sguardo segue la torre dall'alto in basso* (the eyes follow the tower from top to bottom), the tower is equated to a path along which the eyes can move. An interesting aspect of this exploitation is that a cognitive condition must be imposed on the tower for it to be compared to a path (it must have an elongated shape/ surface).

Query seguire 972,267 > GDEX 972,267 (315.99 per million)

Page 1 of 48,614 Go Next | Last

stpauls.it...	si erge maestosa una torre. Lo sguardo la	segue	dall´alto in basso fino a un cortile dove
helpaids.i...	spaventata. Mi chiedi se esista qualche regola da	seguire	. Solo una: trovare insieme a lui, il modo
spazioamic...	certi criminali. D'altra parte stiamo	seguendo	con passione quanto si muove a Palermo.
jobintouri...	dal pay-off Genova 2004. La rotta giusta ,	seguirà	nei prossimi mesi una campagna pubblicitaria
spazioamic...	talento. I dorotei avevano del talento!	Segue	gli umori bestiali della folla indotti
lonelyplan...	ottennero l'indipendenza; il Lussemburgo li	segui	di lì a poco. L'Olanda riuscì a rimanere
eprints.bi...	organizzato in cinque sezioni. All´introduzione	segue	una breve descrizione del sistema logistico
americaont...	enchiladas con concerto di musica mex a	seguire	. L.A. IS MY LADY (Frank Sinatra)
dweb.repub...	Toni Morrison precisa: "Ero costantemente	seguita	, spiata. Ho visto il dossier federale che
cronologia...	di prendere le armi insieme con lui o di	seguirlo	fuori. L'avesse ascoltato! Il conte invece
omero.it	in segreteria. Stesso tono spensierato.	Seguo	la sua voce mentre sfoglia i depliant e
instoria.i...	Non cera stato chi aveva affermato che,	seguendo	quel principio, avremmo dovuto ridisegnare
cattaneo-l...	quanto fa spettacolo. Ieri sera ho	seguito	i due servizi del TG2 sull´accensione di
senzasoste...	tornato dall´emigrazione. Anche negli anni che	seguirono	, che furono contraddistinti dalla feroce
savonanews...	gran direttore d'orchestra a bordo campo ha	seguito	un'impressionante prova di forza delle
ior.it	gli Istituti Ortopedici Rizzoli debbono	seguire	una precisa procedura. Tale procedura è
paginedidi...	assenza è ben nota alle maggiori Potenze, che	seguono	gli avvenimenti con una preoccupazione
denaro.it ...	situazioni a maggiore rischio. Quanti allievi	seguono	i corsi? Ad oggi circa cinquecento. Gli
ristretti....	chi probabilmente non capiva quasi nulla	segui	attentamente i miei movimenti e la convinzione
luisareali...	alle mutilazioni sessuali. E tutte l'hanno	seguita	. Di Gabriele Romagnoli, La Repubblica ,

Page 1 of 48,614 Go Next | Last

Figure 1: Excerpt from the Sketch Engine concordance list of *seguire*

3.3 Comparison with other resources

Following Sinclair's advice, we compared our results with those of existing resources, such as traditional dictionaries and ItalWordNet. What follows are the senses of *seguire* found on ItalWordNet, listed in exactly the same order but labelled according to our convention in order to highlight the similarities:

T1) Synset: (seguire [1])
Gloss: andare dietro a qlcu.

I1) Synset: (seguire [2], succedere [3])
Gloss: accadere successivamente o in conseguenza di qlco.

I3) Synset: (avere_origine [2], conseguire [3], derivare [2], nascere [9], procedere [5], provenire [2], resultare [1], risultare [1], seguire [3], sorgere [6], uscire [11])
Gloss: avere principio, essere causato (fig.); derivare, aver principio, origine, fondamento (fig.).

T2) Synset: (seguire [4], tenere [7])
Gloss: andare per un certo percorso.

T5) Synset: (accompagnare [4], seguire [5])
Gloss: seguire con lo sguardo, con il pensiero, ecc.

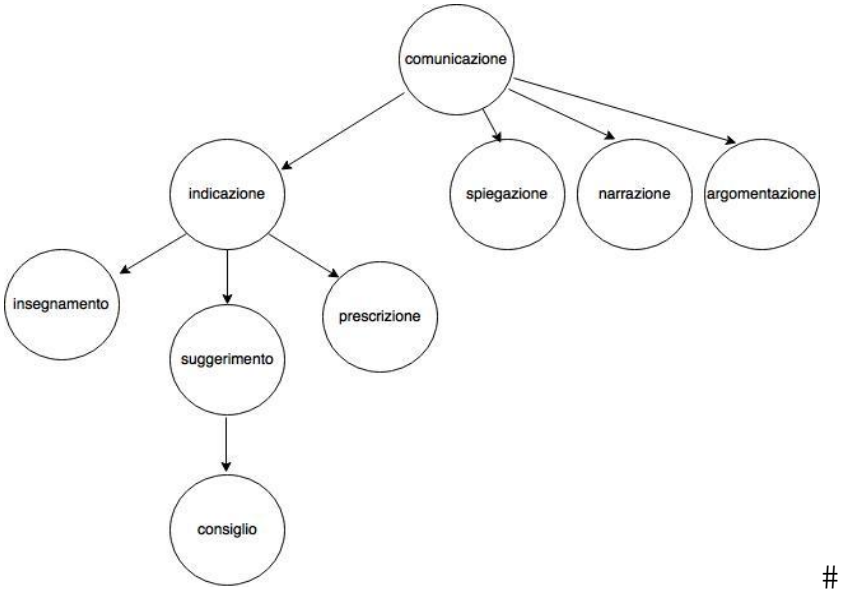
T4) Synset: (conformarsi [1], seguire [6])
Gloss: accettare un'idea, una dottrina e sim. "Seguire l'aristotelismo." "Seguire l'esempio di qc."

I4) Synset: (conseguire [2], seguire [7], susseguire [2])
Gloss: derivare come conseguenza, conseguire.

As aforementioned, the correspondence is remarkable: the main difference is that here senses T3, T6, T7, T8, and T9 seem to be missing. We think that this confirms the validity of our methodology.

As to the dictionaries, all problems lamented by Sinclair about traditional (pre-corpus) lexicography can be attested, e.g., the presence of long lost meanings (like sense C3 in ZINGARELLI: "accadere, avvenire: sono cose che seguono!"), abnormal examples, illogical splitting of meanings (like, in DE MAURO, senses '4a' vs. '5a': "mettere rigorosamente in pratica una regola, una norma, una convenzione" vs. "stare dietro all'evolversi di una tendenza uniformandosi ai suoi dettami"), illogical lumping.

The confirmation of Sinclair's thesis is indeed remarkable, and even the abstract semantic types which we identified are surprisingly robust (cf. Figure 2). The communication types, for example, correspond to Brinker's classification of texts (cf. Brinker, 2005). Furthermore, three out of four subtypes of [Comunicazione] immediately disambiguate to pattern T9, whereas [Indicazione/Descrizione] disambiguates to pattern T2 or T4 (and needs further disambiguation).



#

Figure 2: Communication types

4. Conclusions

As seen in our case study, Sinclair's legacy is more important than ever, most of all in those languages, such as Italian, where the corpus-driven approach is not yet mainstream. This is why the Italian advanced learner's dictionary we are currently designing with Zanichelli (which also aims at bridging an existing gap in Italian learner's lexicography) will be based on Sinclair's patterns of word usage.

The dictionary we are designing will have other important features, which we will introduce in upcoming articles. We will take into account the three mainstream approaches (cognitive linguistics, computational semantics, and lexical pragmatics) to the representation of polysemy in the mental lexicon and to its treatment in lexicography. Based on these, we will propose a user-oriented method for describing and differentiating word meanings. Disambiguators, as key microstructural items, will systematically apply in an ideal top-down procedure: ontological categories will distinguish lemmas and sub-entries (upper-level disambiguation), cognitive principles will determine word sense clusters (middle-level disambiguation) and Sinclair patterns will differentiate main word senses (lower-level disambiguation), whereas pragmatical principles will explain word sub-senses. In the enumeration and grouping of senses, we will prioritize semantic closeness criteria over frequency, since semantic closeness facilitates learning by association and is a key organising principle in our mental lexicon. Definitions will be created for each word sense by coherently employing a restricted defining vocabulary and by avoiding hidden circularities.

5. References

5.1 Dictionaries and lexicons

CALD: *Cambridge Learner's Dictionary*.
Accessed at: http://dictionary.cambridge.org/dictionary/english (22.05.2017)

COBUILD: *Collins COBUILD Advanced Learner's Dictionary* (2014). Harper Collins

COBUILD: *Collins COBUILD English Language Dictionary* (1987). Collins

DE MAURO: *Il Nuovo De Mauro*.
Accessed at: https://dizionario.internazionale.it (22.05.2017)

ItalWordNet. Accessed at: http://www.ilc.cnr.it/iwndb_php (22.05.2017)

LDOCE: Longman Dictionary of Contemporary English for Advanced Learners (2009). Pearson

MEDAL: MacMillan English Dictionary for Advanced Learner's. (2016). MacMillan

MultiWordNet. Accessed at: http://multiwordnet.fbk.eu (22.05.2017)

OALD: *Oxford Advanced Learner's Dictionary*.
Accessed at: http://www.oxfordlearnersdictionaries.com/ (22.05.2017)

PDEV: *Pattern Dictionary of English Verbs*. http://pdev.org.uk (22.05.2017)

TRECCANI: *Vocabolario Treccani*.
Accessed at: http://www.treccani.it/vocabolario (22.05.2017)

WordNet. Accessed at: http://wordnetweb.princeton.edu (22.05.2017)

ZINGARELLI: lo Zingarelli 2017 (2016).
Accessed at: http://dizionari.zanichelli.it/dizionarionline/online.php (22.05.2017)

5.2 Other resources

Bogaards, P. (2008). Frequency in learners' dictionaries. *Proceedings of the XIII EURALEX International Congress*, pp. 15-19.

Brinker, K. (2005). *Linguistische Textanalyse*, Erich Schmidt Verlag.

Carlson, G. (1977). *Reference to kinds of English*. Doctoral thesis, University of Massachusetts.

Carston, R. (2002). Linguistic meaning, communicated meaning and cognitive pragmatics. *Mind & Language*, *17*(1-2), pp. 127-148.

Cehan, A. (2015). The end of meaning-driven dictionaries?. *Proceedings of the International Conference Literature, Discourse and Multicultural Dialogue. Section: Language and Discourse*.

El Maarouf, I. (2013). Methodological Aspects in Corpus Pattern Analysis. *ICAME Journal*, 37, pp. 119-148.

Falkum, I.L. (2011). *The Semantics and Pragmatics of Polysemy: A Relevance-Theoretic Account*. Doctoral Thesis, University College London.

Geeraerts, D. (2001). The definitional practice of dictionaries and the Cognitive Semantic conception of polysemy. *Lexicographica*, *17*, pp. 6-21.

Hanks, P. (2013). Lexical analysis: Norms and exploitations, MIT Press.

Hanks, P. (2008a). The lexicographical legacy of John Sinclair. *International Journal*

of Lexicography, 21(3), pp. 219-229.

Hanks, P. (2008b). Lexical patterns: From Hornby to Hunston and beyond. *Proceedings of the XIII EURALEX International Congress*, pp. 89-129.

Hanks, P. (2004a). Corpus pattern analysis. *Proceedings of the XI EURALEX International Congress*, Vol. 1, pp. 87-98.

Hanks, P. (2004b). The syntagmatics of metaphor and idiom. *International Journal of Lexicography*, 17(3), pp. 245-274.

Hanks, P. (1987). Definitions and Explanations. In Sinclair, J. McH. (ed.), *Looking Up: An Account of the COBUILD Project in Lexical Computing and the Development of the Collins COBUILD English Language Dictionary.* Collins COBUILD, pp. 116-136.

Hanks, P., & Bradbury, J. (2013). Why do we need pattern dictionaries (and what is a pattern dictionary, anyway)?. *DICTIONARY News*, *27.*

Hanks, P. & Jezek, E. (2008). Shimmering lexical sets. *Proceedings of the XIII EURALEX International Congress*, pp. 391-402.

Hanks, P., & Pustejovsky, J. (2005). A pattern dictionary for natural language processing. *Revue Française de linguistique appliquée*, *10*(2), pp. 63-82.

Hanks, P., & Pustejovsky, J. (2004). Common Sense about Word Meaning: Sense in Context. Extended Abstract. In P. Sojka, I. Kopecek, I. & K. Pala (eds.) *Text, speech and dialogue*, Springer, pp. 15-17.

Hanks, P., Pala, K., & Rychlý, P. (2007). Towards an empirically well-founded semantic ontology for NLP. *Workshop on Generative Lexicon,* Paris, France.

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., & Suchomel, V. (2013). The tenten corpus family. *7th International Corpus Linguistics Conference CL*, pp. 125-127.

Jezek, E. (2008). Polysemy of Italian event nominals. *Faits des langues*, *30*, pp. 251-264.

Jezek, E., & Hanks, P. (2010). What lexical sets tell us about conceptual categories. *Lexis. Journal in English Lexicology*, (4), pp. 7-22.

Kilgarriff, A. (1997). I don't believe in word senses. *Computers and the Humanities* 31, pp. 91-113.

Kilgarriff, A. (1997). Putting frequencies in the dictionary. *International Journal of Lexicography*, *10*(2), pp. 135-155.

Kilgarriff, A., Rychly, P., Smrz, P., & Tugwell, D. (2004). Itri-04-08 The Sketch Engine. *Information Technology*, *105*, pp. 116.

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,* pp. 425-431.

Kilgarriff, A., & Gazdar, G. (1995). Polysemous relations. In F.R. Palmer (ed.) *Grammar and Meaning: Essays in Honour of Sir John Lyons*, Cambridge University Press, pp. 1-25.

Kozaki, K, Sunagawa, E., Kitamura, Y. & Mizoguchi, R. (2006). Fundamental consideration of role concepts for ontology evaluation. *Proceedings of EON2006*,

Edinburgh.
Kratzer, A. (1995). Stage-level and individual-level predicates. In G. N. Carlson & F. J. Pelletier (eds.) *The generic book*, University of Chicago Press, pp. 125-175.
Krishnamurthy, R. (2008). Corpus-driven lexicography. *International Journal of Lexicography*, *21*(3), pp. 231-242.
Lakoff, G. (1999). Cognitive models and prototype theory. In E. Margolis & S. Laurence (eds.) *Concepts: Core Readings*, MIT Press, pp. 391-421.
Lakoff, G., & Johnson, M. (2008). *Metaphors we live by*, University of Chicago Press.
Lew, R. (2013). Identifying, ordering and defining senses. In H. Jackson (ed.) *The Bloomsbury companion to lexicography*, Bloomsbury, pp. 284-302.
Louw, B. (1993). Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In M. Baker, G. Francis & E. Tognini-Bonelli (eds.) *Text and Technology: In Honour of John Sinclair*, John Benjamins Publishing, pp. 157-176.
Lyons, J. (1977). *Semantics*, Cambridge University Press.
Moon, R. (1987). The analysis of meaning. In J. McH. Sinclair (ed.) *Looking Up: An Account of the COBUILD Project in Lexical Computing and the Development of the Collins COBUILD English Language Dictionary*, Collins COBUILD, pp. 86-103.
Nerlich, B., Todd, Z., Herman, V., & Clarke, D. D. (eds.). (2003). *Polysemy: Flexible patterns of meaning in mind and language* (Vol. 142), Walter de Gruyter.
Pustejovsky, J. (2001). Type construction and the logic of concepts. In: P. Bouillon & F. Busa (eds.) *The language of word meaning*, Cambridge University Press, pp. 91-123.
Pustejovsky, J. (1995). *The Generative Lexicon*, MIT Press.
Pustejovsky, J., Hanks, P., & Rumshisky, A. (2004). Automated induction of sense in context. *Proceedings of the 20th international conference on Computational Linguistics*, *COLING*, pp. 924-930.
Roventini, A. et al. (2003). ItalWordNet: Building a Large Semantic Database for the Automatic Treatment of Italian. *Computational Linguistics in Pisa, Special Issue XVIII-XIX* Pisa-Roma IEPI. Tomo II, pp. 745-791.
Rundell, M. (2006). More than One Way to Skin a Cat. Why Full-Sentence Definitions Have not Beeen Universally Adopted. *Proceedings of the XI EURALEX International Congress,* pp. 323-337.
Sinclair, J. McH. (2004a). *Trust the text: Language, corpus and discourse*, Routledge.
Sinclair, J. McH. (2004b). In Praise of the Dictionary. *Proceedings of the XI EURALEX International Congress*, pp. 1-12.
Sinclair, J. McH. (2004c). New evidence, new priorities, new attitudes. In Sinclair, J. McH. (ed.), *How to Use Corpora in Language Teaching*, John Benjamins Publishing, pp. 271-299.
Sinclair, J. McH. (2003). *Reading concordances: An introduction*. Pearson Longman.
Sinclair, J. McH. (1998). The lexical item. In E. Weigand (ed.) *Contrastive Lexical Semantics (Current Issues in Linguistics Theory 171)*, John Benjamins

Publishing, pp. 1-24.
Sinclair, J. McH. (1996). The search for units of meaning. *Textus: English Studies in Italy* 9(1), pp. 75-106.
Sinclair, J. McH. (1991). *Corpus, concordance, collocation*. Oxford University Press.
Sinclair, J. McH. (1987). Grammar in the Dictionary. In J. McH. Sinclair (ed.) *Looking Up: An Account of the COBUILD Project in Lexical Computing and the Development of the Collins COBUILD English Language Dictionary*, Collins COBUILD, pp. 104-115.
Sinclair, J. McH. (1986). First throw away your evidence. In G. Leitner (ed.) *The English Reference Grammar,* Max Niemeyer Verlag, pp. 56-65.
Sinclair, J. McH., Jones, S. & Daley, R. (2004). *English Collocation Studies: The OSTI Report,* Continuum, xvii-xxix.
Stewart, D. (2010). *Semantic Prosody: A critical Evaluation*. Routledge.
Tarp, S. (2001). Lexicography and the linguistic concepts of homonymy and polysemy. *Lexicographica*, *17*, pp. 22-39.
Tognini-Bonelli, E. (2001). *Corpus linguistics at work* (Vol. 6), John Benjamins Publishing.
Vossen, P. (2002) (ed.) *EuroWordnet. General Document*, Version 3. Accessed at: http://dare.ubvu.vu.nl/bitstream/handle/1871/11116/EWNG?sequence=1 (22.05.2017).
Vossen, P. et al. (1998). *The EuroWordNet Base Concepts and Top Ontology*. Document LE2-4003, D017, D034, D036, WP5. Accessed at: http://dare.ubvu.vu.nl/bitstream/handle/1871/11130/D017.pdf (22.05.2017).
Whitsitt, S. (2005). A critique of the concept of semantic prosody. *International Journal of Corpus Linguistics* 10*(3)*, pp. 283-305.

The Compilation of an Online Corpus-Based Bilingual Collocations Dictionary: Motivations, Obstacles and Achievements

Adriane Orenha-Ottaiano

São Paulo State University (UNESP), São José do Rio Preto, Brazil
E-mail: adriane@ibilce.unesp.br

Abstract

This paper will discuss the motivations, obstacles and achievements of the building of an Online Bilingual Collocations Dictionary (English–Portuguese Collocations Dictionary and Dicionário de Colocações Portugues–Inglês). It is based on learner, parallel and online corpora and was designed for teachers and learners of English and Portuguese as a foreign language and learner and professional translators, among other users. With a view to fit in with the referred audience's needs and profile, it focuses on all types of collocations (verbal, noun, adjectival and adverbial collocations), so that it may enable them to use collocations more accurately and productively, as well as boosting their collocational competence. The methodology firstly relied on the extraction and analysis of collocations from a Translation Learner Corpus and also of more collocational patterns extracted with the help of Sketch Engine (Kilgarriff et al., 2004) using a selection of the frequency lemma list (only content words, such as nouns, adjectives and verbs) from The Corpus of Contemporary American English (Davies 2008–2012). Being the first online bilingual collocations dictionary in the aforementioned language directions, we hope to incorporate collocational information more quantitatively and qualitatively as well as to achieve the challenge of meeting learners' collocational needs in both languages.

Keywords: collocations dictionary; collocations; corpus-based dictionary; Portuguese as a foreign language; English as a foreign language.

1. Introduction

The aim of this paper is to report on the compilation of the *Online Bilingual Collocations Dictionary* (*English–Portuguese Collocations Dictionary* and *Dicionário de Colocações Português–Inglês*) as well as discussing its motivations, obstacles and achievements. The purpose of developing this collocations dictionary is a lexical pedagogical one. It is intended to help learners achieve collocational competence in the two focused languages, English and Portuguese. In this investigation, collocations are taken as pervasive, recurrent, arbitrary and conventionalized combinations, which are lexically and/or syntactically fixed to a certain degree and may have a more or less restricted collocational range.

The study of collocations has played a substantial role in Lexicology and Phraseology, and also in Lexicography or Phraseography, particularly in Pedagogical Lexicography

or Phraseography in the past few years. Innumerous work has advocated the building of dictionaries with a special focus on collocations or collocations dictionaries (Alonso-Ramos, 2006; Atkins & Rundell, 2008; Moon, 2008; Orenha-Ottaiano, 2013; 2016; Kilgarriff, 2015; etc.), owing to their relevance in learning and mastering a foreign language. As Fontenelle (2008: 12) noted "collocations are a hot topic in linguistics and in lexicography". In addition, the combination of corpus linguistics, the use of corpora and corpus tools has enormously contributed to the investigation of collocations in Corpus Lexicography.

It seems to be a consensus among lexicographers that the quality of monolingual and bilingual dictionaries has improved significantly, due to the methodology provided by corpus linguistics, and the use of corpora has enabled us to identify and extract phraseological units more easily and effectively. However, even though corpus linguistics, and all the approaches and computational tools developed with it, have been of great help and interest to corpus and e-lexicographers, "they were not taken up as routine processes by lexicographers" (Kilgarriff, 2015: 83). According to Kilgarriff (2015: 83–84), "the first impression of a collocation list was a basket of earth with occasional glint of possible gems needing further exploration, and it took long to use them for every word", and the solution to this problem was the creation of *Word Sketch*. Thus, considering that collocations are the central issue in this investigation, the use of Sketch Engine (Kilgarriff et al., 2004) and mainly of *Word Sketch* has become crucial to this investigation.

2. Motivations

The motivation for compiling an *Online Bilingual Collocations Dictionary* as proposed in this investigation lies in the fact that, as professors of a B.A. in Translation and a B.A. in the teaching of English Language, we readily recognize that collocations pose a serious problem, with regards to production (either oral or written), to foreign language learners as well as trainee and professional translators. The difficulties in combining words in order to produce the most frequently used collocations are clearly evident in these groups, mainly among learner translators. After having analyzed Stevick's research (1989) on learner success in learning a foreign language, Wray (2002), sharing Pawley and Syder's views (1983), concludes that the formulaic sequences used by native speakers are not easy for learners to identify and master, and that their absence greatly contributes to learners not sounding idiomatic. In that sense, having a collocations dictionary to provide them with collocational information may be of great help.

Additionally, we also have to take into account the scarcity of collocations dictionaries, especially bilingual ones. There are very good monolingual collocations dictionaries for learners of English as a foreign language, such as *Macmillan Collocations Dictionary for Learners of English* (Rundell, 2010), *Oxford Collocations Dictionary for Students of English* (Mcintosh et al., 2009), *LTP Dictionary of Selected Collocations* (Hill & Lewis,

1999) and *The BBI Combinatory Dictionary of English* (Benson et al., 1997). However, as for bilingual collocations dictionaries, specifically in the English–Portuguese or Portuguese–English directions, the proposed *Online Bilingual Collocations Dictionary* is, to my knowledge, the first.

Another motivation concerns the advantages of compiling an online collocations dictionary over a printed one. As Rundell (2013: 5) claimed "a dictionary accessed on a computer or a mobile device has great advantages over its analogue predecessors". The author quoted Kilgarriff when he described Macmillan's decision to stop printing dictionaries as "A day of liberation from the straitjacket of print", as the researcher regards printed books as a not very efficient medium for reference materials. Rundell also predicted that the trend of online dictionaries was unstoppable, and mentioned that Macmillan's focus only on digital dictionaries was "merely anticipating a move that all dictionary publishers will have to make eventually (and probably sooner than most people think)". And he was right! Four years later, it has been "an even bigger game-changer than the arrival of corpora in the 1980s"! Hence, an Online English Collocations Dictionary will allow users to have access to a wider range of collocations, more examples and will offer the advantage of being constantly updated and revised, which would not be possible in a traditional and printed dictionary.

Besides practical issues regarding the need for compiling a dictionary of collocations, I also have a personal reason for having embarked on this phraseological enterprise. Being a learner of foreign languages, since I first came to know about collocations, I have always seen them as one of the greatest challenges in communicating in a foreign language and achieving fluency. Because of that, I have become a frequent and active user of collocations dictionaries. Nevertheless, I have always wished we had a bilingual collocations dictionary to help me with some of our tasks, especially one from Portuguese into English. The pleasure of compiling and having available online a Bilingual Collocations Dictionary is highly rewarding, mainly because I am sure we are selecting entries which indeed have a significant level of difficulty for Brazilian leaners of English.

3. Methodology for extracting entries and collocations and The Corpus-based Online Bilingual Collocations Dictionary

The *Online Bilingual Collocations Dictionary* is aimed at intermediate, upper intermediate, advanced and proficient learners of English and Portuguese as a foreign language (ranging from B1 to C2, according to Common European Framework of Reference for Languages).

We have been working on this project for over four years. So far, the dictionary contains more than 560 entries and more than 7,500 collocations, both in English and in Portuguese, all of them with their corresponding contexts (examples) (see Figure 3 below). Currently, more data are being extracted and inserted in the dictionary. It is

intended for the dictionary to contain at least 3,000 entries and more than 30,000 collocations by the time it is scheduled to be launched online in 2018, considering that we have now gained more expertise, have selected and managed a stronger team, not to mention the considerable help we have had from the Sketch Engine (Kilgarriff et al., 2004), which has definitely boosted our lexicographical work. We also have the goal of increasing this number from 7,000 to 10,000 in the following years, in order to meet the needs of more advanced learners.

As for the macro and microstructure of the dictionary, the entries (on the left side of Figures 1A and 1B) as well as the collocations (Figure 2) of the *Online Bilingual Collocations Dictionary* are displayed in alphabetical order:

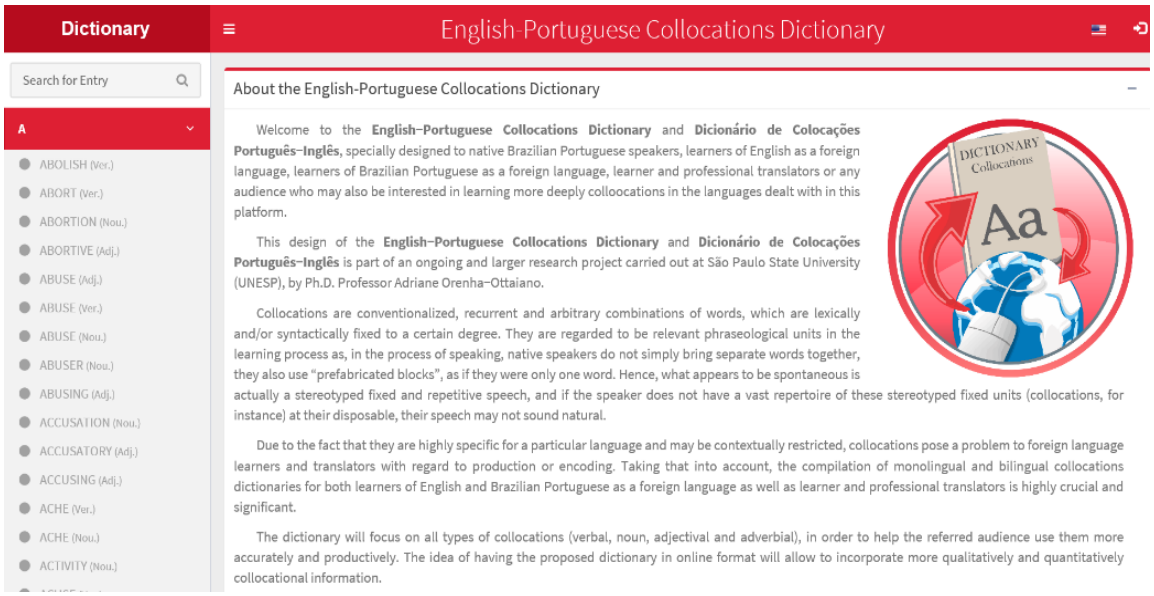


Figure 1A: Entries organized in alphabetical order (English–Portuguese direction)

Figure 1B: Entries organized in alphabetical order (Portuguese–English direction)

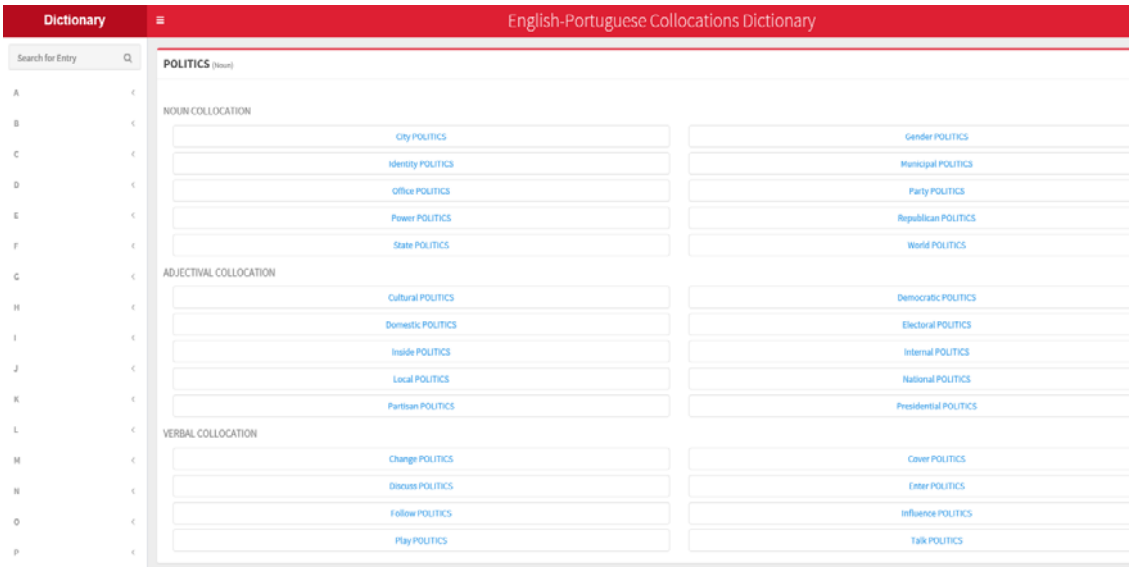


Figure 2: Collocations in alphabetical order

The entries and collocations are being selected on the basis of frequency and were thus chosen:

- From the words with the highest keyness from a *Translation Learner Corpus*, of approximately 100,000 words. This corpus is made up of newspaper articles taken from newspapers and magazines written in Portuguese and translated into English by undergraduates[1] from a BA in Translation course, at São Paulo State University (UNESP), in Brazil. The decision to use a *Translation Learner Corpus* to extract entries and collocations lies in the fact that we wanted to make sure that collocations which are difficult to be produced by Brazilian learners of English as a foreign language would be included in the focused dictionary. The most frequent collocations produced by the translation learners and extracted from the keywords[2] were analyzed and those which were not suitably produced by the students were then replaced by the correct and most frequent ones, extracted with the use of Sketch Engine's English Web 2013 (enTenTen13) corpus and were later included in the dictionary. It is worth mentioning that not only the collocations extracted from the *Translation Learner Corpus* were chosen for the dictionary, but also other patterns that were considered frequent in Sketch Engine's enTenTen13 corpus[3], with the use of Word Sketch[4];

[1] Students' level of English varied from B2 to C1 and their knowledge of language was identified according to the results of the *Oxford Placement Test* (Allan, 2004).

[2] At this point of the research, the computer tool used to generate keywords and concordance lines was WordSmith Tools (Scott, 2008). Currently, the Sketch Engine is used.

[3] The 2013 version of the enTenTen13 corpus contains almost 23 billion tokens.

[4] "The Word Sketch improves on standard collocation lists by using a grammar and a parser to find collocates in specific grammatical relations, and then producing one list of subjects, another of objects, etc, rather than a single grammatically blind list" (Kilgarriff & Tugwell, 2002: 25).

- From the most frequent words selected from *COCA*'s list (only content words). Again, the collocational patterns were extracted from the English Web 2013 corpus, with the use of Word Sketch. In due course, we intend to analyze the keywords from Sketch Engine's English and Brazilian Portuguese corpora;

- From the *New General Service List 1.01* (Browne, Culligan & Phillips, 2016), of approximately 2,800 words. This list was compared to COCA's selected lemma list, so that more collocational patterns were extracted, using Sketch Engine, from the contrastive list, bearing in mind that, if students have access to collocations from the most common vocabulary for learners of English taken from the *New General Service List*, they may have more opportunities to improve collocational competence.

The aforementioned methodological steps to extract keywords and collocations are concerned with the compilation of the entries and collocations for the English–Portuguese direction of the *Online Bilingual Collocations Dictionary.* In what regards to the keywords and collocations of the Portuguese–English direction, the first measure was to translate all entries and collocations from English into Portuguese. At the moment, we are also comparing the translated entries to the keywords from Sketch Engine's Brazilian Portuguese Corpus (*Corpus Brasileiro*), with 1,133,416,757 words.

The organization of entries in a collocations dictionary can be done on two different concepts: either *node* and *collocate* (Sinclair, 1991) or *base* and *collocator* (Hausmann, 1985). In the *Online Bilingual Collocations Dictionary*, we followed the concept of *base* and *collocator*, which means that the lexical entries in this work are the *base*, taking into account that the *base* is usually what we already know and the *collocator* is the element we are looking for, that is to say, what the learners want to find out. We have chosen this approach as we consider it to be more user-friendly in the sense that, as Hausmann claims, users usually know the base and need to find out which words co-occur with it. For example, they may know the adjective *hot*, however, they may want to know which adverbs can be used with it when they mean it is, as Brazilian learners of English say "very, very, very hot": native speakers of English would know they could use *boiling*, *scorching*, *stiflingly*, *unbearably,* etc. Having adjective entries is very productive in a collocations dictionary for Brazilian learners of English as they tend to use *very* for all adjectives, instead of more specific and straightforward ones.

As for the examples of the collocations displayed, if the user clicks on the collocation he or she is interested in learning, he or she will see the context in which this collocation is used. In case he or she wants to see the source of the examples for the selected collocation, the user may click on an icon ⓘ, indicating the source, so that the dictionary gets visually cleaner, as shown in Figure 3, in the right:

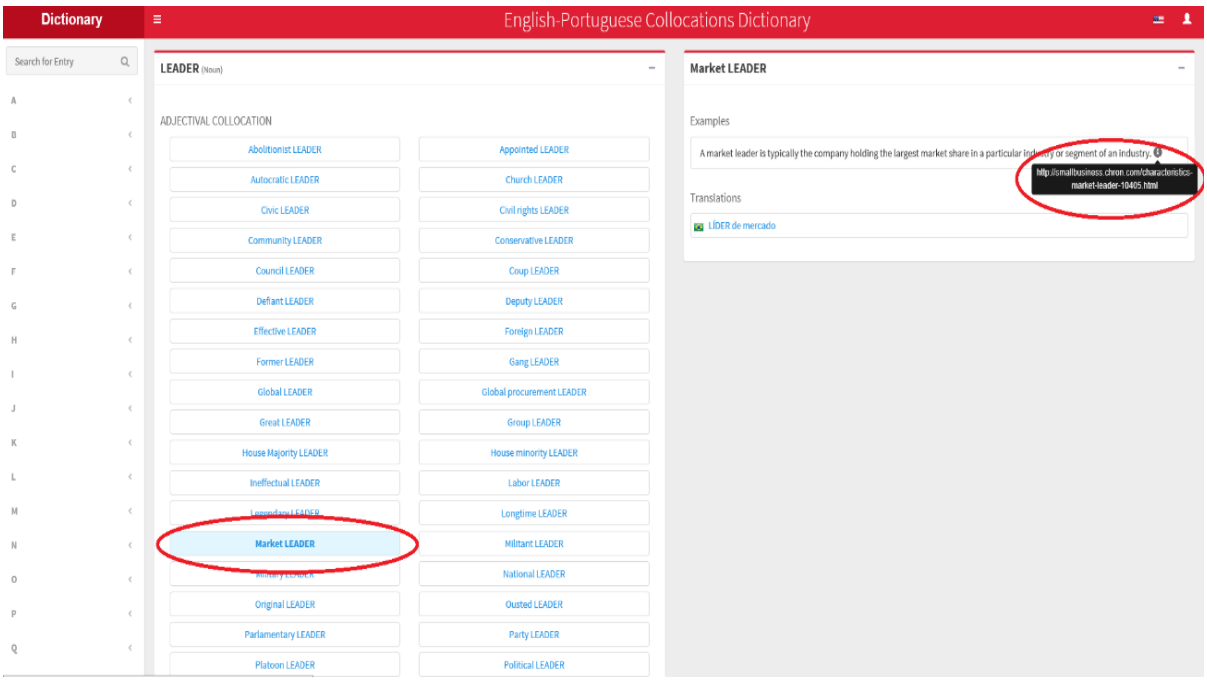


Figure 3: Icon indicating the source of the example

The screenshot from Figure 3 shows that the user clicked on one of the collocations (*market leader*), and an example of this collocation popped up on the right, as well as the equivalent collocation in Portuguese ("líder de mercado").

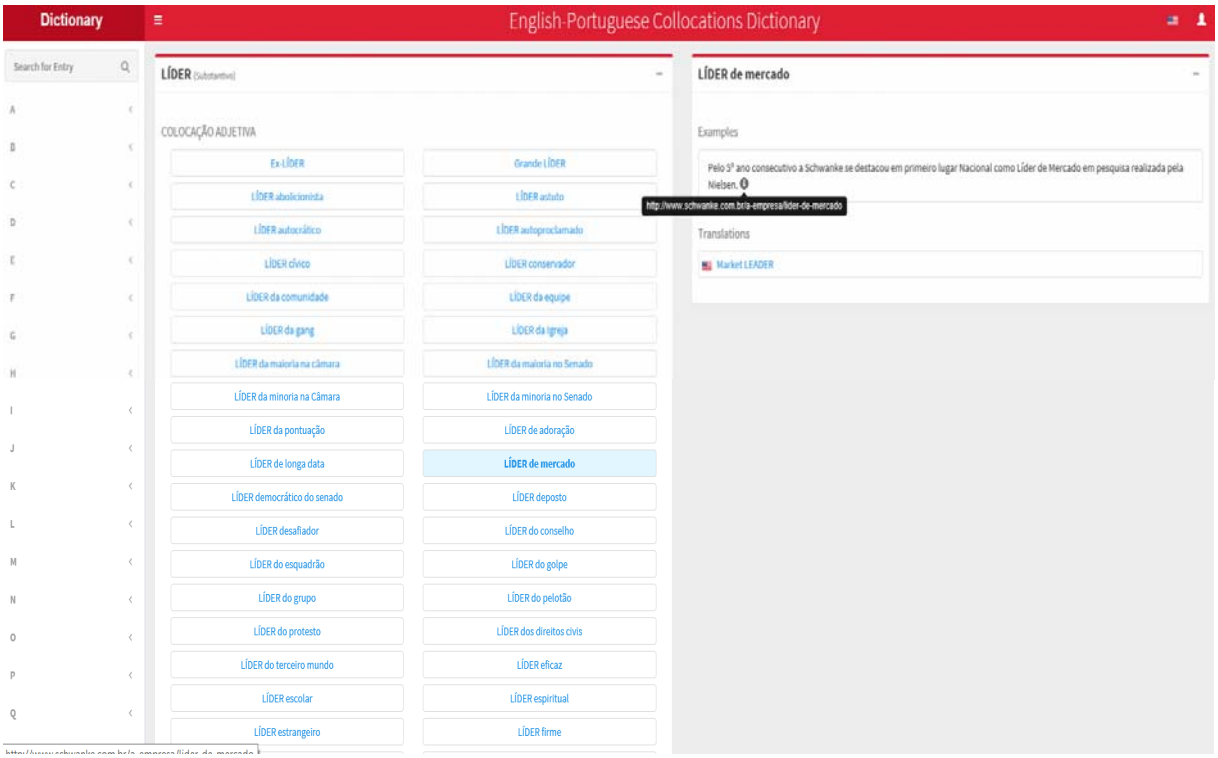


Figure 4: Screenshot of the equivalent collocation

If the user clicks on the translated collocation ("líder de mercado"), he or she will see the example of the equivalent collocation in Portuguese as well as all the possible collocations related to this entry in Portuguese and do the same search in this language:

The user will have the chance to choose the language from which he or she wishes to start, as shown in the top right position of Figure 5:

Figure 5: Language direction of the dictionary

As it can be seen from Figure 5, if you choose to use the Portuguese–English version, the screen will come up in green and blue. If the user chooses the English–Portuguese version, the color of the screen will appear in red and blue (as in Figure 6 below):

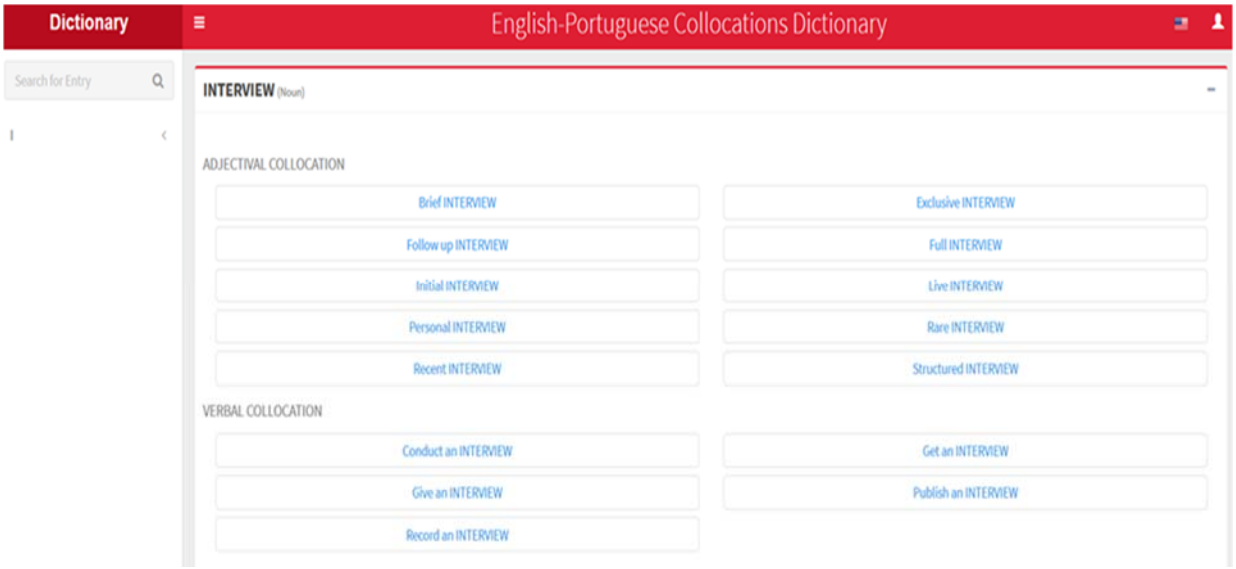


Figure 6: Screenshot of entry selected, types of collocations, and collocations extracted for the entry

According to Figure 6, one can also verify the types of collocations displayed from the entry interview: adjectival and verbal collocations, and the way they are arranged on the platform. As previously mentioned, the dictionary will bring all types of collocations: verbal, adjectival, nominal and adverbial, depending on the entry the user is focusing on. Here it follows the taxonomy of collocations we work with (Orenha-Ottaiano, 2004) expanded from Hausmann's classification:

Verbal Collocations – with four basic structures:

- Verb $_{\text{collocator}}$ + Noun $_{\text{base}}$: acquire shares
- Noun $_{\text{base}}$ + Verb $_{\text{collocator}}$: investments dropped
- Verb $_{\text{collocator}}$ + Preposition + Noun $_{\text{base}}$: dispose of shares
- Verb $_{\text{collocator}}$ + Adverbial Particle+ Noun $_{\text{base}}$: set up a business
- Verb $_{\text{collocator}}$ + Adjective $_{\text{base}}$: grow strong

Nominal Collocations – with two basic structures:

- Noun $_{\text{base}}$ + Noun $_{\text{collocator}}$: share subscription
- Noun $_{\text{collocator}}$ + Preposition + Noun $_{\text{base}}$: holder of shares

Adjectival Collocations – with one basic structure:

- Adjective $_{\text{collocator}}$ + Noun $_{\text{base}}$: bearer shares

Adverbial Collocations – with three basic structures:

- Adverb $_{\text{collocator}}$ + Adjective $_{\text{base}}$: fully eligible
- Verb $_{\text{base}}$ + Adverb $_{\text{collocator}}$: drop dramatically
- Adverb $_{\text{collocator}}$ + Verb $_{\text{base}}$: fully paid; duly appointed

Based on the macro and microstructure presented in this session and the screenshots of The *Online Bilingual Collocations Dictionary*, it it is evident that the dictionary is thus working, but remains under test and is subject to changes.

4. Obstacles and achievements in the compilation of the Online Bilingual Collocations Dictionary

The *Online Bilingual Collocations Dictionary* is the first bilingual English–Portuguese and Portuguese–English collocations dictionary, and this may be seen an achievement, particularly to the Brazilian Portuguese field of lexicography, considering that lexicographic work in Brazil still requires a lot of support, investment and recognition in research.

When we first proposed a bidirectional collocations dictionary, we were not aware of

the difficulties we would encounter. As we mentioned above, we made the decision to translate all the entries and collocations from English into Portuguese. When we finally had the online platform for the dictionary and inserted the translations, we realized we could have results which would conflict with the theoretical framework of a corpus-based work.

If we consider the direction English–Portuguese, there would not be any problem to translate all the entries and collocations from English into Portuguese, as that would be the natural procedure to carry out. However, if we take into account the Portuguese–English direction, as it is already displayed on the platform at this phase of the project, when we first visualize the entries and collocations in Portuguese, as researchers, we may question whether these are frequently used words and collocations in Portuguese, as they came from the translations.

To resolve this problem, we have compared the translated entries to the keywords from Sketch Engine's Brazilian Portuguese Corpus (*Corpus Brasileiro*), as was the original idea, in the sense of having entries and collocations in Portuguese which are frequently used by native speakers of Brazilian Portuguese. By doing so, we intended to make sure learners could master a foreign language by having contact with the lexicon that is considered to be frequently spoken by Portuguese speakers. As for the translated collocations, we have been comparing them to the collocations displayed by Word Sketch from Sketch Engine's Brazilian Portuguese Corpus (*Corpus Brasileiro*).

Having analyzed the online Portuguese–English direction dictionary, in comparison to the online English–Portuguese one, it seems as if they are two different dictionaries. If we delete the entries or collocations which are highly frequent in Portuguese or if we add up new entries or collocations that may be considered to be commonly used in Portuguese, according to the results of the Brazilian Portuguese Corpus, we would have to do the same with the equivalent ones in English. In this respect, we will have to take some measures, so as to better deal with this difficulty. The interface will have to be adjusted with a view to allowing access to an entry from a specific language without conflicting to the entries of the other one. And that is where the problem lies and obstacles have to be overcome. Another possible solution is dividing the dictionary and creating two bilingual dictionaries, in only one direction, instead of a one bidirectional dictionary, as presently proposed: An English–Portuguese Dictionary and a Portuguese–English Dictionary. This way, we would make sure users may have access to the most common collocations in each direction, also taking into account that the translations were carefully given, in the sense that we tried to find the most possible frequent equivalent in the target language.

Another obstacle we faced, mainly in the initial phase of the project, was not having a good team to rely on: and that is crucial when it comes to a phraseographical work. To carry out this project we needed to count on students with advanced or proficient knowledge of the language, so that they could identify collocations. Besides that, we

also needed the help of advanced or proficient learner translators to deal with the translation of the entries and collocations. The greatest problem was recruiting this team, as some students were not committed to research and, resultantly, we could not trust their work and ended up by doing everything ourselves. At the beginning of 2017, we started training a more committed team involving three master's students, one Ph.D. student, six undergraduate students (four of them from B.A in English Language and two from B.A. in Translation). As our aim is also to promote the teaching of collocations in public schools, two high school students from a public school will soon join the project, having applied for a scholarship. However, although we have now built a larger team, it is still necessary for the researcher in charge to undergo considerable revisions and reviewing of hundreds of collocations as well as their translations.

Another point worth mentioning regards the organization of the entries, as it has become a considerable difficulty when extracting collocations. As previously mentioned, the entries were compiled anchored in Hausmann's concept of *base* and *collocator*. When some members of the team were extracting collocational patterns, we realize we could have some problems. For instance, let us consider the keyword *vacancy*. When analyzing some of the collocations extracted, such as *job vacancy*, *senate vacancy*, *vacancy announcement*, and *vacancy rise*, we noticed that, based on Hausmann's concept, *vacancy announcement* and *vacancy rise* would not be inserted under the entry v*acancy,* but on the entries *announcement* and *rise*, as they are considered to be the base. Another very good example refers to the entry *blood*, from which the following collocations were extracted: *blood samples, blood evidence, blood pressure, blood stains, blood pool, blood trail, blood spatter, blood void, blood type, clotted blood*. If we apply Hausmann's theory, the only collocation for the entry *blood* would be *clotted blood*. The other collocations would have new entries: *samples, evidence, pressure, stains, pool, trail, spatter, void, type.*

Thus, we would have to analyse these new candidates for entries (*announcement, rise, samples, evidence, pressure, stains, pool, trail, spatter, void* and *type*), in order to check if they are frequently used, so that they can be included in the dictionary. That has become an obstacle as, besides having to extract collocations from the big list of keywords we have previously mentioned (the *Translation Learner Corpus keyword list*, and the contrastive list resulted from the comparison of COCA's selected lemma list and *New General Service List 1.01)*, we would also have to spend a considerable amount of time extracting more collocations from these new entries and our work may end up by being endless. This way, it would be advisable to reconsider Sinclair's concepts if the change indeed favors the extraction of the most relevant collocations as well as addressing learner needs.

With respect to the translation process of collocations, some collocational types have posed a challenge in translating from English into Portuguese, mainly, and mostly, the adverbial collocations, both the Adverb + Adjective and Verb + Adverb construction.

Adverbial collocations are not as frequent in Brazilian Portuguese as they are in the English language and that is the reason why it turns out to be challenging to translate them or to find a correspondent collocation. Many adverbial collocations presented difficulties in translation, such as *hopelessly naive*, *increasingly political*, *fall sharply*, *rightly fear*, *hard-core unemployed,* etc., indicating that the translation of adverbial collocations may be a challenge in phraseographical work on collocations, in the English–Portuguese direction at least.

For instance, the collocation *increasingly irrelevant* was considered hard to translate into Portuguese. The translation options given were "extremamente irrelevante", "completamente irrelevante", "totalmente irrelevante" and "absolutamente irrelevante". Although the four translation options proved to have a high co-occurrence frequency according to Google statistics, none of them describes the idea of *becoming larger* (*increasingly*). An alternative translation is the use of three words before the noun, "cada vez mais irrelevante", like the dictionarized meaning of *increasingly*: *more and more*, because in Portuguese we have not found a single adverb. Some translations therefore had to be carefully considered as the Word Sketch results for the adjective "irrelevante" were not satisfactory and did not suggest a more suitable equivalent in Portuguese, as shown in Figure 7:


irrelevante (adjective)
Brazilian Portuguese corpus (Corpus Brasileiro) freq = 3,994 (3.52 per million)

modifies		
	706	17.68
detalhe	45	6.69
detalhes irrelevantes		
Internautas	2	6.52
tex-	2	6.50
zos	2	6.49
artifícios	2	6.33
construto	4	6.22
inconsistência	2	6.06
adicional	2	6.01
torna	2	5.78
estímulos	5	5.55
son	2	5.29
concorrência	5	5.29
microestrutura	2	5.27
Alterações	2	5.04
aparte	2	5.02
quantia	4	4.99
pergunta	12	4.90
de perguntas irrelevantes		
parecer	17	4.88
pode parecer irrelevante		
capitão	2	4.82
explicações	2	4.71
é	10	4.63
é irrelevante		
topografia	2	4.50
variância	2	4.37
assunto	16	4.28
assuntos irrelevantes		
informações	9	3.67

y_o		
	5	0.13
próprio	2	9.00
nível	2	6.14

Figure 7: Word Sketch for the search word *irrelevante*

As Figure 7 shows, the result in Word Sketch for the search word *irrelevante* does not present any adverbial collocation patterns in Portuguese. One of the reasons for this may be that as the referred adjective is less frequent in Brazilian Portuguese (3,994 occurrences in the corpus, 3.52 per million) in relation to English (162,325 occurrences in the corpus, 7.14 per million), the number of collocational patterns is lower. Also, the search word "irrelevante" may not be frequently used in the adverb + adjective structure in Brazilian Portuguese and that may be why, out of 1,133,416,757 words which compose the Brazilian Portuguese Corpus (*Corpus Brasileiro*), no adverbial collocations were found. Nevertheless, these findings will have to be investigated more systematically in future studies. Another reason may be related to fact that the Brazilian Portuguese Corpus is not large enough to give the same results in comparison to the enTenTen13 corpus and thus, in order to have more reliable results or to provide more collocational options, it should be expanded. According to our experience in compiling the *Collocations Dictionary*, many search words in Portuguese did not have a satisfactory result in the Word Sketch with regards to collocational patterns, not only for adverbial collocations, but also for noun, adjectival and verbal collocations. Members of the team also reported that they have had difficulties in finding collocational patterns or have not found any successful results in the Brazilian Portuguese Corpus and, on account of that, we believe that expanding the corpus would be greatly welcome.

Another collocation considered hard to translate into an equivalent in Portuguese was *unflagging enthusiasm*. Although it is not an adverbial collocation, but a noun collocation, with the structure Adjective + Noun, it may be given as an example of translation problems. The translation options given, regarding the meaning of *unflagging* (not changing or becoming weaker) were "grande entusiasmo", "crescente entusiasmo", and "constante entusiasmo", the latter being the best translation option. These translations were carefully proposed by the researcher's team, taking into account that the Word Sketch results for the noun "entusiasmo" were also not satisfactory, as Figure 8 illustrates, even though some adjectival collocations for this search word were presented by the tool, differently from the previous search with the node *irrelevant.*

In order to achieve the translation options given, we also had to carry out a careful research using *Google*, checking whether they were frequently used in Portuguese. Besides that, three options were given and we had to decide on the best one to be included in the dictionary. Although some collocations and collocation types in English are hard to translate or may not have a 'perfect' equivalent in Portuguese, we have to take into account that part of the dictionary's target audience is learner and professional translators. Therefore, we have made a great effort to offer them a translation equivalent, so that they may have the chance to enrich their translated texts, as well as easing their work.

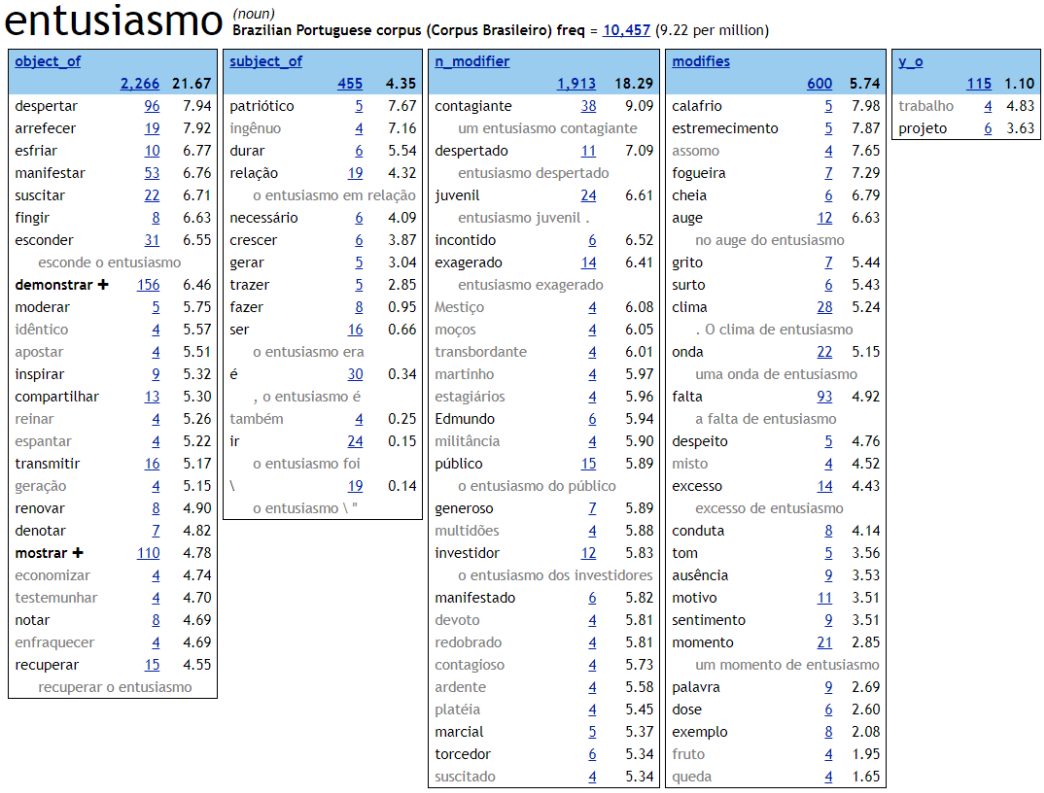


Figure 8: Word Sketch for the search word *entusiasmo*

It is worth mentioning that, having predicted the absence of a collocation in the target language, considering that some collocations in one language may correspond to a single lexical item in another, we created a symbol in the dictionary (☞) which indicates that the dictionary will inform the user whenever there is not a correspondent collocation in the target language.

All in all, compiling a corpus-based bilingual collocations dictionary is a time-consuming and long-term activity, demanding complete dedication and full commitment from the lexicographer. On the other hand, it is an enormously rewarding lexicographical enterprise. Regarding the Sketch Engine, even though it did not help us find a more accurate translation option in Portuguese, it has proved to be an invaluable tool for this work.

5. Conclusion

Although the compilation of a collocations dictionary may provide a lexicographer with a substantial amount of work, the result of it is an extremely useful tool for learners of English and Portuguese as foreign languages, as well as learner and professional translators. The pedagogical result of the lexicographical work may be of great help and value to the referred audience, and finally it can be regarded as a highly rewarding and extremely pleasant investigation and enterprise.

This paper reported on the motivations, as well as obstacles and achievements, in building the *Online Bilingual Collocations Dictionary*. Results have shown that even

though some difficulties and obstacles have arisen in the lexicographical process of compiling the collocations dictionary, the fact of having identified them may help us find more effective solutions so that we may produce a reliable source for the target audience.

Having shed light on the achievements and potential benefits of an online corpus-based bilingual collocations dictionary, and considering it to be the first collocations dictionary in the English–Portuguese and Portuguese–English directions, we hope the publication of the online version of the dictionary will be widely accessed and potentially useful.

6. Acknowledgements

I gratefully acknowledge the partial, but loyal, financial support provided by FAPERP (Fundação de Apoio à Pesquisa e Extensão de São José do Rio Preto).

7. References

Allan, D. (2004). Oxford Placement Tests 2. Oxford: OUP.

Alonso Ramos, M. (ed.) (2006). *Diccionarios y Fraseología*. A Coruña: Universidad A Coruña.

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

Benson, M., Benson, E. & Ilson, R. (1997). *The BBI Combinatory Dictionary of English*. Amsterdam/Philadelphia: John Benjamins.

Browne, C., Culligan, B. & Phillips, J. (2016). The New General Service List. Available at: http://www.newgeneralservicelist.org. [4 Jan 2017].

Kilgarriff, A., Rychly, P., Smrz, P. & Tugwell, D. (2004). The Sketch Engine. In G. Williams & S. Vessier (eds.). *Proceedings of the 11th Euralex International Congress*. Lorient: Universite de Bretagne-Sud, pp. 105-116.

Davies, M. (2008-2012). *The Corpus of Contemporary American English:* 425 million words, 1990-present. [online] Available at: http://corpus.byu.edu/coca. [Nov 20, 2016].

Fontenelle, T. (ed.) (2008). *Practical lexicography*: A reader. New York: Oxford University Press.

Hausmann, F. J. (1985). 'Kollokationen im deutschen Wörterbuch. Ein Beitrag zur Theorie des lexikographischen Beispiels. In H. Bergenholtz & J. Mugdan (eds.). *Lexikographie und grammatik*. Tübingen: Niemeyer.

Hill, J. & Lewis, M. (1999). LTP *Dictionary of Selected Collocations*. Hove, London: Language Teaching Publications.

Kilgarriff, A. (2015). Using corpora as data sources for dictionaries. In H. Jackson (ed) *The Bloomsbury companion to Lexicography*. London: Bloomsbury, pp 77-96.

Kilgarriff, A. & Tugwell, D. (2002). Sketching words. In *Lexicography and Natural Language Processing:* a Festschrift in Honour of B. T. S. Atkins. Marie-Hélène

Corréard (ed.). EURALEX, pp. 125-137.
Mcintosh, C., Francis, B. & Poole, R. (eds.) (2009). *Oxford Collocations Dictionary for Students of English*. 2nd ed. Oxford: Oxford University Press.
Moon. R. (2008). Dictionaries and collocations. In F. Meunier and S. Granger (eds.). *Phraseology in Foreign Language Learning and Teaching*. Amsterdam: John Benjamins, pp. 247-252.
New General Service List 1.01 (2016). Available at: http://www.newgeneralservicelist.org. [25 jan. 2017].
Orenha-Ottaiano, A. (2004). *A compilação de um glossário bilíngue de colocações, na área de jornalismo de Negócios, baseado em corpus comparável*. MA diss., University of São Paulo (USP), Brazil.
Orenha-Ottaiano, A. (2013). The proposal of an electronic bilingual collocational dictionary based on corpora. In *X International School on Lexicography Procceedings*, 2013. Florence, Italy. Life Beyond Dictionaries. Ivanovo: University of Ivanovo, pp. 405-408.
Orenha-Ottaiano, A. (2016). The compilation of an online corpus-based bilingual Collocations Dictionary. In Gloria Corpas Pastor. (ed.). *Computerised and Corpus-based Approaches to Phraseology*: Monolingual and Multilingual Perspectives. Genebra: Editions Tradulex, 2016, pp. 486-493.
Pawley, A. & Syder, F. (1983). Two puzzles for linguistic theory: nativelike selection and nativelike fluency. In J Richards & R. Schmidt (eds.) *Language and Communication*. London: Longman.
Rundell, M. (2010) Macmillan Collocations Dictionary for Learners of English. Oxford: Macmillan Publishers Ltd.
Rundell, M. (2013). Redefining the dictionary: From print to digital. In Kernerman Dictionary News 21. Available at: http://kdictionaries.com/kdn/kdn21.pdf. [4 Jan 2017].
Scott, M. (2008). *WordSmith Tools*, version 5.0. Liverpool: Lexical Analysis Software Ltd.
Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.
Stevick, E. W. (1989). *Success with foreign languages*. Hemel Hempstead: Prentice Hall.
Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Auto-generating Bilingual Dictionaries

Noam Ordan[1], Jorge Gracia[2], Ilan Kernerman[1]

[1] K Dictionaries Ltd, 8 Nahum Hanavi Street, 6350310 Tel Aviv, Israel

[2] Universidad Politécnica de Madrid, Campus de Montegancedo sn, Boadilla del Monte 28660 Madrid, Spain

E-mail: noam@kdictionaries.com, jgracia@fi.upm.es, ilan@kdictionaries.com

Abstract

Inferring a bilingual dictionary from two or more existing bilingual dictionaries is a non-trivial task, as seen in reports of large-scale, computationally-heavy experiments published in recent years. Early works on this have already noted that the main obstacle in such inferences stems from the fact that polysemy is not isomorphic across languages, and often a monosemous lexical item in one language can be polysemous in its corresponding translation into another language. In this paper, we propose an experiment on translation inference across dictionaries, based on a graph-based view of a collection of bilingual dictionaries. The idea is to explore the results and analyze them from a lexicographic point of view to reflect the implications that the issue of anisomorphy introduces in the task, and to illustrate its hurdles and potential benefits.

Keywords: automatic dictionary generation, bilingual lexicography, polysemy, translation

1. Introduction

Dictionaries are a human effort at representing meaning, whether of a language on its own terms, or of one language (L1) vis-à-vis another language (L2). Whereas the task of the translator (human or machine) is to find an ad-hoc solution for substituting the meaning of an L1 item with its equivalent in L2 given some context, this solution may be partial or rare or just good enough for the context at hand, but completely useless otherwise, since estimating and reusing such rare events is not realistic within state-of-the-art machine translation systems. Similarly, a bilingual dictionary that would list all the items in L2 that were ever given as equivalents for a unit in L1 would be an impractical resource, even from the point of view of machines, searching over all possible options is considered NP-complete, i.e., unrealistic computationally (Knight, 1999).

Scaling up the human effort required for compiling bilingual dictionaries into a highly multilingual landscape is an inherently difficult task, due to the combinatorial explosion of pair-wise language comparisons. To alleviate this issue, the automatic generation of bilingual/multilingual dictionaries, based on already existent ones, is a research and practical avenue which merits exploration, with the aim of assisting and complementing human-based dictionary compilation.

Our methodology in the current experiment is computationally straightforward: the algorithm starts with L1 and goes to L2 then L3 (and L4, L5, etc.), and ends with a

translation from the last language in the chain back to L1. By starting with a given sense in L1 and finally retrieving it again as a translation in the last pair of the chain (which we call "closing the loop"), we reinforce the confidence in our selection. In addition, if the loop is not closed in the last chain, we consult another bilingual dictionary (see below).

The rest of the paper is organized as follows: Section 2 describes the data we utilize in our experiments. In Section 3 we present the experiment and in Section 4 the results of the automatically generated translations. Section 5 reports on an additional contribution of our methods, which allows for automatic generation of synonymous and semantically related words, and Section 6 reviews relevant literature, both practical implementations of solutions to the problem and the lexicographic and lexicological obstacles which should be overcome. In Section 7, we conclude with a brief discussion.

2. Dataset

The experiments rely on two subsets of data of K Dictionaries, namely MLDS and KMT:[1]

- MLDS

The Multi-Language Dictionary Series (aka Global Series, cf. Kernerman, 2015), currently contains lexicographic cores for 24 languages. Each consists of approximately 12,000 main entries featuring detailed semantic and grammatical information, including alternative script, word categorization and inflected forms, definitions and examples of usage, word sense disambiguators and various attributes (e.g., synonyms and antonyms, register, sense qualifier), multiword expressions, etc. Several languages have a second level that doubles their size (to about 25,000 entries), and Spanish is quadrupled (50,000 entries). The L1 cores are created from scratch with the idea of minute lexical mapping, and can be used to produce monolingual dictionaries, but serve mainly as a base for integrating translation equivalents and developing bilingual sets. So far, nearly one hundred pairs were developed manually, though their division among L1 cores is unequal: on the one hand, three have no bilingual versions yet, whereas French, on the other hand, is the most extensively translated (into 18 languages). The bilingual versions of a single language core are juxtaposed together, thus forming a multilingual dataset of that L1.

[1] The initial experiments (called Cross-Lingual Automated Common Senses, CLACS) began in-house in 2016, making use of full cross-lingual lexicographic resources. In 2017, the shared task on Translation Inference Across Dictionaries (TIAD) was launched, making available to researchers limited bilingual dictionary resources, with results presented at a workshop held as part of the first Language, Data, Knowledge conference (https://tiad2017.wordpress.com/).

- KMT

The K Multilingual Translators (KMT) (aka MultiGloss, cf. Egorova, 2015; Kernerman, 2015) consist of semi-automatically generated by-products of the English Multilingual Dictionary (KEMD) that include translations in 45 languages. Twenty-two of these translation languages have been reversed and manually edited and refined into detailed bilingual word-to-sense L1 indices to English. Then, the KEMD translations in all the other languages are added to the English equivalents of these bilingual indices, thus producing the multilingual index – linking each sense of the L1 headword via its English counterpart to all the other language translations that are available in KEMD.

3. Procedure

Our graph is rather simple, and we traverse it the following way:

The new bilingual dictionary was generated by using four language pairs from MLDS, as follows:

1. German to Turkish (DE>TR)
2. Turkish to French (TR>FR)
3. French to Brazilian Portuguese (henceforth 'Portuguese', FR>BR)
4. Portuguese (back) to German (BR>DE)[2]

The results were processed with the help of two factors:

1. Check translations from the existing MLDS set from Portuguese to German (i.e., pair 4 above). If a German translation is recognized, we consider it a 'closed' loop – since we begin from a specific sense of a German entry and end up with the same entry.
2. If not found, check for a translation (in 4) in the KMT Portuguese-German resource. Recall that KMT is created semi-automatically, so in terms of confidence we trust more the selection of translations stemming from MLDS. However, we use it as another pivot to validate the inferred translations.

[2] Needless to say, we could reverse the last pair, i.e., Brazilian-Portuguese>German to improve results, but we have self-imposed a restriction to avoid this option so that our study is carried out in lab-clean conditions where only cross-dictionary pivoting is considered.

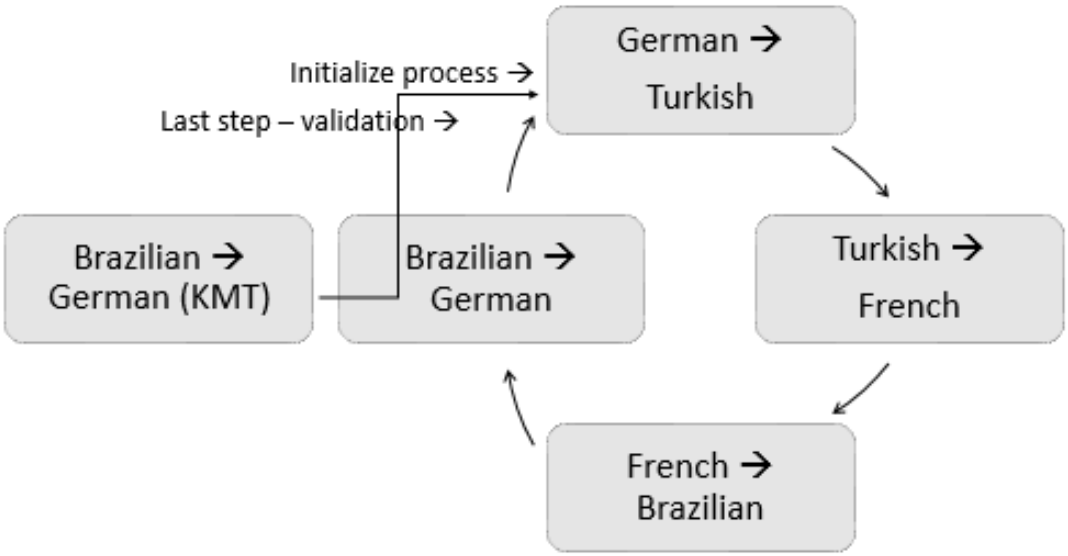


Figure 1: The four language-pair chains with an additional validation step for closing the loop

As we discuss in Section 6, there are many sub-cases of anismorphism across and between languages, and it turns out that more often than not – even though divergence (i.e., one- or few-to-many mapping between L1 and L2) grows exponentially and for each source-language item in German we manage to retrieve a huge number of back-to-German-translations – usually we do not manage to attain the identical German words, although we use two Portuguese>German dictionaries as our final pivot. This is well illustrated in Figure 2. The source-language word *Abkommen* is not found among the eight inferred translations of the last pair that closes the loop, i.e., in the Portuguese to German dictionary.

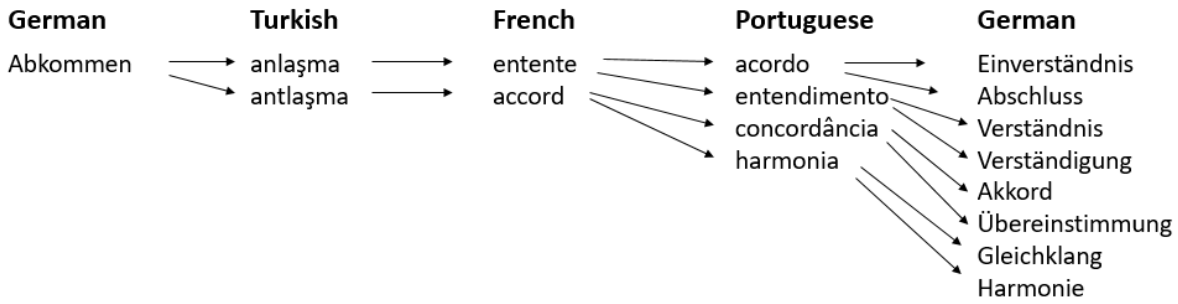


Figure 2: Divergence across languages increasing exponentially
(so the loop does not always close)

4. Findings

We began with 12,000 German entries (from MLDS). The number of entries that were found in both MLDS and KMT (BR>DE) is 5,865. The matches break down according to five quality scores (for the summary statistics see also Table 1). (The total number of matches before closing the loop with KMT was 8,722.)

Quality 1: contains 4,377 entries (74.63%); it is defined operatively as a case-sensitive exact match between the initial headword of the German dictionary plus the part of speech and a translation containing exactly these features (same word, same part of speech) arrived at via a chain of bilingual dictionaries.

Quality 2: contains 44 entries (0.75%); the only difference from the Q1 criterion is a non-match in terms of upper/lower-case letter (which is typical/unique for German).

Quality 3: contains 24 entries (0.41%); the only difference from Q1 and Q2 criteria is the missing article in German, e.g., *Warnung* die *Warnung* (also typical/unique for German).

Quality 4: contains 594 entries (10.13%), where the initial headword is a substring of the final string arrived at in the chain, e.g., *Boden* der *(Erd)Boden*.

Quality 5: contains 826 entries (14.08%), where the initial German headword does not match the final translation arrived at (as regards MLDS), though it does match a translation existing in KMT. For example, *Bestandteil* is potentially synonymous to *Grundbestandteil*. This quality score generates our candidates for synonyms (see Section 5).

quality score	raw frequency	ratio
1	4,377	74.63%
2	44	0.75%
3	24	0.41%
4	594	10.13%
5	826	14.08%
Total	5,865	100.00%

Table 1: Results of automatic matching according to quality

In term of precision, we report ~75% accuracy. This is considerably lower than the 90% accuracy reported in a much more sophisticated algorithm devised in Mausam et al. (2008). However, it should be borne in mind that there are major differences in the setting and the results are not comparable. We suggest that the reasons for the non-

comparability touch on fundamental issues concerning the current task:

1. The number of valid translations for any given word or phrase is much larger than reflected in a bilingual dictionary. Specia and Nunes (2006) estimate that for certain lexical items there are hundreds of possible translations. Our evaluation was done against a medium-sized bilingual dictionary (that had the restriction of offering maximum three translation equivalents), and any item that was automatically inferred and was not found in the dictionary is considered "an error". Mausam et al. (2008), however, sampled hundreds of inferred translations and used crowd-sourcing to decide whether the translations were valid or not. This allows to increase the number of candidates beyond that which is found in a dictionary. Additionally, their evaluation relied on self-proclaimed native speakers, and therefore must be taken with a grain of salt.

2. Given that the number of possible translation is so big, we have no access to the total number of translations for all the entries. This means that recall cannot be calculated, as recall, by definition, is calculated against the total number of relevant/correct items.

5. Synonyms and semantic fields

The lowest Q5 score does not necessarily imply a bad translation, but could indicate potential synonym candidates, or at least semantically-related words of relevance for learners and other users or for computational tasks concerning word-sense disambiguation and information retrieval. We could thus also consider utilizing this architecture to generate *semantic clouds* that surround any word sense in our data.

The intuition behind the generation of these semantic clouds by Q5-scored translations is the following. Consider the scenario we have experimented with: L1>L2>L3>L4>L1. The fourth node, L3>L4, yields a large amount of translations, most of which are noisy. Looking at the resources L4>L1 (MLDS and KMT) is akin to eliciting two judgements, and it stands to reason that if both point back at the same L1 item the chances that the L4 is a good translation candidate for the source-language L1 item increase. However, if one points back at this L1 item (KMT), but the other does not (MLDS), what does it mean? Arguably, and as illustrated in Table 2, both yield valid translations, often synonymous (like *Addresse* and *Anschrifft*).

As we have indicated in Section 3, we preferred to rely more heavily on MLDS as its quality is higher, and therefore preferred to penalize results where MLDS did not have a match and KMT did; however, as can be seen from Table 2, a match in KMT and a non-match in MLDS has three possibilities: (1) a synonym, which can be taken as a valid translation, is yielded; (2) a semantically-related word is retrieved; (3) rarely, a non-related word is retrieved. Our sample space is too small to arrive at statistically meaningful results.

A case of a non-synonymous but closely related words is, for example, *Bestandteil* and *Grundbestandteil*, and it was for this reason that we decided to score differently everything generated through this sub-procedure. Semantically, however, the two words are closely related, meaning *element* and *basic element*, respectively. In other cases, we find that non-matches like these are also related, albeit more vaguely.

Consider the following path:

German		**Turkish**		**French**		**Portuguese**		**German**
Abenteuer	→	*serüven*	→	*aventure*	→	*aventura*	→	*Affäre*

In this case, the match is taken from KMT, so we do have some confidence with respect to the validity of the path/result. The word *Abenteuer* means *adventure*, and *Affäre* means *(love) affair*. This figurative extension indicates a semantic relation.

Headword	**English gloss**	**Synonym candidate**	**English gloss**
Abc	ABC	**Alphabet**	alphabet
Abgrund	precipice, abyss	Welten	worlds
ablehnend	unfavorable	**Negative**	negative
Absatz	paragraph, leap	Sprung	jump
Abschnitt	section	**Absatz**	paragraph
Absicht	intention	**Ziel**	goal
Achtung	danger, esteem	**Wertschätzung**	appreciation
Addresse	address	**Anschrifft**	address
Affe	monkey	Wagenheber	jack (for a car)
Akt	act	**Handlung**	action
autonomy	autonomous	**Unabhängig**	independent
Autonomie	autonomy	**finanzielle Unabhängigkeit**	financial independence
Autoritär	authoritarian	**Diktatorisch**	dictatorial
Backpulver	baking powder	**Hefe**	yeast
Ball	ball (also as in ballroom)	**Fest**	celebration / feast
Bankrott	bankruptcy	**Insolvenz**	insolvency
Barriere	barrier	**Hindernis**	obstacle
Barriere	barrier	**Schranke**	barrier

Table 2: Candidate synonyms generated automatically

Table 2 summarizes 36 more cases for German. As appears below, only three pairs of items are not semantically related, the other 33 (marked in boldface) are either near synonyms or closely related.

6. Related work

Some studies have been reported in the literature. For instance, Tanaka and Umemura (1994) proposed a method to infer indirect translations through a pivot language when constructing bilingual dictionaries. Their pivot is generated by reverse dictionaries and therefore is different than the one suggest here. The loop is thus closed when a two-time inverse consultation is used, such that a set of candidate translations from L1>L2>3 is compared to what they call selection area generated by looking backwards at L3>L2>L1. A modification of their algorithm is suggested by Lim et al. (2011) in the creation of multilingual lexicons.

Other efforts have been done to compile massive multilingual dictionaries automatically, such as Mausam et al. (2008), which rely on the probabilistic exploration of the whole set of translations considered as a graph. Villegas et al. (2016) evolved this notion and moved it to a linked data landscape, applied to the RDF version of the Apertium family of bilingual dictionaries (Gracia et al., 2016). Another method that utilizes corpora and low quality lexicons as seeds is proposed by Shezaf and Rappoport (2010).

These studies illustrate the complexity of the problem, but are hampered by inherent difficulties of the translation process, such as the anisomorphism of languages and lexical gaps. Further, the dictionary compilation process goes far beyond identifying translation candidates. In fact, full-fledged bilingual dictionaries require, among others, selection of L1 lexical items in the first place, division and ordering of their senses, morphological information, usage examples, sense-specific translations in L2 and glosses for lexical gaps, and more. Lexical gaps are just one case out of several that undermines the illusion of a 1:1 mapping between languages; some other cases will be discussed below. As summarized by Adamska-Sałaciak (2006), "a bilingual dictionary cannot perfectly account for meaning, since meaning is always anchored within a particular language"; quoting Zgusta (1971), "the fundamental difficulty of ... a coordination of lexical units [between two languages] is caused by the *anisomorphism* of languages, i.e., by the differences in the organization of designates in the individual languages and by other differences between languages."

Byrd et al. (1987) identify different types of mismatches between languages which violate lexicographic symmetry, notably:

1. Morphologically, some words occur only or mostly as inflected forms, and therefore their lemmatized (uninflected) form should not appear as a translation equivalent. For example, *allege* appears as a lemma in the English to Italian dictionary they used (*Collins English-Italian Dictionary*, CEID), but it is not

provided as a translation in the Italian to English part, since in most cases it is used as a participle.

2. Some languages make specific distinctions not made in another language, and the best way to represent such specific meanings in the target language is with their superordinate equivalents. For example, the French word *fleuve*, which denotes a river that flows into the sea, is normally translated into the more general term *river*, but reversing *river* back to *fleuve* could be a mismatch that should be at least reviewed (it is translatable also to *rivière*, or river that flows into another course of water).

3. In other cases, the opposite case occurs, namely, a more general item is substituted with a more specific one. For example, *book* is translated in CEID into *quaderno* (*notebook*), *bustina* (*of matches*), and *blocchetto* (of tickets). Here the transfer is from a general term to specific ones.

Another form of anisomorphism is lexical gaps.

1. According to Bentivogli and Pianta (2000), "a lexical gap occurs whenever a language expresses a concept with a lexical unit whereas the other language expresses the same concept with a free combination of words". They, too, use (a digital version of) CEID and estimate that around 7.8 percent English entries are translated to Italian with a free combination of words, that is, there are thousands of English words which exist as concepts in Italian but are not lexicalized. It is interesting to note that the least gapped part-of-speech is nouns (5.6%) and the most gapped one is adverbs (19.3%).

It is clear from the above that such discrepancies between languages increase the more pivots are added to a system, L1>L2>L3>L4, etc. But even using just one pivot language, i.e., inferring L1>L3 from L1>L2 and L2>L3, is no simple automatic procedure.

Previous works indicate that new resources can be created automatically, if not completely from scratch then at least based on existing ones. To do it successfully, rigorous lexicographic practice should be applied side by side with automatic methods, which are more error-prone. Admittedly, large-scale automatic generation which report 90% accuracy (Mausam et al., 2008) seems promising, but it is unlikely that (human) users would want to use dictionaries where every tenth word contains an error.

7. Conclusions

Empirically analysing current and new techniques for automatic inference of translations with the aim of integrating them with the more rigorous lexicographic practice has become a necessity. As a step in this direction, an experiment has been devised to explore the potential and hurdles of the task. An outcome of this experience

has been the creation of benchmark data for the comparison of different translation inference techniques.[3]

Our findings indicate that the growth rate is exponential, and that *closing the loop* is a sound method for higher quality assurance, but using it as a sole method – although highly precise – leads to a relatively low recall. We have plugged in another pivot (KMT) as a second source for closing the loop, and have shown that beyond its role as an extra validation step, it can generate synonyms and semantically-related words.

Our initial experiments are promising, as we obtain relatively high-quality translation results as well as open the ground for automatically generating additional useful by-products like synonyms and semantic fields. These findings serve as a first step for further experimentation with the use of pivots (which, how many, and how) and with incorporating additional components of the entry (subject fields, synonyms, etc.).

In the future, we intend to use the full potential graph of MLDS and plug in KMT in each and every node. The use of multiple ways to close the loop raises further problems, some of which are due to an inflation in the number of suggested translations. Some works report pruning algorithms to resolve the problem (Mausam et al., 2008; Gracia et al., 2016). Theoretically, it would be interesting to study the effect language similarity plays in divergence. It stands to reason that the more similar two languages are, the less divergence one could expect. This, too, calls for a future study.

8. References

Adamska-Sałaciak, A. (2006). *Meaning and the bilingual dictionary: The case of English and Polish*. Peter Lang.

Bentivogli, L. & Pianta, E. (2000). Looking for lexical gaps. In U. Heid, S. Evert, E. Lehmann & C. Rohrer (eds.) *Proceedings of the 9th EURALEX International Congress*. Stuttgart, Germany: Institut für Maschinelle Sprachverarbeitung, pp: 663-669.

Byrd, R. J., Calzolari, N., Chodorow, M. S., Klavans, J. L., Neff, M. S. & Rizk, O. A. (1987). Tools and methods for computational lexicology. *Computational Linguistics*, 13(3-4), pp. 219-240.

Egorova, K. (2015). Editing an automatically generated index with K Index Editorial Tool. In I. Kosem, M. Jakubíček, J. Kallas, S. Krek (eds.) *Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference, 11-13 August 2015, Herstmonceux Castle, United Kingdom*. Ljubljana/Brighton: Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd, pp. 268-280. Available at: https://elex.link/elex2015/proceedings/eLex_2015_17_Egorova.pdf

[3] Used for the first time in the context of the TIAD-2017 shared task (https://tiad2017.wordpress.com/)

Gracia, J., Villegas, M., Gómez-Pérez, A., & Bel, N. (2016). The apertium bilingual dictionaries on the web of data. *Semantic Web Journal*.

Kernerman, I. (2015). A multilingual trilogy: Developing three multi-language lexicographic datasets. In I. Kosem, M. Jakubíček, J. Kallas & S. Krek (eds.) *Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference, 11-13 August 2015, Herstmonceux Castle, United Kingdom*. Ljubljana/Brighton: Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd, pp. 372-383. Available at: https://elex.link/elex2015/proceedings/eLex_2015_24_Kernerman.pdf

Knight, K. (1999). Decoding complexity in word-replacement translation models. *Computational Linguistics*, 25(4), pp. 607-615.

Lim, L. T., Ranaivo-Malançon, B. & Tang, E. K. (2011). Low cost construction of a multilingual lexicon from bilingual lists. *Polibits* 43, pp. 45-51.

Mausam, Soderland, S., Etzioni, O., Weld, D, Skinner, M. and Bilmes, J. (2008). Compiling a Massive, Multilingual Dictionary via Probabilistic Inference. In *Annual Meeting of the Association of Computational Linguistics*. ACL.

Shezaf, D. & Rappoport, A. (2010). Bilingual lexicon generation using non-aligned signatures. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL '10*, 98-107. Stroudsburg, PA: Association for Computational Linguistics.

Specia, L. & Nunes, M. G. V. (2006). Exploiting the translation context for multilingual WSD. In *Text, Speech and Dialogue*. Berlin/Heidelberg: Springer, pp. 269-276.

Tanaka, K. & Umemura, K. (1994). Construction of a Bilingual Dictionary Intermediated by a Third Language. In *Proceedings of the 15th Conference on Computational Linguistics, Volume 1*, pp. 297–303. ACL.

Zgusta, L. (1971). *Manual of Lexicography.* Hague – Paris: Mouton.

TEI-Lex0 Guidelines for the Encoding of Dictionary Information on Written and Spoken Forms

Piotr Bański[1], Jack Bowers[2], Tomaž Erjavec[3]

[1] IDS Mannheim, Mannheim, Germany

[2] Austrian Academy, Vienna, Austria

[3] Jožef Stefan Institute, Ljubljana, Slovenia

E-mail: banski@ids-mannheim.de, iljackb@gmail.com, tomaz.erjavec@ijs.si

Abstract

The paper reviews the results of work done in the context of TEI-Lex0, a joint ENeL / DARIAH / PARTHENOS initiative aimed at formulating guidelines for the encoding of retro-digitized dictionaries by streamlining and simplifying the recommendations of the "Print Dictionaries" chapter of the TEI Guidelines. TEI-Lex0 work is performed by teams concentrating on each of the main components of dictionary entries. The work presented here concerns proposals for constraining TEI-based encoding of orthographic, phonetic, and grammatical information on written and spoken forms of the lemma (headword), including auxiliary inflected forms. We also adduce examples of handling various types of orthographic and phonetic variants, as well as examples of handling the representation of inflectional paradigms, which have received less attention in the TEI Guidelines but which are nonetheless essential for properly exposing data content to the various uses that digitized lexica may have.

Keywords: dictionary encoding; TEI XML; TEI-Lex0

1. Introduction

The Text Encoding Initiative (TEI) Guidelines (TEI Consortium, 2016) are the chief deliverable of a project running since the early 1990s and aiming at equipping the scholar with markup suitable for describing the majority of textual forms and analytic approaches and providing extension capabilities to encompass new or infrequently found phenomena. Being a complex toolbox aiming to encode any existing work, the Guidelines provide multiple encoding solutions and have frequently been criticized on this account. The standard response to such criticism and a recommendation for the purpose of ensuring interoperability has been to fully utilize the TEI's modelling and documentation format, ODD ("One document does it all", cf. TEI Consortium, 2013). However, given that tools with the capacity to parse and semantically analyze ODD descriptions are still being developed, a common-sense strategy to secure interoperability is to come up with a lean, transparent format that may not be able to handle all the potential variation, but will instead address "90% of phenomena, 90% of the time". This is the goal of TEI-Lex0, a joint ENeL / DARIAH / PARTHENOS initiative aimed at formulating guidelines for the encoding of retro-digitized dictionaries by streamlining and simplifying the "Print Dictionaries" chapter of the TEI Guidelines and the module defined therein.

The result is not meant to replace that chapter, but rather to serve as baseline encoding against which existing dictionaries can be compared and which could serve as a pivot format for generic querying or visualization tools.

TEI-Lex0 work is performed by teams concentrating on each of the main components of dictionary entries. The main focus of the present paper is on the `form` element, designed to contain orthographic, phonetic, and grammatical information on written and spoken forms of the lemma (headword), including its inflected forms that are sometimes – depending on the source language and established lexicographic practices – used as auxiliary information for the purpose of identifying the entry, or which illustrate inflectional patterns by means of partial or complete paradigms.

Below, we first present the assumptions that underlie the work of TEI-Lex0, and then proceed to review our proposals for constraining the `form` element and its contents. At each point, an illustration is provided, frequently going beyond use types covered by the TEI Guidelines.

1. General Assumptions

This section presents the basic TEI-Lex0 assumptions relevant to the phenomena described in the remainder of the article.

1.1 Abstract models and serialization

A fundamental principle that TEI-Lex0, or virtually any TEI-based dictionary-modelling enterprise, must rely on concerns the nature of the mapping of the physical or "near-physical" (OCR-ed) dictionary structure onto the abstract model of dictionary structure, and the mapping from said model onto its TEI XML serialization.

This is because the TEI vocabulary is heavily restricted and also influenced by some unsystematic historical decisions. The restriction is partially due to the fact that the TEI uses the same elements of the abstract model to serve many kinds of text-modelling tasks, and standardly employs 'features' or 'facets' of these elements to signal differences among them (the features in question are expressed in the XML serialization in the form of attributes, such as, e.g., `@type`). The structural context of these elements often matters as well. The fact that some elements of the serialization have names closely corresponding to what we can customarily find in the dictionary model is more or less a lucky coincidence – it is not a pattern to be expected. A lexicographer coming from outside the TEI should not, therefore, expect their customary terms (names of dictionary objects in the dictionary model) to be straightforwardly reflected in the TEI vocabulary names.

A good illustration is provided by the elements `form` and `sense`, which might be expected to contain information about form (of the headword and related items) and about the sense, respectively. And they do, except they do it in several ways:

```
<entry>
  <form>
     <orth>bray</orth>
     <pron>brei</pron>
  </form>
…
```

Example 1.

Above, the `form` element behaves as expected, but – as exemplified in Section 2.4 below, it can also nest other `form` elements, and then the outer `form` becomes merely a "box" for form-related information. Similarly, with `sense`:

```
<sense>
  <def>cry of an ass; sound of a trumpet</def>
</sense>
```

Example 2.

Above, the element `sense` contains a single definition, but it can also nest other `sense` elements, and then the outer `sense` becomes a "box" for sense-related information within the entry, and its internal structure may reflect the dictionary author's convictions or observations about the relatedness of subsenses, while the ordering of `sense` elements, whether nested or top-level, may express information about the frequency of the given subsense in the base corpus of data (we treat the term "corpus" here to mean the body of data that the lexicographer takes into consideration when creating the dictionary).[1]

The differences in the interpretation of elements such as `form` and other recursive elements make it necessary to adopt in TEI-Lex0 a rule that they may never appear without an accompanying `@type` attribute. Section 3 provides some examples.

1.2 Grammatical Information

In order to determine the complete set of properties of an element constituting a part of a hierarchy of lexicographic objects, onto which a dictionary entry can be mapped, the principle of default inheritance is assumed (cf. Ide et al., 2000; Erjavec et al., 2000). According to this principle, grammatical properties of a form are determined by collecting the sibling `gramGrp` of the ancestor-or-self of the focus element, where the superordinate grammatical properties can be overwritten by the lower-level properties. This principle is relatively straightforward in the case of grammatical properties, but more complex for the word paradigm, especially for variant forms.

The *modus operandi* assumed in the TEI-Lex0 is reductionist: from among the variety of means of encoding the relevant information offered by the TEI, precise guidelines

1 Another relevant example, to which much discussion in the TEI-Lex0 group was devoted, is the `cit` element. Originally, its name derives from "citation", but its semantics has got generalized over time to the point where a more suitable name could be "container-inside-text", given the range of uses and contexts, for and in which it is now applicable.

for the placement and content of the `form` and `gramGrp` elements are proposed, extending to finer-grained elements of the former such as `orth` for orthography and `pron` for pronunciation, and, in the case of the latter, to various subtypes of the `gram` element.

2. Recommendations for Encoding <form>

This section reviews most of the TEI-Lex0 recommendations for the treatment of `form` and dependent elements, including the treatment of `gramGrp`.

2.1 Grammatical information

Grammatical properties of lexical entries should be specified in `entry/gramGrp`.[2] This element will typically specify at least the part-of-speech of the entry, sometimes with some further specifications, such as, for example, transitivity for verbs or gender for nouns. While the TEI has defined a number of specialized elements within `gramGrp`, TEI-Lex0 takes a more generic route in this respect, for reasons of uniformity and sustainability. The former criterion makes it possible to simplify the processing tools and unify the representation. The latter makes the format more resilient to future modifications of the TEI: if, for example, at some point in the future, the TEI defines an element `voice` for grammatical voice, the TEI-Lex0 guidelines will not need to be adjusted – all that will be necessary will be another mapping between, say, `<voice>active</voice>` in the target dictionary and `<gram type="voice">active</gram>` in TEI-Lex0. This last point is also a reminder that TEI-Lex0 is not meant as production format, but rather as the base layer for retro-digitization, and possibly a pivot format to mediate between particular implementations of the "Print Dictionaries" chapter of the TEI Guidelines.

```
<entry xml:lang="en">
  <form type="lemma"><orth>on</orth></form>
  <gramGrp><gram type="pos">prep</gram></gramGrp>
  ...
</entry>
```

Example 3.

Because the part-of-speech property is a property of the entire entry, by the principle of default inheritance mentioned in Section 2.2, it is mandatory to encode it as a direct child of the `entry` element (recall that it is inherited by the `form` element, in the absence of a conflicting specification). In cases reviewed in the following sections, where grammatical properties pertain to the headword alone or to its various inflections, the

2 A `gramGrp` element is a child of an `entry` element. The TEI format is an application of XML, and as such, it follows all the practices, conventions and restrictions that govern XML representations. For the sake of explicitness, we utilize the XPath conventions for referencing fragments of XML structure, and thus "a `gramGrp` element that is contained inside a `form` element bearing an attribute `@type` with the value 'lemma', which in turn is contained within the element `entry`" is concisely expressed as `entry/form[@type="lemma"]/gramGrp`.

`gramGrp` element with appropriate content is placed as a child of `form[@type="lemma"]`, etc.

By the same token, in cases where headwords are distinguished only on the basis of their orthography (e.g., in dictionaries of English which treat conversion pairs of nouns and verbs, such as *run*, as belonging in single entries), `entry/gramGrp` should not be used, because its role is taken over by the individual `sense/gramGrp` elements, which either further specify grammatical properties of the individual sense or override those that pertain to the entire entry.

2.2 Representation of the lemma

The `form` element should always be qualified by its `@type` attribute set to one of the recommended values. The lemma (i.e., headword) should be under `form[@type="lemma"]`. This is illustrated in Example 3 above.

If it is necessary to specify the grammatical properties of the lemma form itself (as opposed to the grammatical properties of entire the entry), the relevant `gramGrp` element should be a child of `form[@type="lemma"]`. This may occur in languages such as Hebrew, where verbs are lemmatized as 3rd Person Masculine (simple) Perfect, or Greek, where verbs are lemmatized as 1st Person Singular (active indicative). In such cases, however, the relevant grammatical information is encoded mostly for the purpose of machine interpretation rather than for direct human consumption, and various project-dependent choices may regulate its actual placement. We will therefore not dwell on such issues here.

2.3 Representation of the inflected forms

Dictionaries often include additional forms next to the lemma. These forms in many cases specify irregular inflectional forms, such as *corpus* / *corpora* or *take* / *took,* while in inflectionally rich languages they enable the user to determine the correct paradigm of the word (e.g., *krava* / *-e* in Slovene or *amo* / *amare* in Latin).

Such inflected forms should be encoded in `entry/form[@type="inflected"]`, e.g.:

```
<entry>
  <form type="lemma"><orth>go</orth></form>
  <form type="inflected">
     <orth>went</orth>
  <gramGrp><gram type="tense">past</gram></gramGrp>
  </form>
...
```

Example 4.

2.4 Paradigms

When several inflected forms can be present next to the lemma, these can be embedded in an `entry/form[@type="paradigm"]` element. The decision of whether to use this extra element depends on the particular dictionary and language.

The other use case for paradigms is when the full inflectional paradigm of the word is embedded in the entry, i.e., when the dictionary also includes all the word-forms of the words covered, which can be useful for example for machine processing.

An `entry` may contain several paradigms, for example a partial one for humans and a full one for machines, or one for each stem of a verb. Each paradigm type should be distinguished by the `form/@subtype` attribute.

```
<entry xml:id="perder" xml:lang="es">
  <form type="lemma">
    <orth>perder</orth>
  </form>
  <gramGrp><gram type="pos">verb</gram></gramGrp>
  <form type="paradigm" subtype="present">
    <form type="inflected">
      <orth>pierdo</orth>
      <gramGrp>
        <gram type="person">1</gram>
        <gram type="number">sg</gram>
        <gram type="mood">indicative</gram>
        <gram type="voice">active</gram>
      </gramGrp>
    </form>
   <!-- other inflected forms (of present indicative) here -->
    <gramGrp><gram type="tense">present</gram></gramGrp>
  </form>
  <form type="paradigm" subtype="preterite">
    <form type="inflected">
      <orth>perdí</orth>
      <gramGrp>
        <gram type="person">1</gram>
        <gram type="number">sg</gram>
        <gram type="mood">indicative</gram>
        <gram type="voice">active</gram>
      </gramGrp>
    </form>
    <gramGrp><gram type="tense">preterite</gram></gramGrp>
  </form>
...
</entry>
```

Example 5.

2.5 Representation of variants

The representation of variation within a form is highly dependent upon the specifics of what exactly varies, and how. As a general principle, variation may be encoded as

`form[@type="variant"]` and embedded within the parent element for which a subordinate feature exhibits variation. Variation within the form can occur with regard to the orthographic representation or the phonetic realization of a given form.

2.5.1 Orthographic Variation

Several kinds of orthographic variation may be distinguished. Below, we present some of the options with the corresponding examples.

The first example addresses spelling variation due to change in a language's orthography conventions.

```
<entry xml:id="Flussschifffahrt" xml:lang="de" type="compound">
  <form type="lemma">
     <orth>Flussschifffahrt</orth>
  <form type="variant">
     <orth>Fluss-Schifffahrt</orth>
  </form>
  <form type="variant">
     <orth notAfter="1996">Flußschifffahrt</orth>
     <usg type="time">Vor 1996 Rechtschreibung Reform</usg>
  </form>
  <gramGrp><gram type="pos">noun</gram></gramGrp>
....
</entry>
```

Example 6.

In the following example, the Hebrew word אֹמֶץ 'courage' can be represented by either the 'dotted' ('vowelized') spelling, or by the full spelling, where vowels are marked as separate characters.

```
<entry xml:id="courage-heb" xml:lang="heb">
  <form type="lemma">
    <form type="variant">
      <orth notation="menukad">אֹמֶץ</orth> <!-- 'dotted' spelling -->
     </form>
     <form type="variant">
      <orth notation="male">אומץ</orth> <!--full spelling -->
     </form>
     <pron notation="ipa">ˈomet͡s</pron>
  </form>
  <gramGrp><gram type="pos">noun</gram></gramGrp>
  <sense> .... </sense>
</entry>
```

Example 7.

Note that in Example 7, the phonetic representation is provided as well, according to the conventions of the International Phonetic Alphabet. The above encoding proposal might be opposed on the grounds of verbosity. However, TEI-Lex0 is primarily meant to be a derived representation format for the purpose of exchange or processing, and the primary stress is on explicitness. A project-internal representation might express

the variation simply by putting two `orth` elements next to one another, within a single `form`. In TEI-Lex0, the otherwise potentially spurious additional `form[@type="variant"]` is a matter of coherence and explicitness.

The next example illustrates a fragment of an American English dictionary in which, due to the lack of official conventions for transliteration of Arabic orthography to the English (Latin) script, the initial vowel in the surname 'Osama Bin Laden' varies between 'O' and 'U'.

```
<form type="lemma">
  <pron notation="ipa">
    <seg xml:id="ousma" corresp="#usma #osma">ow."sa.ma</seg>
    bɪnˈlaːdn̩</pron>
  <form type="variant">
    <orth type="transliterated">
      <seg xml:id="osma" corresp="#usma #ousma">Osama</seg>
      Bin Laden</orth>
  </form>
  <form type="variant">
    <orth type="transliterated">
      <seg xml:id="usma" corresp="#osma #ousma">Usama</seg>
      Bin Laden</orth>
  </form>
</form>
```

Example 8.

Note that the `seg` element is used for the purpose of providing an anchor for linking and at the same time it provides a place for the `@corresp` attribute, used to express the relevant correspondence.

2.5.2 Phonetic Variation

The example entry below contains a single orthographic form as well as phonetic transcriptions of the two roughly equally used variant pronunciations of the word 'caramel' in American English. Since all this information pertains to the lemma, it is contained within a single `form[@type="lemma"]` element.

```
<entry xml:id="caramel-en" xml:lang="en-US">
  <form type="lemma">
    <orth>caramel</orth>
    <form type="variant">
      <pron notation="ipa">'keɹə"mɛl</pron>
    </form>
    <form type="variant">
      <pron notation="ipa">'kaɹmɬ̩</pron>
    </form>
  </form>
  <gramGrp><gram type="pos">noun</gram></gramGrp>
...
</entry>
```

Example 9.

2.5.3 Regional and Dialectal Variation

In the following example from Mixtepec-Mixtec, there is variation in the form of the word for the city of Oaxaca between speakers from the village of Yucanany and the rest of the speakers. Since the Yucanany variety makes up only a small portion of the speakers of the language, this case of variation is represented as an embedded `form[@type="variant"]` within the lemma. Note the use of `usg[@type="geo"]/placeName` to explicitly specify this feature in addition to the use of the private language subtag "mix-x-YCNY" as per BCP 47 (Phillips and Davis, 2009).

```
<entry xml:id="Oaxaca-MIX" xml:lang="mix" type="compound">
  <form type="lemma">
    <orth>Ñuu Ntua</orth>
    <pron notation="ipa">ɲùùndùá</pron>
    <form type="variant" xml:lang="mix-x-YCNY">
      <orth>Ntua</orth>
      <pron notation="ipa">ndùá</pron>
      <usg type="geo">
        <placeName>Yucanany</placeName>
      </usg>
    </form>
  </form>
  <gramGrp>
    <gram type="pos">locationNoun</gram>
  </gramGrp>
...
</entry>
```

Example 10.

3. Summary

TEI-Lex0 focuses on staking a certain consistent path across the variety of choices offered by the TEI Guidelines, with an eye to establishing recommendations for a baseline encoding of the products of retro-digitization and at the same time a certain pivot format that may be further uniformly processed and queried. In this paper, we concentrated on presenting a glimpse of the TEI-Lex0 effort pertaining to encoding information on the parts of entries that specify formal and grammatical features.

We have adduced examples of how orthographic and phonetic variants can be handled, and looked at the representation of inflectional paradigms, which have not received much attention in the TEI Guidelines but which are nonetheless essential for properly exposing data content to the various uses that digitized lexica can have.

4. Acknowledgements

The results presented here have been formulated in the course of work of the “Berlin task force” of the ENeL-DARIAH-PARTHENOS TEI-Lex0 initiative. We are thankful to our colleagues for extensive discussions and brainstorming over the past months. We also stress that the ideas presented here are coloured by our subjective views, and that

TEI-Lex0 is not yet a finished set of coherent guidelines. Some differences between this paper and the final TEI-Lex0 deliverable may be expected, and the blame for any deviations from the eventual specification is solely ours.

5. References

Erjavec, T., Evans, R., Ide, N., Kilgarriff, A. (2000). The CONCEDE Model for Lexical Databases. *Proceedings of the Second Language Resources and Evaluation Conference (LREC)*, Athens, Greece, pp. 355-62. Available at: http://www.lrec-conf.org/proceedings/lrec2000/html/summary/335.htm

Ide, N., Kilgarriff, A., Romary, L. (2000). A Formal Model of Dictionary Structure and Content. *Proceedings of Euralex 2000*, Stuttgart, pp. 113-126. Available at: https://www.kilgarriff.co.uk/Publications/2000-IdeKilgRomary-Euralex.pdf

Phillips, A. and M. Davis (eds). 2009. Tags for Identyfying Languages. BCP 47, RFC 5646. IETF. Available at: https://tools.ietf.org/html/bcp47

TEI Consortium, eds. (2013). *Getting Started with P5 ODDs*. Available at http://www.tei-c.org/Guidelines/Customization/odds.xml

TEI Consortium, eds. (2016). *TEI P5: Guidelines for Electronic Text Encoding and Interchange. [Version 3.1.0]*. [Last updated on 15th December 2016]. TEI Consortium. Available at: http://www.tei-c.org/Guidelines/P5/ ([accessed on 13th February 2017]).

Making 1:N Explorable: a Search Interface for the ZAS Database of Clause-Embedding Predicates

Peter Meyer[1], **Thomas McFadden**[2]

[1]Institut für Deutsche Sprache, R 5, 6-13 D-68161 Mannheim
[2] Leibniz-Zentrum Allgemeine Sprachwissenschaft, Schützenstr. 18, D-10117 Berlin
E-mail: meyer@ids-mannheim.de, mcfadden@leibniz-zas.de

Abstract

We introduce a recently published corpus-based database of German clause-embedding predicates and present an innovative web application for exploring it. The application displays the predicates and the corpus examples for these predicates in two separate tables that can be browsed and searched in real time. While familiar web interface paradigms make it easy for users to get started, the data presentation and the interactive advanced search components for the two tables are designed to accomodate remarkably complex query needs without the need for resorting to a dedicated query language or a more specialized tool. The 1:n relationship between predicates and their examples is exploited in the two tables in that, e.g. the predicate table also shows, for each predicate and each example attribute, all values that occur in the examples for this predicate. An easy-to-use visual query builder for arbitrary boolean combinations of search criteria can optionally be displayed to pre-filter the underlying data presented in both tables. Several options for altering quantifier scope can be activated with simple checkboxes and considerably widen the space of searchable constellations.

Keywords: user interface; lexical data; query building; relational database

1. Introduction

This paper discusses the conceptual underpinnings of a web application user interface for exploring a recently published multilingual corpus-based database of clause-embedding predicates (Stiebels et al., 2017: http://www.owid.de/plus/zasembed2017/main). The relational database represents two basic types of entities that stand in a 1:n relationship: predicates (mostly verbs) and the corpus examples selected for a given predicate. Each of the two types is annotated with its own set of attribute-value pairs (henceforth, example properties vs. predicate properties). In each set, some attributes only apply to a certain subset (e.g. the definiteness attribute only applies to examples with embedded nominalization).

Despite the simplicity of a 1:n relationship, different quantifier/negation scope constellations can give rise to remarkably complex search semantics. In view of this complexity, the user interface was designed with the following goals: to present the user with a simple initial tabular view of the data that has straightforward and easy-to-use real-time filtering and sorting options, in line with accepted search interface design principles (Hearst, 2009; Morville & Callender, 2010; Russell-Rose & Tate, 2012), and to empower advanced users to incrementally construct ever more complex queries without resorting to domain-specific or generic query languages or introducing a separate "advanced search" interface layer.

Section 2 gives a short overview of the linguistic background and history of the ZAS database and introduces its basic data structure. In Section 3, we discuss the requirements that this structure and the target audience impose on a sufficiently elaborate search tool and how this affects the general objectives of the web application. The design decisions that were made to meet these requirements are presented in-depth in Section 4, and we briefly sketch the front-end and back-end technology in Section 5. The closing Section 6 points out a number of limitations of the tool in its current state, discussing some alternatives and competing approaches.

2. The ZAS database of clause-embedding predicates

2.1 Background on the database

The ZAS database of clause-embedding predicates documents how lexical predicates interact with clausal complementation. The examples in (1) give a simple demonstration from English of the kinds of patterns that are of interest.

(1) a. Max believes/knows/hopes [that Sarah works there].
b. Max *believes/knows/*hopes [whether Sarah works there].
c. Max *believes/*knows/hopes [to work there]

While a finite declarative clause is possible as the complement of *believe*, *know* or *hope*, only *know* can introduce finite interrogatives, and only *hope* can take control infinitives. It is thus well known that the properties of specific lexical predicates are important for understanding clausal embedding. In much of the literature, the discussion of complementation types and their licensing has focused on a relatively small number of predicates, taken to be representative of large classes with the same behavior (e.g. *believe*-class vs. *try*-class vs. *want*-class verbs in the discussion of English infinitives). However, it is clear that this oversimplifies matters and fails to capture interesting variation and crucial differences in how specific predicates interact with their syntactic environment. An important demonstration of this point can be found in Levin (1993), which examines in detail the range of distinct classes that can be identified for English verbs based on the structural alternations they engage in.

The ZAS database grew out of the conviction that a similar level of attention to detail is necessary to understand clausal complementation and what properties of lexical predicates are relevant for the behavior of their complements (see also Stiebels, 2011: for more detailed discussion). The methodology chosen to tackle this problem was to build a research tool around an extensive collection of data, illustrating the types of embedded clauses found with a large number of lexical predicates. Whereas a common prior strategy has been to *assume* predicate classes based on external semantic criteria, and then to investigate their behavior, we wanted to make it possible to *identify* classes of predicates based on the properties of the clauses they embed. We would thus collect, for each predicate included, a series of examples of different types of embedding, annotating each for a number of relevant grammatical properties, with the goal of illustrating all of the grammatical embedding possibilities for each predicate.

The database was conceived and initiated by Barbara Stiebels and gradually built up and extended by a team of researchers and student assistants at the Leibniz-Zentrum Allgemeine Sprachwissenschaft (ZAS) in Berlin (http://www.zas-berlin.de) from 2003 onwards. After her 2012 departure for the University of Leipzig, coordination of the project was taken over by Thomas McFadden in 2014. Through most of this period, the focus has been on contemporary German, though significant data have been collected for a number of other languages and older stages of German. It was decided early on that the data should not come from invented examples, but should be naturally occurring sentences extracted from corpora. The two most important corpus sources for the contemporary German portion are the *Digitales Wörterbuch der Deutschen Sprache* (DWDS; http://www.dwds.de), and the *Deutsches Referenzkorpus* (DeReKO; http://

www1.ids-mannheim.de/kl/projekte/korpora/). The embedding types systematically collected are infinitival (2-a) and nominalized (2-b) complements, verb-final finite declarative (2-c) and interrogative (2-d) complements (both polar and *wh*-questions), and embedded verb-second clauses (2-e); coverage of parentheticals and direct speech complements is ongoing work.

(2) a. Der Vorsitzende befahl, den Zeugen aufzurufen
the chair ordered the witness to-call
'The chair ordered the witness to be called' (ZDB 1565: DWDS K-Ge 1910)

b. Sie müssen sich mit dem Verkauf der Wohnungen beeilen.
the must themselves with the sale of-the apartments hurry
'They need to rush the sale of the apartments.' (ZDB 1523: DWDS BZ 1995)

c. Wir machen ab, daß er mich um acht Uhr besucht.
we make off that he me at 8 o'clock visits
'We agree that he will meet me at 8 o'clock.' (ZDB 70: DWDS K-Be 1980)

d. Gib acht, was du dir wünschst!
give attention what you yourself wish
'Be careful what you wish for.' (ZDB 218: IDS wpd 2011)

e. Aber ich ahne, es wird nicht mehr als Blech.
but I suspect it becomes not more than sheet-metal
'But I can tell it's just going to be nonsense.' (ZDB 256: IDS brz 2006)

The idea is that every predicate included should be checked in all relevant meaning variants and valency patterns, with a series of properties relevant for specific complementation types checked systematically. For example, with predicates that can embed finite verb-second clauses like (2-e), we checked for both indicative and subjunctive examples, and with predicates that can embed control infinitives like (2-a), we searched for examples with different types of control. Every example was then coded for these and several additional relevant properties. The guiding principle is that only "surfacy" features should be coded in order to keep the annotation theory-neutral and operationalizable.

In mid-2014, a collaboration was initiated with Carolin Müller-Spitzer and Peter Meyer of the Institut für Deutsche Sprache (IDS) in Mannheim, with the goal of making a version of the database publicly available on the OWID^{plus} platform for lexical-lexicographic data and resources (http://www.owid.de/plus/). A new search interface designed specifically for the ZAS database was then developed by Meyer in consultation with the ZAS team. The current public beta release of the database, which is the focus of this paper, contains only the data from the contemporary German part of the database. Additions are planned, however, for future releases of data on other languages and the historical stages of German.

2.2 The structure of the database and the 1:n relationship

The ZAS-internal version of the database is implemented in MySQL, with an in-house interface written in PHP for entering, editing and searching in the data. It is built primarily around two sets of data and the connections between them: a table of predicates and a table of corpus examples. Each example is associated with one predicate — it demonstrates one particular embedding use of that predicate. The two tables consist of a series

of entries, each of which contains several pieces of information on a single predicate or example. An entry in the predicate table contains the (infinitival) form of the predicate itself, as well as information about its syntactic category and morphological make-up. An entry in the example table contains far more information, with values for up to a dozen properties. This includes the text of the example itself, an indication of the argument structure and realization of the matrix clause, finiteness and word-order properties of the embedded clause, what complementizer it is introduced by (if any), whether it is an interrogative, as well as information about the corpus source and a link to the entry for the predicate. As for the size, the contemporary German version currently contains data on over 1700 distinct lexical predicates, exemplified through nearly 17,000 naturally occurring corpus examples.

While there is some additional complexity in the internal implementation (e.g. to deal with multiple languages and allow for easy extensibility), the system's primary goal is to provide information about examples and predicates, and the public version of the database and the OWID^{plus} search interface are built around this idea. At any given time, the interface displays either an example table view or a predicate table view, and every search query is ultimately interpreted as a search for either predicates or examples fitting certain criteria. At a basic level, this is fairly straightforward, but there are some cases where the interactions between predicate properties and example properties can lead to significant complexity. This arises from the fact that, while each example is tied to exactly one predicate, a given predicate will normally be associated with several examples. Dealing with this 1:n relationship presents interesting challenges for the design of the search interface, and thus will be extensively discussed throughout this paper. To set the stage, it will be useful to go into a bit more detail here about how examples, predicates, and their respective properties interact in the structure of the database itself.

The database is designed to enable sophisticated searches in order to obtain lists of predicates with complex combinations of properties. But it is the examples that constitute the bulk of the data, collected and coded for research use in the database, and what primarily interests us about the predicates is what kinds of embedded clauses are found in the examples associated with them. The information about these clauses is recorded in example properties, and thus to a large extent we search for predicates not by specifying their own properties, but those of the examples they embed. For example, we might want to search for all predicates that can embed subjunctive verb-second clauses, but no infinitives. Of course, it works the other way around as well. When searching for examples, some of the properties we might be interested will actually be properties of the predicate. We could e.g. search for all examples where the embedding predicate is a denominal verb in the hopes of finding out whether this correlates with the control status of embedded infinitives.

The crucial thing is that, despite this parallel, example properties and predicate properties are logically distinct in the way they function, and this is precisely because of the 1:n relationship. Let us consider the predicate properties first, because their status is simpler. In a search for predicates, constraints on predicate properties filter the results in an obvious way. If we specify the value Pt-V for the predicate property "morphology", the search will only return predicates that are morphologically particle verbs. The handling of predicate properties is just as simple in a search for examples, because for every example, there is exactly one predicate. We can thus treat predicate properties as though they

were example properties: we can search for *examples* with the value Pt-V for the property "morphology" (in Section 4.3 we will introduce the term *derived example criterion* for this concept), and rather than a list of all predicates that are morphologically particle verbs, we will get a list of all examples in which the predicate is morphologically a particle verb.

This simple situation does not hold with example properties. Matters are straightforward in an example search, because constraints on example properties simply filter examples. If we specify the value KONJ I for the example property "verb mood", the search will return only examples in which the embedded clause is finite and in the subjunctive I mood. But in a predicate search, the logic of 1:n makes things much more interesting. Because each predicate is associated with many examples, we cannot simply translate an example property into an implicit predicate property. We cannot say "return all predicates with the value KONJ I for the example property "verb mood"", because there is no unique example associated with each predicate. Rather, the relationship between predicates and example properties is more complex and has to be made clear in the search. The basic idea is that you are searching for predicates which are associated with at least one example that is characterized by the example property. We can rephrase this in the terms of the example above as "return all predicates which appear in at least one example which has the value KONJ I for the example property "verb mood"". By itself, this step is not particularly difficult, but it raises interesting questions when it comes to building complex queries involving more than one property.

Imagine now that we are interested in both subjunctives and embedded verb second — two properties that have often been thought to go together in German. There are two different ways to understand a search for predicates that can embed subjunctive clauses and verb-second clauses. A simple conjunction of two queries might return all predicates that have at least one subjunctive I example and at least one verb-second example. In this case, subjunctive I and verb second are independent, and because each predicate is associated with multiple examples, they may both apply to the same example, like (3-a), or they may apply to distinct examples associated with a single predicate, like (3-b) which is subjunctive I and (3-c) which is verb second.

(3) a. Er droht an, er werde nun jemanden befragen.
he threatens he will.SBJI now someone question
'He's threatening that he'll question someone now.' (ZDB 356: DWDS K-Be 1999)

b. Man nahm an, daß Leben ohne Licht unmöglich sei.
one took on that life without light impossible be.SBJI
'It was assumed that life was impossible without light.' (ZDB 624: DWDS TS 1999)

c. Zdenka hat sich ihrerseits in Matteo verliebt und schreibt ihm die
Zdenka has herself her-side in Metteo loved and writes him the
Liebesbriefe, von denen er annimmt, sie stammen von Arabella.
love-letters from which he assumes they come from Arabella
'Zdenka for her part has fallen in love with Matteo and writes him love letters, which he assumes come from Arabella.' (ZDB 629: DWDS K-Wi 1998)

This may indeed be what we want. But a different possibility is that we are interested in predicates that can embed a subjunctive I verb-second clause, i.e. we want single examples like (3-a) in which both properties hold. A task for our tools is thus to make both of these logical combinations of multiple example properties in a single predicate search possible in a way that is as easy as possible to understand. We will discuss the way the interface does this in Section 4.3.

3. Requirements for the UI

The new user interface for the published version of the ZAS database on the OWID^{plus} platform has to meet two broad requirements which we will discuss in turn. First, it has to provide facilities for formulating queries that can take full advantage of the range of data stored in the database and the connections between them. Second, it must be useable, at least at a basic level, for researchers who are interested in the behavior of clause-embedding predicates but have limited experience with databases and sophisticated search tools.

3.1 Semantic requirements for possible searches

The minimum capabilities necessary for the interface to actually reflect the structure of the underlying database are to allow search queries for both examples and predicates, where both types of query can refer to both example criteria and predicate criteria. To really exploit the full capabilities of the database, the interface should additionally provide the means to build complex queries combining multiple example and predicate criteria. The simplest form of this would be to allow the conjunction of criteria, interpreted so that the results returned by a search would be the examples or predicates simultaneously meeting all of the criteria specified. We discuss how the interface manages this in sections 4.1 and 4.2. The 1:n relationship means, however, that even this simple conjunction can potentially involve distinct semantics when a search for predicates involves more than one example criterion. One could design an interface that allows searches with all arbitrary boolean combinations of the different types of criteria, but much of this complexity is unlikely to be particularly useful for the study of lexical effects on clausal embedding, and certain types of combinations are more likely to lead to searches that are difficult to interpret than to allow the posing of typical research questions. We discuss the relevant trade-offs and the design decisions made in Section 4.3.

3.2 UI design and usability requirements

The original ZAS-internal interface was designed for team members working *on* the database. It thus includes facilities for entering and editing data in addition to running searches, and it works on the assumption that users are well acquainted with the structure and workings of the database. The new interface for the public version, however, is intended purely as a way to search and explore the database, for a wide audience of users, including novices. Thus the following considerations guided the design process:

- The interface should be explorable and discoverable for users with various degrees of prior experience.
- It should present an intuitive view of the data, making use of established interface design metaphors familiar from other web applications so that users can easily understand what they are looking at and how they can manipulate it.

- The view should be updated in real time whenever the user takes any action, so that they get immediate feedback and can quickly explore the consequences of different types of input.
- Running a basic query for an example or predicate satisfying some criterion should be extremely easy.
- Running complex queries involving boolean combinations of several criteria, while differentiating the various semantics relating to the 1:n relationship, should be possible.
- Ideally, it should be possible to get from the simplest search to the most complex by adding pieces step-by-step, where each intermediate step is a valid query, so that users can build their way up from novice to expert usage.
- The system should be equipped with extensive documentation, with facilities for accessing relevant parts directly from specific bits of the interface.

4. Design of the UI

4.1 Central concepts and basic search

The design of the user interface reflects the fact that the database is structured around two tables. At any given time, a version of either the example table or the predicate table is presented. The rows correspond to distinct entries — predicates or examples — and the columns display the properties associated with each entry. Every type of user input operates on one of these two views, with the results shown by updating the view in real time. Entering and modifying search queries manipulates restrictions on the entries displayed in the table. Thus there is no dedicated "query entry view" or "search results view", but a single view combining both, allowing users to immediately see the effects of the search criteria they enter and to modify them on the fly in order to test out and craft precise queries.

The two tables are identical in how they work and respond to input and boast the same system of integrated documentation — a detailed User Guide with ⓘ markers adjacent to the various interface elements that link directly to the relevant section of the Guide. The interface also provides facilities for exporting the data currently displayed in either table for local storage and processing. This function is disabled in the current public beta but will be activated in future releases. Both tables also allow for an "advanced search" which adds a more sophisticated query builder to the usual table. Note, however, that the advanced search is not an alternative to the basic table views, but rather an extension. The full functionality of the basic example and predicate tables is still available and works in the same way, just with additional possibilities for filtering the data. This will be described in detail in Section 4.2.

Switching between the example and predicate tables of course radically alters the way in which the data are presented. Even still, in most cases this change does not actually affect which data are presented, only the perspective from which they are viewed. This is because both table views combine information on both examples and predicates and display them in a single table, so that in general the same data can be presented either way — we say that the two tables are 'in sync'. The example table also contains predicate properties, since each example is associated with a predicate, and the predicate table also contains example properties, since each predicate is exemplified by a series of examples. There are, however, circumstances in which the tables can go out of sync, in particular

when searching for predicates based on example properties that can in principle apply to more than one example. This will be discussed in Section 4.3.

The main properties of the basic table views can be seen in Figure 1, with the relevant numbered features explained below.

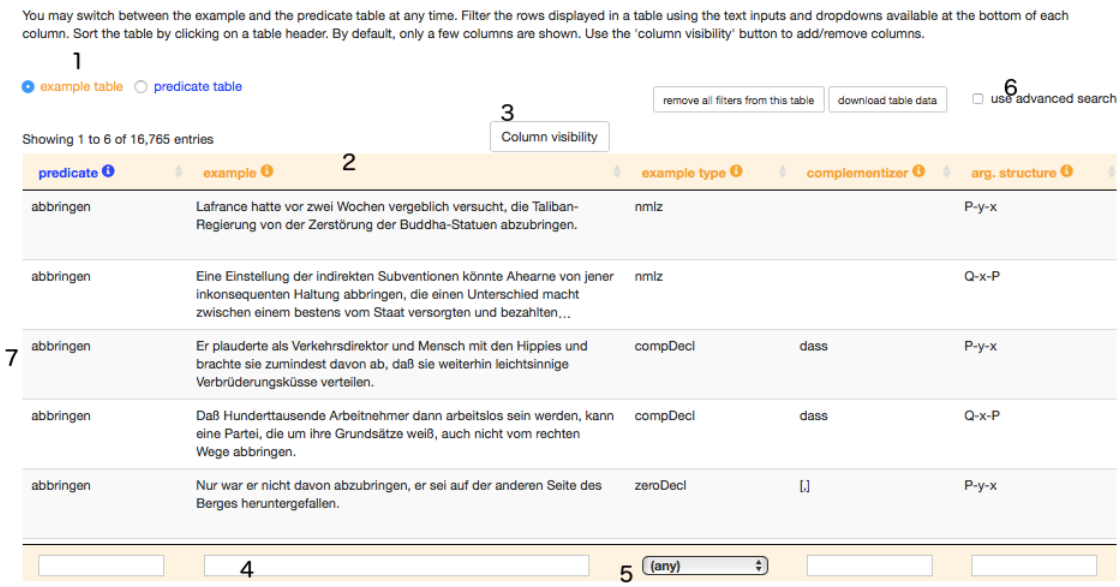


Figure 1: Basic Search

1 Selection of either the example or the predicate table view
2 Headers showing the properties currently displayed, blue for predicate properties, orange for example properties. Clicking on a property sorts the table by its values.
3 Facilities for specifying which properties are displayed as columns
4 Text box for entering a string that should be contained in the value for the relevant property, with autosuggest functionality and regular expressions
5 Pull down, for properties with a small number of permissible values
6 Click here to build an advanced search
7 List of entries in the table that match the current search, updated in real time. Double-clicking a row brings up complete information on its entry.

4.2 Advanced searches with the query builder

The per-column filtering options of the two tables are good for simple, quick and intuitive searches, but they are restricted in the following ways:

- For each example or predicate property, at most one search criterion may be formulated.
- The search criteria cannot be negated.
- The only way the table filtering criteria can be combined is by logical conjunction, such that all criteria must be fulfilled at the same time.

The interface's advanced search option provides an additional layer of search functionality that eliminates these restrictions. The advanced search is not a separate mode of access, i.e. it does not replace the interactive and explorative data presentation in the two tables, but complements it by letting the user restrict the underlying data set that is presented. To be more precise, both tables present the search result for the advanced query, albeit from two different perspectives. The data may then undergo filtering and sorting in a table, which amounts to the logical conjunction of the advanced search criteria and the filtering criteria defined in a specific table. In a sense, the advanced query acts as an additional "super-filter" on both tables.

Advanced search can be activated and deactivated at any time by a simple mouse click. When activated, the advanced query builder is shown above the table. As with the standard table filters, any change a user makes to the advanced query is immediately reflected in both tables. The query builder component itself follows the design of search components in modern operating systems, such as the advanced search offered in the default file manager "Finder" on Apple computers or in the query builder integrated in Microsoft Outlook. It allows the user to formulate an arbitrary number of criteria, even multiple criteria concerning the same property. To this end, the user may add any number of *criterion selectors* using the "+" button. Each criterion selector offers all types of search criteria also available as table filters. Criterion selectors for example properties have the orange background of the example table, while those for predicate properties have the blue one of the predicate table.

Arbitrary boolean combinations are supported as follows: All criterion selectors have an additional drop-down menu for optional negation; in addition, there is a special type of search criterion called "group of conditions" that opens up a *subgroup* of search criterion selectors connected by possibly negated conjunction or disjunction (logical "or", "and", "nor" or "nand"), yielding four types of logical connectivity: "all/none/at least one/not all subgroup criteria is/are true". Subgroups may be nested inside subgroups to any depth. The top-level criterion selectors of the advanced search can be thought of as implicitly contained in a conjunction group. Figure 2 shows a complex advanced search with nested subgroups, yielding examples with a predicate containing the string "sag" that embed an infinitive clause or a subordinate clause with both a complementizer containing "d" and a verb in the subjunctive I.

The table-specific filtering options and the advanced search system show intentional overlap with regard to both functionality and design. Input widgets in tables and in advanced search work exactly the same way. When working with a specific table, the user can freely choose between adding a filter to the table or adding the same condition as an advanced constraint on the top-level of the advanced query builder.

The advanced settings for search semantics, to which we now turn, are somewhat different.

4.3 Advanced search semantics

As discussed above, the 1:n relation between predicates and examples gives rise to potentially complex search questions beyond mere boolean combinations of criteria. The user interface introduced in this paper offers a principled approach to altering the semantics of queries through three user-defined settings. A deeper understanding of these settings

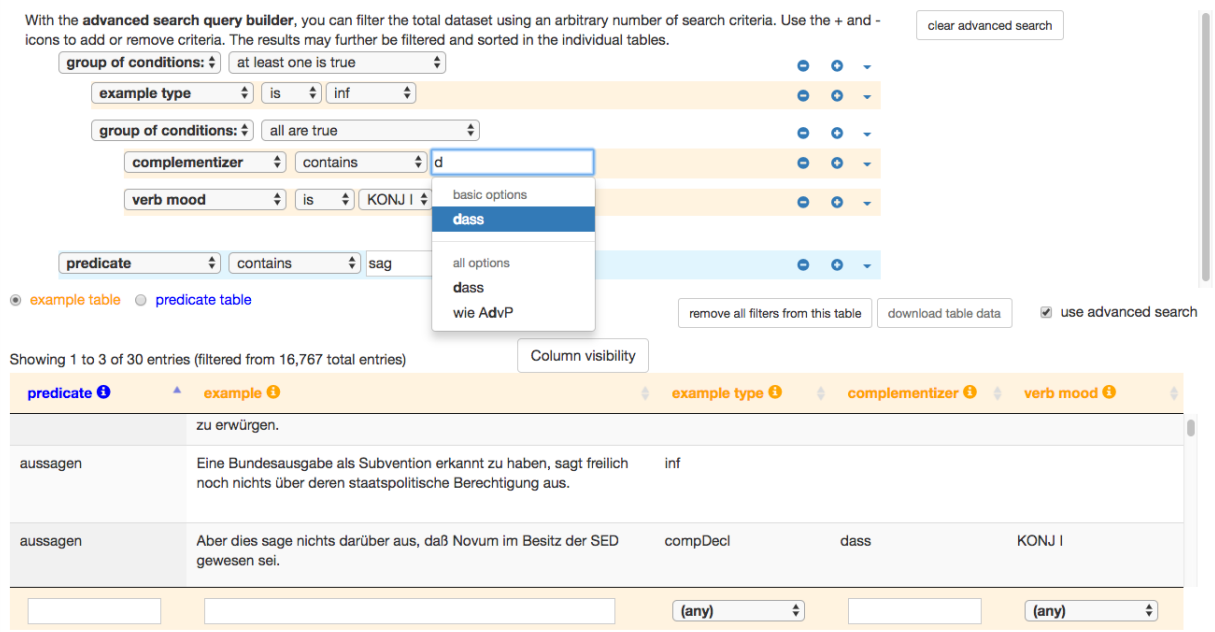


Figure 2: Building an advanced query

requires looking into how aspects of scope and negation interact in complex queries. A tutorial-style and hands-on introduction to these advanced aspects can be found in the online User's Guide. In what follows, we approach the subject from a formal perspective.

Let $\mathbb{E}$ denote the set of examples and $\mathbb{P}$ the set of predicates. Different example criteria will be denoted by E_1, E_2, etc., different predicate criteria by P_1, P_2, etc. An example criterion in the narrow sense (henceforth, *basic example criterion*) can only be applied to examples and puts a user-defined constraint on a specific example property; if the criterion E_1 actually applies to example $e \in \mathbb{E}$, this will be written in predicate-logic fashion as $E_1(e)$. A basic example criterion corresponds to an example property filter in the example table. Correspondingly, a predicate criterion in the narrow sense (henceforth, *basic predicate criterion*) can only be applied to predicates and puts a user-defined constraint on a specific predicate property; if the criterion P_1 actually applies to predicate $p \in \mathbb{P}$, this will be written as $P_1(p)$. A basic predicate criterion corresponds to a predicate property filter in the example table. Almost all search criteria can be used in negated form; we will use the superscript bar, as in $\overline{X}$, to signal user-defined negation of a criterion X, such that $\overline{X}(x) \Leftrightarrow \neg X(x)$. We write $pred(e)$ to denote the predicate exemplified by example e. To make the formulas a bit shorter and more legible, the letters e and p will always be used in such a way that $e \in \mathbb{E}$ and $p \in \mathbb{P}$ are implied.

In what follows, we assume that a search query is formulated in the advanced search query builder and we pose the question of how exactly these criteria define a search result set with respect to the two tables.

Queries concerning the example table are the easier part of the picture. For each predicate criterion P_j we define a *derived* example criterion E^{P_j} such that $E^{P_j}(e)$ is true iff $P_j(pred(e))$. If P_j, for instance, means "is a verb", then E^{P_j} stands for "is an example whose predicate is a verb". A derived example criterion corresponds to a predicate property filter in the example table. A single search criterion (regardless of whether we are dealing with table filters or with advanced query criteria) as applied to the exam-

ple table is either a basic or a derived example criterion. Since the example table shows the result of applying the conjunction of table filters and advanced search criteria to the entire database data set, searching the example table always means applying a boolean combination of basic and derived example criteria. To simplify formal exposition, we will stick to a sample advanced query consisting of a conjunction of two basic example criteria E_1 and E_2 and one basic predicate criterion P_3; extension to the general case is straightforward. When applied to the example table, an example e is shown in the table if and only if $E_1(e) \wedge E_2(e) \wedge E^{P_3}(e)$, which is equivalent to formula (1):

$$E_1(e) \wedge E_2(e) \wedge P_3(pred(e)) \tag{1}$$

Figure 3 shows a concrete case of this kind of query on the example table, in a search for examples with a particle verb (Pt-V) [criterion P] that embed an interrogative clause [criterion E_1] with a verb in subjunctive I mood [criterion E_2].

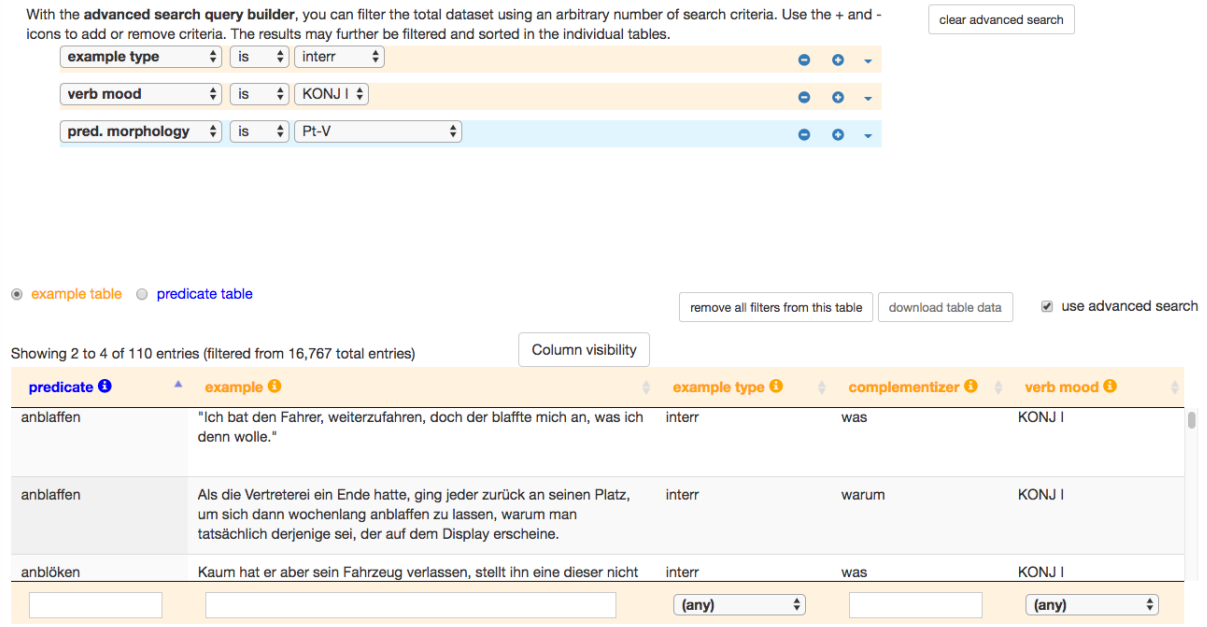


Figure 3: Sample advanced query with results in the example table

At this point, we will briefly discuss the semantics of example criteria derived from *negated* predicate criteria. The relevance of this interlude will emerge later. It is easy to see that $\overline{E^{P_j}}(e) \Leftrightarrow E^{\overline{P_j}}(e)$. In our previous example case, $\overline{P_j}$ would mean "is not a verb"; correspondingly, $E^{\overline{P_j}}$ represents the property "is an example whose predicate is *not* a verb", which is trivially coextensive with $\overline{E^{P_j}}$ "is *not* an example whose predicate is a verb". In other words, it is not necessary to separately define derived example criteria for negated predicate criteria; one can always negate the derived criterion instead.

Let us now turn to the way searches apply to the predicate table. When our sample advanced query with two basic example criteria E_1 and E_2 and one basic predicate criterion P_3 is applied to the predicate table, then, with default settings, a predicate p is shown in the table if and only if the basic predicate criterion applies to p and there is at least one example e for predicate p such that *both* example criteria apply to e. Formally, it

is required that $\exists e\,(p = pred(e) \wedge E_1\,(e) \wedge E_2\,(e)) \wedge P_3\,(p)$, which is equivalent to formula (2):

$$\exists e\,(p = pred(e) \wedge E_1\,(e) \wedge E_2\,(e) \wedge P_3\,(pred\,(e))) \tag{2}$$

In other words, the default semantics for searches in the predicate table requires that all basic example criteria be met simultaneously by (at least) one example e for p. The reason why the default settings for predicate table searches are defined like this becomes apparent from a comparison of formulas (1) and (2). It is easy to see that, with our sample search, a predicate p will appear in the predicate table if and only if at least one example for p appears in the example table. We say that the two tables are *in sync* in this case; this is a formalization of the intuitive notion, mentioned earlier, that both tables represent the same underlying set of data. This is, of course, also the ultimate reason why it is legitimate to have one advanced search applying to two different tables. Figure 4 shows the results of applying the sample query of 3 on the predicate table, in a search for particle verb (Pt-V) predicates [criterion P] for which there is at least one example that embeds an interrogative clause [criterion E_1] with a verb in subjunctive I mood [criterion E_2]

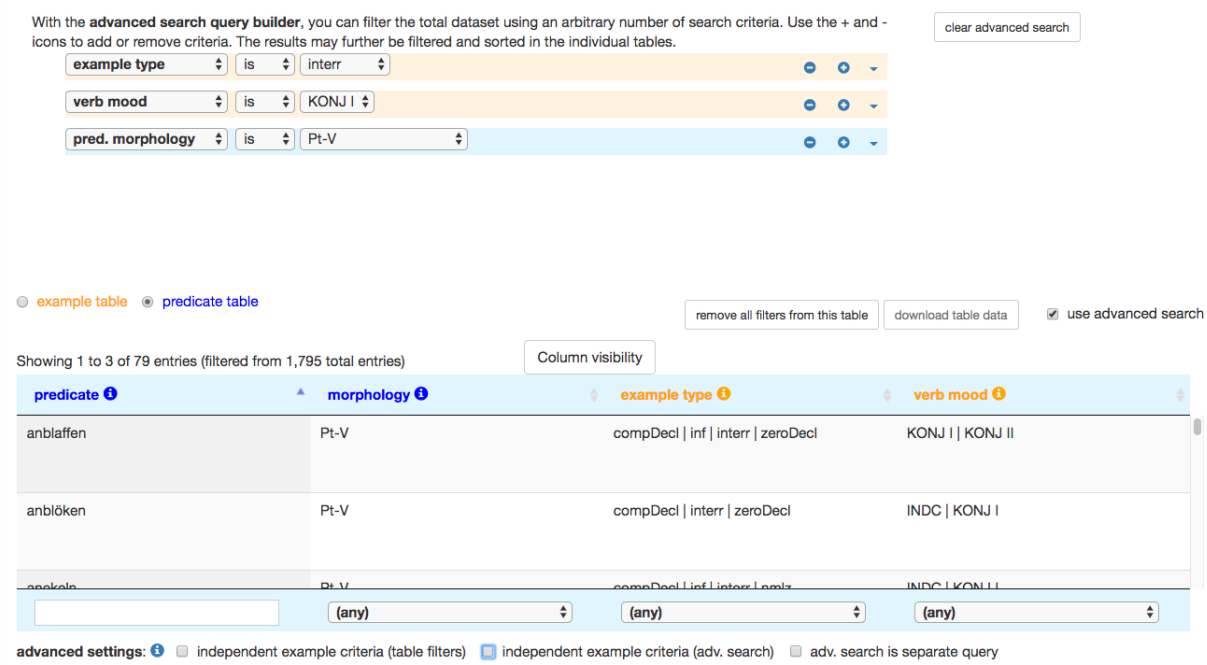


Figure 4: Sample advanced query with results in the predicate table

In many cases, the default semantics for the predicate table is not sufficient to meet users' needs. In our sample advanced query, a user might be interested in seeing all predicates p fulfilling criterion P_3 for which there is

- at least one example e_1 fulfilling criterion E_1 and
- at least one example e_2 (possibly, but not necessarily identical to e_1) fulfilling criterion E_2.

To handle this case, the user can specify what we (for want of a better term) call *independent example semantics* for the advanced query builder by ticking the "independent example criteria (adv. search)" checkbox appearing under the predicate table. With this semantics turned on, the search logic for the predicate table re-interprets the basic example criteria as *derived predicate criteria*. Clearly, deriving predicate criteria from example criteria has to be done differently than the other way around. We define, for each basic example criterion E_i, a *derived* predicate criterion P^{E_i} that holds of a predicate p iff $\exists e\,(p = pred\,(e) \wedge E_i\,(e))$. If E_i, for example, means "has an embedded infinitive clause", then P^{E_i} stands for "is a predicate with *at least one* example that has an embedded infinitive clause". With our sample query and independent example semantics turned on, a predicate p appears in the predicate table if and only if (3) holds:

$$P^{E_1}\,(p) \wedge P^{E_2}\,(p) \wedge P_3\,(p) \tag{3}$$

In formula (3), the two derived predicate criteria induce, by definition, two separate existential quantifications over the set $\mathbb{E}$ of examples, whereas the standard query semantics of (2) puts both example criteria in the scope of one existential quantifier. Figure 5 shows how the sample query of 3 is applied to the predicate table, this time with "independent example semantics", returning a list of particle verb (Pt-V) predicates [criterion P] for which there is at least one example that embeds an interrogative clause [criterion E_1] and at least one example with a verb in subjunctive I mood in the embedded clause [criterion E_2].

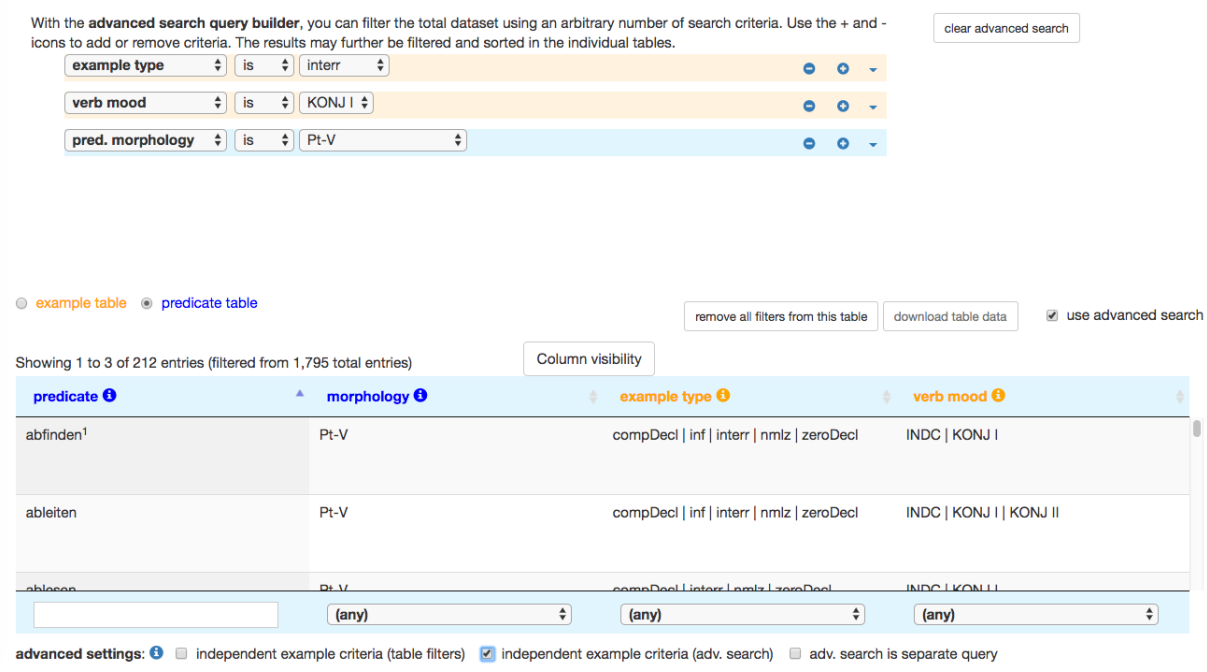


Figure 5: Sample advanced query with "independent example semantics"

The behavior of negation in derived predicate criteria is more complicated than with example criteria. It is easy to prove that $\overline{P^{E_i}}\,(p) \not\Leftrightarrow P^{\overline{E_i}}\,(p)$. In our previous example, $\overline{E_i}$ would mean "does not have an embedded infinitive clause"; correspondingly, $P^{\overline{E_i}}$ represents the property "is a predicate with at least one example *not* embedding an infinitive clause",

which is obviously not the same as $\overline{P^{E_i}}$ "is *not* a predicate with at least one example embedding an infinitive clause". This implies that P^{E_i}, $P^{\overline{E_i}}$, $\overline{P^{E_i}}$ and $\overline{P^{\overline{E_i}}}$ are, in general, four different criteria, because we have two logically different levels of negation. As far as the user interface is concerned, this implies that, for every derived predicate search criterion, two separate negation options would be needed to cover all possible cases. We decided to only offer one of these negation options: negating an example criterion always means negating the predicate criterion derived from it, as this seems to be the more intuitive and linguistically more relevant choice. In particular, it makes the formulation of queries such as the one in Figure 6 more plausible: With independent example semantics, this

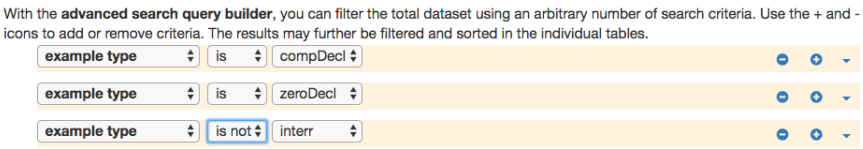


Figure 6: Advanced query with three example criteria, one of which is negated

query makes the system look for predicates whose examples exhibit the example types **compDecl** and **zeroDecl**, but not **interr**. If the negation on the third criterion were to be interpreted as pertaining to the underlying example criterion, then the query would read as follows: "Look for predicates that have at least one example with example type **compDecl**, at least one example with example type **zeroDecl**, and at least one example where the example type is not **interr**." Obviously, the third criterion would be redundant in this interpretation. Overall, the design of the system ensures that the conjunction of a criterion and its negation always yields an empty result set.

A major complication with independent example semantics is the fact that it puts the two tables "out of sync"; that is, they do not represent answers to the same query anymore. The query of Figure 6 produces zero results in the example table since no single example can fulfill all three conditions at the same time.

Independent example semantics can also be chosen for predicate table filters with the "independent example criteria (table filters)" checkbox, such that this semantics can be turned on and off separately for the two search components of the interface. If "independent example semantics" is activated neither for the advanced query builder nor for the table filters, all user-defined example criteria of both search components are, by default, in the scope of the existential quantifier of formula (2). This can be changed through a third checkbox "adv. search is separate query". If this setting is activated, the two components (criteria in the table filters vs. in the advanced builder) are treated separately and generate two separate searches according to formula (2). The two formulas are then joined with logical AND, returning the intersection of the two result sets. This is useful if a user looks for examples for predicates that have examples with multiple example properties A, B, C, ... and (possibly different) examples with multiple example properties D, E, F, ... Figure 7 shows the "separate query" setting in a query for predicates that can embed subjunctive 1 interrogative clauses and subjunctive II finite declarative clauses without a complementizer.

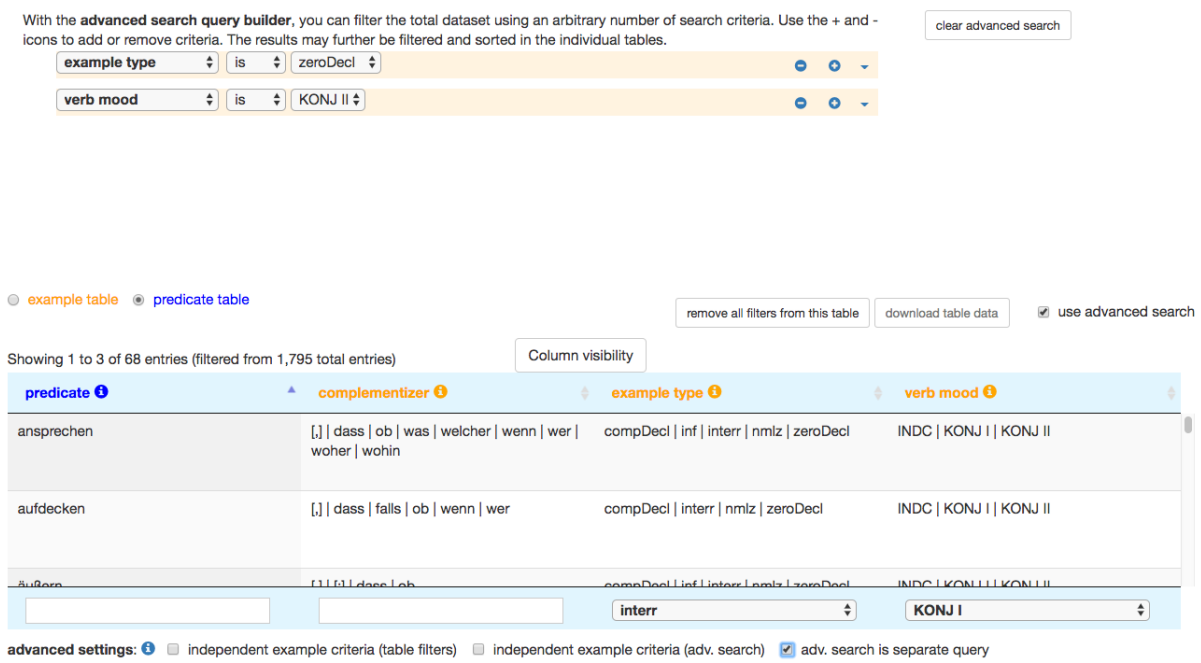


Figure 7: Advanced query for predicates with "separate query" setting turned on

5. Software architecture

Here we briefly sketch the software and data modeling strategy used to ensure that even complex search results can be presented in the form of potentially very long tables in real time in the browser, instantaneously adapting to every change a user makes in a search criterion. Response times have to be very short as each change in one of the search components, such as adding or deleting a single letter in a text field, generates a new server request.

5.1 Backend and database design

The version of the ZAS database of clause-embedding predicates published on OWID^{plus} is a self-contained web application running in a standard Java Servlet container (based on the Sparkjava framework, sparkjava.com) with an embedded relational database (H2, h2database.com). For the purposes of the online version, a snapshot of the 14 original MySQL database tables is systematically denormalized to construct a database of just two flat tables (examples vs. predicates) where each attribute, including each "inherited" property, is represented as a separate column, very similar to what is actually presented to the users in the interface. This "pivoting" procedure, though not strictly necessary, greatly reduces programming and execution overhead and minimizes the need for joins in SQL queries.

5.2 Frontend (browser) technology

An important feature of the user interface is that all data is presented in the form of scrollable tables. Loading up to 17,000 rows with more than 15 columns would take too long, however. Our solution to this problem consists in virtualizing the table (in our case, using the DataTables plugin, datatables.net). Only those data that are currently visible

in the browser (plus some spare rows) are loaded from the server; on scrolling, further rows are fetched using AJAX (XHR) requests.

5.3 Evaluation

For the amount of data available in the ZAS database, the approach outlined above delivers satisfactory response times even for complex queries involving joins, regular expressions etc. Preliminary tests show that the application scales well only up to some 100,000 data rows in the example table. By changing to an in-memory database, this limit can be pushed considerably; however, datasets with millions or even billions of rows would require a more elaborate way of indexing the data and, possibly, limiting the application of regular expressions.

6. Discussion and prospects

The user interface presented in this paper attempts to strike a middle ground between the availability of complex search options and ease of use. The tool deliberately resorts to a powerful combination of two familiar and easily accessible types of interactive interface components, viz. tables with sorting and filtering options and hierarchical query builders. In addition, a set of three yes/no settings can be used to alter the behavior of scope and negation, resulting in an amazing range of possible searches. Through the concepts of inherited columns, derived search properties and the default in-sync setting of search semantics, the 1:n relation between the two tables is exploited as much as possible.

On the other hand, it is self-evident that the query system is not relationally complete in the sense of Codd (1972) and, a fortiori, not equivalent in expressive power to standard SQL. The discussion in subsections 4.2 and 4.3 already pointed to several areas where the range of possible queries could easily be extended. The possible enhancements listed below are under consideration for future versions of the interface.

- With "independent example semantics" turned on for the predicate table, the advanced search criterion input widgets for predicate criteria derived from example criteria could offer both kinds of negation mentioned in subsection 4.3, such as "{at least one | no} example: example type {is | is not} {compDecl | zeroDecl | ...}".
- Instead of having one global, all-or-nothing setting for "independent example semantics", the interface could offer a choice to activate this semantics (separate quantification over example set) for each individual example property, e.g. through a checkbox available on all predicate criteria widgets. The downside is that it would be easy to build advanced queries whose precise meaning is difficult to understand for human users (and therefore not likely to be useful for pursuing typical research questions).
- An even more general approach to multiple quantifications on the example set $\mathbb{E}$ in the predicate table would be to explicitly introduce a mechanism of "scope subgroups" in the query builder. All example criteria within a scope subgroup would be under the scope of a separate existential quantifier on $\mathbb{E}$. Interpreting such queries can, again, be a demanding task for inexperienced users. On the technical side, the more scope subgroups are defined in a query, the more SQL joins appear in the database query on the server side, possibly impairing performance.

Finally, we compare our tool against other approaches. A textbook strategy for online presentation of two tables in a one-to-many relation would be to show the two tables on different web pages and to take account of the relational character of the data in the following way:

- create hyperlinks on the 'many' side (in our case linking the predicates mentioned in the example table to the corresponding row in the predicate table);
- create a master-detail view option on the 'one' side (in our case showing all examples for a given predicate upon, e.g., double-clicking a row in the predicate table).

Our solution does provide master-detail views for both tables, but interweaves both data presentation and search options for the tables to a much higher degree: each table includes as much information from the other table as possible; standard advanced search works in a cross-table way; "independent example semantics" options give more search power.

At the other end of the spectrum, a full-blown visual query tool for relational databases could be used to provide the user with the full expressive power of modern SQL. An important early example of a relational query language with a graphical interface is *Query By Example* (Zloof, 1977; cf. Ramakrishnan & Gehrke, 2002: chapter 6, pp. 177ff.). The most widely known visual query system today is probably the one found in Microsoft Access; a large number of interface components and full-blown web applications work in a similar way. However, such a system would not be friendly for the casual user and has a much steeper learning curve than the immediate interaction with tables. The case of "independent example semantics" shows how quickly query formulation can get very abstract: for each example property included in a predicate search with this semantics, an additional join with the example table must be created, i.e. a new "instance" of the example table must be added to the visual representation.

7. Acknowledgements

The work of the second author was supported by the Bundesministerium für Bildung und Forschung (BMBF, Grant Nr. 01UG1411). We would like to thank Barbara Stiebels and the ZAS database team, in particular Kerstin Schwabe, Torgrim Solstad, Livia Sommer, Katarzyna Stoltmann, Noemi Geiger, Gediminas Schüppenhauer and Sybille Kiziltan.

8. References

Codd, E.F. (1972). Relational completeness of data base sublanguages. In R. Rustin (ed.) *Data Base Systems, Proceedings of 6th Courant Computer Science Symposium*. New York: Prentice-Hall, pp. 65–98.

Hearst, M.A. (2009). *Search User Interfaces*. Cambridge University Press, 1st edition.

Levin, B. (1993). *English Verb Classes and Alternations*. University of Chicago Press.

Morville, P. & Callender, J. (2010). *Search Patterns: Design for Discovery*. O'Reilly Media, 1 edition edition.

Ramakrishnan, R. & Gehrke, J. (2002). *Database Management Systems*. Mcgraw-Hill, 2nd edition edition.

Russell-Rose, T. & Tate, T. (2012). *Designing the Search Experience: The Information Architecture of Discovery*. Morgan Kaufmann.

Stiebels, B. (2011). Von den Herausforderungen des lexikalischen Reichtums. In V. der Geisteswissenschaftlichen Zentren Berlin e.V. (ed.) *Bericht über das Forschungsjahr 2010*. pp. 51–72.

Stiebels, B., McFadden, T., Schwabe, K., Solstad, T., Kellner, E., Sommer, L. & Stoltmann, K. (2017). ZAS Database of Clause-embedding Predicates, release 0.2 (Public Beta). In *OWID^{plus}*. Institut für Deutsche Sprache, Mannheim. URL http://www.owid.de/plus/zasembed2017.

Zloof, M.M. (1977). Query-by-example: A data base language. *IBM systems Journal*, 16(4), pp. 324–343.

KBBI Daring: A Revolution in The Indonesian Lexicography

Ian Kamajaya[1], David Moeljadi[2], Dora Amalia[3]

[1]ASTrio Pte Ltd, Singapore
[2] Nanyang Technological University, Singapore
[3] Badan Pengembangan dan Pembinaan Bahasa, Indonesia
E-mail: ian@astriotech.com, D001@e.ntu.edu.sg, dora.amalia@kemdikbud.go.id

Abstract

Kamus Besar Bahasa Indonesia (KBBI) is the official dictionary of the Indonesian language, published by Badan Pengembangan dan Pembinaan Bahasa (The Language Development and Cultivation Agency) or Badan Bahasa, under the Ministry of Education and Culture, Republic of Indonesia. The current, fifth edition of KBBI (Amalia, 2016) was launched on 28 October 2016 and contains more than 100,000 entries and 120,000 senses. It is available in three formats: printed, online, and offline mobile applications. The online version, called KBBI Dalam Jaringan or KBBI Daring (kbbi.kemdikbud.go.id), is categorized as Dictionary Writing System (DWS) (Atkins & Rundell, 2008). Through it, we invite online public participation to make proposals to add and to edit entries, senses, and examples. We are changing our workflow from manual to computerized work which has greatly reduced the time needed to make a dictionary. KBBI Daring greatly expands the database which was previously made for the fourth edition of KBBI (Sugono, 2008) using the data in Microsoft Excel and Word files (Moeljadi et al., 2017), fitting to its online usage. This paper describes our efforts in building the KBBI Daring which has revolutionized both the way people use a dictionary and the lexicographical workflow of the editorial staff in Indonesia.

Keywords: online dictionary; Dictionary Writing System; Indonesian language; electronic lexicography; lexicographical workflow

1. Introduction

Kamus Besar Bahasa Indonesia (KBBI) is the official dictionary of the Indonesian language,[1] published by Badan Pengembangan dan Pembinaan Bahasa (The Language Development and Cultivation Agency) or Badan Bahasa, under the Ministry of Education and Culture, Republic of Indonesia. Up until present, KBBI is the most comprehensive and the most authoritative reference for the Indonesian language. Its first edition, published in 1988, has 62,000 entries. The number of entries increased to 72,000 or about 10,000 entries over three years in the second edition (1991). Its third edition, published in 2001, contains 78,000 entries and seven years later, the number of entries in the fourth edition increased to more than 92,000. Its latest, fifth edition was released for the first time in 2016 in three formats: printed, online, and offline mobile versions.[2]

The online KBBI before 28 October 2016, launched in 2006, used the data from the third edition of KBBI and allowed searches only by headwords. For example, to look up *mengacang*, a user must first look up the root word *kacang*, as shown in Figure 1. The sound assimilation process (a morphophonemic process) in *meN-* prefix makes the first

[1] Indonesian (ISO 639-3: ind), called *bahasa Indonesia* (lit. "the language of Indonesia") by its speakers, is a Western Malayo-Polynesian language of the Austronesian language family. Within this subgroup, it belongs to the Malayic branch with Standard Malay in Malaysia and other Malay varieties (Lewis, 2009). It is spoken mainly in the Republic of Indonesia as the sole official and national language and as the common language for hundreds of ethnic groups living there (Alwi et al., 2014: 1–2). In Indonesia it is spoken by around 43 million people as their first language and by more than 156 million people as their second language (2010 census data). The lexical similarity is over 80% with Standard Malay (Lewis, 2009). It is written in Latin script. Morphologically, Indonesian is mildly agglutinative, compared to Finnish or Turkish. It has a variety of prefixes, suffixes, circumfixes, and reduplication.

[2] The printed version and the online application of KBBI were launched on 28 October 2016. The offline Android and iOS mobile applications were launched on 17 November 2016.

sound /k/ in the root word *kacang* become a nasal sound /ŋ/ (orthographically written as <ng>). This may present some difficulties if the user is not familiar with Indonesian morphological rules. In addition, it was not designed to support targeted lookups, such as search by word classes and entry types. Furthermore, it was not built to support addition or modification of any of its data – neither by editorial staff nor by other users. In short, it was solely built for the purpose of searching dictionary entries by headwords.

The new online KBBI called "KBBI Dalam Jaringan" or "KBBI Daring"[3] was launched on 28 October 2016. Compared to its predecessor, it is designed with richer features which allow targeted and flexible, rather than exact, lookups. Besides, and more importantly, it serves as a unified and reliable platform for Indonesian lexicographers—which include not only professional lexicographers, but also common, non-professional Indonesian language users—across the world to enrich and to edit KBBI, increasing the efficiency of the editorial workflow. Although this concept is by no means new, it is a revolutionary work in Indonesia (or rather, globally, for Indonesian language users). Prior to KBBI Daring, editorial workflow in Indonesia was greatly scattered, unreliable, marked by many loss of editorial requests (such as enrichment or correction request from Indonesian language users – oftentimes sent by letter or by raising the issues during Indonesian language seminars), and slow response from Badan Bahasa to edit and to enrich KBBI against the emerging globalization that introduces a lot of new cultures, concepts, and technologies which demand a vast amount of new words and concepts to be included in KBBI. Figure 2 shows the main page of KBBI Daring. This paper describes our efforts in building KBBI Daring and is organized as follows: Section 2 describes how the editorial workflow was before KBBI Daring was built. Section 3 explains the features of KBBI Daring which are intended to deal with the situations described in Section 2. Section 4 reports the receptions from users and editorial staff. Section 5 concludes and mentions some future works.

2. Dictionary Use and Lexicographical Work before KBBI Daring

Before KBBI Daring was launched on 28 October 2016, we had manual lexicographical work, less public participations, and inconsistencies in the dictionary format.

2.1 Lexicographical Workflow

Editorial work for KBBI includes adding new entries, checking the accuracy of spellings, definitions, and examples, as well as formatting and layout. Before KBBI Daring was launched, the editorial staff worked manually. The data collectors collected some data sources, such as magazines, newspapers, and books in order to find new words not listed in KBBI by looking up those words in the paper dictionary. They recorded the new words on small pieces of card. In the first and second editions of KBBI, the cards were manually sorted in alphabetical order and placed in a special catalog cabinet. Other staff inputted the data of the new words into a file in a computer and did the formatting and layout. After the dictionary was printed, the editors wrote some notes on the pages if they found some errors and other staff fixed them in the file.

[3] https://kbbi.kemdikbud.go.id/

Figure 1: Screenshot of the online KBBI before 28 October 2016

Figure 2: Screenshot of the main page of KBBI Daring

2.2 Dictionary Use and Public Participations

Indonesian has a vast number of speakers (see Footnote 1). They rely on printed KBBI throughout Indonesia as the most authoritative source, while the printed dictionaries can be accessed only by limited users, i.e. those who can afford to buy them. In addition, Indonesia's vast geographical condition also makes the distribution more difficult. Public contributions were also very limited. The public submitted proposals via post, email, and direct personal communication. They did not know whether their proposals had been processed or not. They did not even know whether their proposals had arrived at the editorial staff or not. There was no channel to inform them of the status of their proposals. Due to this limited access, the proposals received were also very small in number. Furthermore, the proposals submitted were checked by the editorial staff in Badan Bahasa without any help from experts outside the editorial team which might speed up the editorial work, mainly due to the limited access of the experts to the editorial requests.

2.3 Dictionary Format

Since KBBI was compiled and formatted manually by hand, errors (such as typos and inconsistencies in the formatting) are inevitable. In addition to some sporadic errors found and mentioned in Moeljadi et al. (2017), other errors in KBBI that were detected by our diagnostic tools in KBBI Daring (see Section 3.4) are, for example, compounds having word class labels (or contrarily, root words or derived words not having word class labels), duplicate entries, examples not containing the headwords used, and root words or derived words not having syllabifications.

3. KBBI Daring Features

KBBI Daring was built to deal with the issues described in Sections 1 and 2. This section explains some major features of KBBI Daring.

3.1 Dictionary Data Structure

KBBI Daring uses KBBI Database (Moeljadi et al., 2017). The database file is an SQLite file. The data structure of KBBI consists of four types of data: *entry*, *sense*, *example*, and *category*. The relationship between *entry* and *sense*, as well as the one between *sense* and *example* are one-to-many. The *category* is a list of descriptions or a metadata for *entry*, *sense*, and *example*. Figure 3 illustrates the KBBI data structure. An entry can be a fixed expression (*ungkapan*)[4] or a root word (*kata dasar*). A fixed expression should have at least one sense and one example. In this case, one fixed expression may have one to multiple senses and one sense may have one to multiple examples. A root word should have at least one cross-reference, one sense, one compound, or one derived word. In this case, one root word may have zero to multiple senses and one sense may have zero to multiple examples. A root word may also have variant(s), proverb(s), and idiom(s). A proverb or an idiom should have at least one sense. A compound should have at least one cross-reference or one sense. One sense may have zero to multiple examples. Similar to the root word, a derived word should have at least one cross-reference, one sense, or one compound. It may also have variant(s), proverb(s), and idiom(s). The root word can be in the form of compound if it can be affixed and have derived word(s).

[4] Fixed expressions are commonly used foreign phrases in Indonesian written works, such as *ad hoc*.

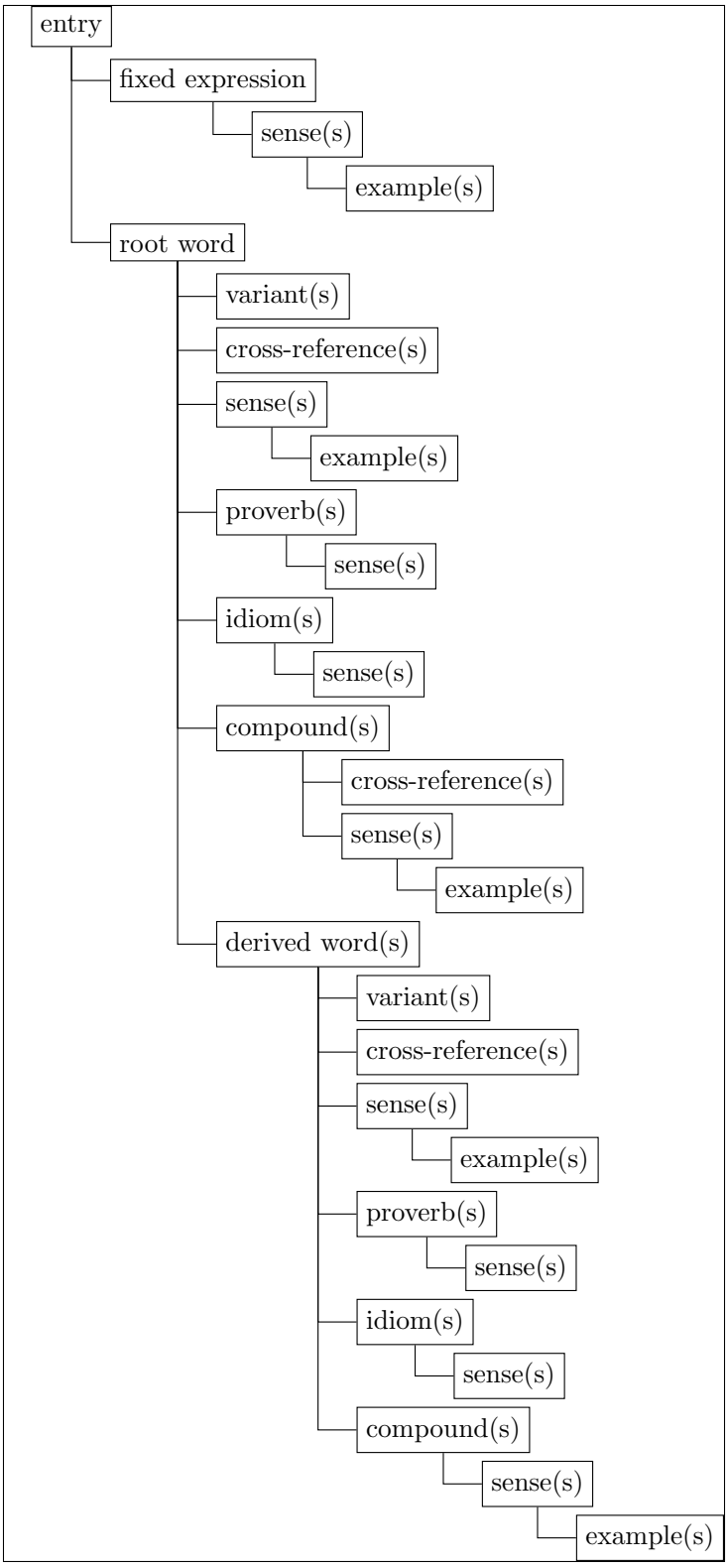


Figure 3: The KBBI data structure

3.2 User Groups and Privileges

User groups or user roles in KBBI Daring are primarily designed to represent the actual lexicographical workflow.[5] Each user group is the realization of a certain group of Indonesian language users in the real-world counterpart. Therefore, KBBI Daring implements six groups of users: non-registered users, registered users, registered editors, main editors, validators, and one main administrator. Table 1 shows the privileges of each user group. A user in a higher group has all the privileges (features) of the lower groups. Figure 4 shows KBBI Daring system and access rights.

User group	**(Additional) Features**
Non-registered users	• Basic search tool
Registered users	+ Criteria-based search tool + Basic proposal tools
Registered editors	+ Advanced search tools + Advanced proposal tool + Basic editorial tools + Basic diagnostic tool
Main editors	+ Advanced editorial tools + Lexicographical tools
Validators	+ Validation privilege + Advanced diagnostic tool
Main administrator	+ Mass diagnostic and recovery tools + Printing tool

Table 1: (Additional) features for each user group in KBBI Daring

3.2.1 Non-registered Users

Non-registered users are those who use the KBBI Daring without registering their email addresses. They are only allowed to search using a basic search feature and are not allowed to make any proposal.

3.2.2 Registered Users

Registered users are those who use the KBBI Daring after registering their email addresses. They are given two additional privileges: to search based on predefined criteria and to make proposals to add or to edit the dictionary data. The criteria-based search tool is available on the registered user's main page. The users can search entries by their initial letters, word classes, styles, entry types, languages, and domains. Figure 5 shows some

[5] Albeit in a website setting, where an additional administrator group is needed and non-registered users are given limited access to prevent anonymous stealing of the data.

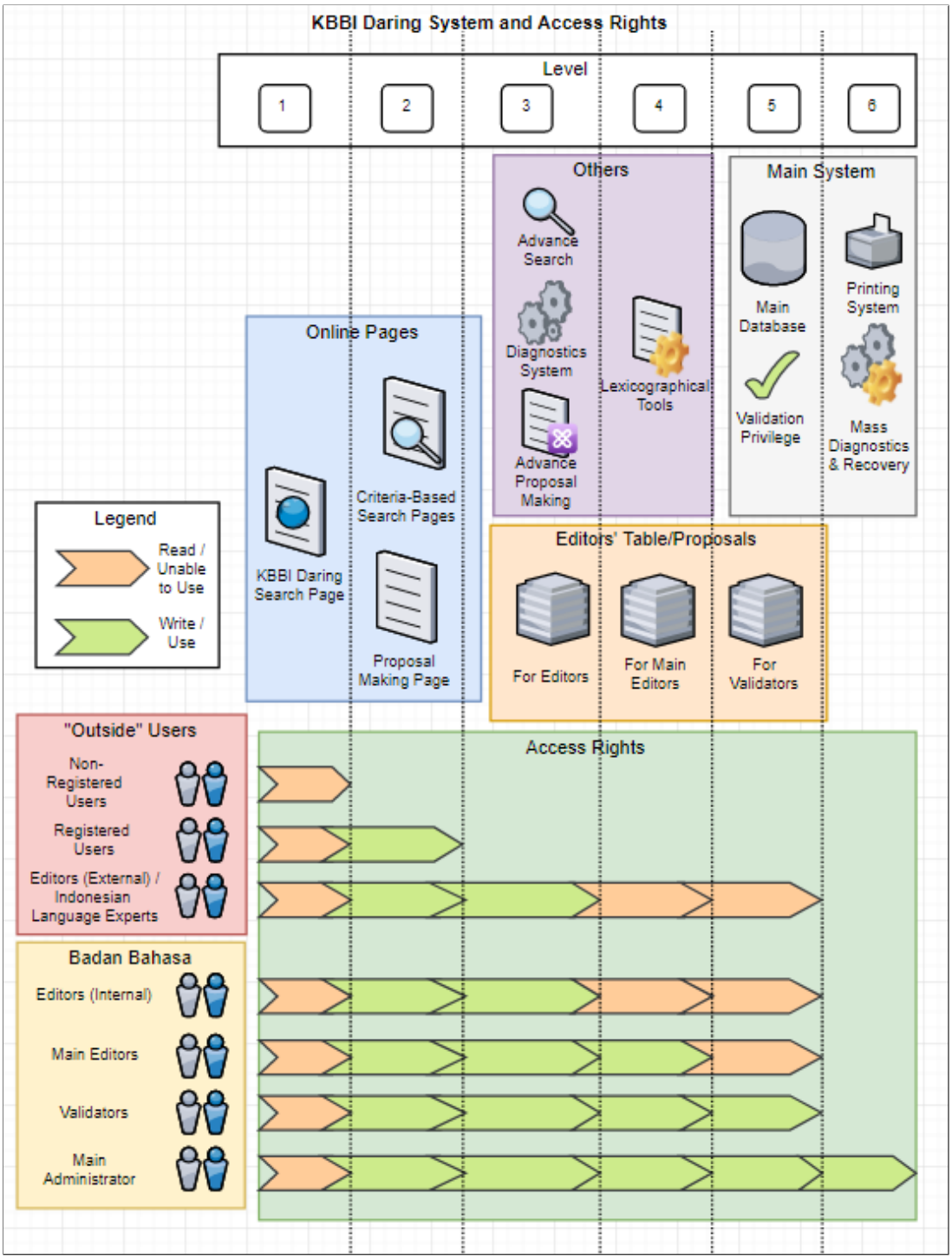


Figure 4: KBBI Daring system and access rights

proposal tools in the registered users' search result. The proposal system will be explained in Section 3.3. Figure 6 shows an example of a proposal page. Figure 7 shows an account management page where the users can check the status of their proposals.

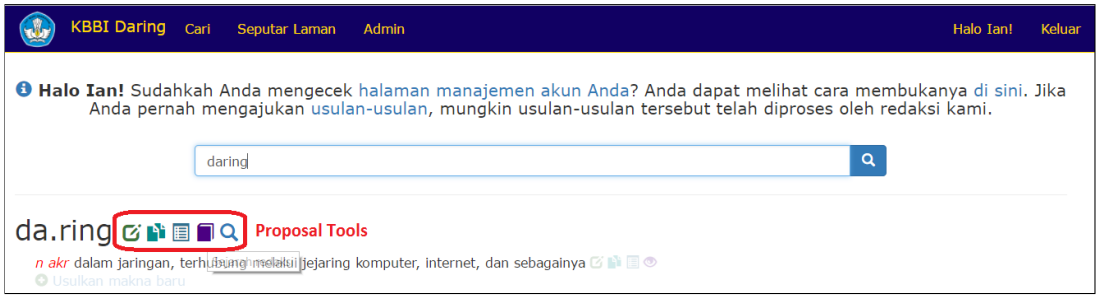


Figure 5: Screenshot of a search result for registered users

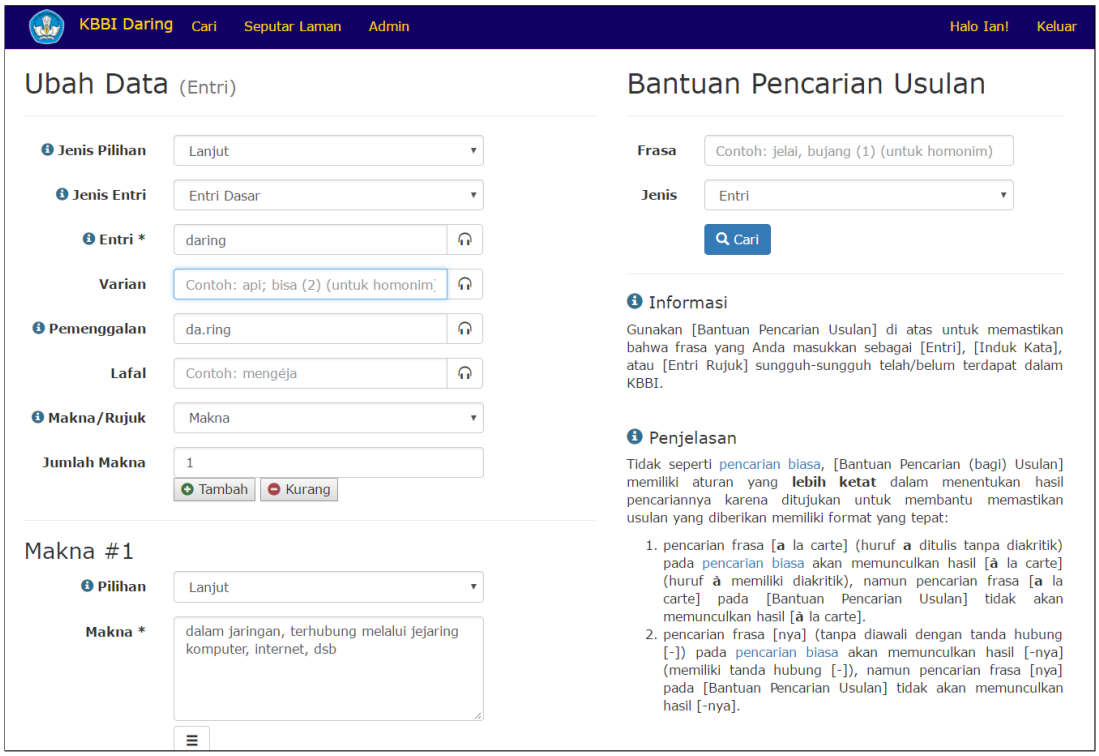


Figure 6: Screenshot of a proposal page

3.2.3 Registered Editors

Registered editors are the registered users who are granted privileges as KBBI editors because they understand linguistic issues and lexicographical theory and have attended lexicography workshops. They are given additional privileges to access advanced search tools, advanced proposal tools, basic diagnostic tools, and basic editorial tools. The advanced search tools allow them to search in the dictionary data structure (entry, sense, example, or category).

Figure 7: Screenshot of an account management page

The advanced proposal tool allows registered editors to make a deactivation proposal. The basic diagnostic tool allows registered editors to diagnose an entry to check if it has a formatting error. Figure 8 shows a search result page with the advanced proposal tool and the basic diagnostic tool.

The basic editorial tools allow registered editors to review the submitted proposals proposed by the registered users. A registered editor may accept and pass it up to the main editors, or take it over/change it, or return it to the registered user with comments, or reject it. Proposals which are taken over/changed become the registered editor's proposals. Proposals which are returned can be resubmitted but proposals which are rejected cannot. Figure 9 shows a list of proposals submitted by registered users to be further processed/reviewed by registered editors.

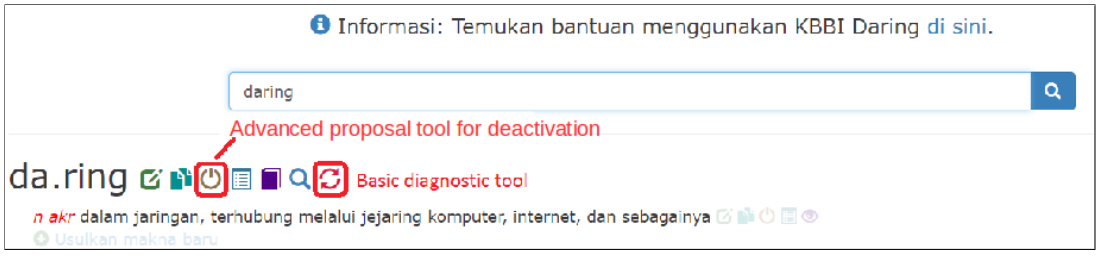


Figure 8: Screenshot of a search result with the advanced proposal tool and the basic diagnostic tool

3.2.4 Main Editors

The main editors are the ones inside Badan Bahasa whose main responsibility is to create KBBI. Compared to the registered editors, they have two additional privileges: access to

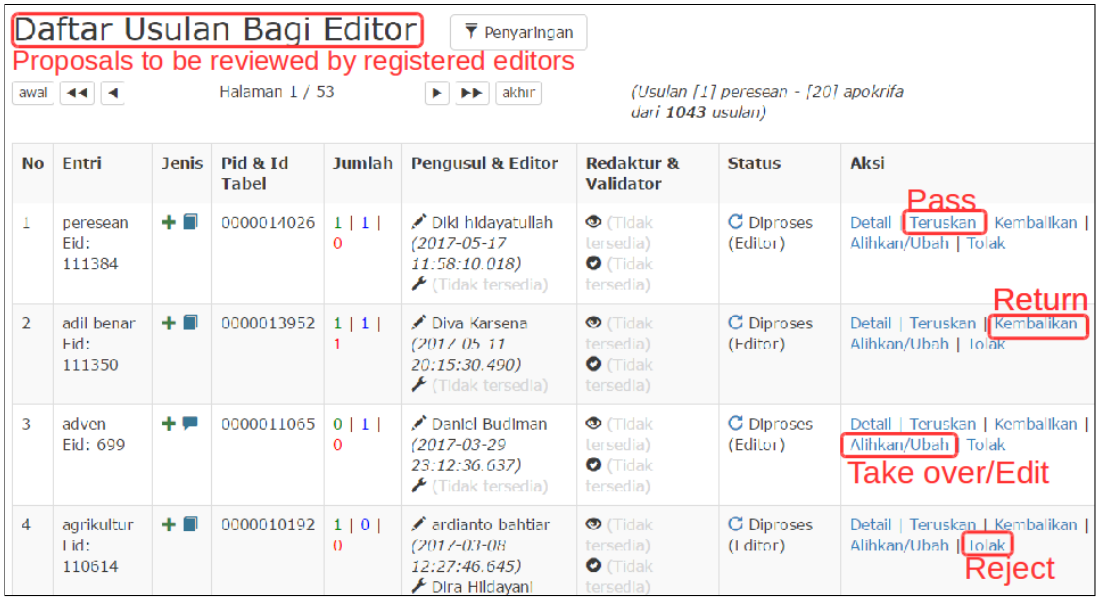


Figure 9: List of proposals to be reviewed by registered editors

advanced editorial tools and access to lexicographical tools. The advanced editorial tools give the main editors additional options to pass down, to archive, or to abort a proposal. A proposal which is passed down is returned to the registered editors for reviewing. A proposal which is archived is not aborted, but cannot be further processed. A proposal which is aborted is essentially deleted – not only rejected or returned. The lexicographical tools allow the main editors to reorder the polysemies (senses) of an entry and to "redirect" or "reattach" a sense to a different entry. The polysemy reordering is particularly useful to determine which senses should appear first or later on a search result page and on paper. The advanced editorial tools and lexicographical tools are only given to the main editors and above. Figure 10 shows the (additional) advanced editorial tools for processing proposals in the main editors' table. Figure 11 shows two lexicographical tools for the main editors to reorder the polysemies and to "redirect" a sense to a different entry.

3.2.5 Validators

Validators are the editors who have a right to decide whether a proposal should be accepted and are the last examiners of the proposals. Validators have two additional privileges compared to the main editors: validation privilege and advanced diagnostic tool. Due to the validation privilege, once a validator accepts a proposal, the change will be reflected in the website. The advanced diagnostic tool allows the validators to diagnose multiple elements at the same time. Figure 12 shows a validator's option to accept a proposal. Figure 13 shows the advanced diagnostic tool to diagnose multiple elements.

3.2.6 Main Administrator

The main administrator is an additional role to the actual editorial roles. It is designed, however, to allow a single most privileged user to use the printing application in the website as well as to do mass diagnostics and to recover the data. Since these features

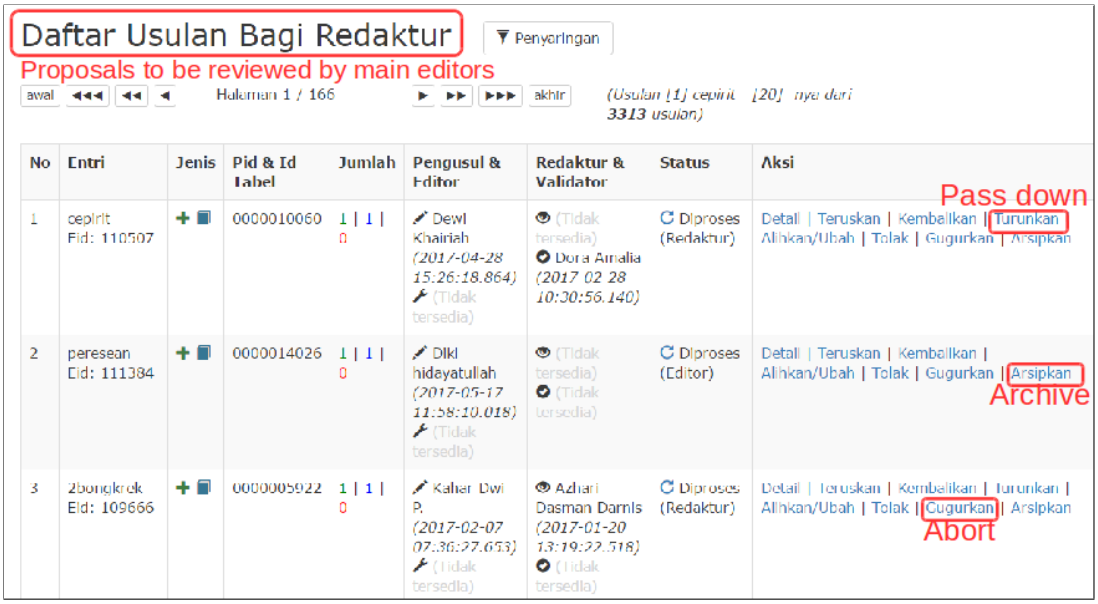


Figure 10: Advanced editorial tools

Figure 11: Lexicographical tools

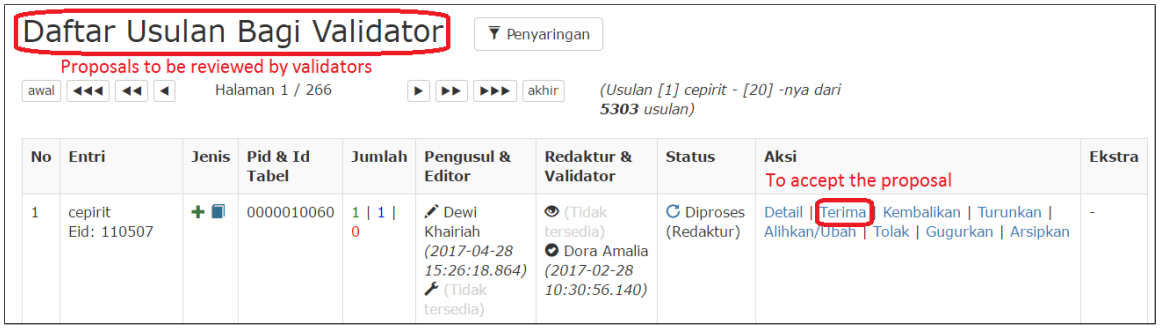


Figure 12: Validator's option to accept a proposal

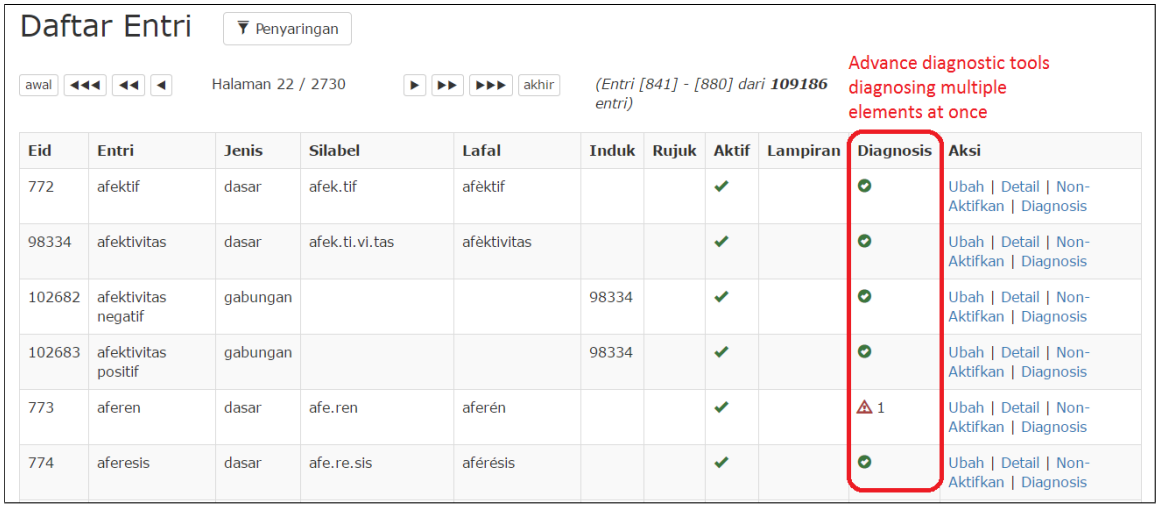


Figure 13: Advanced diagnostic tool

will consume high resources in the server hardware, they are given only to one main administrator.

3.3 Proposal System

The proposal system in KBBI Daring is a guided, non-anonymous, transparent, crowdsourcing system. Its design is closely tied with the user groups and their privileges in order to transform the traditional lexicographical workflow to its current, public-friendly, transparency-imbued, web-based form – making it a lot more accessible for crowdsourcing while having an official body to guide the overall process.

Though most KBBI Daring users are non-registered users, the privilege to use the proposal system is only given to registered users and above. The reason is rather obvious: anonymous contribution for the official and the most authoritative reference for the Indonesian language is hardly a promising idea. Furthermore, such a feature may do more harm than good as it can be exploited anonymously to send "junk" proposals. Forcing registered accounts to access the proposal tools would limit the number of proposals generated from anonymous sources. People are required to register and to verify their email addresses for their accounts to be registered on the KBBI Daring website. Furthermore, fake and temporary email domains are filtered by the account registration system in KBBI Daring, leaving mostly only valid email addresses to be registered.

The proposal can target any one of the following three data types: entry, sense, or example. Additionally, the proposal made must be one of the following types: add, edit, or deactivate (only for registered editors and above). In the end, the validator must decide whether the proposal is acceptable or not. When a proposal is accepted, the targeted item in the database is replaced with the proposed item, the proposal data are logged (for historical references), and the changes are immediately reflected in the website. The contributor's names, i.e. the proposal maker, the editors, and the validator, are shown in the editorial history. Figure 14 shows the editorial history of entry *Yesus Kristus* "Jesus Christ". The editorial history, which is accessible to registered users and above, contains all essential information: details of the proposal, the proposer, the registered editor, the main editor,

and the validator of the proposal – each with their respective explanation or comment, the acceptance date and time, as well as the revision number. Thus, it enforces transparency of all the items added or changed in KBBI Daring through proposals.

The registered user group is designed to be the major group (in terms of number) in KBBI Daring which is given the privileges to participate in the enrichments and corrections of KBBI by creating reasonable proposals. It is the main crowdsourcing group which is designed to represent the "common" Indonesian language users (compared with different user groups in the following paragraphs) who are willing to contribute. Anyone who registers can immediately become a contributor. However, between the registered users and the validators, there are two groups: registered editors and main editors.

Registered editors review the proposals from registered users. They "recommend" good proposals to be passed up to the main editors, edit potential proposals with few flaws to be more acceptable, return and guide registered users to make better-formed proposals, or immediately reject the proposals when they are considered unacceptable in the first place (for example, if a user proposes a duplicate entry or a duplicate sense). The registered editor group is designed to represent the experts who want to participate in the lexicographical workflow of KBBI. They consist mostly of adept people in lexicography and linguistics. Consequently, the registered editors are expected to create good proposals and to be another major contributing group.

Main editors and validators consist exclusively of people inside Badan Bahasa who are responsible for KBBI, e.g. the chief editor of KBBI and the head of the lexicography subdepartment. Technically, they can make proposals, but they are not expected to do it as their prime task. Instead, using their official positions, their main task is to verify the acceptableness of proposals created by the registered users and the registered editors, and to focus more on the lexicographical work.

3.4 Mass Diagnostic and Recovery Tools

KBBI Daring also helps the editors find potential errors inherited from the previous editions, as mentioned in Section 2.3, or created by the users, and helps correct them. It is designed with mass diagnostic and recovery tools for that purpose. Figure 15 shows an example of mass diagnostic results.

Some errors, e.g. pronunciation containing letters other than é, ê, and è[6] or definitions having certain words in the lengthened forms instead of the shortened forms (i.e. the word *seperti* "like, as" should be written in its shortened form *spt* in KBBI), are errors with a definite (single) solution. They are automatically correctable. On the other hand, other errors, e.g. duplicate entries with different senses or absence of syllables in root words or derivative words, are errors with non-definite solutions. They can only be corrected by humans. Table 2 shows the list of errors diagnosable using the mass diagnostic tool.

The recovery tool is designed to correct errors which have a definite (single) solution. It is not designed to correct errors with multiple viable solutions. Instead, the diagnostic tool

[6] The pronunciation field in KBBI only deals with entries having the letter <e>. The Indonesian language has the sounds [e], [ə], and [ɛ]. However, they are not orthographically distinguished in the current spelling system, all of them are written as <e> (Alwi et al., 2014). The pronunciation field in KBBI indicates them as é, ê, and è respectively.

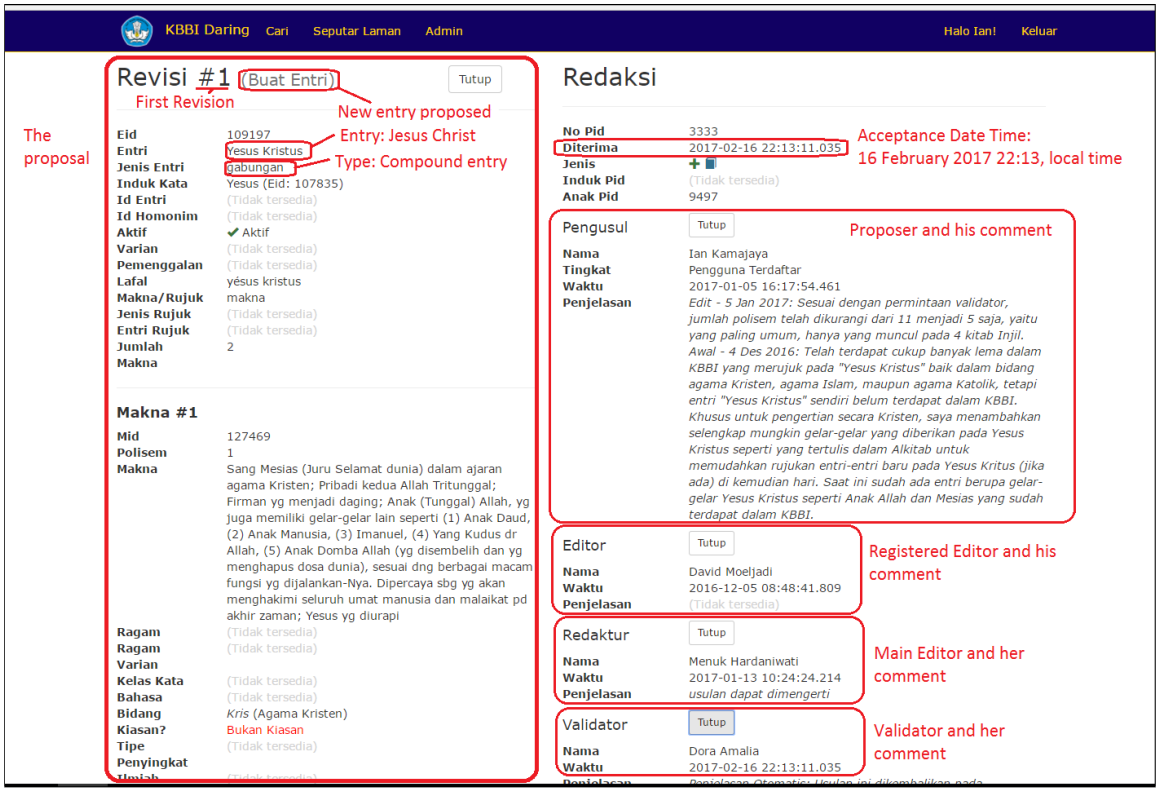


Figure 14: The editorial history of an entry

is designed to detect such cases, to help human editors identify and correct the mistakes. Table 3 shows the list of errors which are correctable by the recovery tool.

3.5 Printing Tool

The printing tool is integrated as a part of KBBI Daring and is only accessible to the main administrator. The printing tool is built to make consistent formatting effects for the printed version. The major contribution is to eliminate human errors in the formatting. The major challenge is the complex and potentially-growing formatting effects. However, manual editing has advantages over machines because it can handle exceptions in the formatting rules more flexibly, especially if the exceptions are only small parts of the dictionary, yet varying enough to be handled non-uniformly with ease using the printing tool. Therefore, the printing tool is primarily designed to create the base version of the printed version – solid enough as not containing formatting errors, but flexible enough for human agents to do a little finishing touch before being sent to the printing company in a ready-to-print format.

The printing tool is designed with Microsoft-created `.Net` dynamic link library (`.dll`) `Microsoft.Office.Interop.Word` to produce the base version in `.doc` format. The setting page is provided in KBBI Daring to allow the main administrator to choose the format and the range of the printing. Afterwards, the printing command is given to the server and the server starts to write the `.doc` document for printing. Once finished, the resulting document is made available in KBBI Daring, accessible only by the main administrator. It can be downloaded and further processed by human agents to make the ready-to-print format of KBBI.

Data type	List of diagnoses
Entry	• Does it have duplicates with different senses? • Does it have identical variants? • Does it have syllabification (where it should not have and vice versa)? • Does it have pronunciation (where it should not have and vice versa)? • Does the pronunciation contain letters other than é, ê, and è? • For an entry having a root word, does it refer to an active root word? • For a cross-reference entry, does it have a correct reference item?
Sense	• Does it refer to a particular entry? • Does it have a word class (where it should not have and vice versa)? • Does it contain repeated scientific names (binomial names)?
Example	• Does it refer to a particular entry? • Does it refer to a particular sense? • Does it contain the corresponding headword? • Does the headword in the example have the same spelling? • Does it have misplaced spaces or punctuations?
Entry, Sense, and Example	• Does it have "odd" or "missing" IDs? • Does it contain shortened form of certain words while they should be written in their lengthened form (and vice versa)?

Table 2: Errors diagnosable by the mass diagnostic tool

Data type	Error	Correction
Entry	should be without syllabification but has syllabification	remove the syllabification
	does not contain 'e' but has pronunciation	remove the pronunciation
	contains 'e' but does not have pronunciation	give the default pronunciation ê
Sense	a compound which has a word class	remove the word class
	contains repeated scientific names (binomial names)	remove one of the scientific names
Entry, Sense, and Example	contains shortened form of certain words while they should be written in their lengthened form	lengthen the word
	contains lengthened form of certain words while they should be written in their shortened form	shorten the word

Table 3: Some errors and the corrections using the recovery tool

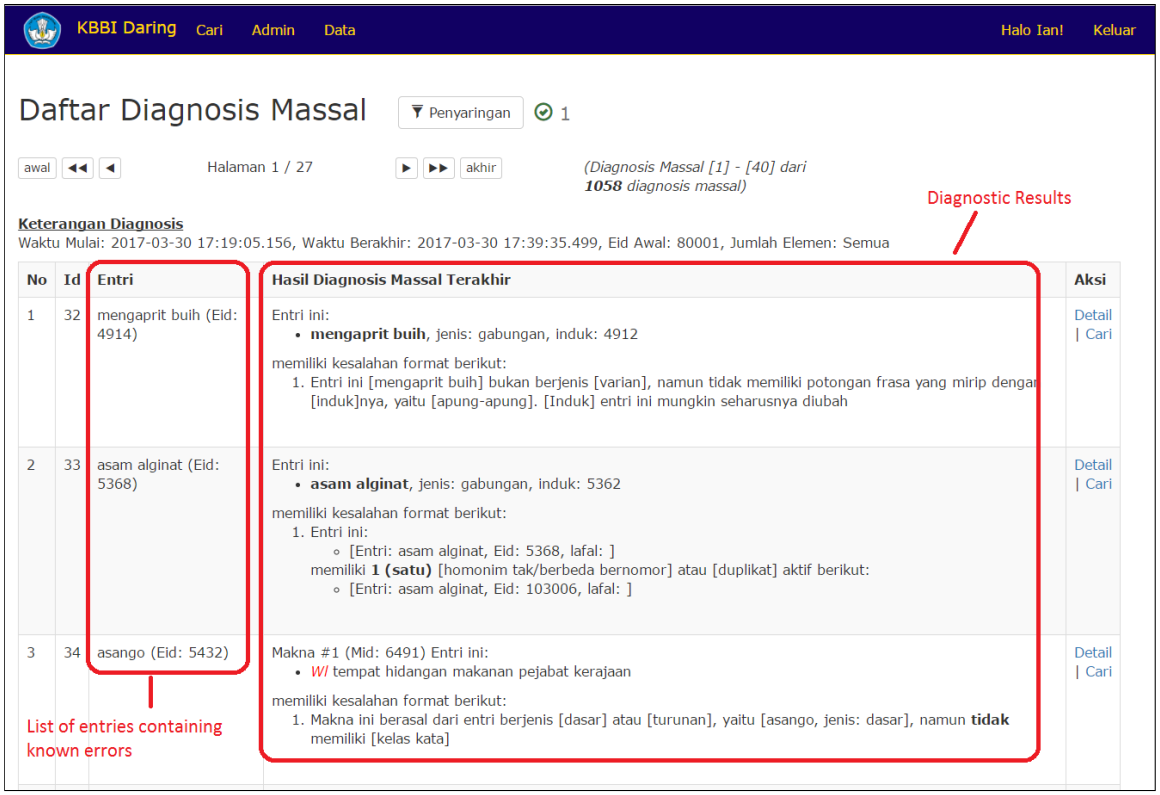


Figure 15: Mass diagnostic results

The base version of the KBBI printed version generated by the printing tool is a near ready-to-print version with the formatting effects specified by Badan Bahasa – except for the word hyphenation which occasionally needs to be corrected by human editors. Unlike LaTeX, the Microsoft Word 2013 tool used to generate the auto-hyphenation does not have a list of exceptions for correcting wrongly generated hyphenations.[7] Word is used due to its popularity and its ease of integration with Adobe InDesign file (`.indd` format), which is used by Badan Bahasa for the finishing touch on the document and is occasionally required by major printing companies in Indonesia.

3.6 Others

Besides the features mentioned above, KBBI Daring is also equipped with a customized security system to protect the data from web crawlers.

4. KBBI Daring Impacts and Receptions

KBBI Daring made the editorial work more efficient. Automatic notifications of errors are shown by the machine and the editorial staff can focus only on the notified errors. For some formatting errors, the changes can even be done solely by the machine. The changes allow the team to focus more on the substantial issues, such as the accuracy of definitions. The conversion process from the database to the near ready-to-print version is automatically done, including the header on each page. However, formatting issues such

[7] As of May 2017, LaTeX-based printing tool is still being built as an alternative version of the currently used Word-based printing tool because it can handle hyphenation better and saves more time.

as the 'widow' and 'orphan' line as well, as an adjustment of the columns on the last page of each letter section, need to be handled manually.

Dissemination was held to introduce KBBI Daring to the public (students, teachers and lecturers, journalists, translators and writers) for the Indonesian vocabulary enrichment program. The dissemination event has been held in 15 provinces in Indonesia in 2016. As of May 2017, it was held in 15 other provinces in 2017. In the event, the participants are encouraged to use KBBI Daring and to make proposals. The number of proposals received from the participants varies, ranging from the fewest (fewer than 10) to the largest (more than 50). We expect that the participants will continue using KBBI Daring.

As of 26 May 2017, KBBI Daring has been used to search entries more than 3.4 million times and has accumulated more than 9,800 proposals for the dictionary's enrichments and corrections. Alexa site[8] shows the domain as the 81st most searched domain in Indonesia and the first (most searched) among all the domains ending with `go.id`.

5. Conclusion and Future Works

We have described our work in building KBBI Daring which revolutionized the lexicographical workflow, helping the editorial staff work more efficiently, and involves more public participations in enriching and improving the dictionary. It also minimizes formatting errors in the printed version which are inevitable in the previous editions. In the future, we will add etymological information and connect KBBI Daring to corpora and lexical databases such as Wordnet Bahasa (Bond et al., 2014). We plan to publish supplements for the printed version every six months in order to provide the users with the current lexicon and other lexicographical information which reflect the language used by the society over time.

6. Acknowledgements

Thanks to Ardianto Suhendar for creating some base code for the printing tools using `Elistia.DotNetRtfWriter`. Thanks to elistia[9] and Serg-Norseman[10] for creating and making it available for public use in the GitHub. Thanks to Ivan Lanin for his valuable examples of the main page and search results for KBBI Daring. Thanks to Francis Bond for his ideas for KBBI Daring features, especially the deactivation feature and the feature to create new proposals based on the already existing items.

7. References

Alwi, H., Dardjowidjojo, S., Lapoliwa, H. & Moeliono, A.M. (2014). *Tata bahasa baku bahasa Indonesia*. Jakarta: Balai Pustaka, 3 edition.

Amalia, D. (ed.) (2016). *Kamus Besar Bahasa Indonesia*. Jakarta: Badan Pengembangan dan Pembinaan Bahasa, 5 edition.

Atkins, B. & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

[8] www.alexa.com

[9] https://github.com/elistia

[10] https://github.com/Serg-Norseman

Bond, F., Lim, L.T., Tang, E.K. & Riza, H. (2014). The combined Wordnet Bahasa. *NUSA: Linguistic studies of languages in and around Indonesia*, 57, pp. 83–100.

Lewis, M.P. (ed.) (2009). *Ethnologue: languages of the world*. Dallas: SIL International, 16 edition.

Moeljadi, D., Kamajaya, I. & Amalia, D. (2017). Building the Kamus Besar Bahasa Indonesia (KBBI) Database and Its Applications. In *Proceedings of The 11th International Conference of the Asian Association for Lexicography*. Guangzhou, pp. 64–80.

Sugono, D. (ed.) (2008). *Kamus Besar Bahasa Indonesia Pusat Bahasa*. Jakarta: PT Gramedia Pustaka Utama, 4 edition.

E-VIEW-alation – a Large-scale Evaluation Study of Association Measures for Collocation Identification

Stefan Evert[1], Peter Uhrig[1], Sabine Bartsch[2], Thomas Proisl[1]

[1]Friedrich-Alexander-Universität Erlangen-Nürnberg
[2]Technische Universität Darmstadt
E-mail: stefan.evert@fau.de, peter.uhrig@fau.de, bartsch@linglit.tu-darmstadt.de, thomas.proisl@fau.de

Abstract

Statistical association measures (AM) play an important role in the automatic extraction of collocations and multiword expressions from corpora, but many parameters governing their performance are still poorly understood. Systematic evaluation studies have produced conflicting recommendations for an optimal AM, and little attention has been paid to other parameters such as the underlying corpus, the size of the co-occurrence context, or the application of a frequency threshold.

Our paper presents the results of a large-scale evaluation study covering 13 corpora, eight context sizes, four frequency thresholds, and 20 AMs against two different gold standards of lexical collocations. While the optimal choice of an AM depends strongly on the particular gold standard used, other parameters prove much more robust: (i) small co-occurrence contexts are better than larger spans, and the best results are usually obtained from syntactic dependencies; (ii) corpus quality is more important than sheer size, but large Web corpora prove to be a valid substitute for the British National Corpus; (iii) frequency thresholds seem to be unnecessary in most situations, as the statistical AMs successfully weed out rare and unreliable candidates; (iv) there is little interaction between the choice of AM and the other parameters.

In order to provide complete evidence for our observations to readers, we created an interactive Web-based application that allows users to manipulate all evaluation parameters and dynamically updates evaluation graphs and summaries.

Keywords: collocations; association measures; evaluation; multiword expressions; visualization

1. Introduction

Traditionally, the identification of collocations and other types of lexicalized multiword expressions (MWE) has been based on co-occurrence data quantified by statistical association measures (AM). A typical extraction pipeline obtains co-occurrence counts (within a span of n words, within a sentence, or in a direct syntactic dependency relation) from a given source corpus. Candidates are then ranked according to their association scores, optionally filtered by various criteria, and finally presented to lexicographers or domain experts for manual validation (Evert, 2008).

Recent work has focused on complementing AMs with other indicators for the non-compositionality (Katz & Giesbrecht, 2006; Kiela & Clark, 2013; Yazdani et al., 2015), non-modifiability (Villada Moirón, 2005; Nissim & Zaninello, 2013; Squillante, 2014) or non-substitutability (Pearce, 2001; Farahmand & Henderson, 2016) of candidate expressions; on combining different information sources using machine learning techniques (Ramisch et al., 2010; Tsvetkov & Wintner, 2014); or on the extraction of a specific subtype of MWE (Baldwin, 2005; Tu & Roth, 2011; Smith, 2014).

Statistical association remains an important component in virtually all of these approaches, but our understanding of the properties of different AMs and of other parameters such as the size of the co-occurrence context is still incomplete. Previous evaluation studies on collocation identification (cf. Section 3) leave a number of important gaps: (i) most studies evaluate only a small range of AMs (except for Pecina, 2005); (ii) the evaluation typically focuses on a specific subtype of MWE, so that different studies often report contradictory results; (iii) to date there has been no systematic analysis of the influence of source corpus, co-occurrence context and frequency threshold.

In this paper, we present the results of a large-scale evaluation study aiming to fill these gaps. Since we believe that AMs should not be tuned to a particular subtype of MWE, but rather capture a general "attraction" between words that may then be combined with more specific indicators such as syntactic flexibility, our gold standard is based on the broad and intuitive notion of lexical collocations (see Section 2). We draw on two different English collocation dictionaries in order to assess the robustness of evaluation results. We evaluate 20 association measures, 13 corpora, eight co-occurrence contexts and four frequency thresholds against the two collocation dictionaries. In order to be able to deal with the complexity of $20 \times 13 \times 8 \times 4 \times 2 = 16{,}640$ parameter combinations, we introduce an interactive Web-based viewer for evaluation graphs.[1]

2. Lexical collocations

Lexical collocations – salient co-occurrences of two lexical items (for a full definition and literature review, see Bartsch, 2004) – form a subtype of the larger family of lexicalized MWE and are notoriously difficult to delineate due to the fuzzy nature of the linguistic relation between their constituent words (which is sometimes described as a "habitual" combination, or simply defined mechanistically in terms of recurrence; e. g. Firth, 1957; Sinclair, 1966). In contrast to many other types of MWEs, lexical collocations are more susceptible to regular syntactic alternations. They are, furthermore, semantically transparent to a large degree, although many collocations carry additional, often domain-specific meanings. Examples of lexical collocations are *argue* + *plausibly*, *attempt* + *thwart* and *measure(s)* + *coercive*.

Our evaluation operationalizes lexical collocations as combinations of two lexical words. We assume that larger combinations such as *in a certain measure* can easily be recognized based on a two-word nucleus (*measure* + *certain*) by a lexicographer working with a corpus-based list of candidates, or generated by an automatic MWE extraction pipeline from the same nucleus.

Since the early days (Sinclair, 1966), the automatic identification of lexical collocations has relied primarily on the co-occurrence frequency of the words in question within a given context window. This window is typically defined as a surface span of 3 to 5 words to the left and right, but other span sizes have been employed in collocation studies ranging from one-word spans to entire sentences. Some authors define lexical collocations as a syntactic phenomenon (Bartsch, 2004), which suggests a co-occurrence context based on direct syntactic dependency relations, requiring a parsed corpus. After data extraction, researchers often apply a frequency threshold (e. g. $f \geq 5$) to filter the co-occurrence data. Finally, candidates are ranked according to a statistical association measure based on the joint and marginal frequencies of each word pair; more than 50 different measures have already been proposed in the literature (Pecina, 2005).

3. Related work

A typical approach to assessing the quality of a collocation extraction method is to extract a ranked list of collocation candidates and to manually identify the number of true

[1] Since some parameter combinations are not feasible (e. g. because a high frequency threshold does not leave enough candidates for the evaluation), the actual number of evaluation settings in our experiments and in the viewer is 12,860.

collocations among the n highest ranking candidates. This methodology is adopted, for example, by Seretan & Wehrli (2008) who compare their syntax-based extraction method with a window-based approach by manually annotating 250 candidates taken from the top 0%, 1%, 3%, 5% and 10% of the candidate lists for each of the four languages and two approaches they are looking at. Disadvantages of this evaluation methodology are that it is impossible to determine recall and that it is difficult to add new approaches or association measures to the evaluation since that would require additional manual annotation of the new candidate lists (consequently, Seretan & Wehrli, 2008 only report precision and focus on a single association measure, log-likelihood).

Another approach, introduced by Evert & Krenn (2001), focuses on a fixed set of true collocations and on the one hand allows us to determine precision and recall for arbitrarily large n-best lists of candidates and on the other hand makes it easy to add new association measures or extraction strategies to the evaluation. Results for this approach to evaluation of collocation extraction are usually given in the form of precision-recall curves. This is the approach taken, for example, by Pearce (2002) whose evaluation is based on 4,152 multiwords from the New Oxford Dictionary of English or by Pecina (2005) who evaluates a wide range of AMs based on more than 2,500 collocational dependency bigrams. Pecina & Schlesinger (2006) and Pecina (2010) also calculate the mean average precision for recall values between 0.1 and 0.9 to arrive at a single evaluation score. Kilgarriff et al. (2014) do not use precision-recall curves but report precision, recall and F_5-scores (giving more weight to recall) for different combinations of parameter settings such as AM, size of the n-best candidate lists or frequency thresholds based on 5,327 collocations for 102 headwords for English and 4,854 collocations for 100 headwords for Czech.

A related approach to evaluation treats collocation extraction as a classification task and uses a test set consisting of true collocations and non-collocations, reporting the usual metrics of precision, recall and F-score. This is the approach taken, for example, by Karan et al. (2012) who evaluate machine learning models for collocation extraction for Croatian based on a test set of 84 collocations and 450 non-collocations.

Finally, there are also approaches that focus on a qualitative evaluation instead of a quantitative one. Wermter & Hahn (2006), for example, compare ranked candidate lists by looking at the true positives and true negatives in the upper and lower half of the candidate lists.

Most of these studies focus on a particular system for collocation or MWE identification, on the comparison of different AMs and the effect of linguistic filters, or on optimizing extraction quality with the help of machine learning. To our knowledge, no systematic comparative study of the influence of source corpus and co-occurrence context has been published so far.

4. Data and methods

4.1 Gold standard

We adopt the evaluation methodology of Evert & Krenn (2001) and Pecina (2005), using precision-recall graphs in order to visualize and compare the distribution of true positives in candidate lists ranked according to different AMs. As has been explained in Section 2, lexical collocations are operationalized as pairs of lexical words (nouns, verbs,

adjectives and adverbs). Since most such collocations are combinations of lexemes rather than specific word forms, all word pairs are lemmatized. We do not distinguish between homographs with different parts of speech (e. g. the noun *attempt* vs. verb *to attempt*) because one of the two sources for our gold standard does not provide POS information.[2]

Because of the wide scope of our study and the large number of parameter combinations to be considered, manual annotation of candidate sets extracted from the corpus – as recommended by Evert & Krenn (2001) – is not feasible. Instead, we follow Pearce (2002) in using a fixed set of known collocations as a gold standard. We obtained this gold standard from two specialized collocation dictionaries:

BBI = The BBI Combinatory Dictionary of English (Benson et al., 1986);
OCD = Oxford Collocations Dictionary for students of English, 2nd edition (McIntosh et al., 2009).

Since BBI is not available in machine-readable form, we selected a set of 203 node words based on various criteria (words sampled from different frequency bands, words known to have interesting collocational patterns, at least 4 collocates in the two dictionaries). For each of the 203 nodes, all lexical words were manually transcribed from the corresponding entries in BBI and lemmatized.

measure I *n.* 1. a cubic; dry; liquid; metric ~ 2. a tape ~ 3. in a certain ~ (in large ~) 4. (misc.) for good ~ ('as smt. extra'); made to ~ ('custom-made'); to take smb.'s ~ ('to evaluate smb.') (see also **measures**)
measure II *v.* 1. (d; tr.) to ~ against (to ~ one's accomplishments against smb. else's) 2. (P; intr.) the room ~s twenty feet by ten
measures *n.* 1. to carry out, take ~ 2. coercive; compulsory; draconian; drastic, harsh, stern, stringent; emergency; extreme, radical; preventive, prophylactic; safety, security; stopgap, temporary ~ 3. ~ to + inf. (we took ~ to insure their safety) 4. ~ against (to take ~ against smuggling)

Figure 1: BBI entries corresponding to the node lemma *measure* in our gold standard

Consider the lemma *measure* as an example. Since we do not distinguish between different POS categories, collocates are collected from three entries in the BBI dictionary (for the noun *measure*, the verb *measure* and the plural noun *measures*), as shown in Figure 1. Our annotators identified 26 lemmas of lexical words in these entries, yielding the following collocates of *measure* in the BBI gold standard: *carry, certain, coercive, compulsory, cubic, draconian, drastic, dry, emergency, extreme, good, harsh, liquid, make, metric, preventive, prophylactic, radical, safety, security, stern, stopgap, stringent, take, tape, temporary.*

The corresponding OCD collocations were extracted from an electronic version of the dictionary, using the same strategy as Uhrig & Proisl (2012). In this way, we found a total of 2,845 lexical collocations for our 203 node lemmas in the BBI, and 18,545 in the OCD. We refer to these sets as the BBI and OCD gold standard below.

[2] A second reason is that the Web1T5 n-gram database does not include POS tagging; application of an off-the-shelf tagger is impossible because the underlying text corpus is not publicly available.

BBI was selected in a previous study (Bartsch & Evert, 2014) as a dictionary dating from the pre-corpus age. Unlike more recent collocation dictionaries, it can safely be assumed to be free of any bias in favour of a particular corpus or collocation extraction method. There are some limitations – due to the time of its compilation, its relatively small size and scope, as well as the heterogeneity of entries[3] – which have to be taken into consideration when interpreting the evaluation results.

4.2 Corpus data and parameters

We extracted co-occurrence data from the 13 corpora listed in Table 1, ranging in size from small, relatively clean corpora such as the British National Corpus (BNC) of 100 million words to huge Web corpora of up to 16 billion words (joint Web corpus = ENCOW + WebBase + ukWaC + Wackypedia). The corpora cover a wide diversity of text types: a balanced sample (BNC), movie subtitles (DESC), newspaper data (Gigaword), encyclopaedia articles (Wackypedia), Web corpora (ukWaC, WebBase, UKCOW, ENCOW). In addition, we included n-gram databases derived from Web text (Web1T5) and scanned books (Google Books), which can also be used to obtain co-occurrence data (Evert, 2010). All corpora except for Web1T5 include POS tagging and lemmatization.

Corpus	**Size**
British National Corpus (BNC)	0.1 G
English movie subtitles (DESC)	0.1 G
Wackypedia subset (WP500)	0.2 G
Wackypedia (Wiki)	1 G
ukWaC	2 G
Gigaword newspaper corpus	2 G
WebBase	3 G
UKCOW	4 G
ENCOW	10 G
Joint Web	16 G
Google Books BrE	50 G
Google Books	500 G
Google Web 1T5	1000 G

Table 1: Source corpora for the evaluation study. Sizes are specified in billion tokens

We extracted candidate collocations for the 203 node words using different co-occurrence contexts:

- direct syntactic relations;
- surface span of 1, 2, 3, 5 and 10 words;[4]
- sentence context.

We used the efficient and robust C&C parser (Clark & Curran, 2004) to extract syntactic dependencies from all corpora. For Google Books, we used the dependency bigrams

[3] In addition to lexical collocations proper, the BBI entries include phenomena ranging from fixed multiword units to combinations that might rather be described as colligations.

[4] Following Evert (2008), we denote these spans as L1/R1, L2/R2, etc. For example, a L2/R2 span includes two words to the left and two words to the right of each occurrence of the node word.

included in the database; syntactic context is not available for the Web1T5 n-grams. For surface spans, care was taken to obtain valid co-occurrrence counts and marginal frequencies as mandated by Evert (2008), using the UCS toolkit.[5] Note that 5- and 10-word spans are not available for the Google Books and Web 1T5 n-grams. In order to keep the amount of data manageable, potential collocates were restricted to a set of 37,437 general English words.[6] Even so, sets of up to five million candidate pairs were obtained for the 203 node lemmas, depending on corpus and context size (cf. Table 2). Optionally, frequency thresholds were used to pre-filter the candidates.

Candidate sets were then ranked according to 20 different association measures. In addition to measures recommended by Evert (2008), we included the asymmetric ΔP that has recently become popular in the corpus linguistics community (Gries, 2013). We evaluated the "forward" $\Delta P_{2|1}$ and the "backward" $\Delta P_{1|2}$ version of the measure, as well as two symmetrical variants. See Appendix A for a complete listing with equations and references.

4.3 Evaluation methodology

Like Evert & Krenn (2001) and Pecina (2005), we pool the candidate collocations extracted for all 203 nodes into a single set (for a given combination of corpus, co-occurrence context and frequency threshold), which is then ranked according to one of the 20 AMs. In addition, candidates are marked as true positives (TP) or false positives (FP) by comparison with either the BBI or the OCD gold standard.

After setting a cutoff threshold to obtain an n-best list of highest-ranked candidates, we compute precision (P, the percentage of TPs among the n candidates) and recall (R, the percentage of all TPs in the gold standard found in the n-best list) as quantitative evaluation criteria. The number n of candidates is chosen arbitrarily to trade off between high precision (short n-best lists) and high recall (long n-best lists). As proposed by Evert & Krenn (2001), we visualize this trade-off by plotting precision against recall for all possible n. An example can be seen in Figure 2 for the BNC corpus, syntactic context, and BBI as gold standard. Such P/R graphs allow a direct and detailed comparison of different AMs. For example, the solid blue line in Figure 2 shows that a ranking according to t-score (t) achieves a recall of 10% of the BBI gold standard (i. e. 285 of the 2,845 BBI collocations have been found) at a precision of 20% (i. e. one in five candidates in the n-best list is a true positive). The coverage of 91.6% shown at the top of the plot is the proportion of BBI collocations found among the full set of 374,239 candidates extracted from the BNC; this coverage corresponds to the highest recall value that can be reached on this data set.

The "higher" a P/R graph is located in the plot, the better the ranking achieved by the corresponding association measure. However, sometimes P/R graphs of different measures intersect (e. g. $\Delta P_{2|1}$ and log-likelihood G^2 in Figure 2), making it difficult to determine an unambiguous ranking. A related problem of P/R graphs is that they allow a straightforward comparison of different association measures, but not of other parameters such

[5] http://www.collocations.de/software.html

[6] This word list comprises the lexical nodes and collocates found in BBI and OCD entries as well as all lexical words from the CUVplus dictionary (http://ota.ox.ac.uk/headers/2469.xml). Inflected forms were lemmatised using a heuristic mapping derived from the British National Corpus.

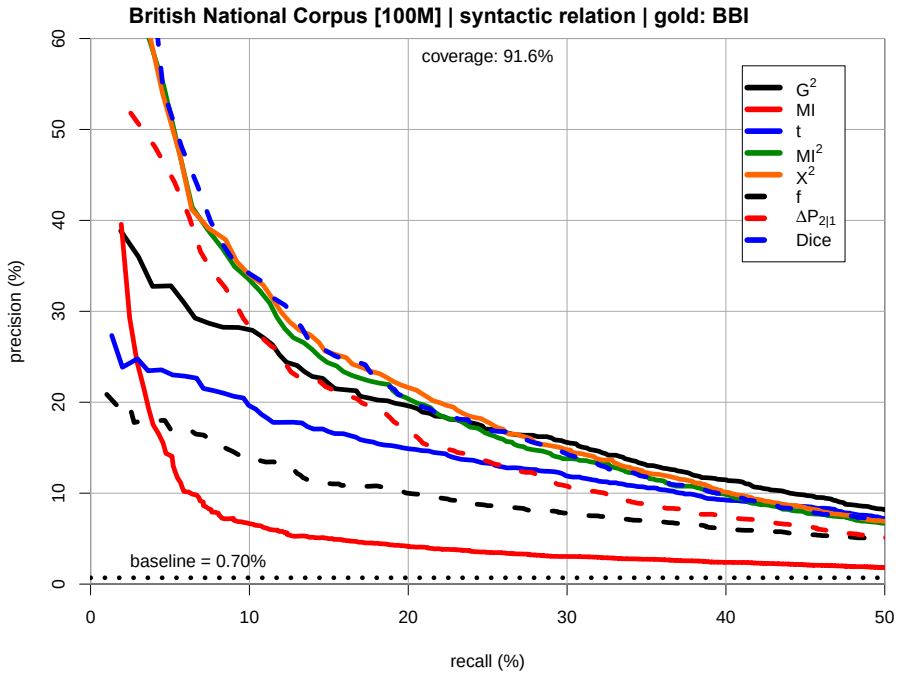


Figure 2: Precision-recall graphs for selected association measures evaluated against the BBI gold standard (British National Corpus, syntactic co-occurrence context, $f \geq 1$)

as source corpus and co-occurrence context (unless a single fixed association measure is chosen *a priori*).

For these reasons, it is desirable to introduce a composite evaluation criterion that summarizes the complete P/R graph into a single score. Following Pecina & Schlesinger (2006), we use average precision – corresponding to the area under a P/R graph – as a composite measure. Since recall points above 50% can only be achieved with unrealistically long n-best lists, we average precision values only up to 50% recall and refer to this composite measure as AP50.

5. Results

Figure 2 shows striking differences between association measures. Neither log-likelihood (G^2), which is popular in computational linguistics, nor t-score (t), which is popular in computational lexicography, achieve convincing performance. Mutual Information (MI) can only be described as abysmal, partly due to the lack of a frequency threshold for this data set.[7] The best – and almost indistinguishable – results are obtained by Pearson's chi-squared test (X^2), a heuristic variant of Mutual Information (MI^2) and the Dice coefficient.[8] In the composite ranking of association measures, X^2 takes first place with AP50 = 24.2%, followed by Dice with 24.0%. This is particularly surprising given the widely-accepted claim that G^2 is vastly superior to X^2 for collocation identification (Dunning, 1993).

A second striking observation is how much the evaluation results depend on which collocation dictionary is used as a gold standard, even though both are targeted at the same type

[7] As we will see below, frequency thresholds have little impact on the best-performing AMs, so it makes sense to present the basic findings here without a frequency threshold (i. e. $f \geq 1$).

[8] This is particularly relevant for users of the SketchEngine (Kilgarriff et al., 2004) which uses (a rescaled version of) the Dice coefficient for word sketches (Rychlý, 2008).

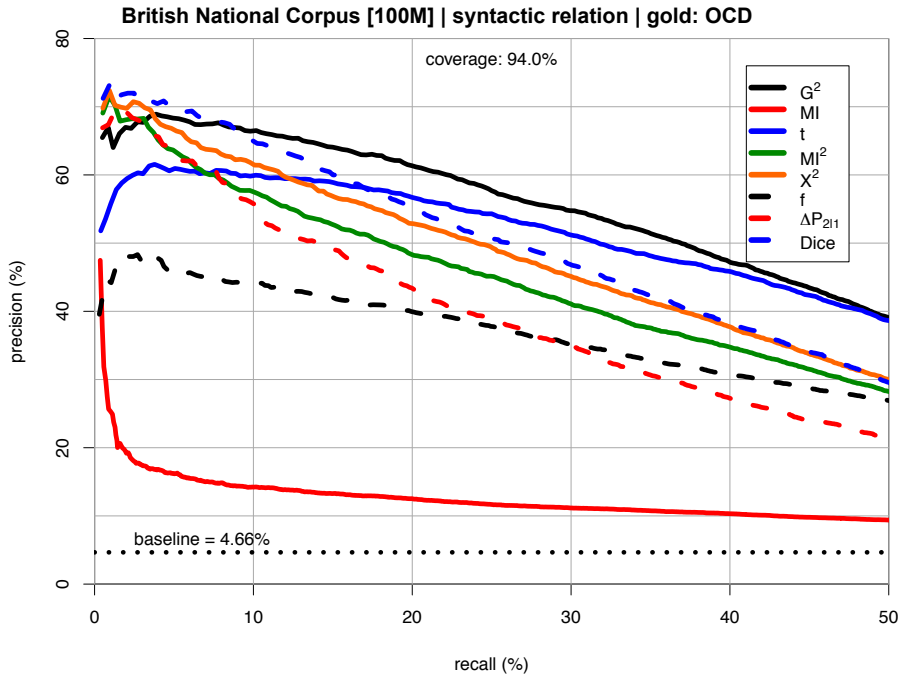


Figure 3: Precision-recall graphs for selected association measures evaluated against OCD gold standard (BNC, syntactic context, $f \geq 1$)

of users, i. e. foreign and second language learners. Figure 3 shows an entirely different ranking of the association measures, even though corpus and co-occurrence context are the same as in Figure 2: best results are now obtained by log-likelihood (G^2, AP50 = 56.8%) and t-score (t, AP50 = 52.5%).[9] These differences presumably reflect the more focused notion of lexical collocations underlying OCD, but also its bias towards the particular association measures used in the compilation of the dictionary.

Using AP50 as a composite evaluation criterion, we can now study the effects of the other parameters. For every combination of source corpus, co-occurrence context and frequency threshold, we selected the best performing association measure and used its AP50 value as an overall score. The left-hand panel of Figure 4 compares different co-occurrence contexts on the British National Corpus ($f \geq 1$). For both gold standards, smaller contexts achieve considerably better performance, and the best results are achieved if candidate pairs must occur in a direct syntactic relation. Similar plots for other corpora and frequency thresholds (not shown for reasons of space) reveal the same pattern, except for minimal differences (e. g. L1/R1 might be slightly better than L2/R2 if a frequency threshold is applied).

The right-hand panel of Figure 4 compares results obtained on different source corpora for the same two-word co-occurrence span (which is available for all 13 corpora), again without frequency threshold ($f \geq 1$). This chart shows a more intricate pattern. Summarizing, we find that:

1. Size matters: larger corpora of the same kind (WP500 vs. full Wiki; Web corpora) perform better. However, the corpus size has to be scaled up by a factor of 10 in order to achieve a notable improvement.

[9] AP50 values are also much higher overall for OCD than for BBI. This is to be expected, though, simply because of the much larger number of TPs in the OCD gold standard (6.5× as many as in BBI).

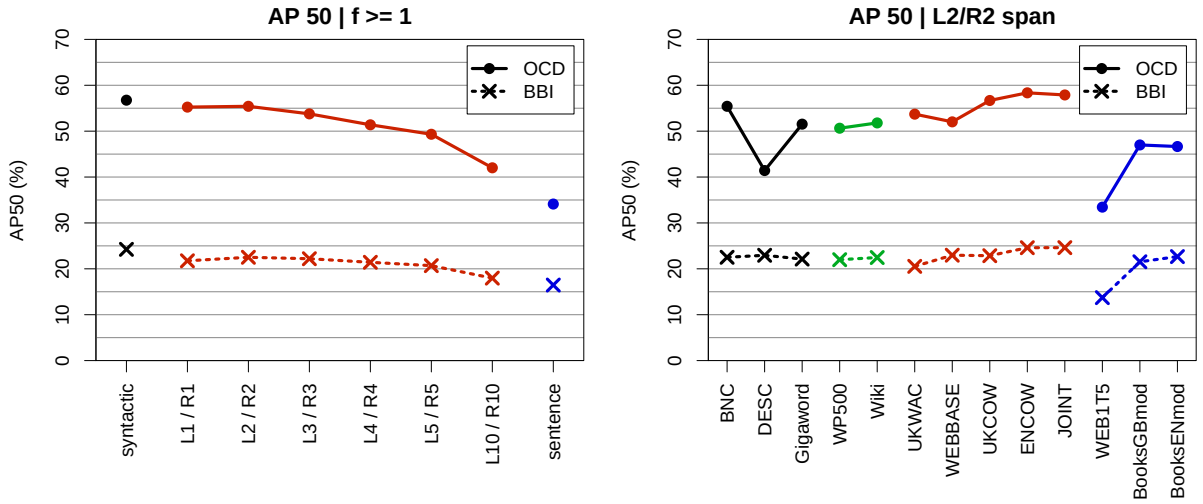


Figure 4: Left panel: Best AP50 scores achieved on the British National Corpus for different co-occurrence contexts. Right panel: Best AP50 scores achieved on different corpora with two-word co-occurrence span (L2/R2). In each case, the optimal AM has been selected

2. Clean, balanced samples (BNC) are better than large, messy Web corpora of the same size. The biggest Web corpora outperform the BNC, but this requires almost 100 times as much data (ENCOW: 10G words vs. BNC: 100M).
3. Movie subtitles (DESC), which are closer to spoken language and match psycholinguistic observations (New et al., 2007), perform better than the BNC against the BBI gold standard, but much worse when evaluated against OCD.[10]
4. Even though n-gram databases have been compiled from huge corpora (from 50 billion words for British GoogleBooks to 1 trillion words for Web1T5), they appear to be unsuitable for collocation identification.
5. There are some differences between the two gold standards, but the main observations hold equally well for BBI and OCD.

Again, similar plots for other co-occurrence contexts and frequency thresholds (not shown) always reveal the same pattern.

Figure 5 shows that there is virtually no interaction between the choice of AM and the other parameters (co-occurrence context and source corpus); similar patterns hold for the OCD gold standard and the other 15 AMs. The only exception is the combination of a frequency threshold with a small corpus, which improves the performance of MI (right panel). This has little practical relevance, though, because MI never comes close to the best-performing measures.

One of the most surprising results of our evaluation is the negligible impact of frequency thresholds: apparently, the statistical measures successfully weed out unreliable low-frequency candidates. Figure 6 compares a wide range of frequency thresholds on the BBI gold standard. The top panel shows that thresholds up to $f \geq 10$ only lead to a tiny

[10] One possibility is that OCD in particular is focused on British English as represented in the BNC, which provided the empirical basis for the first edition of the dictionary. British films account for only 10% of the DESC corpus and the subtitle files consistently use American spelling. This would also explain the lower performance of Gigaword (mostly U.S. newspapers) and WebBase (a Web corpus compiled in the U.S., while ukWaC and UKCOW only include Web pages from `.uk` domains).

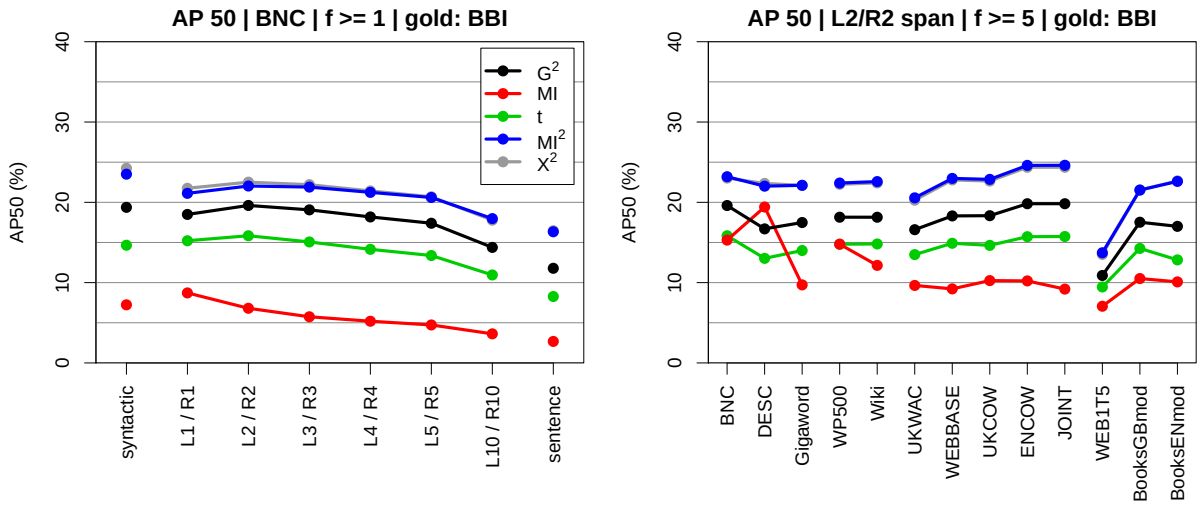


Figure 5: Co-occurrence context (left panel) and source corpus (right panel) do not interact with the choice of association measure. Illustrated for the BBI gold standard, the British National Corpus with $f \geq 1$ (left panel) and a two-word co-occurrence span with $f \geq 5$ (right panel)

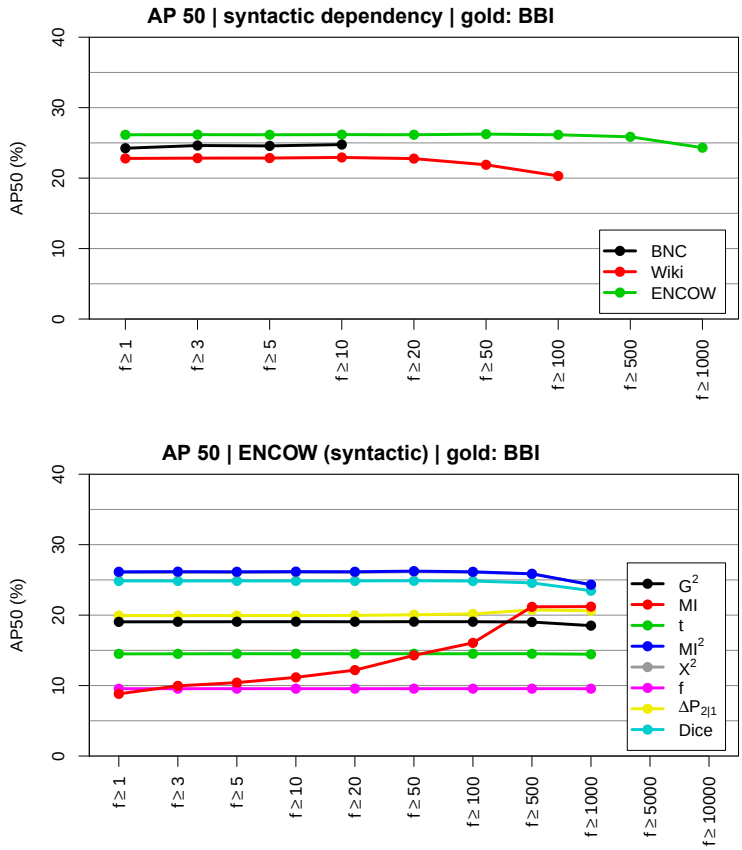


Figure 6: Effect of frequency thresholds on various corpora (top panel) and AMs (bottom panel), for syntactic context and BBI gold standard

$f \geq 1$		BBI				OCD			
corpus	n_{cand}	context	AM	AP50	coverage	context	AM	AP50	coverage
BNC	0.5M	syntactic	X^2	24.2	91.6	syntactic	G^2	56.8	94.0
DESC	0.3M	syntactic	MI_{conf}	24.6	80.9	syntactic	MI_{conf}	44.0	72.8
Gigaword	1.2M	L2/R2	X^2	22.1	97.6	L1/R1	G^2	52.3	95.6
WP500	0.5M	syntactic	X^2	22.6	92.2	L2/R2	G^2	50.6	92.8
Wiki	1.0M	syntactic	MI^2	22.8	97.0	L2/R2	G^2	51.8	97.4
ukWaC	1.4M	syntactic	MI^2	22.8	98.7	L1/R1	G^2	56.5	97.5
WebBase	1.7M	syntactic	MI^2	25.1	99.2	syntactic	G^2	54.2	99.5
UKCOW	1.9M	syntactic	MI^2	24.6	99.3	L1/R1	G^2	58.0	98.1
ENCOW	2.5M	syntactic	MI^2	26.1	99.7	L1/R1	G^2	**59.7**	98.7
Joint	2.8M	syntactic	MI^2	**26.4**	99.8	L1/R1	G^2	59.5	99.4
Web1T5	1.8M	L1/R1	MI^2	15.5	97.5	L1/R1	MI^3	37.1	97.9
BooksGB	0.9M	syntactic	MI^2	21.7	95.4	L1/R1	G^2	47.9	93.0
BooksEN	1.5M	syntactic	MI^2	22.8	96.1	syntactic	G^2	48.6	96.9

Table 2: Overview table of best evaluation result for each corpus against the BBI and OCD gold standard (coverage indicates the highest recall point that can be achieved by a given parameter combination)

improvement for the smallest corpus (BNC) and have no effect at all for larger corpora. The bottom panel shows that thresholds mainly help to counteract the low-frequency bias of the MI measure. All other AMs are unaffected, and even with a high threshold, MI remains well below the best-performing measures.

A detailed overview of the evaluation results is shown in Table 2. For each source corpus, the AP50 score achieved by the optimal co-occurrence context and association measure is shown, as well as the coverage of the respective gold standard. In order to indicate the amount of data processed, the second column (n_{cand}) shows how many million word pairs were extracted from each corpus for a two-word surface span (L2/R2).

6. An interactive viewer

Any paper-length treatment of association measures is faced with the problem that the large number of parameter settings makes it impossible to give the reader a full overview of their influence in all possible combinations. For example, in Section 5 we showed the influence of source corpus and co-occurrence context based on AP50 values achieved by the best AM in each case. Such summary charts hide important details of the trade-off between precision and recall (e. g. some applications may prefer a measure that achieves very high precision even if recall is only 10%); they also cannot show whether the overall shape of a P/R graph remains stable across different parameter settings. Even so, space constraints make it impossible to provide comprehensive evidence for all our observations within this paper (e. g. the similar effect of parameters for both gold standards, and in particular the consistently small impact of frequency thresholds). Figures 2, 3 and 5 can only show a small selection of the 20 association measures included in our evaluation. While correlations between the rankings for different association measures (Figure 7) provide an objective criterion for a principled selection – each group of almost perfectly correlated measures (Dice and Jaccard; chi-squared and z-score; MI, relative risk and two variants of the odds ratio) can be represented by one member – there are only few such strong correlations. Moreover, even measures with correlation $\rho > .99$ (e. g. log-likelihood,

t-score and chi-squared) sometimes achieve substantially different results in the evaluation (cf. Figures 2 and 3) and should not be grouped together.[11]

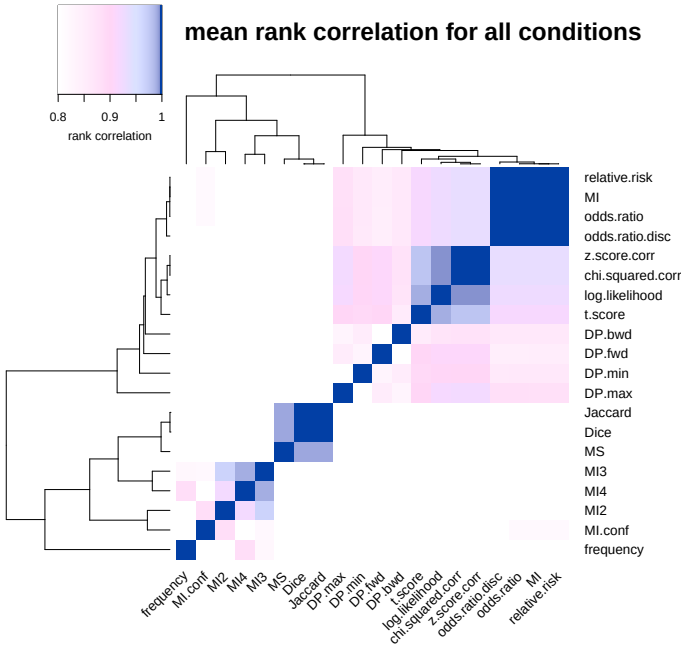


Figure 7: Spearman rank correlation of different association measures, averaged over all experimental conditions

In order to remedy these problems, an interactive viewer was created to complement the present paper and allow the reader to explore the influence of the parameters discussed above as well as their interactions.

Since the extraction of collocations candidates from large corpora is a very time-consuming process,[12] all evaluation graphs have been pre-computed using the statistical software R and exported as a set of JSON files. These files are processed further, filtered and served through a REST API with the help of Perl scripts. The front-end of the viewer is written in JavaScript and provides a set of sliders and buttons to modify the following parameters:

1. gold standard (BBI vs. OCD2);
2. corpus (see Section 4);
3. co-occurrence context (syntactic relation, various spans, whole sentence);
4. frequency threshold ($f \geq 1, 5, 50, 1000$);[13]
5. association measures (select measures to be displayed at the same time).

[11] We believe that this surprising observation is connected to the fact that rank correlations were computed over very large data sets comprising a million candidate pairs and more. Crucial differences between the rankings of the relatively small number of TPs, which affect the evaluation scores directly, are lost among the rankings of many irrelevant FPs. This example shows clearly how difficult and counter-intuitive the interpretation of correlation coefficients can be.

[12] The extraction procedure ran for several weeks on a high-end server (16 cores and 256 GiB RAM).

[13] Since the sizes of the corpora used in this study vary by several orders of magnitude, the range of thresholds is quite wide. Keep in mind that a threshold of $f \geq 5$ in the BNC (100M words) corresponds to a threshold of $f \geq 500$ in UKCOW (10G words). It might be profitable to explore thresholds relative to corpus size in future work.

The full P/R graphs for the chosen parameter settings are displayed to the user and dynamically updated as the sliders are moved. Additionally, coverage and composite AP50 scores are shown. The viewer software will be made available under an open-source license, including the R code for exporting suitable JSON data. An online version for the evaluation reported here can be accessed at http://www.collocations.de/eviewalation/.

7. Conclusion

The systematic evaluation of different association measures, source corpora, co-occurrence contexts and frequency thresholds in a collocation extraction tasks fills important gaps in the current state of research into AMs and MWE identification.

We were able to show that the carefully sampled British National Corpus is superior to comparably-sized messy Web corpora for the identification of lexical collocations. However, sufficiently large Web corpora (close to 10 billion words) achieve similarly good or even better results than the BNC. Concerning the co-occurrence context, it was shown that small spans deliver more accurate information than larger contexts and the most restricted context, i. e. syntatic dependency, is almost always the best choice. Contrary to widespread assumptions, frequency thresholds have very little effect except to counteract the low-frequency bias of the MI measure.

The choice of an optimal AM is a more intricate problem, which depends not only on the type of MWE to be identified (lexical collocations in our case) but also on the specific definition of this MWE type, embodied by the two different collocation dictionaries (BBI and OCD) in our study. For BBI, Pearson's chi-squared statistic (X^2) and MI^2 yield the best results; for OCD, log-likelihood (G^2) is the optimal AM. Fortunately, performance differences between AMs do not interact with the other parameters: in all cases, very large Web corpora and small co-occurrence contexts produce the best results. It is thus valid to optimize AMs independently of these parameters in future research.

Since the present evaluation builds entirely on English data, no conclusions regarding other languages can be drawn and further research is required. Nonetheless, it is to be expected that collocation extraction for languages with a richer morphology and/or a freer word order, e. g. German or Russian, will benefit from larger window sizes and in particular from dependency parsing. This would be in line with the results by Ivanova et al. (2008) and Ambati et al. (2012).

8. References

Ambati, B.R., Reddy, S. & Kilgarriff, A. (2012). Word Sketches for Turkish. In N. Calzolari, K. Choukri, T. Declerck, M.U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis (eds.) *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, TR: European Language Resources Association, pp. 2945–2950. URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/585_Paper.pdf.

Baldwin, T. (2005). Deep lexical acquisition of verb-particle constructions. *Computer Speech and Language*, 19, pp. 398–414.

Bartsch, S. (2004). *Structural and Functional Properties of Collocations in English*. Tübingen: Narr.

Bartsch, S. & Evert, S. (2014). Towards a Firthian Notion of Collocation. In A. Abel & L. Lemnitzer (eds.) *Vernetzungsstrategien, Zugriffsstrukturen und automatisch ermittelte Angaben in Internetwörterbüchern*, number 2/2014 in OPAL – Online publizierte Arbeiten zur Linguistik. Mannheim: Institut für Deutsche Sprache, pp. 48–61. URL http://ids-pub.bsz-bw.de/frontdoor/index/index/docId/2402.

Benson, M., Benson, E. & Ilson, R. (1986). *The BBI Combinatory Dictionary of English: A Guide to Word Combinations*. Amsterdam, New York: John Benjamins.

Church, K., Gale, W.A., Hanks, P. & Hindle, D. (1991). Using Statistics in Lexical Analysis. In *Lexical Acquisition: Using On-line Resources to Build a Lexicon*. Lawrence Erlbaum, pp. 115–164.

Church, K.W. & Hanks, P. (1990). Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, 16(1), pp. 22–29.

Clark, S. & Curran, J.R. (2004). Parsing the WSJ using CCG and Log-Linear Models. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*. Barceona, Spain, pp. 104–111.

Daille, B. (1994). *Approche mixte pour l'extraction automatique de terminologie: statistiques lexicales et filtres linguistiques*. Ph.D. thesis, Université Paris 7.

Dunning, T.E. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1), pp. 61–74.

Evert, S. (2008). Corpora and collocations. In A. Lüdeling & M. Kytö (eds.) *Corpus Linguistics. An International Handbook*, chapter 58. Berlin, New York: Mouton de Gruyter, pp. 1212–1248.

Evert, S. (2010). Google Web 1T5 N-Grams Made Easy (but not for the computer). In *Proceedings of the 6th Web as Corpus Workshop (WAC-6)*. Los Angeles, CA, pp. 32–40.

Evert, S. & Krenn, B. (2001). Methods for the Qualitative Evaluation of Lexical Association Measures. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*. Toulouse, France, pp. 188–195. URL http://www.aclweb.org/anthology/P01-1025.

Farahmand, M. & Henderson, J. (2016). Modeling the Non-Substitutability of Multiword Expressions with Distributional Semantics and a Log-Linear Model. In *Proceedings of the 12th Workshop on Multiword Expressions*. Berlin, Germany, pp. 61–66.

Firth, J.R. (1957). A synopsis of linguistic theory 1930–55. In *Studies in linguistic analysis*. Oxford: The Philological Society, pp. 1–32.

Gries, S.T. (2013). 50-something years of work on collocations: What is or should be next .... *International Journal of Corpus Linguistics*, 18(1), pp. 137–165.

Ivanova, K., Heid, U., Schulte im Walde, S., Kilgarriff, A. & Pomikalek, J. (2008). Evaluating a German Sketch Grammar: A Case Study on Noun Phrase Case. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis & D. Tapias (eds.) *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Marrakech, MA: European Language Resources Association, pp. 2101–2107. URL http://www.lrec-conf.org/proceedings/lrec2008/pdf/537_paper.pdf.

Johnson, M. (1999). Confidence intervals on likelihood estimates for estimating association strengths. Unpublished technical report.

Karan, M., Snajder, J. & Basic, B.D. (2012). Evaluation of Classification Algorithms and Features for Collocation Extraction in Croatian. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012), Is-*

tanbul, Turkey, May 23-25, 2012. pp. 657–662. URL http://www.lrec-conf.org/proceedings/lrec2012/summaries/796.html.

Katz, G. & Giesbrecht, E. (2006). Automatic Identification of Non-Compositional Multi-Word Expressions using Latent Semantic Analysis. In *Proceedings of the ACL Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties (MWE 2006)*. Sydney, Australia: Association for Computational Linguistics, pp. 12–19.

Kiela, D. & Clark, S. (2013). Detecting Compositionality of Multi-Word Expressions using Nearest Neighbours in Vector Space Models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*. Seattle, WA, pp. 1427–1432.

Kilgarriff, A., Rychlý, P., Jakubícek, M., Kovár, V., Baisa, V. & Kocincová, L. (2014). Extrinsic Corpus Evaluation with a Collocation Dictionary Task. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014*. pp. 545–552. URL http://www.lrec-conf.org/proceedings/lrec2014/summaries/52.html.

Kilgarriff, A., Rychlý, P., Smrz, P. & Tugwell, D. (2004). The Sketch Engine. In G. Williams & S. Vessier (eds.) *Proceedings of the 11th EURALEX International Congress*. Lorient, FR: Université de Bretagne-Sud, Faculté des lettres et des sciences humaines, pp. 105–115.

McIntosh, C., Francis, B. & Poole, R. (eds.) (2009). *Oxford Collocations Dictionary for students of English*. Oxford University Press, 2nd edition.

New, B., Brysbaert, M., Véronis, J. & Pallier, C. (2007). The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics*, 28, pp. 661–667.

Nissim, M. & Zaninello, A. (2013). Modeling the Internal Variability of Multiword Expressions Through a Pattern-based Method. *ACM Transactions on Speech and Language Processing*, 10(2), pp. 7:1–7:26.

Pearce, D. (2001). Synonymy in Collocation Extraction. In *Proceedings of the NAACL Workshop on WordNet and Other Lexical Resources*. Pittsburgh, PA.

Pearce, D. (2002). A Comparative Evaluation of Collocation Extraction Techniques. In *Proceedings of the Third International Conference on Language Resources and Evaluation, LREC 2002, May 29-31, 2002, Las Palmas, Canary Islands, Spain*. URL http://www.lrec-conf.org/proceedings/lrec2002/pdf/169.pdf.

Pecina, P. (2005). An Extensive Empirical Study of Collocation Extraction Methods. In *Proceedings of the ACL Student Research Workshop*. Ann Arbor, MI, pp. 13–18.

Pecina, P. (2010). Lexical association measures and collocation extraction. *Language Resources and Evaluation*, 44(1–2), pp. 137–158. URL http://dx.doi.org/10.1007/s10579-009-9101-4.

Pecina, P. & Schlesinger, P. (2006). Combining Association Measures for Collocation Extraction. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL 2006), Poster Sessions*. Sydney, Australia: ACL, pp. 651–658.

Pedersen, T. & Bruce, R. (1996). What to Infer from a Description. Technical Report 96-CSE-04, Southern Methodist University, Dallas, TX.

Ramisch, C., Villavicencio, A. & Boitet, C. (2010). mwetoolkit: a Framework for Multiword Expression Identification. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*. Valetta, Malta: European Language Resources Association.

Rychlý, P. (2008). A Lexicographer-Friendly Association Score. In P. Sojka & A. Horák (eds.) *Proceedings of Recent Advances in Slavonic Natural Language Processing (RASLAN 2008)*. Brno: Masaryk University, pp. 6–9.

Seretan, V. & Wehrli, E. (2008). Multilingual collocation extraction with a syntactic parser. *Language Resources and Evaluation*, 43(1), pp. 71–85. URL http://dx.doi.org/10.1007/s10579-008-9075-7.

Sinclair, J.M. (1966). Beginning the Study of Lexis. In C.E. Bazell, J.C. Catford, M.A.K. Halliday & R.H. Robins (eds.) *In Memory of J. R. Firth*. London: Longmans, pp. 410–430.

Smith, A. (2014). Breaking Bad: Extraction of Verb-Particle Constructions from a Parallel Subtitles Corpus. In *Proceedings of the 10th Workshop on Multiword Expressions (MWE)*. Gothenburg, Sweden, pp. 1–9.

Squillante, L. (2014). Towards an Empirical Subcategorization of Multiword Expressions. In *Proceedings of the 10th Workshop on Multiword Expressions (MWE)*. Gothenburg, Sweden, pp. 77–81.

Tsvetkov, Y. & Wintner, S. (2014). Identification of Multiword Expressions by Combining Multiple Linguistic Information Sources. *Computational Linguistics*, 40(2), pp. 449–468.

Tu, Y. & Roth, D. (2011). Learning English Light Verb Constructions: Contextual or Statistical. In *Proceedings of the ACL 2011 Workshop on Multiword Expressions: From Parsing and Generation to the Real World*. Portland, OR.

Uhrig, P. & Proisl, T. (2012). Less hay, more needles – using dependency-annotated corpora to provide lexicographers with more accurate lists of collocation candidates. *Lexicographica*, 28(1), pp. 141–180.

Villada Moirón, M.B. (2005). *Data-driven identification of fixed expressions and their modifiability*. Ph.D. thesis, Rijksuniversiteit Groningen.

Wermter, J. & Hahn, U. (2006). You Can't Beat Frequency (Unless You Use Linguistic Knowledge) – A Qualitative Evaluation of Association Measures for Collocation and Term Extraction. In *ACL 2006, 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*. Sydney, Australia. URL http://aclweb.org/anthology/P06-1099.

Yates, F. (1934). Contingency tables involving small numbers and the χ^2 test. *Supplement to the Journal of the Royal Statistical Society*, 1, pp. 217–235.

Yazdani, M., Farahmand, M. & Henderson, J. (2015). Learning Semantic Composition to Detect Non-compositionality of Multiword Expressions. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*. Lisbon, Portugal, pp. 1733–1742.

A. Association measures

The listing below details the complete list of statistical association measures included in our evaluation. Equations are specified using the notation of Evert (2008):

expected frequencies

	collocate	¬collocate
node	$E_{11} = \frac{R_1 C_1}{N}$	$E_{12} = \frac{R_1 C_2}{N}$
¬node	$E_{21} = \frac{R_2 C_1}{N}$	$E_{22} = \frac{R_2 C_2}{N}$

observed frequencies

	collocate	¬collocate	
node	O_{11}	O_{12}	$= R_1$
¬node	O_{21}	O_{22}	$= R_2$
	$= C_1$	$= C_2$	$= N$

O_{ij} = contingency table of observed frequencies
O_{11} = observed co-occurrence frequency
E_{ij} = contingency table of expected frequencies
E_{11} = expected co-occurrence frequency
R_i = row sums of the contingency table
R_1 = marginal frequency of node
C_j = column sums of the contingency table
C_1 = marginal frequency of collocate
N = sample size

- **log-likelihood** (Dunning, 1993)

$$G^2 = 2 \sum_{ij} O_{ij} \log \frac{O_{ij}}{E_{ij}}$$

- **chi-squared** test (with Yates's correction)

$$X^2 = \frac{N\left(|O_{11}O_{22} - O_{12}O_{21}| - \frac{N}{2}\right)^2}{R_1 R_2 C_1 C_2}$$

- **t-score** (Church et al., 1991)

$$t = \frac{O_{11} - E_{11}}{\sqrt{O_{11}}}$$

- **z-score** (with Yates's (1934) correction)

$$z = \frac{O_{11} - E_{11} \pm \frac{1}{2}}{\sqrt{E_{11}}}$$

- co-occurrence **frequency**

$$f = O_{11}$$

- **mutual information** (Church & Hanks, 1990)

$$\mathrm{MI} = \log_2 \frac{O_{11}}{E_{11}}$$

- **MI**^{k} (Daille, 1994)

$$\mathrm{M}^k = \log_2 \frac{(O_{11})^k}{E_{11}} \quad \text{for } k = 2, 3, 4$$

- **conservative MI** (Johnson, 1999)

$$\mathrm{MI}_{\text{conf},\,\alpha} = \log_2 \min \left\{ \mu > 0 \;\middle|\; e^{-\mu E_{11}} \sum_{k=O_{11}}^{\infty} \frac{(\mu E_{11})^k}{k!} \geq 10^{-5} \right\}$$

- **Dice** coefficient

$$\mathrm{Dice} = \frac{2O_{11}}{R_1 + C_1}$$

- **Jaccard** coefficient

$$\mathrm{Jaccard} = \frac{O_{11}}{O_{11} + O_{12} + O_{21}}$$

- **minimum sensitivity** (Pedersen & Bruce, 1996)

$$\mathrm{MS} = \min \left\{ \frac{O_{11}}{R_1}, \; \frac{O_{11}}{C_1} \right\}$$

- log **odds ratio** (with optional discounting)

$$\log \theta = \log \frac{O_{11} O_{22}}{O_{12} O_{21}}$$
$$\log \theta_{\text{disc}} = \log \frac{(O_{11} + \frac{1}{2})(O_{22} + \frac{1}{2})}{(O_{12} + \frac{1}{2})(O_{21} + \frac{1}{2})}$$

- log **relative risk**

$$r = \log \frac{O_{11} C_2}{O_{12} C_1}$$

- forward or backward **Delta P** (Gries, 2013)

$$\Delta P_{2|1} = \frac{O_{11}}{R_!} - \frac{O_{21}}{R_2}$$
$$\Delta P_{1|2} = \frac{O_{11}}{C_!} - \frac{O_{12}}{C_2}$$

- **symmetrical Delta P**

$$\Delta P_{\min} = \min \left\{ \Delta P_{2|1}, \; \Delta P_{1|2} \right\}$$
$$\Delta P_{\max} = \max \left\{ \Delta P_{2|1}, \; \Delta P_{1|2} \right\}$$

B. Set of node lemmas

The following 203 lemmas were used as node words in our evaluation experiments: *abortion, accountant, achievement, act, advantage, affair, allocation, amusement, appetite, argue, art, artery, assault, attempt, authority, back, bag, balance, ban, basket, battery, battle, beach, bean, beat, beef, beg, bend, bent, biology, blast, bomb, bone, boot, break, broth, brother, bulb, bulletin, burst, cancer, carbon, care, cell, chain, chance, change, character, check, chess, chief, child, citizen, claim, clean, cleaner, cliff, close, cold, collaboration, commitment, confinement, consequence, cooking, cord, cotton, crime, criminal, cry, cupboard, cut, decision, deny, diet, director, door, draft, dressing, drunk, earth, elbow, enforce, environment, error, examination, executive, fee, feedback, fellowship, fever, fin, finger, fist, fitness, flow, fly, force, forgive, foundation, fund, funeral, garlic, gas, gender, gene, get, go, goal, gown, harm, havoc, head, health, heater, heating, heaven, heed, hernia, high, hotel, humanity, hygiene, injury, inmate, insight, intercourse, jam, juice, kick, know, lapse, letter, light, line, majority, malice, maniac, measure, measurement, meat, mechanic, membrane, minister, mother, move, nail, negligence, open, paint, pan, pardon, pay, pie, pipe, place, plague, plant, plantation, plead, pool, power, prime, problem, progress, query, question, quilt, race, radio, range, remark, representation, resuscitation, right, sauce, say, sentence, set, shake, shotgun, shoulder, soda, spirit, state, steel, storm, syllable, take, thirst, time, toss, trample, trial, triangle, tune, ulcer, universal, vacuum, vein, way, weapon, wiper, wire*

Toward Linked Data-native Dictionaries

Jorge Gracia[1], Ilan Kernerman[2], Julia Bosque-Gil[1]

[1]Universidad Politécnica de Madrid, Campus de Montegancedo sn, Boadilla del Monte 28660 Madrid, Spain
[2] K Dictionaries Ltd, 8 Nahum Hanavi Street, 6350310 Tel Aviv, Israel
E-mail: jgracia@fi.upm.es, ilan@kdictionaries.com, jbosque@fi.upm.es

Abstract

The ways in which dictionaries are compiled and used have evolved dramatically in recent years owing to the processes of digitization. This evolution has found in the Web an optimal means to empower the visibility and usability of dictionaries. In this context, we witness nowadays increasing interest in the interoperability of linked data (LD) technologies for the development and representation of lexicographic data on the Web.
In this paper we propose the notion of *LD-native dictionaries* as a natural next step in the evolution of lexicography. These future dictionaries could be LD-native and, as such, graph-based. Their nodes are not dependent on any internal hierarchy and are uniquely identified at a Web scale. We analyze the advantages of such an approach and identify its possible impact on the dictionary representation, compilation, and usage processes. Some challenges related to interoperability and data aggregation issues are also discussed.

Keywords: linguistic linked data; linked data-native dictionaries; e-lexicography

1. Introduction

The dictionary concept has been evolving over the last generation alongside the advent of technology and digitization of modern life, both as regards the lexicographic compilation process and the dictionary's media, dissemination and forms of usage. In the first wave of the electronic era (1990's), dictionaries usually remained little more than that same old 'book of words' in new e-dress(es), but gradually more e-features were introduced, such as advanced search modes, dictionary as corpus, morphological connections, integrating with other language software, embedding audio and images, and so on.

This logical evolution has found in the Web an optimal means to empower the visibility and usability of dictionaries. In particular, we witness nowadays increasing interest in the interoperability with linked data (LD) technologies to develop and represent lexicographic data on the Web. LD refers to a set of best practices for exposing, sharing and connecting data on the Web (Bizer et al., 2009). In short, the LD paradigm requires that *resources* be represented on the Web via URIs (Unique Resource Identifiers) and that, once a resource is accessed via its URI, useful information can be obtained, along with links to other resources. The basic mechanism that enables this is the Resource Description Framework (RDF),[1] which follows the *subject-object-predicate* pattern. The result is a vast graph of linked resources on the Web, whose nodes can be practically anything, including documents, people, physical objects and abstract concepts (such as lexical entries or any other entity that lexicography needs to model).

Some of the advantages of using LD to represent lexicographic content have already been reported in the literature (e.g., Klimek & Brümmer, 2015; Declerck et al., 2015; Bosque-Gil et al., 2016a) and the number of initiatives applied toward the conversion of proprietary dictionary formats to LD continues to grow (e.g., Bosque-Gil et al., 2016b; Parvizi et al., 2016). Also the community of ontology lexica has shown interest in LD for lexicography and started discussing best practices and modelling issues on this topic (Bosque-Gil et al., 2017) in the context of the W3C Ontolex community group.[2]

[1] http://w3.org/TR/rdf11-primer/
[2] https://www.w3.org/community/ontolex/

As a natural next step, we envisage dictionaries that are born and evolve dynamically on the Web. These will not be (only) the result of transforming lexicographic data from previous electronic formats into LD, but will ensue from compiling dictionaries as LD from scratch. Thus, such future dictionaries are LD-native and, as such, graph-based. Their nodes are not dependent on any internal hierarchy and are uniquely identified at a Web scale. This will enable the enhancement of a vast network of interconnected linguistic elements through semantically well-defined lexical, syntactic, pragmatic, etc. relations, through which lexicographers and users navigate to edit, query, or aggregate data. Links to other lexical resources, including other dictionaries, would thus be quickly and naturally established.

In this paper we analyze this vision and its advantages as compared to a more traditional tree-based view of lexicographic data. We also explore its impact on the editorial process, both on the content itself and on the way lexicographers work. Some challenges about interlinking and data aggregation are discussed as well.

The rest of the paper is organized as follows. In Section 2 the vision of LD-native dictionaries is presented. Then in Section 3 the impact of the LD-native dictionaries notion on the editorial process is discussed. Some challenges related to data integration are presented in Section 4. Finally, our conclusions are presented in Section 5.

2. The vision of LD-native dictionaries

Several experiences have been reported in the literature related to the conversion of different types of dictionaries as LD (e.g., Klimek & Brümmer, 2015; Declerck et al., 2015; Bosque-Gil et al., 2016b; Gracia et al., 2016), which illustrate the growing interest for LD in lexicography. Nevertheless, the idea of developing dictionaries as LD in a native way, rather than converting already existent ones from their proprietary formats into LD, has received little attention so far.

2.1 LD in lexicography

There are, of course, a number of advantages in using LD in lexicography (Bosque-Gil et al., 2016b) that do not depend on whether the dictionary data have been converted from previous formats or have been built as LD from scratch. For instance, the main models developed for representing linguistic information as LD (e.g., OntoLex-lemon[3]) do not make claims on the structure of our mental lexicon, being agnostic of the particular linguistic theory underlying the lexicographic data. Thus, LD constitutes an ideal common representation framework for dictionaries that have been built by following different practical and theoretical perspectives, while retaining all the benefits related to interoperability, visibility and NLP-services compliance. Another evident advantage is the fact that LD enables a seamless integration with other internal and external resources (via links among entities, expressed for example in RDF), allowing for a natural graph-based representation of dictionary data on the basis of Web standards.

These and other benefits have been reported as a result of the initial experiences of converting already existent dictionaries to LD format. We envision, however, a situation

[3] https://www.w3.org/2016/05/ontolex/

in the near future when dictionaries will be developed natively as LD, that is, by compiling them from scratch in an RDF-based environment and directly following the LD principles. This will have an impact on the process of dictionary compilation, representation, and interoperation with other resources.

2.2 Issues of tree-like dictionary structures

In modern electronic dictionaries, entries are typically represented as a tree (usually encoded in XML), following a hierarchical data structure where every element has at most one parent. As discussed by Měchura (2016), this choice of data structure makes some aspects of the lexicographer's work unnecessarily difficult, from deciding where to place multiword items to reversing an entire bilingual dictionary. This is a consequence of the fact that dictionary writing, although assisted by computing methods, still tends to replicate what lexicographers would be doing on paper or with a word processor. This raises a number of issues. Although we are not exhaustive in describing them (see, e.g., Bosque-Gil et al., 2016a, for a more detailed analysis) we illustrate them through a couple of examples. First we can mention the problem of headword selection for multiword phrasemes (Měchura, 2016), e.g., under which entry to place *bow and scrape* (meaning *to be overly polite*), *bow* or *scrape*? Ideally, it should be placed under both entries. However, in a tree-like representation, this obliges the lexicographer to copy the same information in both places, which makes the data more difficult to be maintained or updated (changes in one place need to be propagated into other places). Of course, clever search mechanisms can be built to work around this problem, as modern digital dictionaries do, in which a lemma is provided just once and the system is able to search it wherever it appears. However, that does not solve the problem at source, and the search mechanism is not able to infer the particular sense or homograph of the parent entries that should be associated to the phraseme. For instance, our previous example *bow and scrape* would be associated to the sense of *bow* that corresponds to the action of *inclining to show respect*.

Another example of an issue caused by tree-based view of the dictionary information is that cross-references typically depend on the order of appearance of lexical entries or senses, being usually indicated by a superscript in numeric form in printed or electronic format, e.g., *bow*^{2}, meaning, for instance, the second homograph of the entry *bow*. The problem of this approach is that the introduction of new elements in the middle of the sequence obliges to review and redefine all the involved cross-references across the dictionary, making this modelling technique very sensitive to any change in the ordering criteria. Techniques such as the latter are prone to errors and might result in the collision of identifiers. Again, mechanisms have been implemented that reduce such a problem, although they do not solve it at source.

2.3 Building a graph-based reusable structure

A key aspect of an LD-based dictionary is that every lexical element (headword, sense, written form, grammatical attribute, etc.) is treated as a first-class citizen, being identified by its own URI at a Web scale, and being attached to its own descriptive information and linked to other relevant elements through RDF statements. That allows for a graph-based view of the lexicographic information where the above referred issues can be easily avoided.

Continuing with the example cited above, in an LD-native dictionary the *bow and scrape* multiword expression will be a headword on its own with its own URI, and links will be drawn to relate it to the two parent entries *bow* and *scrape*, directly pointing to their suitable senses or homographs whenever appropriate. In that way, changes will be done in a single place, avoiding the need for copying information and reducing the risk of bad maintenance. This implies that an idiom or collocation, for instance, will not be encapsulated under the container of the entry in which it was originally defined, but will be related to it with the suitable property. Since the idiom now becomes a node, we are able to link it to any other node from any other entry in the dictionary.

Similarly, LD solves the issue of maintaining cross-references. Since entries and senses are now uniquely identifiable throughout the dictionary data and graphs are not actually ordered, cross-references can be direct pointers to the entry or sense to which they are referring. Cross-references will not (only) be manual annotations for human consumption but real links between nodes in the dictionary graph.

Differently from other graph-based approaches for representing lexicographical information (Miller, 1995; Polguère, 2014), LD is based on Web standards, has interoperability as its main focus, and is agnostic of the particular lexicographic theory underlying the dictionary data.

Of course, the conversion of already existent XML-based dictionary data into LD might solve the aforementioned issues, and other similar ones, at the modelling level, but still not at the source. We argue that, by solving such issues at source, LD-native dictionaries will make lexicographers' work more efficient and will make the *consistency of lexicographic data* easier to maintain, given that redundancies are more easily avoided.

LD-native dictionaries will maximise *re-usability of lexical knowledge* during the lexicographic compilation process. For instance, a lexical entry can be characterised by synonyms. In a hierarchical arrangement, such synonyms are nested under their associated entry and there is no guarantee of their existence as lexical entries for themselves. In an LD set-up, each synonym is designed as a new node in the graph and then linked to the initial lexical entry through a synonymy relation. Such a new lexical entry only needs to be defined once, no matter the number of times it appears in the dictionary. External re-use of lexical knowledge is also enhanced via link declarations (in RDF) to other LD sources. That enables, for instance, the re-use of grammatical categories already defined in external catalogues (e.g., LexiInfo[4]), the import of additional semantic descriptions from encyclopaedic resources such as BabelNet[5] or DBpedia (Auer et al., 2007), or the connection of different LD-based dictionaries.

Conceiving a dictionary as LD from scratch has also another advantage. In previous XML to LD conversion experiences, it was necessary to preserve as much information content as possible in order to keep the process reversible. This has led to the propagation of superfluous information into RDF, such as internal dictionary identifiers of the lexical elements or information related to how lexicographic data are displayed in a user interface. In the latter case, we argue that such information should be maintained apart from the purely lexicographic graph. In the former case, the definition of URIs for every lexical element makes the internal identifiers redundant. Further, well designed URIs will avoid

[4] http://www.lexinfo.net/ontology/2.0/lexinfo.owl
[5] http://babelnet.org/

collision of identifiers when integrating several dictionaries, which might be a risk if only internal dictionary identifiers were used.

3. Impact on the editorial process

The lexicographic compilation process generally attempts to represent language in a faithful and authoritative manner, whether inscriptively or descriptively, author-created or corpus-based, for reception or production purposes, and to present the results of the lexicographer's investigation and analysis in one dictionary format or another, as considered to be the most suitable for that editorial concept and most beneficial to the user. The entry microstructure is determined accordingly, to best reflect the items of linguistic knowledge selected by the lexicographer, and is arranged in some hierarchical system, whether historically or by order of frequency, with or without definitions, descriptions or translation equivalents, and examples of usage or citations, accompanied or not by relevant attributes such as synonyms and antonyms, register and geographical or dialect information, grammatical, usage or etymological notes, etc. The dictionary can thus resemble a closed world, with each element minutely selected and designed by the creator, and the end result expressing that mastermind and vision.

Overall, this approach is still valid today for lexicography at the wake of the LD era, even though the resources in service of the lexicographer are tremendously multiplied. At this stage, we are only starting to reveal and get acquainted with the new possibilities and horizons offered by LD lexicography, alongside its related requirements and priorities. In this section we analyze the impact LD-native dictionaries will have on the work of lexicographers in several aspects and their related challenges.

Modelling. For a dictionary to be created in LD, we first have to select the kind of information its entries will cover, and make sure this information is indeed representable as LD by available mechanisms. Once the information that an entry will capture is decided upon (syntatic, semantic, pragmatic, phonetic, etc.), the selection of available vocabularies, and the models to represent them, will proceed in order to create the model that will be the backbone of the editing tool that the lexicographer will be using to generate the data. Modelling challenges include the representation of the sense hierarchy, translations, examples, inflections, homographs or multimedia content in a way that stays true to the lexicographer's view and maximizes re-usability according to the LD principles. However, as mentioned above, the major shift that lexicography would experience involves a transition from a hierarchical ordering of the information recorded in a dictionary entry into a graph structure with its nodes uniquely identified by URIs, whose form should be also determined by the editor. The lexicographer will be required to identify the precise nature of the relation between two pieces of information by using ontological properties rather than unbounded textual descriptions. This echoes the difference between compiling dictionaries with only the human as target or creating them for both humans and computers.

Basic knowledge on LD. Even though expert knowledge of RDF and SPARQL[6] should ideally not be required on the lexicographer's part, he or she would need to assimilate the principles of LD lying at the heart of lexicographic compilation. By doing so, the editor will be able to unlock the potential of both using different URI naming strategies and linking to diverse external or internal resources to enrich his or her own data, for example.

[6] https://www.w3.org/TR/rdf-sparql-query/

Technical needs. Even though developing dictionaries natively in LD would allow them to be integrated into bigger knowledge systems and consumed by LD-aware NLP applications from the very beginning, the daily tasks that human users perform with the help of dictionary data should not be relegated to the background. In this respect, a clear challenge that we must face as we envision this ecosystem is the lack of a well-established and solid mechanism for everyday dictionary users to query LD resources without the need to rely on Semantic Web and LD knowledge. In order to build LD-native dictionaries, tools for graph editing and visualization would be called for to enable the compilation without expert knowledge of Semantic Web formalisms. Natural language and guided interfaces on top of SPARQL would evolve into essential tools for the editor to query the different LD versions created during the editorial process and thus control the project's progress. A paradigm-shift in lexicography would involve reconsidering the skills that are required both from lexicographers and editors as well as from the potential users of such linked dictionaries. Just as new natural language or guided interfaces will be called for in order for non-experts to query the datasets, their maintenance in terms of modification, enrichment and quality control on part of the editor will require new mechanisms as well.

Quality control. As reported in recent surveys on LD quality (Zaveri et al., 2016), there are aspects concerning data quality that are original to LD and therefore will need to be taken into account in LD-driven lexicography. Quality can be assessed through different dimensions, ranging from availability, licensing and security (accessibility dimensions) to data accuracy, consistency, etc. (intrinsic dimensions) and reputation and verifiability, among others (trust dimensions) (see Zaveri et al. (2016) for a state-of-the-art account on LD quality). Although each dictionary data provider may define its own criteria, they all share a common goal with respect to the intrinsic data, namely to provide lexical information that is semantically and syntactically correct, compact (i.e, without redundant data), complete (gathering all available data concerning an entry), and logically consistent (without contradictions or conflicting values). Processes aimed at evaluating the quality of the ontology in support of the dictionary editing phase, as well as for assessing the quality of the generated instances would need to take place as part of the regular lexicographic workflow.

4. Making the graph grow

LD technologies enable the vision of an ecosystem of linked lexicographical resources in the form of a giant cloud of lexicographic data at a Web scale. This heterogeneous cloud could consist of several hubs of dictionaries, each containing data from the same dictionary family or type. This does not mean that all the information must be open and publicly accessible. Different licences and exploitation schemes could be supported,[7] including public and free dictionary data, data with conditional access (e.g., accessible under payment) or closed data internal to a company. Lexical resources are conceived from different theoretical backgrounds and with dissimilar goals and use cases, so that not all of them are equally integrable into a single dictionary. Such an ecosystem will be explorable along several dimensions (language, grammatical information, granularity level). In that sense the traditional notion of a dictionary is diluted because different views or aggregations of data are possible depending on the user's needs (Spohr, 2012).

[7] Declared by means of specialised vocabularies such as http://purl.oclc.org/NET/ldr/ns

4.1 Dictionary data as an asset for LD

As a result of our initial experiences in adapting existing, pre-LD, data into LD, the first thing we discover is that regardless of how fine and well-structured such data may be, and how successful its conversion from e.g. XML to RDF format is, there is a fundamental difference stemming from how such data were originally conceived. Basically, what we look for are the best points of automatic connection to other linguistic data sources and among any sets of data, which can be optimized by further annotation for its use in NLP applications such as word sense disambiguation and induction. Moreover, that, in turn, might lead us deeper into standardization, which facilitates such linking. Our principle observations from this experience so far can be summed as follows:

Metalanguage. The metalanguage that is part of the lexicographic editorial process (e.g., names of attributes, parts of speech, language tags, etc.) is an asset for LD'fying the content, as it helps to uniformize the names of the entry components and their various bits of information, and thus to enhance the communication with other datasets.

Free text. Some of the texts that are written freely by the lexicographer as additional semantic, syntactic or pragmatic information besides the predefined labels seem to be the least valuable for LD, as it is harder to relate them to specific and precise details in other sources. This does not concern definitions and examples of usage, which often contain semantic categories, semantic relations, collocates and so on, which may be useful for sense disambiguation and thus for LD.

Subject field. Tagging the 'domain' of each sense of an entry tends to generate the most accurate *sense-to-sense* linking to other data resources. Unless the specific sense is tagged appropriately, we perform general *word-to-word* linking and might obtain poor results for polysemous lemmas. Different resources do not necessarily use the same 'subject field' tag, for example the monetary aspect of *bank* can be labelled *finance* in one place and *economics* or *commerce* in others, but the relation between these domains is fairly simpler to establish. There is no standard list of domains that is applied universally, not even borrowed from the world of terminology. One of the most highly regarded domain lists is that of the Library of Congress,[8] which is more complex and detailed than lists often used in dictionaries, but its system of sub-classification (e.g. *Art* includes *Painting*, *Sculpture*, *Architecture*, etc.) makes it more precise and suitable for LD.

Attributes. Various types of attributes that can be very helpful for word sense disambiguation in lexicography play a minor role for LD. For example, the register and geographical groups are not relevant enough, nor is grammatical information and patterns in general, such as 'range of application' and inflected forms. Synonyms and antonyms form a group of their own, though failing to offer full one-to-one linking, they serve to expand the semantic field of a word or a sense and may be helpful for indirect linking (surprisingly enough, antonyms tend to be more precise than synonyms in carrying relevant information, and could therefore be more useful for word sense disambiguation and thus for LD). This perhaps accentuates the function of LD as a vehicle for Semantic Web technologies, which must nourish primarily on semantic information.

[8] http://loc.gov/catdir/cpso/lcc.html

4.2 Challenges of interconnecting LD dictionaries

In the rest of this section we discuss some challenges related to making LD-native dictionaries grow and interconnecting them. In particular, we discuss aspects related to interlinking and data integration.

Interlinking. As a first step, lexical or general conceptual resources would need to be identified as suitable linking targets (Villazón-Terrazas & Corcho, 2011; Vila-Suero et al., 2014). Among the numerous datasets already available in the cloud of linguistic linked open data,[9] BabelNet[10] and DBpedia (Auer et al., 2007) emerge as the conceptual encyclopaedic resources with the highest in-degree of links, thus acting as pivotal elements among multiple language datasets. LD-based systems aimed to support the automatic discovery and validation of such relations among language resources would be required to assist the lexicographer at this stage.

Data integration. Services should be developed on top of the LD-based ecosystem that, given a query, aggregate data from the different entries and offer users a unified representation. In this way, the system acts as a 'single dictionary' that is actually the sum and combination of many of them, which are in turn managed separately and developed independently. The major challenge that we would meet here is the fact that information about the same dictionary entry would be sometimes repeated and scattered throughout the cloud of linked dictionaries. Each dictionary would be likely to show some differences in its underlying schema even though elements of the de facto standards had been re-used, especially if the editorial choice involved the use of a custom ontology. Some of the tasks that we would face in this stage have been already addressed in the literature in the LD integration context (Bleiholder & Naumann, 2009; Knap & Michelfeit, 2012), namely: schema matching, duplicate detection, and data fusion.

Schema matching refers to the detection of equivalent schema elements in the different sources (Bleiholder & Naumann, 2009). Proprietary schemas developed for the compilation of a dictionary often have equivalent counterparts in linguistic data category registries, such as LexInfo, but this is not always the case: mismatches between proprietary schema values for a specific DTD tag and individuals of an homologous class in an already available linguistic vocabulary can occur as well. Mappings between the dictionary editorial's custom ontology and other models thus become crucial for overcoming these difficulties.

Duplicate detection is the task of detecting equivalent resources to integrate data into one single and consistent representation (Bleiholder & Naumann, 2009). This means that information repeated across different linked dictionaries, e.g. the part of speech of a lexical item, should be presented only once in the answer to a query on the datasets. The problem arises when dissimilar values are extracted from different dictionaries and conflicts need to be resolved as part of the **data fusion** step. Compatible values which however are different in granularity (e.g. *noun* and *common noun*) would need to be distinguished from different and contradictory ones for the same dictionary entry (e.g. *common noun* and *proper noun*). As reported in the literature (Bleiholder & Naumann, 2009; Knap & Michelfeit, 2012), these conflicts would need to be either avoided (in our *proper noun* and *common noun* example, no information about the part of speech would be given), ignored

[9] http://linguistic-lod.org/llod-cloud
[10] http://babelnet.org/

(both values are presented as parts of speech), or resolved with a set of conflict handling strategies, which, for example, identify some sources as more trustworthy than others.

5. Conclusions

LD is generating a rising interest in the area of lexicography, and many dictionaries have been already converted into LD. In this paper, however, we have focused on what constitutes a step beyond by introducing the notion of LD-native lexicography. That is, dictionaries that will be compiled as LD from scratch. We have analyzed the main advantages of this networked approach in contrast with the more traditional tree-oriented view. We have also discussed its potential impact on the lexicographic data and on the work of lexicographers.

In this current intermediate phase between traditional and LD-driven lexicography, the observations described in this paper prompt us to revise existing lexicographic resources with LD in mind, and prioritize and emphasize certain ingredients, such as the subject field, and modify the entry structure. At the next stage, in the aim of being instantly understood by machines as part of *machine-to-machine* communication for the benefit of human beings, future LD-native lexicography will be considerate of these and other factors from its very conception and inception and throughout its compilation and usage. Although the LD format is displayed within an equalized non-hierarchical graph, its linking points are absolutely crucial. Metaphorically, while all the skin of our body is a living organ and sensitive to touch, it has a few points that serve most commonly for touching, like the fingertips. LD-native lexicography will attribute special attention to any such fingertips, as its precious communicative agents.

6. Acknowledgements

Supported by the LDL4HELTA EUREKA project,[11] the Spanish Ministry of Economy and Competitiveness through the ReTeLe Excellence Network (TIN2015-68955-REDT) and the Juan de la Cierva program, and by the Spanish Ministry of Education, Culture and Sports through the FPU program.

7. References

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R. & Ives, Z. (2007). Dbpedia: A nucleus for a web of open data. *The semantic web*, pp. 722–735.

Bizer, C., Heath, T. & Berners-Lee, T. (2009). Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 5(3), pp. 1–22.

Bleiholder, J. & Naumann, F. (2009). Data fusion. *ACM Computing Surveys (CSUR)*, 41(1), p. 1.

Bosque-Gil, J., Gracia, J. & Gómez-Pérez, A. (2016a). Linked data in lexicography. *Kernerman Dictionary News*, pp. 19–24. URL http://kdictionaries.com/kdn/kdn24.pdf#page=19.

Bosque-Gil, J., Gracia, J. & Montiel-Ponsoda, E. (2017). Towards a Module for Lexicography in OntoLex. In *Proc. of the LDK workshops: OntoLex, TIAD and Challenges for Wordnets at 1st Language Data and Knowledge conference (LDK 2017), Galway, Ireland*. CEUR-WS.

[11] https://ldl4.com/

Bosque-Gil, J., Gracia, J., Montiel-Ponsoda, E. & Aguado-de Cea, G. (2016b). Modelling Multilingual Lexicographic Resources for the Web of Data: the K Dictionaries case. In *Proc. of GLOBALEX'16 workshop at LREC'15, Portoroz, Slovenia*.

Declerck, T., Wandl-Vogt, E. & Mörth, K. (2015). Towards a Pan European Lexicography by Means of Linked (Open) Data. In I. Kosem, M. Jakubíček, J. Kallas & S. Krek (eds.) *Proceedings of eLex 2015. Biennial Conference on Electronic Lexicography (eLex-2015), electronic lexicography in the 21st century: Linking lexical data in the digital age, August 11-13, Herstmonceux, United Kingdom*. Ljubljana/Brighton: Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd.

Gracia, J., Villegas, M., Gómez-Pérez, A. & Bel, N. (2016). The Apertium Bilingual Dictionaries on the Web of Data. *Semantic Web Journal*. URL http://www.semantic-web-journal.net/system/files/swj1419.pdf.

Klimek, B. & Brümmer, M. (2015). Enhancing lexicography with semantic language databases. *Kernerman DICTIONARY News*, 23, pp. 5–10.

Knap, T. & Michelfeit, J. (2012). Linked Data Aggregation Algorithm: Increasing Completeness and Consistency of Data. *Provided by Charles University*.

Miller, G.A. (1995). WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11), pp. 39–41.

Měchura, M. (2016). Data structures in lexicography: from trees to graphs. In *Proc. of 10th Workshop on Recent Advances in Slavonic Natural Language Processing (RASLAN 2016)*. URL http://www.lexiconista.com/raslan2016.pdf.

Parvizi, A., Kohl, M., González, M. & Saur'ı, R. (2016). Towards a Linguistic Ontology with an Emphasis on Reasoning and Knowledge Reuse. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*. URL http://www.lrec-conf.org/proceedings/lrec2016/summaries/523.html.

Polguère, A. (2014). From Writing Dictionaries to Weaving Lexical Networks. *International Journal of Lexicography*, 27(4), pp. 396–418. URL https://hal.archives-ouvertes.fr/hal-01097112/document.

Spohr, D. (2012). *Towards a multifunctional lexical resource: Design and implementation of a graph-based lexicon model*, volume 141. Walter de Gruyter.

Vila-Suero, D., Gómez-Pérez, A., Montiel-Ponsoda, E., Gracia, J. & Aguado-de Cea, G. (2014). Publishing linked data: the multilingual dimension. *Towards the Multilingual Semantic Web*, pp. 101–118.

Villazón-Terrazas, B. & Corcho, O. (2011). Methodological guidelines for publishing linked data. *Una Profesión, un futuro: actas de las XII Jornadas Españolas de Documentación: Málaga*, 25(26), p. 20.

Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J. & Auer, S. (2016). Quality assessment for linked data: A survey. *Semantic Web*, 7(1), pp. 63–93.

On-the-fly Generation of Dictionary Articles for the DWDS Website

Alexander Geyken, Frank Wiegand, Kay-Michael Würzner

Berlin-Brandenburg Academy of Sciences and Humanities, Jägerstraße 22/23, 10117 Berlin, Germany
E-mail: {geyken|wiegand|wuerzner}@bbaw.de

Abstract

We present a method for generating on-the-fly dictionary articles for the DWDS website (https://www.dwds.de). The DWDS website contains electronic versions of large legacy dictionaries as well as very large corpora. On-the-fly articles are a fallback solution for user queries that cannot be matched with dictionary headwords or one of its inflected forms on the website. They depend on an automatic morphological analyser that segments complex words into parts that formally match existing dictionary headwords in a reliable way. On-the-fly articles are a useful mechanism for increasing the number of headwords with minimal manual effort. They are particularly useful for compounding languages like German. The generation method described in this article is fully integrated into the DWDS website.

Keywords: automatic creation of dictionary content; compound recognition; German morphology

1. Introduction

A major challenge for (monolingual) online dictionaries is to guarantee exhaustive vocabulary coverage, a goal that is time consuming, labour intensive and therefore generally considered as impossible to achieve. This is even more true for languages such as German, a language well known for its very large and theoretically even unlimited number of compounds. Therefore additional methods have to be developed to provide users with lexical information for as many words as possible with minimal manual intervention.

In this article we show how "out-of-headword-range" user queries, i.e. queries that cannot be directly matched to headwords in the dictionary, are dealt with in the Digital Dictionary of German language (DWDS), a comprehensive lexical information system of contemporary German. The problem of "out-of-headword" queries is a major practical problem for the DWDS system since there are numerous morphologically complex words (compounds and derived forms) in German that are not lexicographically described in neither of the largest monolingual dictionaries of New High German, including Duden (1999), Wahrig (Wahrig-Burfeind, 2011) and DWDS (Klein & Geyken, 2010; Geyken, 2015). These "hand-crafted" dictionaries have a size of between 150,000 and 200,000 headwords whereas the number of German words occurring in corpora is estimated as being well above five million (Klein, 2013). Even though many of those words may not require a full description from a lexicographer's point of view, they are nevertheless targeted by regular user queries and therefore need to be handled by the lexical information system.

We propose a solution to this kind of user query by providing—wherever possible—dynamically generated dictionary articles on the DWDS platform with automatic methods. These articles generated "on the fly" are presented in the same way as dictionary articles compiled by lexicographers. Nevertheless, both automatic and hand-crafted articles are labeled as such. Thus, the dictionary user is provided with lexicographic information for many of those compounds that are not contained in the hand-crafted dictionaries.

The remainder of the article is organized as follows: in the next section the DWDS lexical information platform is presented. Section 3 briefly describes the quality management of DWDS platform that is used to identify missing entries as well as incomplete or false

information of existing entries. Section 4 briefly introduces mechanisms of morphological productivity, shows how automatic morphological analysers deal with the problem of segmenting complex words and applies these methods on the problem addressed in this article, namely to relate "out-of-headword" compounds to headwords in the DWDS dictionary. Automatic morphological analysis is at the basis of the generation of "on-the-fly" dictionary articles. Its different components are presented in Section 5. Morphological analysis is just one mechanism to deal with "out-of-headword" queries. Section 6 shows how the automatic morphological analysis is combined with other fallback mechanisms dealing with queries that are commonly used to deal with "out-of-headword" user queries. The method presented here is fully integrated into the DWDS platform. In Section 7 some results together with an evaluation on the basis of DWDS user queries are presented. The article ends with a short conclusion (Section 8).

In this paper the following terminology is adopted. The term "headword" is used to denote the lemma string of a dictionary entry. The term "dictionary entry" refers to the lexicographic description of a headword that consists of a form and a sense description. The term "dictionary article" is used for aggregated information, including the dictionary entry as well as information from automatically extracted information from corpora or from external lexicographic resources.

2. The DWDS platform

The *Digital Dictionary of the German Language* (DWDS, *Digitales Wörterbuch der deutschen Sprache*) is a long term project of the Berlin-Brandenburg Academy of Sciences and the Humanities (BBAW, Berlin-Brandenburgische Akademie der Wissenschaften). The goal of the DWDS project is to compile a large aggregated word information system based on large legacy dictionaries, large corpora, word statistics and automated methods to speed up the process of updating and amending the existing lexical resources (Geyken, 2014). The platform integrates an automatic collocation extractor and a good example finder (Didakowski & Geyken, 2014). Furthermore, the DWDS draws on large corpora with a size of 12.5 billion running words (as of May 2017) that cover the period between 1600 and now. The DWDS website with all the data and functions described in the article can be consulted under https://www.dwds.de/. The dictionary component of the DWDS draws mainly on two legacy dictionaries: the Dictionary of the German Contemporary Language (Klappenbach & Steinitz, 1964–1977), a synchronic dictionary of 4,800 pages in six volumes with 120,000 keywords, compiled between 1961 and 1977 at the GDR Academy of Sciences, and second, a subset of about 70,000 articles of the Duden GWDS (Scholze-Stubenrecht, 1999), the largest printed dictionary of contemporary German. Articles from Duden were chosen for cases where the WDG articles are missing, incomplete or outdated. In addition to these entries in WDG and Duden, another 45,000 entries were selected by corpus-based methods (Geyken & Lemnitzer, 2012) and integrated as entries with minimal morphological information into the DWDS dictionary plattform. Since 2013, a team of six lexicographers edits new articles and revises the existing entries. The goal of the DWDS project is to obtain a coherent and up-to-date lexicographic description of the present German language at the end of the project in 2025.

3. Quality management within the DWDS platform

The revision process of the legacy dictionaries requires a check of all entries for their correctness and up-to-dateness on all lexicographic levels. This process is feasible only by

a distributed effort, and it goes without saying that this revision process is too complex to be done without digital assistance. To this end we use *MantisBT*,[1] an open source, web-based issue tracker that is easy to install and requires only little time for users to familiarize with the system. Users of the issue management system can report either missing entries or inconsistencies on any type of lexicographic information, including spelling, morphology, sense, collocation, phraseology. Furthermore, we use the field *Tags* to provide the reported issue with additional workflow information such as 'for this word, a basic entry is sufficient', 'provide definition only', 'word should become a full entry'. Those Tag values can be used as a flag to be displayed on the DWDS platform. As of 22nd May 2017 more than 18,500 issues have been submitted by a group of 30 people, the majority of them are employees of the BBAW. According to the summary page of the *MantisBT* the top three issues are: *missing entry* (11,500), *missing/wrong meaning* (4,850), and *grammar or word formation errors* (870).

It is important to note here that only those words are submitted to the issue management system as "missing entries" where major additional and manual lexicographic description is deemed necessary. However, as stated in the introduction, due to the very large number and the high productivity of (new) German compounds it is not possible to manually compile full lexicographic entries for all compounds. Therefore automatic methods are used to generate basic dictionary entries (cf. Section 4) that form one component of the aggregated dictionary article that is used for the DWDS platform (cf. Section 5).

4. Automatic morphological segmentation as a building block for dynamic dictionary articles

The idea of this section is to use automatic morphological analyses in order to split complex words which are not in the dictionary into less complex components for which dictionary entries exist. More precisely, we are looking for the least complex decomposition that corresponds best to the word formation of the complex word. In the remainder of this section, we briefly mention linguistic aspects of German word formation (4.1), we summarize the relevant aspects of automatic morphological analysers for German (4.2), and we present a method to map complex words to the appropriate headwords in the DWDS dictionary.

4.1 German word formation

The term *word formation* subsumes operations to create novel (complex) words[2] based on existing linguistic units (i.e. words and affixes). Together with *lexical borrowings* and *semantic shifts* it is one of the means to cover the need for "new" words. Word formation operations are usually distinguished in terms of their operands: The combination of two words is called *compounding* while the combination of a word and an affix is called *derivation*.[3]

The German language is not only known for its rich productivity of compounding. It has also some very productive affixes that can be used to form new compounds. Example (1)

[1] MantisBT: https://www.mantisbt.org/

[2] Note that this is the principal difference to *inflection* which does not result in novel words.

[3] *Conversion*, i.e., the covert changing a word's category may be treated as a special case of derivation involving an invisible affix.

below illustrates this combinatorial process. The noun *Vollstreckbarkeit* (engl. 'enforceability') is derived from the verb *strecken* by subsequently adding the verbal prefix *voll-*, the suffix *-bar*, and the suffix *-keit*:

$$\Big(((voll_{\texttt{P}}\,(streck_{\texttt{V}}))_{\texttt{V}}\,bar_{\texttt{S}})_{\texttt{A}}\,keit_{\texttt{S}}\Big)_{\texttt{NN}} \tag{1}$$

In addition to such iterated derivation operations, German, in contrast to e.g. English or French, knows "non-spaced" compounding: compounds are realized as a continuous sequence of characters optionally agglutinated with non-empty linking elements such as *-s* or *-er*; the subparts may very well be complex words again:

$$\Big(((voll_{\texttt{P}}\ streck_{\texttt{V}})_{\texttt{V}}\,bar_{\texttt{S}})_{\texttt{A}}\,keit_{\texttt{S}}\Big)_{\texttt{NN}}\,s_{\texttt{Link}}\,((er_{\texttt{P}}\ kl\ddot{a}r_{\texttt{V}})_{\texttt{V}}\,ung_{\texttt{S}})_{\texttt{NN}} \tag{2}$$

The sequence of operations leading to a complex word is called its *derivational history*. A fundamental problem of (word-based) morphological analysis is ambiguity; often, multiple analyses for a single word are available. Lemnitzer & Würzner (2015) distinguish four types of ambiguities:

segmentation ambiguities A complex word may be split into several morpheme sequences: `Musik<NN>Erleben<+NN>` ('musical experience') vs. `Musiker<NN>Leben<+NN>` ('a musician's life').

categorial ambiguities A word belongs to more than one category: *weiß* (adj. 'white' vs. verb '[I] know').

lexical ambiguities Multiple lexemes are realized with the same word: *Bank* as financial institution and as seating-accommodation.

morpho-syntactic ambiguities Multiple forms of the same morphological paradigm have an identical realization: *übe* ('practice') as first person singular indicative active as well as imperative singular.

Complex morphological processes must therefore be employed to generate one or more plausible segmentations of a complex word, and eventually, to link these segments to existing dictionary entries. This is discussed in more detail in the next section.

4.2 Automatic morphological analysis

The overall goal of the morphological analysis of a (possibly) complex word form is its decomposition into smaller segments consisting of a combination of affixes and stems together with symbols marking segment separators. It can thus be understood as the identification of operations and operands which led to formation of that complex word.

Finite-state morphology is a technique to implement the analysis of productive word formation processes using a set of *rational rules* (cf. Lawson, 2003) over a finite alphabet. It is a very popular model in computational morphology and has been applied to a large number of languages (cf. Beesley & Karttunen, 2003). Rational rules can be efficiently represented and applied using (weighted) finite-state transducers. There are several finite-state

morphologies available for German, most notably GERTWOL (Haapalainen & Majorin, 1995), TAGH (Geyken & Hanneforth, 2006) and SMOR (Schmid, 2004). While GERTWOL is not freely available for large-scale testing and application, TAGH and SMOR have a comparable coverage of German word formation. SMOR allows for a segmentation into atomic morphemes whereas TAGH regroups morphemes to larger units. Since we need a 1:1 mapping of automatically analysed morphemes onto headwords of the DWDS dictionary, SMOR is more flexible and therefore better suited for the task at hand. Figure 1, as an example, shows the output of SMOR for the German compound *Kürzungen* ('shortages', 'cuts'):

```
> Kürzungen
Kür<NN>:<>Z:zunge<+NN>:<><Fem>:<><>:n<Nom>:<><Pl>:<>
Kür<NN>:<>Z:zunge<+NN>:<><Fem>:<><>:n<Gen>:<><Pl>:<>
Kür<NN>:<>Z:zunge<+NN>:<><Fem>:<><>:n<Dat>:<><Pl>:<>
Kür<NN>:<>Z:zunge<+NN>:<><Fem>:<><>:n<Acc>:<><Pl>:<>
k:Kürze:<>n:<><V>:<>ung<SUFF>:<><+NN>:<><Fem>:<><>:e<>:n<Nom>:<><Pl>:<>
k:Kürze:<>n:<><V>:<>ung<SUFF>:<><+NN>:<><Fem>:<><>:e<>:n<Gen>:<><Pl>:<>
k:Kürze:<>n:<><V>:<>ung<SUFF>:<><+NN>:<><Fem>:<><>:e<>:n<Dat>:<><Pl>:<>
k:Kürze:<>n:<><V>:<>ung<SUFF>:<><+NN>:<><Fem>:<><>:e<>:n<Acc>:<><Pl>:<>
```

Figure 1: SMOR analyses for *Kürzungen*

Notation	**Meaning**
`<NN>`	morpheme category: normal noun
`<V>`	morpheme category: verb
`<A>`	morpheme category: adjective
`<+`*x*`>`	denotes the category of the word (part of speech)
`<SUFF>`	suffix
`<Fem>`	feminine gender
`<Nom>`, `<Gen>`, `<Dat>`, `<Acc>`	grammatical case
`<Pl>`	plural
`<>`	empty string (epsilon)
x : *y*	mapping from lemma to word-form level

Table 1: SMOR syntax[4]

A number of strategies have been proposed to deal with the aforementioned ambiguity phenomena, usually employing the context of a word's occurrence. In our use-case, i.e., the analysis of dictionary queries, context is not available. We therefore make use of a simple heuristic which goes back to Volk (1999) in order to reduce the number of analyses. Each word formation operation is assigned a specific cost (e.g., 2.5 for suffixation and 5 for compounding). From the two possible analyses for *Kürzungen* (i.e., `Kür<NN>Zunge<+NN>n`

[4] Note that the analysis contains the lemma as well as the word-form level. Differences between the two are denoted by the colon symbol. Symbols only present on the lemma level are mapped onto the empty string. For details, the reader is referred to Schmid (2004).

vs. `kürzen<V>ung<SUFF><+NN>en`), the latter is 'cheaper' and thus considered to be more likely. In addition, we increase the total cost of a segmentation by the edit-distance between the lemmas associated with the segmentation and the input word. Favoring orthographically closer analyses helps for example resolving ambiguities introduced by the optional dative suffix *-e* in cases like *Hängebuche* ('hanging book' or 'weeping beech') with analyses `hängen<V>Buch<+NN>e` and `hängen<V>Buche<+NN>`.

4.3 Mapping morphological analysis to dictionary entries

After performing the morphological analysis of the queried word and the ranking of the resulting analyses according to the weighting sketched above, only the best (i.e., cheapest) analyses are considered as candidates for linkage. Instead of simply linking to the entries of the identified (atomic) morphemes, we try to be as specific as possible by linking to the most complex available dictionary entries. This is done by constructing the set of all possible derivational histories leading from the morphemes to the complex word form for each remaining analysis. Derivational histories can be depicted as trees:

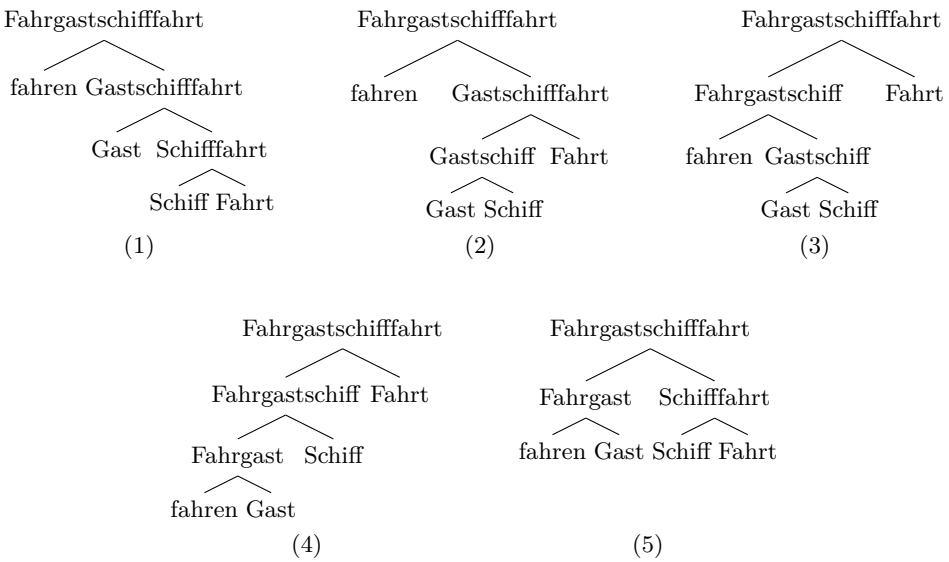


Figure 2: Derivational histories for the word *Fahrgastschifffahrt* depicted as trees

The components of each level in each tree are looked up in the dictionary. The least complex segmentation is used for the mapping, i.e., the highest in a tree where each segment matches a dictionary headword. For *Fahrgastschifffahrt* ('passenger shipping') the selected segmentation is on level one of tree number five since both *Fahrgast* ('passenger') and *Schifffahrt* ('shipping') are listed in the DWDS dictionary.

5. Components of dynamically generated dictionary articles

Dynamically generated dictionary articles consist of all components of the DWDS system which can be generated automatically for a given word from various resources, including information about its form (spelling, grammar, word formation), word frequency, thesaurus information (synonyms, antonyms, hyponyms, and hyperonyms) re-

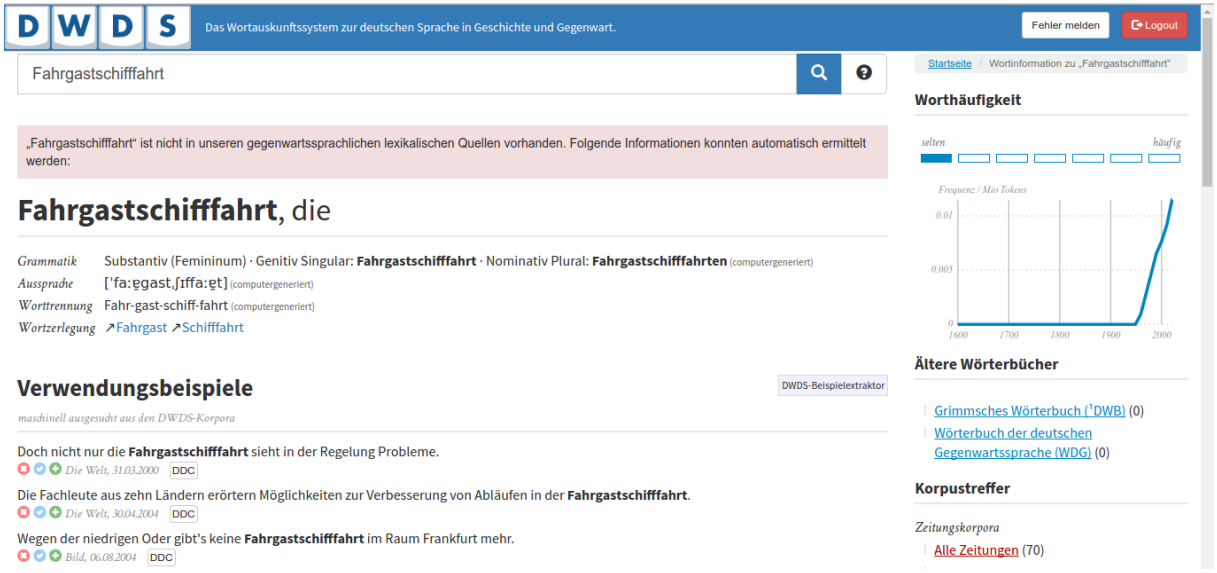


Figure 3: Generated article for *Fahrgastschifffahrt* on the DWDS website

trieved from the OpenThesaurus[5] dataset as well as automatically selected usage examples from DWDS corpora by using the DWDS-Beispielextraktor[6] and collocations by the DWDS-Wortprofil.[7] The extraction of usage examples and collocations are described in more detail elsewhere (cf. Section 2). Therefore this section focuses on the description of frequency and form information.

Using the DDC search engine[8] indices, we can provide information about the word frequencies within the DWDS corpora. Using a level meter, a value between one and seven on a logarithmic scale shows how often the requested lemma occurs within the corpus texts.[9] Since all corpus documents are marked with reliable metadata (including its date of publication), a graph of the distribution of the word frequencies from 1600 until today can be computed. This graph is shown on the website below the frequency meter.[10] The graph image is linked to an extended version of our corpus search plotting tool. In addition, hyperlinks to occurrences of the keyword in the public searchable corpora are provided as well.

The form part of the dynamically generated article consists of several parts: Information about the word's pronunciation[11] and hyphenation is provided by the gramophone webservice.[12] The grammatical information (i.e. inflection and the Part-of-Speech tag, more precisely the mapping of an STTS tag to the principal word classes of the dictionary such as nouns, verb, adjective and adverb) is obtained via the SMOR analysis. If applicable, morphological segmentation is displayed and all components are linked to their respective dictionary articles.

[5] OpenThesaurus: https://www.openthesaurus.de/
[6] DWDS-Beispielextraktor: https://www.dwds.de/d/beispielextraktor
[7] DWDS-Wortprofil: https://www.dwds.de/d/ressources#wortprofil
[8] DDC (DWDS/Dialing Concordance), the search engine used in the DWDS project: https://www.dwds.de/d/suche
[9] https://www.dwds.de/d/api#frequency
[10] DWDS-Wortverlaufskurve: https://www.dwds.de/d/plot
[11] Only for users with a DWDS user account.
[12] http://kaskade.dwds.de/~kmw/gramophone.py (Würzner & Jurish, 2015).

6. Combination with other fallback mechanisms

In Section 4 it was shown how user queries corresponding to "out-of-headword" compounds can be correctly mapped to headwords in the DWDS dictionary. However, this is only one way to handle query strings that do not directly match dictionary entries. In the current implementation of the DWDS platform the following fallback mechanisms take effect:

1. If the query string can be morphologically analysed via SMOR then
   (a) if the query string corresponds to an inflected form of a dictionary headword, the user query is redirected to the dictionary article of that headword.
   (b) else if SMOR provides a valid segmentation into two or more morphologically valid segments and if all components of the word are itself valid dictionary entries in the DWDS system, an aggregated dictionary article is generated "on-the-fly".
2. If the morphological analysis fails, a "Did you mean?" function is triggered. It aims to refer the user to orthographically close (defined in terms of edit distance) dictionary entries.
3. If the "Did you mean?" function fails, i.e. no close dictionary headword can be identified, the user is referred to a corpus search and corpus concordances for the query string are offered.

7. Results and evaluation

The method for generating on-the-fly articles presented here is fully integrated into the DWDS platform. The results in Table 2 are based on an evaluation of the user queries for a period of one month from 23rd April to 23rd May 2017. The logfile for that period contains a total of 190,554 unique lexical queries (types), i.e. only those queries that consist of "bare words" without special characters. Among those queries, 17% (i.e. 33,134) do not have a direct match with a dictionary headword of the DWDS dictionary.

A quick evaluation of the 100 most frequent of these queries led to the classification in Table 2, which shows that 35% of these "out-of-headword" queries correspond to inflected forms of existing dictionary entries and for another 20%, an on-the-fly article can be dynamically generated. For another 28% it was possible to identify candidates via a "Did you mean?" function. Only for 17% of the "out-of-headword" queries the user had to be redirected to a corpus query.

Fallback method	**% of total**	**% correct**
1. Inflected input, redirected to lemma entry	35%	91%
2. On-the-fly dictionary article generated	20%	95%
3. Suggestions "Did you mean?"	28%	68%
4. Redirection to corpus search	17%	n/a

Table 2: Proportion of processed user queries with no direct match for a DWDS dictionary headword

The correctness of this classification is displayed in the last column of Table 2. It shows that more than 91% of the entries were lemmatised correctly and for even 95% a correct dictionary article was generated on-the-fly.

Since the main topic of this paper is on the dynamic generation of dictionary articles, we will focus on a discussion of the second fallback method. Figure 4 lists various examples and how they are dealt with by our approach. A main observation is that the ambiguity problem of automatic morphological analysers is solved remarkably well in our case. This is due to the fact that wrong segmentations can be eliminated in general because at least one of their segments does not have a match with a dictionary headword. This is illustrated by the Examples (1)–(3). Ambiguities due to linking elements can often be solved with the least weight method of the morphological analysis (cf. Section 4.2) as in Examples (4) and (5). Much more difficult is the mapping to the correct word category. Example (6) is a case where the mapping of the morphological analyser works correctly whereas in Example (7) it is incorrect.

(1) *Angsthasenpolitik* ('politics of cowardice'): correct segmentation is found: `Angsthase<>:n<NN>P:politik`, but not `Angst<NN>H:hase<>:n<NN>P:politik<+NN>`.
(2) *Autobahnmeisterei* ('highway maintenance area'): correct segmentation is `Autobahn<NN>M:meisterei<+NN>` (*Meisterei* is `Meister<N>ei<SUFF><+NN>`) and not `Meister<NN>E:ei<+NN>`.
(3) *Krötenlaubfrosch* ('tree frog'): correct segmentation is `Kröte<>:n<NN>L:laubfrosch<+NN>`, but not `Kröte<>:n<NN>L:laub<NN>F:frosch<+NN>` or `Kröte<>:n<NN>L:laub<NN>F:frosch<+NN>`.
(4) *Reiseabschnitt* ('travel segment'): correct segmentation is `Reise<NN>A:abschnitt<+NN>`, not `Reis<>:e<NN>A:abschnitt<+NN>`.
(5) *Arbeitsamtsbericht* ('job center report'): correct segmentation is `Arbeitsamt<NN>B:bericht<+NN>`, and not `Arbeit<>:s<NN>A:amt<NN>B:bericht<+NN>` or even `Arbeit<NN>S:samt<>:s<NN>B:bericht<+NN>`.
(6) *Treibschnee* ('drift snow'): correct expansion is `t:Treibe:<>n:<><V>S:schnee<+NN>`.
(7) *Grillfest* ('barbecue party'): automatic analysis `G:grill<NN>fest<+A>`, whereas the correct segmentation `g:Grille:<>n:<><V>F:fest<+NN>` is not found.
(8) *Arbeitsstellenleiter* ('work place leader', masc., or 'work place ladder', fem.).
(9) *Ballbesitzfußball* ('football game with possession of the ball'): wrong plural.
(10) *Schweinsteiger* (a family name which should not be segmented).

Figure 4: Various examples for correct or incorrect compound segmentations

There are also cases where the ambiguity is undecidable. An example for this case is the homography of *Leiter* ('ladder', fem. vs. 'leader', masc.) as in Example (8). In this case two dictionary entries are generated and the decision about the correctness is left to the user. Another problem for the automatic morphological analyser is the correct generation of inflected forms. For example, in the case of ambiguities between count nouns and non-count nouns, the system has to decide if a plural is possible (count noun sense) or not (non-count noun sense), see Example (9). Finally, the ambiguity between a proper noun and a common noun is a difficulty for our method. This is generally true for all family names that can be segmented into two or more common nouns like in Example (10).

8. Conclusion

We have presented a method to generate on-the-fly articles for the DWDS platform as a fallback solution for user queries that cannot be directly matched with dictionary headwords of the DWDS system. The strategy of generating dynamic dictionary articles on the fly is closely related to the activities in the issue management system: cases of wrong or insufficient articles generated "on the fly" can be reported to the system and eventually a full lexicographic dictionary entry can be compiled manually.

An evaluation of the logfiles of the DWDS platform for a one month period shows that approximately one out of six queries corresponds to a "out-of-headword" query. For 20% of those queries a DWDS dictionary article can be successfully generated on-the-fly. Thus the method presented in this article proves to be useful to augment the number of—actually used—headwords of the DWDS dictionary system.

9. References

Beesley, K.R. & Karttunen, L. (2003). *Finite State Morphology*. Stanford, CA: CSLI.

Didakowski, J. & Geyken, A. (2014). From DWDS corpora to a German word profile–methodological problems and solutions. *OPAL – Online publizierte Arbeiten zur Linguistik*, 2/2014, pp. 39–47.

Geyken, A. (2014). Methoden bei der Wörterbuchplanung in Zeiten der Internetlexikographie. *Lexicographica*, 30(1), pp. 77–111.

Geyken, A. (2015). Recent developments in German lexicography. In *Kernerman Dictionary News*, volume 23. pp. 16–19.

Geyken, A. & Hanneforth, T. (2006). TAGH: A Complete Morphology for German based on Weighted Finite State Automata. In *Finite State Methods and Natural Language Processing. 5th International Workshop, FSMNLP 2005, Helsinki, Finland, September 1-2, 2005. Revised Papers*, volume 4002. Springer, pp. 55–66.

Geyken, A. & Lemnitzer, L. (2012). Using Google Books Unigrams to Improve the Update of Large Monolingual Reference Dictionaries. In *Proceedings EURALEX 2012*. Oslo, pp. 362–366.

Haapalainen, M. & Majorin, A. (1995). GERTWOL und morphologische Disambiguierung für das Deutsche. In *Proceedings of the 10th Nordic Conference of Computational Linguistics*. University of Helsinki, Department of General Linguistics.

Klappenbach, R. & Steinitz, W. (eds.) (1964–1977). *Wörterbuch der deutschen Gegenwartssprache (WDG)*. Berlin: Akademie-Verlag.

Klein, W. (2013). Von Reichtum und Armut des deutschen Wortschatzes. In *Reichtum und Armut der deutschen Sprache. Erster Bericht zur Lage der deutschen Sprache*. Berlin/Boston: De Gruyter Mouton, pp. 15–56.

Klein, W. & Geyken, A. (2010). Das 'Digitale Wörterbuch der Deutschen Sprache DWDS'. In *Lexicographica*, volume 26. pp. 79–96.

Lawson, M.V. (2003). *Finite Automata*. CRC Press.

Lemnitzer, L. & Würzner, K.M. (2015). Das Wort in der Sprachtechnologie. In U. Haß & P. Storjohann (eds.) *Handbuch Wort und Wortschatz*. De Gruyter, pp. 297–319.

Schmid, H.e.a. (2004). SMOR: A German Computational Morphology Covering Derivation, Composition and Inflection. In *Proceedings of LREC*.

Scholze-Stubenrecht, W. (ed.) (1999). *Duden – Das große Wörterbuch der deutschen Sprache in 10 Bänden*. Mannheim: Bibliographisches Institut, 3. edition.

Volk, M. (1999). Choosing the right lemma when analysing German nouns. In *Multilinguale Corpora: Codierung, Strukturierung, Analyse. 11. Jahrestagung der GLDV.* Frankfurt a. M., p. 304–310.

Wahrig-Burfeind, R. (2011). *Wahrig, Deutsches Wörterbuch. Mit einem Lexikon der Sprachlehre.* Gütersloh/München: wissenmedia in der inmedia ONE] GmbH.

Würzner, K.M. & Jurish, B. (2015). A hybrid approach to grapheme-phoneme conversion. In *Proceedings of the 12th International Workshop on Finite State Methods and Natural Language Processing.* URL http://www.aclweb.org/anthology/W/W15/W15-4811.pdf.

Exploiting a Corpus to Compile a Lexical Resource for Academic Writing: Spanish Lexical Combinations

Margarita Alonso-Ramos, Marcos García-Salido and Marcos Garcia

Universidade da Coruña, Grupo LyS
Departamento de Letras, Facultade de Filoloxía
E-mail: lxalonso@udc.es, marcos.garcias@udc.gal, marcos.garcia.gonzalez@udc.gal

Abstract

This paper provides insight into ongoing research focusing on the exploitation of Spanish academic corpora in order to build up a lexical tool addressed to novice writers of academic texts. The object of the lexical tool is what we call *academic lexical combinations* (ALC). By ALC we mean recurrent segments of words that may or may not be semantically compositional and fulfill rhetorical functions such as giving examples, concluding, expressing emphasis, etc. These functions are particularly prominent in academic discourse. ALCs comprise from collocations to idioms as well as formulas, as they are understood in the Meaning-Text Theory (Mel'čuk, 2012). The procedure adopted for the extraction of the ALC from the corpus is described along with how we combine statistical information and native speakers' intuition. Even if corpora play a leading role in the construction of our lexical tool, we need to filter out corpus output with phraseological criteria, which makes human intervention necessary. Finally, we specify the architecture of the lexical tool and we show different prototype lexicographical entries.

Keywords: academic language; collocation; idiom; formula; lexical bundle; corpus

1. Introduction

In today's knowledge society, the written text plays a primary role, especially in the academic context. When students get into the university, they have to face a new discourse genre and need tools that allow them not only to understand academic texts, but also to produce them. Whereas languages such as English are relatively well provided in this respect (McCarthy & O'Dell, 2008; Swales & Feak, 2012; Lea et al., 2014), no resource of this kind exists for Spanish so far. Although academic writing is a multifaceted phenomenon, we believe that the gist of acquiring academic writing skills resides in learning what we call *academic lexical combinations* (ALCs). By ALC we mean recurrent segments of words that may or may not be semantically compositional and that fulfill rhetorical functions such as giving examples, concluding, expressing possibility or certainty, etc. These functions are particularly prominent in academic discourse. ALCs comprise from collocations (*extraer conclusiones* 'to draw conclusions') to idioms (*en conclusió*n 'in conclusion') as well as formulas that traditionally do not have a place in the phraseological spectrum (roughly *lexical bundles*, Biber et al. (2004), as *como se ha dicho previamente* 'as stated previously').

The literature on English ALC is extensive, especially on lexical bundles (Biber et al., 1999, 2004; Cortes, 2004; Hyland, 2008; Verdaguer & Salazar, 2013; Salazar, 2014). The reason for this growing interest lies mainly in the predominant role of English as an international academic language. Therefore, there is a need to build up lexical resources helping principally non-native English speakers to write research articles or, more generally, academic texts. In recent years, some initiatives to compile academic lexical resources in languages other than English have been undertaken as well. Though the following is not an exhaustive list, we can mention some projects on European languages. For French an extensive academic corpus has been compiled around the project Scientext, which has served as the basis for a considerable amount of research into French phraseology (Tutin & Grossmann, 2014; Cavalla & Loiseau, 2014; Tutin, 2010, 2014). Likewise, academic Brazilian Portuguese, especially that found in article abstracts, has been the focus of a research

team based at the Universidade do Rio Grande do Sul (Krause Kilian & Dias Loguercio, 2015). Similar projects in other languages are less advanced. For instance, there is a joint multi-disciplinary Scandinavian project aimed at developing three new academic lexical resources based on corpora consisting of texts from Swedish, Norwegian and Danish academic settings (Johansson Kokkinakis et al., 2012), but, to the best of our knowledge, there is not yet published research deriving from this corpus.

As far as Spanish is concerned, the interest on academic discourse has not a long tradition. The pioneering project was ADIEU (Vázquez, 2001), more focused on Spanish as a second language and including a collection of transcribed texts of oral presentations and master classes. The main interest has been in the differences between academic genres (Regueiro Rodríguez & Sáez Rivera, 2013; Sanz Álava, 2007; Perea Siller, 2013). In the studies on academic genres, the research around the School of Valparaíso stands out (Parodi, 2010). This team has compiled an academic corpus PUCV-2006 (http://www.elgrial.cl/) gathering texts form the academic and professional areas of four domains: industrial chemistry, construction, engineering, social work, and psychology. However, this corpus has not yet been used for the research of lexical phenomena. In the same vein, the reference handbook on academic and professional writing in Spanish, edited by Montolío Durán (2014), does not include any chapter entirely devoted to phraseology. The only previous work on academic lexical combinations in Spanish comes mainly from researchers who conducted contrastive studies in Spanish and English; that is, research focusing on non native speakers of English and dealing with the differences between academic lexical combinations in these two languages (see, among others, Tracy-Ventura et al., 2007; Cortes, 2008; Perales-Escudero & Swales, 2011; Pérez-Llantada, 2014).

Even if English is gaining ground in Spanish universities, Spanish is still the most used language in academic texts by university students. However, to be a native speaker of a language does not guarantee to be academically competent in this language: there is no native speaker of academic language and, therefore, the competence in academic writing has to be learnt. Despite of writing in their native language, the academic writings of university students show often certain deficiencies, many of which come from a poor knowledge of ALC. If the difficulty is considerable for students who write in their L1, the challenge is still bigger in an L2. The growing number of foreigner students in the Spanish universities has shown the need of lexical resources which help them in their academic writing. Furthermore, these resources could also serve to improve the academic writing of experts researchers, since, due to internationalization, their academic L1 begins to be damaged (Johansson Kokkinakis et al., 2012).

The research presented in this paper forms part of a project that intends to fill that gap: we aim to build a combined dictionary-corpus tool in accordance with the current trends in lexicography, where resources provide lexical information in the form of a concordance program exploiting language corpora, instead of doing so only in the form of a dictionary (Asmussen, 2013; Paquot, 2012). Our focus is the discourse and phraseological conventions of academic Spanish in different domains, as we will explain later. In order to build up a useful resource, we need also to study the academic writing of students and examine the differences between the command of novice and expert writing with respect to ALC. The research questions behind the whole project are very similar to those presented by Cortes (2004):

1. Which are the most frequent ALC in published academic writing?
2. How are these ALC classified in phraseological and functional terms? Are there more collocations, idioms or formulas? Can the functional classifications thought for English lexical bundles by Biber et al. (1999) or Hyland (2008) serve for Spanish ALC?
3. Are there any significant differences of these ALC between disciplines?
4. Are the ALC used by university students? Are there differences in Bachelor's degree and Master's degree students? In different disciplines?

To answer these questions we simultaneously take two perspectives: Corpus Linguistics and Phraseology. Corpus Linguistics provides us with tools (frequency and other measures of lexical association) useful to identify ALC candidates. Phraseology allows for selecting among these candidates by applying some criteria issued mainly from the Meaning-Text Theory (MTT) (Mel'čuk, 2015), keeping in mind that the final aim is to build up a useful tool for writing academic texts in Spanish.

This paper is structured as follows. Section 2 focuses on the different types of ALCs and tries to establish distinctions among the messy characterization of phraseological expressions present in the literature. Section 3 provides a description of the methodology we are using, along with a presentation of the expert academic corpus that we are studying and of the compilation of the student corpus. Section 4 is devoted to the description of the tool's design. There, we present how the corpus and the lexical database are intertwined and we provide some samples of prototype entries for different kinds of ALC. Finally, in Section 5, we draw some conclusions on the presented work and give future lines of research.

2. Academic phraseology: defining ALC

It is well known that there is not an established terminology to distinguish between different multiword units. Depending on different linguistic schools or traditions, what is a collocation for an author is a free phrase for another (e.g. *the results suggest*) and what is an idiom (*locución* in Spanish) from one perspective is considered a discourse marker from another (e.g. *in conclusion*), which is not contradictory. It is not only an issue of using different terms for the same concept, but also of labeling different concepts by means of the same term. The disagreement on the taxonomies of multiword units is not specific of research in the academic genre, but is common in phraseological inquiries, regardless of textual type.

In order to determine the phraseological nature of multiword sequences and to adscribe them to a phraseological category, we will adopt the tenets of Meaning-Text theory (Mel'čuk, 2015). Within this theoretical framework, two criteria are of paramount importance to ascertain whether a certain lexical combination is phraseological: its *compositionality* (not to be confused with its *transparency*) and the free or conditioned choice of its components. Compositionality is a property whereby the meaning of a given expression is the result of adding up the meanings of its constituent parts. Compositionality, which is production-oriented, should not be confused with transparency, which has to do with the understandabilty of an expression. Thus, an expression that is fully transparent is necessarily compositional, but the inverse is not true; for example, if a speaker does not know what the verb *respectar* means, he cannot guess the meaning of the expression *en*

lo que respecta a 'concerning X', even if this expression is fully compositional. Therefore, a compositional expression can be non-transparent.

If an expression is fully compositional, it could still be considered phraseological, as long as its components are not *freely* chosen or combined. When a phrase is *free*, each of its lexical components is selected strictly due to its meaning, independently of the lexical identity of other components (Mel'čuk, 2012, 33). The adjective *free* must be then understood strictly as allowing the selection of one lexical unit independently of the other lexical components of the same expression (Mel'čuk, 2012, 33). Thus, in the Spanish phrases *la probabilidad de que* ('the probability that') or *al revisar la selección* ('when reviewing the selection'), each of their lexical components is selected because of its meaning and combinatorial properties in conformity with the corresponding rules of Spanish (Mel'čuk, 2015, 59).

In contrast, a non-free phrase (*lexical phraseme*, in MTT terminology) is not constructed out of its lexical components by selecting each individually and arranging them according to the standard rules of L. Other non-standard rules specify a non-free phrase as a whole. The constraints that operate in the production of a non-free phrase can take place at different levels. Depending on compositionality and the type of constraint, our theoretical framework distinguishes several types of lexical phrasemes. The following pages focus on three types: idioms, collocations and formulas, which are the ALCs that our lexical tool will include. In what follows we are going to present each in turn.

2.1 ALC: idioms

We consider an idiom any non-free phrase if it is non compositional. An idiom is selected as a whole: from its semantic representation, a special rule maps its meaning to a single lexical node in a syntactic representation. Thus, for example, *en conclusión* (or its English equivalent, *in conclusion*) is one lexical unit, yet made up of two words. It should be the headword of its own lexicographical entry with its definition, its part of speech, and all relevant combinatorial information.

Idioms are very frequent in academic prose, especially those considered discourse markers from other perspectives: *en consecuencia* ('consequently'), *al contrario* ('on the contrary'), *por otra parte* ('on the other hand'), etc., although we encounter other types, such as verbal idioms, like *llevar a cabo* ('carry out'), *dar lugar* ('bring about') or *tener en cuenta* ('take into account'), nominal idioms such as *punto de vista* ('point of view') and — fewer — adjectival idioms.

There is also an overlap between idioms and lexical bundles; for instance, *en relación con* ('in relation with') is traditionally included in Spanish dictionaries as a prepositional idiom.

2.2 ALC: collocations

Unlike idioms, collocations are compositional. They are composed of two lexical units: the *base*, the selection of which is semantically-driven and the *collocate*, which is chosen not only on semantical, but also on lexical grounds (Mel'čuk, 1996, 37). Thus, in the verbal collocation *sacar conclusiones* ('draw conclusions'), the base *conclusión* conditions

the choice of the collocate *sacar* (lit. 'pull out'). If the base were *decisión* ('decision'), the choice of the support verb would be different: namely, *tomar* (lit. 'take'). Even if collocations are compositional, they are phraseological because the choice of one of its components is constrained by the other. The lexicographical description of each collocation should be made under the entry of the base. We intend that the user of our lexical tool will be able to recover information on collocates by means of an inverse search (see Section 4).

In academic prose, we focus on verbal collocations with the syntactic pattern verb-object, and also subject-verb, as *problema* and *estribar* in e.g. *el problema estriba* ('the problem lies'). Adjectival collocations are also object of our interest: *conclusión correcta*, *obvia, lógica*, *contraria*.

2.3 ALC: formulas

Formulas (*formulemes* in terms of MTT) are also compositional: *en otras palabras* ('in other words'), *es bien conocido que* ('it is well known that'), *no hay que olvidar que* ('we should not forget that'), *como se ha señalado previamente* ('as previously stated'), etc. However, both the meaning of a formula and its lexical implementation are constrained. Mel'čuk (2015) points out that if a speaker has the intention: 'I will now express the same content I have just expressed, but using different words', he cannot select the meaning 'I signal that the following fragment of my speech means the same as the preceding fragment' (the meaning of expressions such as *in other words* or *to put it differently* has more to do with the notion of 'rephrasing' than with that of 'repeating ideas'). From the former meaning, the speaker is not free to select any fairly synonymous expression, such as *using some different expressions* or *I say this in a different way*, because these expressions are not natural in English. The same happens in Spanish. From the same semantic representation, a Spanish speaker could produce *en otras palabras* and *dicho de otro modo/otra forma/otra manera*, but not *por ponerlo diferente* (cf. Eng. 'to put it differently').

As shown, formulas are doubly constrained. However, they do not need a lexicographic definition because a formula means exactly what it says. They need, in contrast, a description of its discourse function, especially in academic discourse (Cortes, 2004, 241). Thus, users of a lexical tool as the one proposed could obtain, for instance, different ways to emphasize a statement; e.g., *hay que destacar* ('it is necessary to stand out'), *es importante subrayar* ('it is important to emphasize'), *mención especial merece* ('it is worth mentioning'), etc.

Even though academic texts swarm with formulas, their theoretical status is not sufficiently clear. English dictionaries collect formulas such as *in other words*, (*and*) *what's more*, etc., but the Spanish dictionaries do not. For example, *en lo que respecta* ('in what concerns') appears under the headword *respectar* but this verb is defective and is used only in this expression with the variant (*por lo que respecta*). Other formulas are perceived as having less "lexical entity". Thus, recurrent sequences of academic discourse such as Engl. *the aim of this work is*, *the results suggest*, *this study has shown that*, among others, are not collected as phrases in any academic English dictionary, although they appear in lists of lexical bundles.

This third category is perhaps the one having more in common with the concept of *lexical bundle*, which has gained increasing acceptance in current research in academic

discourse. However, in contrast to our formulas, lexical bundles are not defined on account of the choice or their components and their compositionality, but on purely distributional terms: *lexical bundles* are contiguous word sequences or n-grams that display a minimum frequency (usually from 10 to 40 occurrences per million words) and a minimum dispersion in corpora (cf. Biber et al., 1999). Apart from the theoretical differences, it could be relatively safely stated that all formulas are lexical bundles, but the opposite is not always the case. For example, *la probabilidad de que* ('the probability that') can be considered as a lexical bundle by virtue of its recurrence and dispersion, but from our perspective this multiword sequence is not phraseologically relevant. As we will explain in the next section, the techniques developed to identify and extract lexical bundles are useful for our research, but lexical bundles themselves are, so to speak, raw materials that have to be processed before being included in our lexical tool.

2.4 Recapitulation

The limits between the three different ALCs are not always completely clear. The compositionality draws a boundary between idioms, on the one hand, and collocations and formulas, on the other. When one of the components is a grammatical word, the distinction is less obvious. For instance, *sin duda* ('without a doubt') seems to be compositional because its meaning includes 'without' and 'doubt'. However, its meaning includes also a discourse semantic component that emphasizes speaker's statements.

ALCs can merge sometimes. This happens, for instance, when a formula contains a collocation. In academic prose it is frequent to encounter formulas such as *la pregunta que nos tenemos que formular* ('the question we should ask'), which includes the verbal collocation *formular una pregunta* ('to ask a question').

In our lexical tool, all formulas and some idioms will receive a discourse function. Collocations will be included in the entries of their respective bases and will not be associated to any specific discourse function, since arguably those are associated to specific sequences of words. E.g., the lexical entry for the base *pregunta* will include all its collocates, but its collocations will not have discourse function because this one is associated only to a concrete sequences of words.

3. A not so radical corpus-driven approach to academic phraseology

Our methodology is corpus-driven, but not as radical as the one adopted by Biber (2009, 281). Even if corpora play a leading role in the construction of our lexical tool, we need to filter out corpus output with phraseological criteria, which makes human intervention necessary. This section describes the corpora used for our study and the methodology applied to extract information from them.

3.1 Corpus description

We need two types of corpora: first, an expert academic corpus in order to obtain the list of ALCs for our lexical tool. The corpus used is the Spanish part of the Spanish–English Research Article Corpus (SERAC 2.0), a 5.7-million word compilation of 1,056

research articles (RAs). It includes 360 L1 RAs in Spanish published by Spanish scholars in peer-reviewed journals targeted at a national-based scholarly readership (Pérez-Llantada, 2014). The corpus contains about two million running words. It is divided into four sections, namely: Arts and Humanities, Biological and Health Sciences, Physical Sciences and Engineering, Social Sciences and Education.

Second, we have begun to compile a novice academic corpus for Spanish with a view to building a resource similar to BAWE (Gardner & Nesi, 2013) or MICUSP (Römer & O'Donnell, in preparation) for English. We are compiling Bachelor's and Master's degree theses of Spanish university students in the same areas as the expert corpus. The identification of student's difficulties with ALC in this corpus will be key for the design of the lexical tool that we project.

3.2 Quantitative approach

Currently, we have completed the compilation of a list of academic Spanish words and the extraction of academic collocations, formulas and idioms is in progress.

The Spanish Academic Word List (SAWL) consists of about 1,000 lemmas of content words (nouns, verbs, adjectives and adverbs) and has been extracted following two criteria (similar to Coxhead (2000) or Paquot (2010), among others): (a) the *keyness* of the forms extracted and (b) their dispersion. The keyness of the lemmas has been determined by comparing their distribution in the expert corpus and in a contrast corpus (the narrative part of the LEXESP corpus, Sebastián-Gallés et al., 2000) by means of the Wilcoxon-Mann-Whitney test (cf. Kilgarriff, 2001; Lijffijt et al., 2014). We retained those items with a significance of $p < 0.001$. To avoid vocabulary specific of only a certain thematic field, we have controlled for dispersion using Gries's DP coefficient (Gries, 2008) by including only those items with a value of 0.5 or less (cf. Durrant, 2014).

This vocabulary list will be further manually filtered assessing the collocational and the discourse productivity of its items: if a word of the SAWL is productive as a basis of many collocations and it is a member of formulas with discourse functions, it will be candidate to be part of the macrostructure of the lexical database. Collocations will be extracted by using dependency parsing and measures of lexical association. Such extraction procedure in all probability will yield combinations with different phraseological status (e.g. collocations such as *extraer conclusiones* and idioms such as *tener en cuenta*) that will have to be manually sorted out.

The extraction of recurring n-grams seems a strategy more suitable to extract formulas and certain types of idioms such as prepositional or adverb phrases, made up of contiguous word sequences (e.g., *a través de, sin embargo, no obstante*; cf. Tutin & Kraif, 2017). A preliminary analysis has put into question the suitability of keyness when filtering lists of n-grams for our current purposes: such filtering yields poor recall values, since a considerable amount of phraseologically interesting multiword chains do not reach significance thresholds. Likewise, while frequency thresholds conventionally used for retrieving lexical bundles produce acceptable results with 4-grams, additional measures seem to be necessary in order to get rid of 2-grams and 3-grams of dubious interest (backwards transition probability as proposed in Appel & Trofimovich (2015), seem to get the best results in our n-gram list).

3.3 Filtering and enriching raw data

We adopt a mixed-method approach similar to Simpson-Vlach & Ellis (2010) or Ackermann & Chen (2013), who also combined statistical information and human judgment when compiling their respective lists of academic lexical combinations for English. After obtaining n-gram lists by using statistical measures, we will apply phraseological criteria to discriminate between idioms, collocations and formulas. The classification is necessary because each type of ALC requires a different lexicographic description, as we will show in Section 4. Only idioms and formulas are enriched with discourse functions.

The typology of discourse functions is being obtained following a bottom-up approach. Even if we start from previous classifications of lexical bundles in English (Hyland, 2008; Simpson-Vlach & Ellis, 2010; Salazar, 2014), their taxonomy cannot be directly imported to Spanish ALCs. Most of them distinguish between three main functions: 1) describing research, 2) organizing text and 3) conveying the author's stance and interacting with reader (Salazar et al., 2013, 45). Each function has a long list of subfunctions that are not always easily interpretable for a potential user of a lexical tool. For instance, the function "framing", used by Hyland (2008) or Salazar (2014), groups together lexical bundles such as *with respect to*, *with the exception of*. The function framing serves to "situate arguments by specifying limiting conditions" (Salazar, 2014, 52). Even if the cited bundles fit within this definition, it might be useful to provide the user with more specific information about when to use each one. A similar objection can be raised against putting together *it should be noted that*, *see Figure 1, as seen in* under the function "address readers directly" (Salazar, 2014, 52). These formulas do indeed address readers directly, but they do not have the same discourse function in an academic text: the first one boosts the statement that follows, whereas the other two point out specific fragments of the text.

We aim to build a typology with the main discourse functions in academic writing more oriented to the user, following Gilquin et al. (2007) and Prat Ferrer & Peña Delgado (2015) with simple and clear headings (e.g., "how to begin", "changing subject", "presenting conclusions", etc. see Figures 1–4). We adopted the following process: first a sample of articles included in the expert corpus has been examined in order to put forward a list of discourse functions. We are studying which formulas and idioms fulfill these functions by checking the contexts where they occur. It is likely that the typology of discourse functions devised after the qualitative revision of the mentioned sample will be improved during this process. The final product will be a database of ALCs associated with discourse functions, rather than a corpus annotated with discourse functions.

The assignment of discourse functions cannot be made automatically, save perhaps some exceptions. Thus, a formula such as *esta es la principal conclusion* ('this is the main conclusion') is not necessarily used to conclude. Its context must be examined in order to verify whether, for instance, it is mainly employed to introduce the conclusions of a paper or to refer to the research of other authors, as in *esta es la principal conclusión a la que llega el estudio X*, etc. We are aware that this manual assignment is slow, but we project to get complete products by working discourse functions. In this way, we can obtain finished descriptions in different phases of our project, such as "ALC which serve to conclude", "ALC which serve to emphasize", etc.

4. Design of lexical tool HARTA[1]

We aim to build a combined dictionary-corpus tool in accordance with the current trends in Lexicography, where resources can provide lexical information by means of concordances coming from corpora, ins addition of doing so only in the form of a dictionary (Asmussen, 2013; Verlinde & Peeters, 2012). The corpus is intertwined with the lexical database, because, in many cases, user queries are more easily answered by showing examples of a given ALC in corpus, rather than by offering a whole lexicographic description. In the last few years, several authors recommend to expose both L2 learners and novice writers to corpus-based evidence (Cortes, 2013; Pérez-Llantada, 2014; Cotos, 2014). More recently, Laso & John (2017) have taken a step beyond awareness-raising by investigating the influence of corpus consultation on the written production.

4.1 Macrostructure of HARTA

The macrostructure of HARTA is only partially based on the list SAWL. The selected headwords must fulfill a discourse function or be part of an ALC fulfilling one. There will be two kinds of lexicographical entries: proper entries for single and multiword lexical units, with all the information an entry is supposed to contain in an MTT framed dictionary (semantic, syntactic and combinatorial), and *ad hoc* entries for formulas. As explained above, many formulas are not properly a lexical entity, but it is useful for the user to access them through their discourse function. Thus, for instance, the noun *resultado* ('result') is chosen to be part of the macrostructure and will receive a whole entry because this noun is part of several formulas fulfilling discourse functions (*estos resultados sugieren/indican que*). Likewise, the idiom *punto de vista* ('point of view') will receive an entry because it is part of several formulas used to cite or to convey the author's perspective (*desde nuestro punto de vista*). Some idioms are used to serve a discourse function as a whole, such as *en conclusión* ('in conclusion') and, therefore, they will be provided with a proper entry also. For formulas we will choose a canonical form on criteria similar to those employed by Salazar (2014) to establish *prototypical bundles*.

4.2 Microstructure of HARTA

The whole entry includes information of two types: 1) the core information, consisting of semantic and combinatorial information about the lexical unit and 2) the usage information, including frequency, disciplines in which the unit occurs, etc., and access to the corpora (see Figure 1).

The entry for a formula contains the following fields (see Figure 2):

1. Discourse functions. A formula can have more than one function: e.g. as Salazar et al. (2013, 46) point out *these results suggest* serves to draw conclusions, but involves also the function of hedging due to the use of mitigating verb *suggest*.
2. Disciplines where the formula appears. Some disciplines are more prone than other to use a lofty style. Thus, a formula such a *mención especial merece* will probably be less frequent in Sciences than in Literature research.

[1] HARTA stands for *Herramienta de Ayuda a la Redacción de Textos académicos* ('tool of help for writing academic texts').

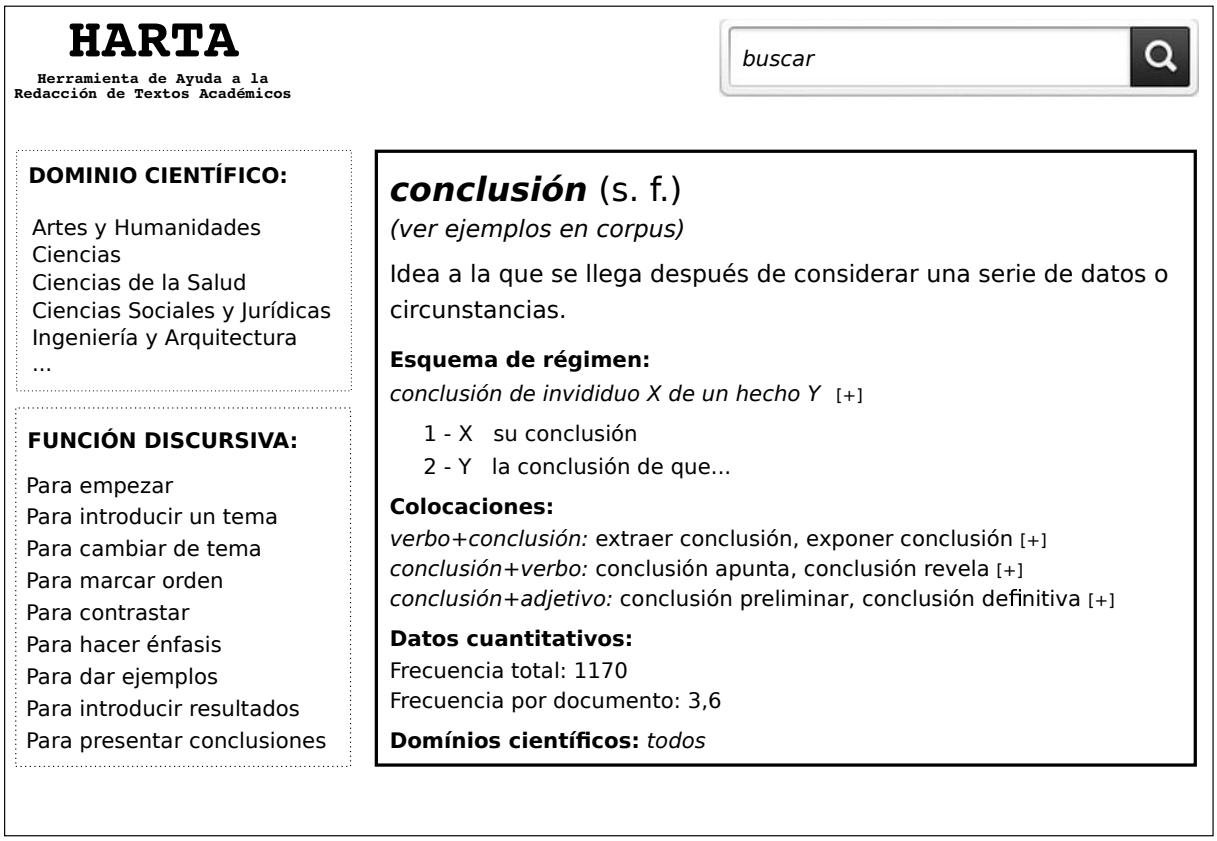


Figure 1: Entry for the noun *conclusión*

3. Frequency of co-occurrence. It is useful information for the user to know if the formula expressing a given discourse function is more or less productive than others.
4. The sections of the research article where the formula appears. We have marked up the sections of the text included in the expert corpus (abstract, introduction, body (method, result, discussion), conclusion). As Salazar et al. (2013, 49) pointed out, the discourse function can vary according to the section of the text. For instance, the formula *in accordance with* has the function of describing a procedure in the Methods section, whereas it is used to present the results from previous studies in the Discussion section.

Any lexical component of a formula will have a hyperlink to its own entry or trigger another kind of search. E.g., for the formula in Figure 2, there would be a hyperlink to the information associated to the idiom *tener en cuenta* ('to take into account').

4.3 Different access to the information

There will be two main search types: 1) the discourse function search and 2) the word search.

In the discourse function search, the user will be able to select a given function and get all the formulas fulfilling this function. In Figure 3 the user clicks on *para hacer énfasis* ('to emphasize'), and the tool provides a list of formulas (in their canonical form) which can be ordered alphabetically or by frequency. If the user clicks on each formula, he sees the entry (see Figure 2).

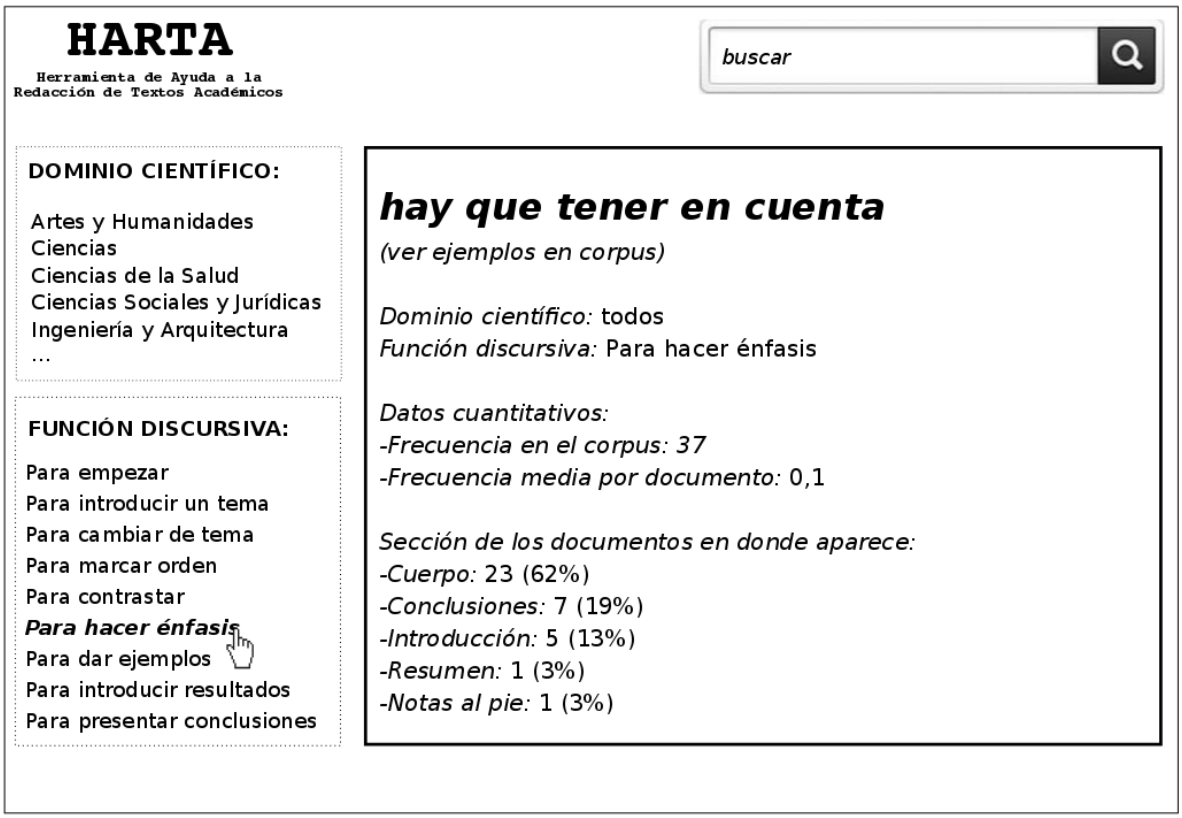


Figure 2: Entry for a formula

Furthermore, information will be accessible through word search (Figure 4). For example, if the user wants to obtain information on the noun *resultado* ('result'), the interface will provide access to its entry, if there is one, or to the formulas, idioms and collocations where it occurs. The entry for the noun *resultado* displays links to its collocations. More information will be found when clicking on the entry (see Figure 1 where you can see the proper entry for *conclusion*).

If the queried word has no proper entry, the interface will provide the formulas and the collocations in which it occurs. For instance, if an user looks up for the verb *sugerir* ('suggest'), the inferface would provide the formulas and all nouns which are the subject of this verb in collocations: *autor*, *análisis, dato, experimento*, *resultado*, etc. It should be noted that this information is what a search on a collocational database returns, not the static information included in an entry. In our theoretical framework we claim that collocational information must be described in the base's entry but should be recoverable both through the base and the collocate.[2]

5. Conclusions

This paper has presented an ongoing research on academic lexical combinations in Spanish with the aim of building a lexical resource accessible on the web. In contrast to other

[2] This is the policy that we use in the compiling of the Spanish collocation dictionary DiCE (http://www.dicesp.com/). We will build entries for bases, but information for collocates will be recoverable through special searches (Alonso-Ramos, 2016; Alonso-Ramos et al., 2010).

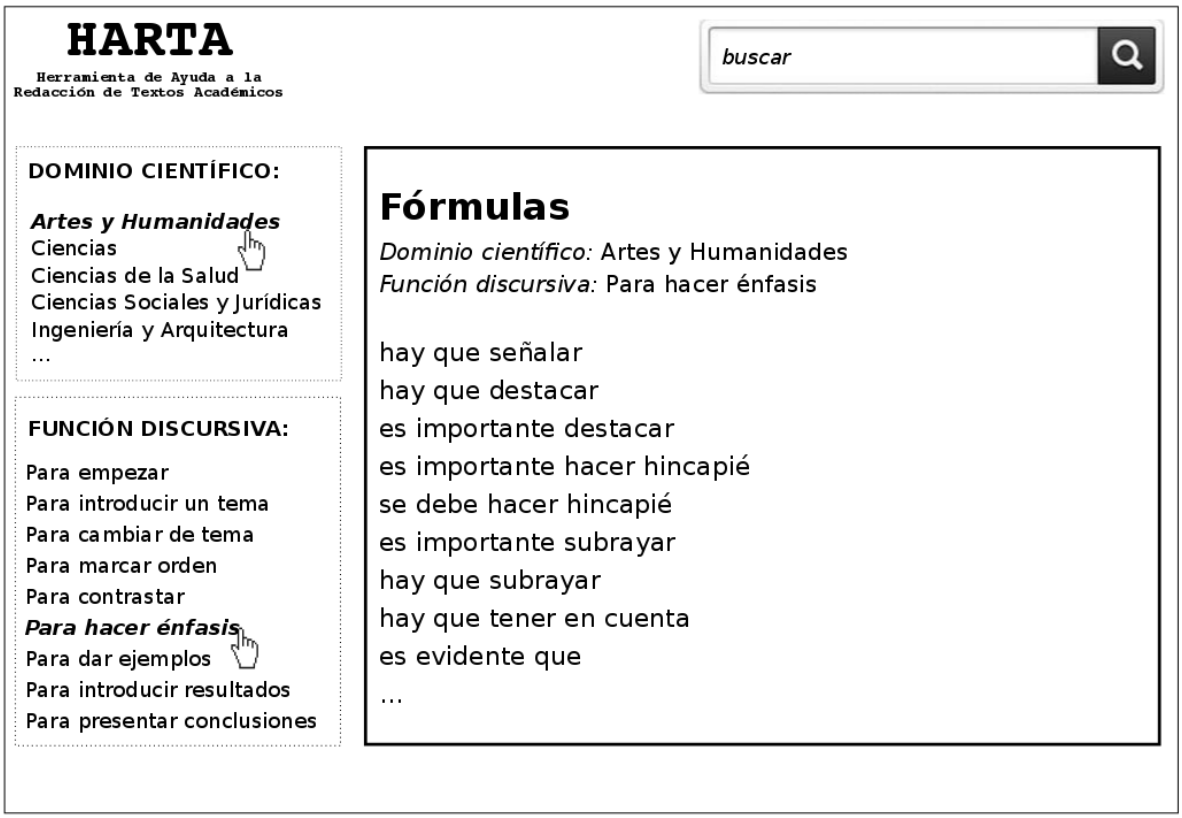


Figure 3: Discourse function search

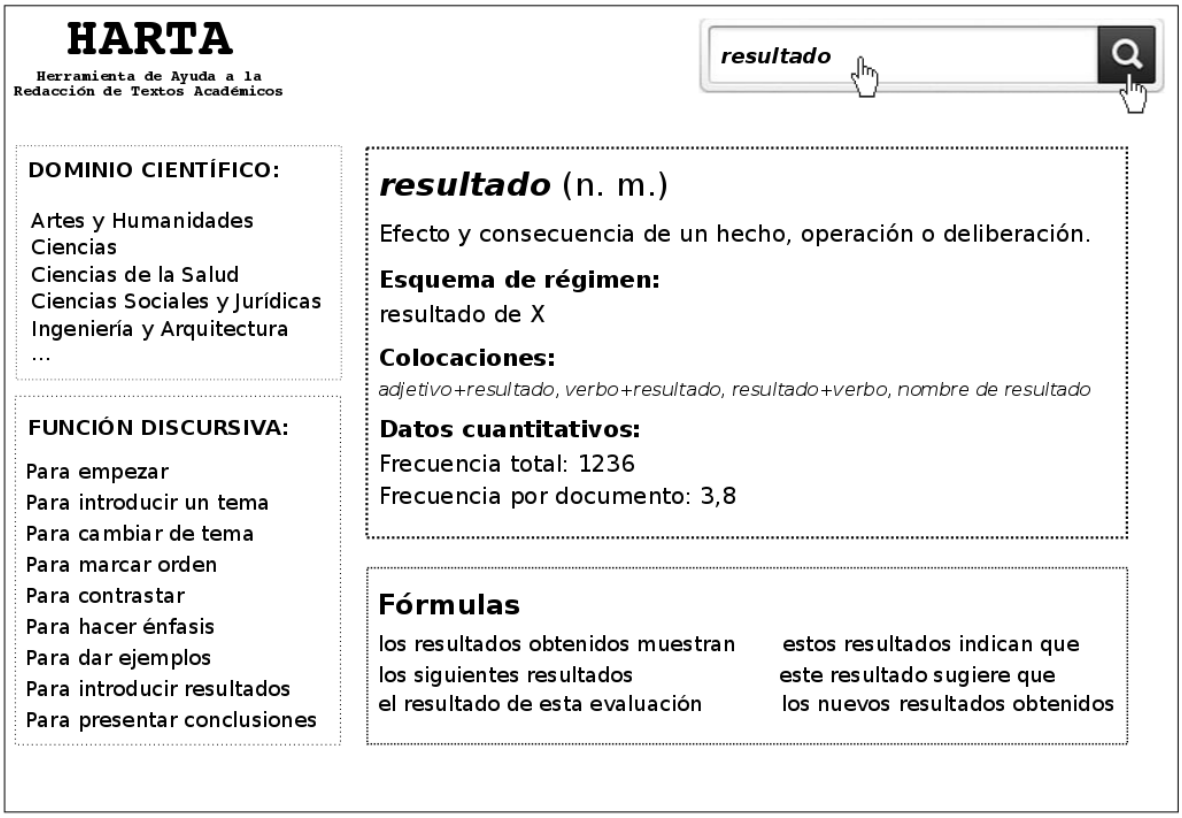


Figure 4: Word search

similar tools, such as LEAD (Paquot, 2012) or ScieLex (Verdaguer & Salazar, 2013), we intend a finer classification of phraseological units, since we rely on a theoretical framework that provides the necessary theoretical tools for the endeavour. We are aware that such distinctions involve a longer process. However, we project to get a product of increasing completeness along the successive stages of our research by devising an exhaustive classification of discourse functions. We believe that the final user will appreciate more a rich entry than lists of lexical bundles organized by mere frequency. In the meantime, access to the expert corpus will be profitable for any user.

We will better adapt to user needs when we have analyzed the student corpus. Differences in frequency of use between expert and novice writers will provide clues as to the difficulties faced by the latter and, accordingly, the type of information that should be given priority in the different entries. This analysis can also provide teaching material devoted to novice writers such as Salazar (2014) proposes.

6. Acknowledgements

The work presented in this paper has been partially supported by the Spanish Ministry of Economy and Competitiveness (MINECO), by the FEDER Funds of the European Commission under the contract number FFI2016-78299-P, by a postdoctoral fellowship granted by the Galician Government (POS-A/2013/191), and by a *Juan de la Cierva formación* grant (FJCI-2014-22853).

7. References

Ackermann, K. & Chen, Y.H. (2013). Developing the Academic Collocation List (ACL) – A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes*, 12, pp. 235–247.

Alonso-Ramos, M. (2016). Learning resources for Spanish collocations: From a dictionary towards a writing assistant. In B.S. Vilas (ed.) *Collocations Cross-Linguistically. Corpora, Dictionaries and Language Teaching, volume C of Mémoires de la Société Néophilologique de Helsinki*. Helsinki, Finland: Société Néophilologique de Helsinki, pp. 65–95.

Alonso-Ramos, M., Nishikawa, A. & Vincze, O. (2010). DiCE in the web: An online Spanish collocation dictionary. In S. Granger & M. Paquot (eds.) *eLexicograpy in the 21st century: New Challenges, New Applications. Proceedings of eLex 2009*. Louvain-la-Neuve: Presses universitaires de Louvain, pp. 367–368.

Appel, R. & Trofimovich, P. (2015). Transitional probability predicts native and non-native use of formulaic sequences. *International Journal of Applied Linguistics*, 27, pp. 24–43.

Asmussen, J. (2013). Combined Products: Dictionary and Corpus. In R. Gouws, U. Heid, W. Sheweickard & H. Wiegand (eds.) *Dictionaries. An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography*. Berlin/Boston: De Gruyter Mouton, pp. 1081–90.

Biber, D. (2009). A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics*, 14(3), pp. 275–311.

Biber, D., Conrad, S. & Cortes, V. (2004). If you look at...: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, 25(3), pp. 371–405.

Biber, D., Johansson, S., Leech, G., Conrad, S. & E., F. (1999). *Longman Grammar of Spoken and Written English*. Harlow: Pearson.

Cavalla, C. & Loiseau, P. (2014). Scientext comme corpus pour l'enseignement. In Tutin & Grosman (eds.) *L'écrit Scientifique: Du Lexique Au Discours*. Rennes: Presse universitaire de Rennes, pp. 163–180.

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes*, 23(4), pp. 397–423.

Cortes, V. (2008). A comparative analysis of lexical bundles in academic history writing in English and Spanish. *Corpora*, 3, pp. 43–58.

Cortes, V. (2013). The purpose of this study is to: Connecting lexical bundles and moves in research article introductions. *Journal of English for Academic Purposes*, 12, pp. 33–43.

Cotos, E. (2014). Enhancing writing pedagogy with learner corpus data. *ReCALL*, 26, pp. 202–224.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, 34(2), pp. 213–238.

Durrant, P. (2014). Discipline and level specificity in university students' Written vocabulary. *Applied Linguistics*, 35(3), pp. 328–356.

Gardner, S. & Nesi, H. (2013). A classification of genre families in university student writing. *Applied Linguistics*, 34(1), pp. 1–29.

Gilquin, G., Granger, S. & Paquot, M. (2007). Writing sections. In M. Rundell (ed.) *Macmillan English dictionary for advanced learners*. Oxford: Macmillan Education, 2 edition, pp. 1–29.

Gries, S. (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics*, 13(4), pp. 403–437.

Hyland, K. (2008). As can be seen: lexical bundles and disciplinary variation. *English for Specific Purposes*, 27, pp. 4–21.

Johansson Kokkinakis, S., Sköldberg, E., Henriksen, B., Kinn, K. & Bondi Johannessen, J. (2012). Developing Academic Word Lists for Swedish, Norwegian and Danish – a Joint Research Project. In R.V. Fjeld & J.M. Torjusen (eds.) *Proceedings of the 15th EURALEX International Congress*. Oslo: Department of Linguistics and Scandinavian Studies. University of Oslo, pp. 563–569.

Kilgarriff, A. (2001). Comparing Corpora. *International Journal of Corpus Linguistics*, 6(1), pp. 97–133.

Krause Kilian, C. & Dias Loguercio, S. (2015). Fraseologias de gênero em resumos científicos de Linguística, Engenharia de Materiais e Ciências Econômicas. *Tradterm*, 26, pp. 241–267.

Laso, N. & John, S. (2017). The pedagogical benefits of a lexical database (SciE-Lex) to assist the production of publishable biomedical texts by EAL writers. *Ibérica*, 33, pp. 147–172.

Lea, D., Bull, V. & Webb, S. (eds.) (2014). *OLDAE: Oxford Learner's Dictionary of Academic English*. Oxford: Oxford University Press.

Lijffijt, J., Nevalainen, T., Saily, T., Papapetrou, P., Puolamaki, K. & Mannila, H. (2014). Significance testing of word frequencies in corpora. *Digital Scholarship in the Humanities*, 31(2), pp. 374–397.

McCarthy, M. & O'Dell, F. (2008). *Academic Vocabulary in Use: 50 Units of Academic Vocabulary Reference and Practice ; Self-study and Classroom Use*. Cambridge: Cambridge University Press.

Mel'čuk, I. (1996). Lexical functions: a tool for the description of lexical relations in the lexicon. In L. Wanner (ed.) *Lexical Functions in Lexicography and Natural Language Processing*. Amsterdam: John Benjamins, pp. 37–102.

Mel'čuk, I. (2012). Phraseology in the language, in the dictionary, and in the computer. *Yearbook of Phraseology*, 3(1), pp. 31–56.

Mel'čuk, I. (2015). Clichés, an Understudied Subclass of Phrasemes. *Yearbook of Phraseology*, 5, pp. 35–50.

Montolío Durán, E.d. (2014). *Manual de escritura académica y profesional*. Barcelona: Ariel.

Paquot, M. (2010). *Academic vocabulary in learner writing: from extraction to analysis*. London/New York: Continuum.

Paquot, M. (2012). The LEAD dictionary-cum-writing aid: an integrated dictionary and corpus tool. In S. Granger & M. Paquot (eds.) *Electronic Lexicography*. Oxford University Press, pp. 163–185.

Parodi, G. (ed.) (2010). *Academic and Professional Discourse Genres in Spanish*. Amsterdam/Philadelphia: John Benjamins.

Perales-Escudero, M. & Swales, J. (2011). Tracing convergence and divergence in pairs of Spanish and English research article abstracts: The case of Ibérica. *Ibérica*, 21, pp. 49–70.

Perea Siller, F.c. (2013). *Comunicar en la Universidad. Descripción y metodología de los géneros académicos*. Córdoba: Universidad.

Pérez-Llantada, C. (2014). Formulaic language in L1 and L2 expert academic writing: Convergent and divergent usage. *Journal of English for Academic Purposes*, 14, pp. 84–94.

Prat Ferrer, J. & Peña Delgado, A. (2015). *Manual de escritura académica*. Madrid: Ediciones Paraninfo.

Regueiro Rodríguez, M. & Sáez Rivera, D. (2013). *El Español Académico. Guía Práctica Para La Elaboración de Textos Académicos*. Madrid: Arco Libros.

Römer, U. & O'Donnell, M.B. (in preparation). *MICUSP: A Corpus Resource for Exploring Proficient Student Writing across Disciplines*. Amsterdam: John Benjamins.

Salazar, D. (2014). *Lexical Bundles in Native and Non-native scientific writing*. Amsterdam/Philadelphia: John Benjamins.

Salazar, D., Verdaguer, I., Laso, N., Comelles, E., Castano, E. & Hilferty, J. (2013). Formal and functional variation of lexical bundles in biomedical English. In J.L. Isabel Verdaguer N. & D. Salazar (eds.) *Biomedical English: A corpus-based approach*, volume 56 of *Studies in Corpus Linguistics*. John Benjamins Publishing Company, pp. 39–54.

Sanz Álava, I. (2007). *El Español Profesional y Académico en el aula universitaria. El discurso oral y escrito*. Valencia: Tirant Lo Blanch.

Sebastián-Gallés, N., Martí Antonín, M., Carreiras Valiña, M. & Cuetos Vega, F. (2000). *LEXESP: Léxico informatizado del español*. Barcelona: Edicions de la Universitat de Barcelona.

Simpson-Vlach, R. & Ellis, N. (2010). An Academic Formulas List: New Methods in Phraseology Research. *Applied Linguistics*, 31(4), pp. 487–512.

Swales, J. & Feak, C. (2012). *Academic Writing for Graduate Students: Essential Tasks and Skills*. Michigan series in English for academic & professional purposes. Ann Arbor: University of Michigan Press.

Tracy-Ventura, N., Cortes, V. & Biber, D. (2007). Lexical bundles in speech and writing. In G. Parodi (ed.) *Working with Spanish Corpora*. London: Continuum, pp. 217–231.

Tutin, A. (2010). Showing phraseology in context: Onomasiological access to lexicogrammatical patterns in corpora of French scientific writings. In S. Granger & M. Paquot (eds.) *eLexicography in 21st century. New Challenges, new applications*. Louvain-la-neuve: Presses universitaires de Louvain, pp. 313–324.

Tutin, A. (2014). La phraséologie transdisciplinaire des écrits scientifiques: des collocations aux routines sémantico-rhétoriques. In A. Tutin & F. Grossman (eds.) *L'écrit scientifique: du lexique au discours. Autour de Scientext*. Rennes: PUR, pp. 24–44.

Tutin, A. & Grossmann, F. (2014). *L'écrit scientifique: du lexique au discours. Autour de Scientext*. Rennes: Presse universitaire de Rennes.

Tutin, A. & Kraif, O. (2017). Comparing Recurring Lexico-Syntactic Trees (RLTs) and Ngram Techniques for Extended Phraseology Extraction: a Corpus-based Study on French Scientific Articles. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017) at the European Chapter of the Association for Computational Linguistics Conference (EACL 2017)*. Valencia: Association for Computational Linguistics, pp. 176–180.

Vázquez, G. (2001). *Guía didáctica del discurso académico escrito: ¿cómo se escribe una monografía?* Madrid: Edinumen.

Verdaguer, Laso, N. & Salazar, D. (eds.) (2013). *Biomedical English. A corpus-based approach*. Amsterdam/Philadelphia: John Benjamins.

Verlinde, S. & Peeters, G. (2012). Data access revisited: the Interactive Language Toolbox. In S. Granger & M. Paquot (eds.) *Electronic lexicography*. Oxford: Oxford University Press, pp. 147–162.

The OntoLex-Lemon Model: Development and Applications

John P. McCrae[1], Julia Bosque-Gil[2], Jorge Gracia[2], Paul Buitelaar[1], Philipp Cimiano[3]

[1]Insight Centre for Data Analytics, National University of Ireland Galway
[2]Ontology Engineering Group, Universidad Politécnica de Madrid, Spain
[3]Cognitive Interaction Technology Excellence Cluster, Bielefeld University
E-mail: john@mccr.ae, jbosque@fi.upm.es, jgracia@fi.upm.es, paul.buitelaar@insight-centre.org, cimiano@cit-ec.uni-bielefeld.de

Abstract

The *lemon* model has become the primary mechanism for the representation of lexical data on the Semantic Web. The *lemon* model has been further developed in the context of the W3C OntoLex community group, resulting in the new OntoLex-Lemon model, recently published as a W3C report. In this paper, we describe the development and future outlooks for this model as well as briefly review some of its current applications. The recent evolution of *lemon* into OntoLex-Lemon, in the context of the community group, has led to improvements on the model that further extends its application domain from formal applications such as question answering and semantic parsing to the representation of general machine-readable dictionaries, including WordNet and digitized versions of existing dictionaries.
We look at two use cases of the OntoLex-Lemon model: in representing dictionaries and in the WordNet Collaborative Interlingual Index. Finally, we consider the future of the OntoLex-Lemon model, which we intend to continue to develop and have recently identified areas that increase the applicability and value of the model to more users.

Keywords: linked data; lexicography; ontologies; Semantic Web; ontology-lexicon interface

1. Introduction

The use of ontologies has become an increasingly important method for modelling domains and representing data in a variety of forms, most notably the Semantic Web. However the existing standards for ontologies, in particular the Web Ontology Language (McGuinness & van Harmelen, 2004: OWL), provide little support for the representing information about how a word is expressed in language beyond a simple string. In order to close this gap, the *lemon* model (McCrae et al., 2012) was proposed, which created a separate lexicon that could describe how an ontological concept was lexicalized in more detail. This builds on the paradigm of the ontology-lexicon interface, as well as existing models for lexicography including LMF (Francopoulo et al., 2006) and the EAGLES[1] and ISLES projects[2], where the expression of a concept in natural language and its formal description in the ontology are kept separated. This has several advantages, most notably in that by separating the ontological and the lexical layer we can easily switch an ontology from one language to another by changing its lexicon.

The *lemon* model was adopted by a number of projects (Ehrmann et al., 2014; Navigli & Ponzetto, 2012; Sérasset, 2015; Eckle-Kohler et al., 2015) and several authors have proposed modifications, improvements or changes (Khan et al., 2014; Chavula & Keet, 2014; Bosque-Gil et al., 2016a; Gracia et al., 2014) to the model. In order to accommodate these changes, it was decided that the model should be further developed under an open forum and some of the authors of this paper founded the OntoLex Community Group.[3]

[1] http://www.ilc.cnr.it/EAGLES/home.html
[2] http://www.ilc.cnr.it/EAGLES96/isle/ISLE_Home_Page.htm
[3] https://www.w3.org/community/ontolex

This group was part of the World Wide Web Consortium's Business and Community group program, a new initiative to support the development of emerging standards on the Web. The results of this group's work was the publishing of an updated version of the model in May 2016, namely the OntoLex-Lemon model.

The new OntoLex-Lemon model has already been applied in a number of cases. In this paper we will examine some of these use cases, in particular looking at the expanded use case of the model for representing existing dictionaries and the conversion of several existing commercial and free dictionaries. Secondly, we will consider the use of the OntoLex model in the recently proposed Global WordNet Interlingual Index (Vossen et al., 2016; Bond et al., 2016), whereby the model is used as a foundation for creating a truly interlingual concept index.

Finally, in this paper we will also provide an outlook of the next steps we aim to achieve for the model, in particular in terms of the new modules that we aim to create in order to address concerns raised in the community. Thus, we briefly sketch four modules on morphology, lexicography, etymology (and diachronicity) and lexical categories.

2. The OntoLex Community Group

The OntoLex Community Group was founded in December 2011 to support the development of a model for the representation of lexical information relative to ontologies. The group provided a number of tools for collaboration on this task including a wiki and a public mailing list for discussion of topics. Moreover, the group chaired by the authors of this paper organized public telephone conference calls, of which over 70 have taken place between 2012 and 2016. The group developed the model firstly by collecting relevant use cases,[4] and then distilling this into a set of essential requirements[5] for the model. Then, the development of the model took place in two stages: firstly the *core* model was defined, which incorporates the basic elements that it was assumed that all applications of the model would use and then in the second stage, four extra modules were defined: Syntax and Semantics, Decomposition, Variation and Translation, and Metadata (Lime). Finally, these models were combined and documented in a final report that was published by the W3C (Cimiano et al., 2016) along with technical model files in OWL.

One significant difference in the creation of this standard, in contrast to the processes of other standard organizations, was the degree of openness in the development of this model. The community group has over 100 members from a very diverse number of institutes and this is due to the fact that admission to the group was dependent only on assenting to a short agreement that any contributions would be open. Moreover, many issues of the model were decided by open conversation or votes and all of these contributions are available publicly in the form of wiki contributions and mailing list posts, all of which are archived on the Web and accessible to anyone.[6]

3. The OntoLex-Lemon Model

Here we provide a brief summary of the OntoLex-Lemon model, for a more complete description please see Cimiano et al. (2016). The OntoLex-Lemon model is based around

[4] https://www.w3.org/community/ontolex/wiki/Specification_of_Use_Cases
[5] https://www.w3.org/community/ontolex/wiki/Specification_of_Requirements
[6] https://lists.w3.org/Archives/Public/public-ontolex/

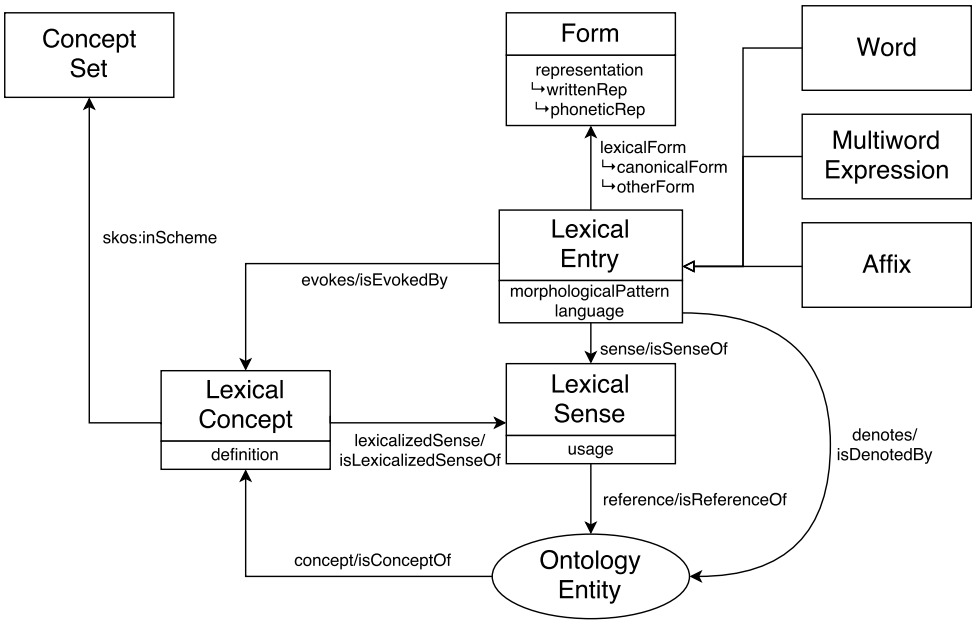


Figure 1: The Core OntoLex-Lemon Model

the core module, as depicted in Figure 1. The primary element of this is the *lexical entry* which represents a single word and thus collects together all morphological expressions of that word, which correspond to *forms* in the model, and all possible concepts in the ontology it can refer to, which correspond to *lexical senses* in the model. It is important to note that the actual meaning of a word is given by reference to an ontological concept and *lexical senses* represent only the mapping from a word to a concept. In contrast to the previous *lemon* model, a third semantic element called the *lexical concept* was introduced that allows for a meaning to be defined independently of an ontology. For example, the verb 'to die' may refer to different ontological properties such as `deathDate` and `deathPlace` while still referring to a single concept of *Dying*. The model also supports some other features including marking the canonical form (lemma), whether an expression is a multiword expression and giving a *usage* condition describing when a particular word expresses a given concept (for example the register), which is annotated on the lexical sense showing its role in giving a mapping between concepts.

In addition to the core, there are four modules defined by the specification:

Syntax and Semantics The syntax and semantics module describes how particular lexical constructs, e.g., verb frames, can be mapped to constructs in the ontology. As a simple example, this concerns how a transitive verb frame such as 'X knows Y' can be mapped to the subject and object arguments of a property such as `A foaf:knows B`. In this case there are only two options based on whether the grammatical subject (X) refers to the property's subject (A) or object (B), however more complex multi-argument structures are also covered.

Decomposition The decomposition module allows for multiword lexical entries to be decomposed into individual words, which are also represented as *components*. Components are allowed to be marked with their own grammatical properties and are said to *correspond to* either a lexical entry (i.e., for the word), an argument in a frame or another frame (to model phrasal arguments).

Variation and Translation The variation and translation model represents relationships between words at three levels: (purely) lexical, sense (lexico-semantic) and conceptual. These correspond to the levels of the model, with lexical relations being between lexical entries and as such not considering the meaning of a word and only its syntactic properties. Similarly, a conceptual relationship occurs between concepts and does not consider the lexical form and hence language of a relation. Sense relations require knowledge of both the word form and the meaning and translation is thus considered a special case of a sense relation. The module also allows technical modelling of a relation either as a single triple or as a dereferenceable entity in itself, which allows for further annotation of metadata about the link. This module integrates previously proposed extensions to *lemon* such as the translation module (Gracia et al., 2014).

Metadata (Lime) The Linguistic Metadata (Lime) module (Fiorelli et al., 2013) adds modelling for grouping sets of lexical entries together into a lexicon and providing simple metadata such as the number of entries, senses, etc. Note that it is intended for linked data to be published together on the Web, the necessity to have all words grouped into a lexicon is no longer core, but remains a useful feature.

4. Use cases

4.1 Representing dictionaries with OntoLex

In the past few years, the linguistic linked data community has showed a growing interest in the publication of dictionaries as linked data. The benefits of representing lexicographic content as linked data (LD) are twofold: on the one hand, LD resources are easily reused, gain in visibility and accessibility at a Web scale, their content can be seamlessly aggregated with content from external lexical resources (not necessarily dictionaries), as well as integrated and exploited by LD-aware Natural Language Processing (NLP) tools (Klimek & Brümmer, 2015; Gracia et al., 2016). On the other hand, LD offers several advantages to the modeling of the macro and micro-structure of a dictionary (Bosque-Gil et al., 2016a): moving beyond traditional cross-references, dictionary entries and each of their components are uniquely identified at a Web scale and become internally reusable thanks to URIs; hierarchical ordering of information is replaced by graph structures, where each node becomes a potential entry point to traverse the whole graph, and any relation between two elements is typed and defined in a vocabulary over which previous consensus was reached. The dictionary allows thus for an interpretation as a vast interoperable typed network of lexical elements, as opposed to the more traditional list-inspired view of it.

Initiatives such as the European Network for e-Lexicography (ENeL),[7] Linked data lexicography for high-end language technology application (LD4HELTA)[8] or the Linked Open Dictionaries (LiODi) project[9] foster the conversion of dictionaries to linked data as part of the adoption of the new technological advances in the Semantic Web by digital humanities.

As *lemon* and OntoLex gradually become widespread models for the conversion of lexical resources to linked data, dictionaries represented with them can be easily integrated with other resources previously converted to RDF without any remodelling efforts. This

[7] http://www.elexicography.eu/
[8] http://www.eurekanetwork.org/project/id/9898
[9] http://acoli.cs.uni-frankfurt.de/liodi/home.html

means, in turn, that in many cases dictionaries go from a proprietary data model to one widely accepted by the community. In fact, dictionary conversion to linked data was already receiving much attention prior to OntoLex, and several contributions put forward LD-versions of dictionaries based on *lemon*. Examples of these are the family of bilingual dictionaries Apertium RDF (Gracia et al., 2016), the Germ monolingual dictionary in K Dictionary's Series (Klimek & Brümmer, 2015), sentiment lexica (Vulcu et al., 2014), the Parole-Simple lexica (Villegas & Bel, 2015), the Pattern Dictionary of English Verbs (El Maarouf et al., 2014), the classical Al-Qamus dictionary (Khalfi et al., 2016) and DBpedia lexicalizations such as DBlexipedia (Walter et al., 2015), just to mention a few. Some of these efforts, e.g. Dbnary (Sérasset, 2015), called for the definition of new properties that at that time were not covered by *lemon* (e.g. `dbnary:isTranslationOf`) and that nowadays have a counterpart in OntoLex or by extension vocabularies such as LexInfo (Cimiano et al., 2011). Recently, the interest is being directed towards the transformation of dictionaries which contain a variety of rich annotations and which are developed both for NLP purposes and human users. These include multilingual (Bosque-Gil et al., 2016b), dialectal (Declerck & Mörth, 2016), etymological (Abromeit et al., 2016), and ancient Greek (Khan et al., 2016) dictionaries, among others (Declerck et al., 2015). The work on these resources and the dictionaries mentioned above has lead to the proposal of extensions and modifications to OntoLex to account for specific information ranging from etymological annotations, translations of examples, groups of inflections and temporal information to the sense-subsense hierarchy in a dictionary entry.

4.2 The Collaborative Interlingual Index

Princeton WordNet (PWN, Fellbaum (2010)) is the most widely used lexicographic resource for natural language processing, but yet is only available for English. There have been many versions of wordnets for other languages and these have been collected together in the Open Multilingual WordNet (Bond & Foster, 2013); however they have primarily been created by the *extend approach*, where existing synsets from PWN have been translated and then new synsets are added for words which do not exist in English. Unfortunately, this has led to a degree of fragmentation, where certain concepts may be independently defined by different wordnets. In order to address this issue, it has been proposed that all wordnets contribute to a single index of concepts (Pease et al., 2008). This has recently been realized by the Collaborative Interlingual Index (CILI; Bond et al., 2016), in which all wordnets are converted to a common format and linking is made between the synsets. In order to do this, it is assumed that each concept must have both an English definition and a link to a synset already defined in the CILI.

In order to implement this, it has been necessary to define a common format for the definition of wordnets.[10] This format allows for three forms: XML, JSON and RDF, all of which can be converted without any loss of information. The XML format is based on the existing Lexical Markup Framework (Francopoulo et al., 2006) and in particular on the WordNet-LMF variant (Soria et al., 2009). Both the JSON and RDF formats are based on the OntoLex model described in this paper, and the RDF version of this format is considered a limited *profile* of the OntoLex model, suited particularly for the case of representing wordnets. The JSON version more precisely defines its semantics by means of a JSON-LD context (Sporny et al., 2014). An example of this is given in

[10] https://globalwordnet.github.io/schemas

```
{
  "@context": "http://globalwordnet.github.io/schemas/wn-json-context-1.0.json",
    "@graph": [{
      "@context": { "@language": "en" },
      "@id": "example-en",
      "@type": "ontolex:Lexicon",
      "label": "Example wordnet (English)",
      "language": "en",
      "email": "john@mccr.ae",
      "rights": "https://creativecommons.org/publicdomain/zero/1.0/",
      "version": "1.0",
      "entry": [{
          "@id": "w1",
          "lemma": { "writtenForm": "grandfather" },
          "partOfSpeech": "noun",
          "sense": [{
              "@id": "example-en-10161911-n-1",
              "synset": "example-en-10161911-n"
          }]
      }],
      "synset": [{
          "@id": "example-en-10161911-n",
          "ili": "i90287",
          "partOfSpeech": "noun",
          "definition": [{
              "gloss": "the father of your father or mother"
          }],
          "relations": [{
              "relType": "hypernym",
              "target": "example-en-10162692-n"
          }]
      }]
  }]
}
```


Figure 2: An example WordNet in the Global WordNet JSON-LD format

Figure 2, in which the term "grandfather" is defined. In this example, a number of required standard metadata properties are defined using widely-used vocabularies, namely Dublin Core (Weibel et al., 1998) and Schema.org.[11] Then the file contains two sections `entry` and `synset`, which define the *lexical entries* and *lexical concepts* in this lexicon. They both have a part-of-speech property, with specific values defined in a custom WordNet ontology.[12] The senses of the model correspond to the *lexical senses* of the OntoLex model. For synset and sense relations the variation modules are used that enable relationships between senses to be further described with metadata.

The use of linked data to represent the interlingual index has a number of advantages, most specifically that each ILI identifier is associated with a unique URL, where further information about the term can be found. For example, information about the resource `i1234` can be obtained at http://ili.globalwornet.org/ili/i1234, including the definition of the concept in English as well as links to the PWN and other wordnets which have contributed their links to the ILI. The URL thus allows for a stable identifier that can be referred to unambiguously as opposed to the current method of referring to offsets in release files.

5. Extensions and Future Plans

The OntoLex Community Group released its "final report" on 10th May 2016, however the work of the group has not yet stopped and the group has an ambition to develop more modules in response to critical analysis and novel uses case (such as Chavula & Keet (2014)). In particular, the group has recently aimed to develop four new modules in order to further extend the applicability of the model:

Morphology The first *lemon* proposal contained a module for "inflectional and agglutinative morphology", which primarily defined morphological processes by means of regular expressions. This methodology was very simple to implement in any programming language that support Perl-like regular expressions, however does not very accurately represent the phonological process that occur in word morphology. As such, under this model certain regular cases like the plural of 'leaf' to 'leaves' would be modelled as distinct morphological paradigms even though it is generally considered part of the normal paradigm of pluralization in English. Thus the original model was not included in the OntoLex model and has been made available as a standalone ontology called LIAM (Lemon Inflectional Agglutinative Morphology).[13] There have since been a number of new proposals for morphology and in particular the group is discussing the adoption of the MMoOn Ontology of (Klimek et al., 2016; Klimek, 2017), which enables the documentation of the morphological data of any inflectional language in RDF.

Lexicography Previous experience in the representation of dictionaries using OntoLex-Lemon, as those described in Section 4.1, have led to a number of issues[14] (Bosque-Gil et al., 2017). In particular, these issues include associating senses with forms and syntactic information such as grammatical gender, adding examples, geographic information and ordering senses in terms of importance, along with other aspects of dictionary information that are not always explicitly covered in the core OntoLex-Lemon

[11] https://webmasters.googleblog.com/2011/06/introducing-schemaorg-search-engines.html
[12] http://globalwordnet.github.io/schemas/wn
[13] http://lemon-model.net/liam
[14] For more details see: https://www.w3.org/community/ontolex/wiki/Lexicography

model. As such, the OntoLex community has perceived the necessity of adding extra modelling to cover such issues. To this end, a new OntoLex Lexicography module will be built targeted at the representation of dictionaries and which will address structures and annotations commonly found in lexicography. Such a module is intended to be compatible with other modules in OntoLex (e.g., Etymology and Diachronicity) and should constitute a viable mechanism for lexicographers in the development of dictionaries as linked data.

Etymology and Diachronicity Some authors (Khan et al., 2014; Abromeit et al., 2016; Khan et al., 2017) have proposed using the OntoLex model to represent dictionaries of historical languages, and moreover many dictionaries contain some etymological information. As such, the ability of a dictionary to represent the change of lexical items over time is important. Thus, it has been recognized that the development of a module to capture the meaning of words over time is a key use case of the model.

Lexico-syntactic categories The OntoLex model follows a principle of avoiding prescriptive modelling, for example allowing individual applications to define their own categories. This is helpful as in the example of part-of-speech values in wordnets discussed above, where this approach allows the resource to define categories that may not be accepted by other lexicographers.[15] However, the definition of standard categories greatly helps interoperability between resources and the LexInfo ontology (Cimiano et al., 2011) has been used by a number of authors for this purpose (Buitelaar et al., 2013; Villegas & Bel, 2015; Sérasset, 2015). This resource, originally derived from the ISOcat (Kemps-Snijders et al., 2008) categories, is currently maintained as a single OWL file. As such, the group aims to re-evaluate this model and establish a procedure for adding new categories to a single ontology. This will still only be a suggestion for data categories and we expect particular communities to define their own ontologies.

6. Conclusion

The OntoLex model has been developed under an open process and as such represents one of the most significant open models for the representation of electronic lexicographic resources. While the model as proposed retains aspects of the proposal of (McCrae et al., 2012), it has also been significantly innovated in order to allow new use cases. In particular, the application of the model beyond the Semantic Web community has required new modelling, in particular the introduction of *lexical concepts* and dereferenceable relations. These developments have seen the model adapted to a wider community and as such have consequently lead to requests for new features. The group remains committed to developing the model and new use cases in morphology and diachronic lexicography will further show the flexibility of this linked data based model.

7. Acknowledgements

We would like to thank all members of the OntoLex group for their contributions and discussion of the model. This work was supported in part by the Science Foundation Ireland under Grant Number SFI/12/RC/2289 (Insight), by the Spanish Ministry of Economy and Competitiveness through the ReTeLe Excellence Network (TIN2015-68955-REDT) and the Juan de la Cierva program, and by the Spanish Ministry of Education, Culture

[15] In particular, PWN defines 'adjective satellite' as a distinct category to 'adjective'

and Sports through the FPU program (UPM), and the CITEC excellence initiative funded by the DFG (Deutsche Forschungsgemeinschaft).

8. References

Abromeit, F., Chiarcos, C., Fath, C. & Ionov, M. (2016). Linking the Tower of Babel: Modelling a Massive Set of Etymological Dictionaries as RDF. In *5th Workshop on Linked Data in Linguistics: Managing, Building and Using Linked Language Resources*. p. 11.

Bond, F. & Foster, R. (2013). Linking and Extending an Open Multilingual Wordnet. In *Proceedings of the 2013 Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 1. pp. 1352–1362.

Bond, F., Vossen, P., McCrae, J.P. & Fellbaum, C. (2016). CILI: the Collaborative Interlingual Index. In *Proceedings of the Global WordNet Conference 2016*.

Bosque-Gil, J., Gracia, J. & Gómez-Pérez, A. (2016a). Linked data in lexicography. *Kernerman Dictionary News*, pp. 19–24.

Bosque-Gil, J., Gracia, J. & Montiel-Ponsoda, E. (2017). Towards a Module for Lexicography in OntoLex. In *Proceedings of the LDK workshops: OntoLex, TIAD and Challenges for Wordnets*. pp. 1–11.

Bosque-Gil, J., Gracia, J., Montiel-Ponsoda, E. & Aguado-de Cea, G. (2016b). Modelling multilingual lexicographic resources for the Web of Data: The K Dictionaries case. In *GLOBALEX 2016 Lexicographic Resources for Human Language Technology Workshop Programme*. pp. 65–72.

Buitelaar, P., Arcan, M., Iglesias Fernandez, C.A., Sánchez Rada, J.F. & Strapparava, C. (2013). Linguistic linked data for sentiment analysis. In *Proceedings of the 2nd Workshop on Linked Data in Linguistics*. pp. 1–8.

Chavula, C. & Keet, C.M. (2014). Is lemon Sufficient for Building Multilingual Ontologies for Bantu Languages? In *OWLED*. Citeseer, pp. 61–72.

Cimiano, P., Buitelaar, P., McCrae, J.P. & Sintek., M. (2011). LexInfo: A declarative model for the lexicon-ontology interface. *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(1), pp. 29–51.

Cimiano, P., McCrae, J.P. & Buitelaar, P. (2016). Lexicon Model for Ontologies: Community Report. W3C Community Group Final Report.

Declerck, T. & Mörth, K. (2016). Towards a Sense-based Access to Related Online Lexical Resources. In G.M. Tinatin Margalitadze (ed.) *Proceedings of the 17th EURALEX International Congress*. Tbilisi, Georgia: Ivane Javakhishvili Tbilisi University Press, pp. 660–667.

Declerck, T., Wand-Vogt, E. & Mörth, K. (2015). Towards a Pan European Lexicography by Means of Linked (Open) Data. In *Proceedings of eLex 2015. Biennial Conference on Electronic Lexicography (eLex-2015), electronic lexicography in the 21st century: Linking lexical data in the digital age, August 11-13, Herstmonceux, United Kingdom*. Ljubljana/Brighton: Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd., pp. 342–355.

Eckle-Kohler, J., McCrae, J.P. & Chiarcos, C. (2015). lemonUby - a large, interlinked, syntactically-rich lexical resource for ontologies. *Semantic Web*, 6(4), pp. 371–378.

Ehrmann, M., Cecconi, F., Vannella, D., McCrae, J.P., Cimiano, P. & Navigli, R. (2014). Representing Multilingual Data as Linked Data: the Case of BabelNet 2.0. In N.C. Chair, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno,

J. Odijk & S. Piperidis (eds.) *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), pp. 401–408.

El Maarouf, I., Bradbury, J. & Hanks, P. (2014). PDEV-lemon: a Linked Data implementation of the Pattern Dictionary of English Verbs based on the Lemon model. In *Proceedings of the 3rd Workshop on Linked Data in Linguistics (LDL): Multilingual Knowledge Resources and Natural Language Processing at the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland*. pp. 88–93.

Fellbaum, C. (2010). WordNet. In *Theory and applications of ontology: computer applications*. Springer, pp. 231–243.

Fiorelli, M., Pazienza, M.T. & Stellato, A. (2013). LIME: Towards a Metadata Module for Ontolex. In *2nd Workshop on Linked Data in Linguistics: Representing and Linking lexicons, terminologies and other language data*. Pisa, Italy, pp. 18–27.

Francopoulo, G., George, M., Calzolari, N., Monachini, M., Bel, N., Pet, M., Soria, C. et al. (2006). Lexical Markup Framework (LMF). In *Proceedings of 5th International Conference on Language Resources and Evaluation (LREC)*. pp. 233–236.

Gracia, J., Montiel-Ponsoda, E., Vila-Suero, D. & Aguado-de Cea, G. (2014). Enabling Language Resources to Expose Translations as Linked Data on the Web. In *Proc. of 9th Language Resources and Evaluation Conference (LREC'14), Reykjavik (Iceland)*. European Language Resources Association (ELRA), pp. 409–413.

Gracia, J., Villegas, M., Gómez-Pérez, A. & Bel, N. (2016). The Apertium Bilingual Dictionaries on the Web of Data. *Semantic Web Journal*.

Kemps-Snijders, M., Windhouwer, M., Wittenburg, P. & Wright, S.E. (2008). ISOcat: Corralling Data Categories in the Wild. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*. pp. 887–891.

Khalfi, M., Nahli, O. & Zarghili, A. (2016). Classical dictionary Al-Qamus in lemon. In M.E. Mohajir, M. Chahhou, M.A. Achhab & B.E.E. Mohajir (eds.) *2016 4th IEEE International Colloquium on Information Science and Technology (CiSt)*. IEEE, pp. 325–330.

Khan, F., Bellandi, A., Boschetti, F. & Monachini, M. (2017). The Challenges of Converting Legacy Lexical Resources to Linked Open Data using OntoLex-Lemon:The Case of the Intermediate Liddell-Scott Lexicon. In *Proceedings of the LDK workshops: OntoLex, TIAD and Challenges for Wordnets*. pp. 1–8.

Khan, F., Boschetti, F. & Frontini, F. (2014). Using lemon to Model Lexical Semantic Shift in Diachronic Lexical Resources. In *3rd Workshop on Linked Data in Linguistics*. pp. 50–54.

Khan, F., Díaz-Vera, J.E. & Monachini, M. (2016). Representing polysemy and diachronic lexicosemantic data on the semantic web. In *Proceedings of the Second International Workshop on Semantic Web for Scientific Heritage co-located with 13th Extended Semantic Web Conference (ESWC 2016), Heraklion, Greece*, volume 1595. pp. 37–46.

Klimek, B. (2017). Proposing an OntoLex - MMoOn Alignment: Towards an Interconnection of two Linguistic Domain Models. In *Proceedings of the LDK workshops: OntoLex, TIAD and Challenges for Wordnets*. pp. 1–16.

Klimek, B., Arndt, N., Krause, S. & Arndt, T. (2016). Creating Linked Data Morphological Language Resources with MMoOn - The Hebrew Morpheme Inventory. In *Proceedings of the 10th edition of the Language Resources and Evaluation Conference (LREC)*. pp. 892–899.

Klimek, B. & Brümmer, M. (2015). Enhancing lexicography with semantic language databases. *Kernerman Dictionary News*, 23, pp. 5–10.

McCrae, J.P., de Cea, G.A., Buitelaar, P., Cimiano, P., Declerck, T., Gómez-Pérez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D. & Wunner, T. (2012). Interchanging lexical resources on the Semantic Web. *Language Resources and Evaluation*, 46(6), pp. 701–709.

McGuinness, D.L. & van Harmelen, F. (2004). OWL Web Ontology Language Overview. W3C Recommendation.

Navigli, R. & Ponzetto, S.P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193, pp. 217–250.

Pease, A., Fellbaum, C. & Vossen, P. (2008). Building the global WordNet grid. In *Proceedings of the 18th International Congress of Linguists (CIL18)*. pp. 1–4.

Sérasset, G. (2015). DBnary: Wiktionary as a lemon-based multilingual lexical resource in RDF. *Semantic Web*, 6(4), pp. 355–361.

Soria, C., Monachini, M. & Vossen, P. (2009). Wordnet-LMF: fleshing out a standardized format for wordnet interoperability. In *Proceedings of the 2009 International Workshop on Intercultural Collaboration*. ACM, pp. 139–146.

Sporny, M., Longley, D., Kellogg, G., Lanthaler, M. & Lindström, N. (2014). JSON-LD 1.0. W3C Recommendation.

Villegas, M. & Bel, N. (2015). PAROLE/SIMPLE 'lemon' ontology and lexicons. *Semantic Web*, 6(4), pp. 363–369.

Vossen, P., Bond, F. & McCrae, J.P. (2016). Toward a truly multilingual Global Wordnet Grid. In *Proceedings of the Global WordNet Conference 2016*. pp. 419–426.

Vulcu, G., Buitelaar, P., Negi, S., Pereira, B., Arcan, M., Coughlan, B., Sánchez, Fernando, J. & Iglesias, C.A. (2014). Generating Linked-Data based Domain-Specific Sentiment Lexicons from Legacy Language and Semantic Resources. In *5th International Workshop on Emotion, Social Signals, Sentiment and Linked Open Data, co-located with LREC 2014*. pp. 1–4.

Walter, S., Unger, C. & Cimiano, P. (2015). DBlexipedia: A Nucleus for a Multilingual Lexical Semantic Web. In H. Paulheim, M. van Erp, A. Filipowska, P.N. Mendes & M. Brümmer (eds.) *NLP-DBPEDIA@ISWC*, volume 1581 of *CEUR Workshop Proceedings*. CEUR-WS.org, pp. 87–92.

Weibel, S., Kunze, J., Lagoze, C. & Wolf, M. (1998). Dublin Core Metadata for Resource Discovery. RFC 2413.

Automatic Extraction of TEI Structures in Digitized Lexical Resources using Conditional Random Fields

Mohamed Khemakhem[1,2,4], **Luca Foppiano**[1], **Laurent Romary**[1,2,3]

[1]Inria – ALMAnaCH, 2 Rue Simone IFF 75012, Paris
[2] Centre Marc Bloch, Friedrichstrasse 191 10117, Berlin
[3] Berlin-Brandenburgische Akademie der Wissenschaften, Jaegerstrasse 22-23 10117, Berlin
[4]University Paris Diderot, 5 Rue Thomas Mann 75013, Paris
E-mail: mohamed.khemakhem@inria.fr, luca.foppiano@inria.fr, laurent.romary@inria.fr

Abstract

This paper presents an open source machine learning system for structuring dictionaries in digital format into TEI (Text Encoding Initiative) encoded resources. The approach is based on the extraction of overgeneralised TEI structures in a cascading fashion, by means of CRF (Conditional Random Fields) sequence labelling models. Through the experiments carried out on two different dictionary samples, we aim to highlight the strengths as well as the limitations of our approach.

Keywords: automatic structuring; digitized dictionaries; TEI; machine learning; CRF

1. Introduction

An important number of digitized lexical resources remain unexploited due to their unstructured content. Manually structuring such resources is a costly task given their multifold complexity. Our goal is to find an approach to automatically structure digitized dictionaries, independently of the language or the lexicographic school or style. In this paper we present a first version of GROBID-Dictionaries,[1] an open source machine learning system for lexical information extraction.

2. Approach

By observing how the lexical information is organised in different paper dictionaries, it is clear that the majority of these lexical resources share the same visual layout to represent the same categories of text information. That served as our starting point to develop our approach for dismantling the content of digitized dictionaries. We tried to build cascading models for automatically extracting TEI (Text Encoding Initiative; Budin et al., 2012) constructs and make sure that the final output is aligned with current efforts to unify the TEI representations of lexical resources. To be easily adaptable to new dictionary samples, we chose machine learning over rule-based techniques.

2.1 Cascading extraction models

We followed a divide-and-conquer strategy to dismantle text constructs in a digitized dictionary, based initially on observations of their layout. Main pages (see Figure1) in almost any dictionary share three blocks: a header (green), a footer (blue) and a body (orange). The body is, in its turn, made of several entries (red). Each lexical entry can be further broken down (see Figure 2) into: form (green), etymology (blue), sense (red) or/and related entry.

Layout features become less relevant when the segmentation process reaches a deeper information level and we consequently give them up for the corresponding models. The

[1] https://github.com/MedKhem/grobid-dictionaries

Figure 1: First and second segmentation levels of a dictionary page

same logic could be applied further for each extracted block, as long as the finest TEI elements are not yet reached. But in the scope of this paper, we focus just on the first six models, details which are given below.

Such a cascading approach ensures a better understanding of the learning process' output and consequently simplifies the feature selection process. Limited exclusive text blocks per level help significantly to diagnose the cause of prediction errors. Moreover, it would be possible to detect and replace early on any irrelevant selected features that can bias a trained model. In such a segmentation, it becomes more straightforward to notice that, for instance, the token position in the page is very relevant to detect headers and footers but has almost no relevance for capturing a sense in a lexical entry, which is very often split over two pages.

2.2 Towards a more unified TEI modelling

Our choice for TEI, as the encoding format for the detected structures, is based on its widespread use in lexicographic projects, as well as on some technical factors which will be detailed in the following section. The domination of the lexicographic landscape by TEI is

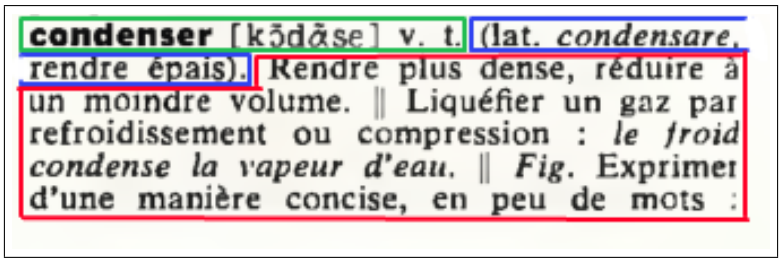


Figure 2: Example of the segmentation performed by the Lexical Entry model

due to the fact that this initiative has provided the lexicographic community with diverse alternatives for encoding different kinds of lexical resources, as well as for modelling the same lexical information. However, the flexibility that this standard ensures has led to an explosion of TEI schemes and, consequently, limited the possibilities for exchange and exploitation.

Our cascading models are conceived in a way to support the encoding of the detected structures in multiple TEI schemes. But to avoid falling into the diversity trap, we are adopting a format that generalises over existing encoding practices. The final scheme has not yet been finalised, but we are continuously refining our guidelines as we move deeper with our models and apply them to new dictionary samples. We are aiming to ensure a maximal synchronisation with existing research efforts in this direction, by collaborating with COST ENeL and ISO committee TC 37/SC 4.

Presenting the details of our encoding choices is beyond the scope of this paper, since we are still shaping them, especially for fine grained information. But we aim to highlight some constraining decisions we made for the upper levels, to give an idea about our modelling direction. A lexical entry, for instance, is always encoded using <entry> exclusively, which means we do not make use of any possible alternatives, such as <superEntry> and <entryFree>. The semantic loss is not important in this case, since the nature of the entry could be inferred from the elements it contains. As for lexical entries, they can be completely encoded using five main elements: <form> for morphological and grammatical information of the whole entry, <etym> for etymological information, <sense> for semantic and syntactic information, <re> for related entries and <dictScrap> for any text that does not belong to the previous elements. Note here that we are trying to use the more generic elements to encode the lexical information in each level, which will be more refined in the following levels.

3. GROBID-Dictionaries

To implement our approach, we took up the available infrastructure from GROBID (Lopez & Romary, 2015) and we adapted it to the specificity of the use case of digitized dictionaries.

3.1 GROBID

GROBID (GeneRation Of Bibliographic Data) is a machine learning system for parsing and extracting bibliographic metadata from scholar articles, mainly text documents in PDF format. It relies on CRF (Lavergne et al., 2010) to perform a multi-level sequence

labelling of text blocks in a cascade fashion which are then extracted and encoded in TEI elements. Such an approach has been very accurate for that use case and the system's Java API has been one of the most used by bibliography research platforms and research bodies worldwide.

We have been struck by the analogy between the structures that can be extracted by GROBID, in the case of full scientific articles, and the actual constructs we wanted to extract from a digitized dictionary. At its first extraction level, GROBID detects the main blocks of a paper such as the header, the body, the references, annexes, etc. These main parts will be further structured at the following level, like the references which will be extracted in separate items and then parsed one by one to detect the titles, the authors and the other publication details. By recalling the segmentation steps presented in the previous section, there is a clear analogy between the case of a reference in a scientific document and a lexical entry in a dictionary.

This correspondence is reinforced by the fact that GROBID relies on layout, as well as text features, to perform the supervised classification of the parsed text and generates a TEI compliant encoding where the various segmentation levels are associated with an appropriate XML tessellation.

3.2 GROBID-Dictionaries

Due to the above-mentioned similarities, we undertook the adaptation of GROBID for the case of digitized dictionaries in order to build a system, which uses the core utilities of GROBID and applies them for lexical information processing. In building GROBID-Dictionaries, we faced several challenges, the three major ones being detailed in the following.

3.2.1 TEI cascade modelling

After having fully encoded a lexical entry, the task became more specific and more challenging when it comes to defining the TEI structures to be extracted by each model. It is a question of finding the appropriate mapping between the TEI elements and the labels to be set for the models that share the task of structuring the text in cascade. In addition, the process is at the same time constrained by the need to avoid having structures from different hierarchy levels being extracted at once. In fact, the CRF models, as they could be used from GROBID core, do not allow the labelling of nested text sequences. We clarify this technical point by explaining how the sequence labelling process works in the case of segmenting a lexical entry.

The following matrix represents the set of feature vectors corresponding to the lexical entry *condenser*, which will be labelled by a first version of the "Lexical Entry" model. The latter has the task of detecting the five main blocks in a lexical entry, if they exist. For the sense information, the model has been trained to extract each parsed text sequence representing a sense.

Each vertical column is a specific feature for all the tokens of the lexical entry and each horizontal line corresponds to all the features of each token. For this model, a set of features is going to be assigned to each token based on criteria we chose in the feature

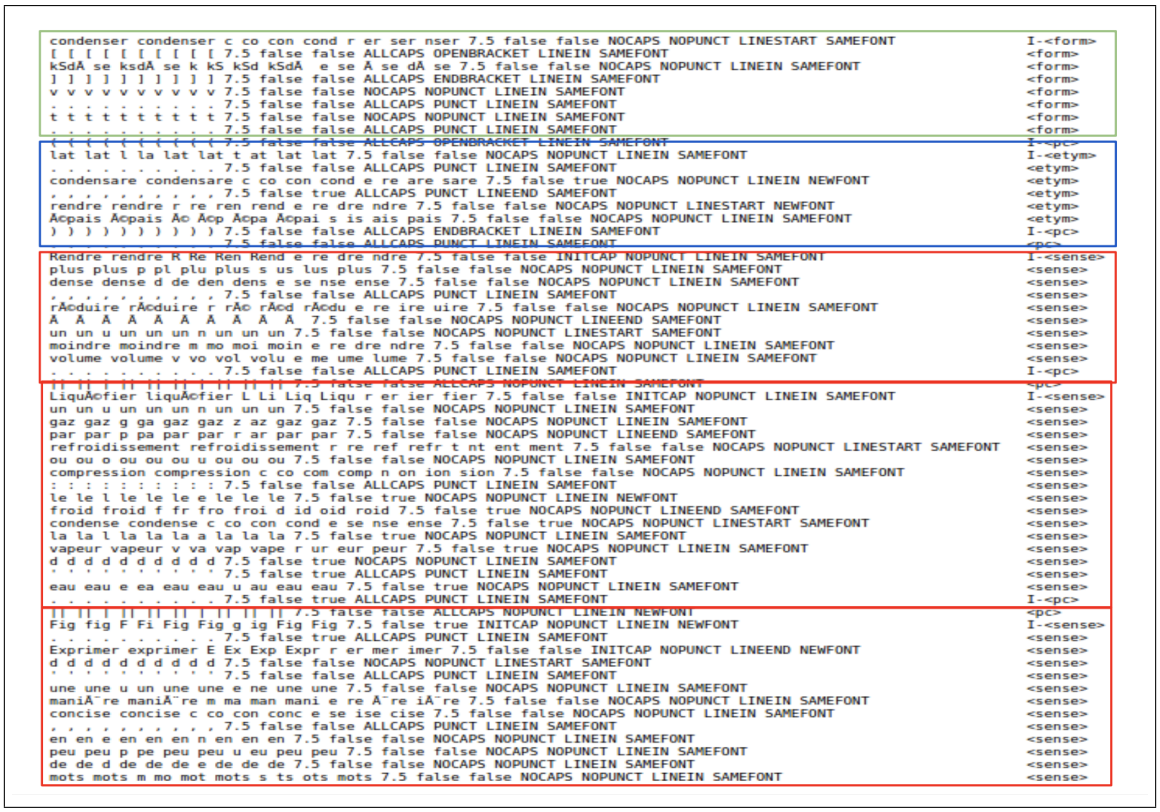


Figure 3: Sequence labelling using a first version of the "Lexical Entry" segmentation model

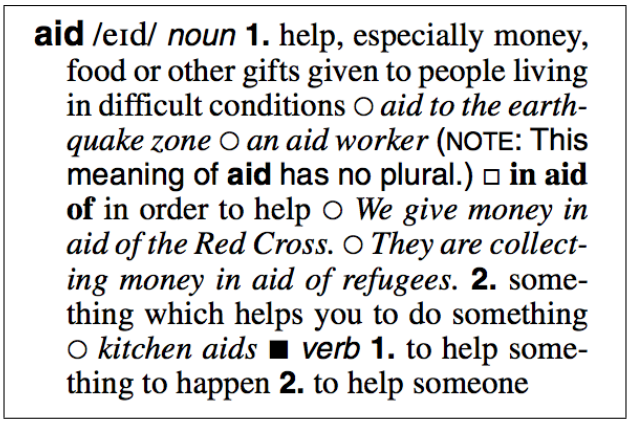
selection process. In the second phase, comes the role of the trained model to give a prediction of a suitable label for each token, based on all its feature values. A structure corresponds then to the sequence of tokens having the same label, where the *I-Label* marks the beginning of a new sequence. Following this technique, it is obviously not possible in this model to structure the example *"le froid condense la vapeur d'eau"*(see Figures 2 and 3) in the sense, since just one label is allowed per token. Therefore, the segmentation of the examples should be delegated to another model that follows the current one.

3.2.2 Sample annotation

This is the phase where the previous rules will be applied on different instances, to annotate data for training the models. An adjustment of the directives is necessary to make the models more general, as soon as new instances appear to show the modelling limits of our current guidelines. To illustrate such a case, we could take the example of the previously defined "Lexical Entry" model and apply it to the lexical entry *aid*.

The TEI encoding for this entry with the "Lexical Entry" model is the following (see Figure 5):

We could notice that the model presented in Figure 3 is no longer valid to perform the segmentation of senses aggregated by part of speech (POS), with respect to avoiding nested constructs. This issue could be fixed by having a first model that does not find the boundaries of the senses of a part of speech in this level.



aid /eɪd/ *noun* **1.** help, especially money, food or other gifts given to people living in difficult conditions ○ *aid to the earthquake zone* ○ *an aid worker* (NOTE: This meaning of **aid** has no plural.) □ **in aid of** in order to help ○ *We give money in aid of the Red Cross.* ○ *They are collecting money in aid of refugees.* **2.** something which helps you to do something ○ *kitchen aids* ■ *verb* **1.** to help something to happen **2.** to help someone

Figure 4: Lexical entry having more than one POS

```
<entry>
    <form>aid /e□d/ noun</form>
    <sense>1. help, especially money, <lb/>food or other gifts given to people living <lb/>
        in difficult conditions aid to the earth-<lb/>quake zone an aid worker (NOTE: This <lb/>
        meaning of aid has no plural.) in aid <lb/>of in order to help We give money in <lb/>
        aid of the Red Cross. They are collect-<lb/>ing money in aid of refugees</sense><pc>. </pc>
    <sense>2. some-<lb/>thing which helps you to do something <lb/>kitchen aids</sense> i
    <sense>verb 1. to help some-<lb/>thing to happen</sense>
    <sense>2. to help someone</sense>
</entry>
```

Figure 5: Structured output of the "Lexical Entry" model's primary version

This segmentation of main POS-aggregated senses should be performed by a second model, called "Sense" for example, to find the limits of each sense as well as any grammatical information, if any exists.

The labelling and extraction of the TEI structures should be performed further for the other blocks, by following the same approach. For the case of the *aid* entry, a dedicated model should be used to segment the <form> block by extracting the morphological and grammatical information and decide about of the parent of the latter. In the current case, the <gramGrp> will be the direct child node of the entry, since it carries information about the sense of the entry given a POS, and not about the lemma. The <gramGrp> block will, in its turn, have another specific model to structure its content. Figure 8 shows the final output generated by our cascading model tree.

Annotation guidelines seem to be mandatory here to guide the process since an annotator, especially with a linguistic or lexicographic background, could be easily biased by the TEI practices and tags which are used differently in our cascading approach but will converge in the final output. We noticed this issue after having lexicographers annotate a few samples and we therefore, defined a first version of the guidelines,[2] which we are actively maintaining.

3.2.3 Feature selection

In this phase, the cumulated data will be used for generating features that will be used by the models to discriminate between their labels. For the first model, we kept the

[2] https://github.com/MedKhem/grobid-dictionaries/wiki/How-to-Annotate%3F

```
<entry>
    <form>aid /e□d/ noun</form>
    <sense>1. help, especially money, <lb/>food or other gifts given to people living <lb/>
in difficult conditions aid to the earth-<lb/>quake zone an aid worker (NOTE: This <lb/>
meaning of aid has no plural.) in aid <lb/>of in order to help We give money in <lb/>
aid of the Red Cross. They are collect-<lb/>ing money in aid of refugees.2. some-<lb/>
thing which helps you to do something <lb/>kitchen aids</sense> í
    <sense>verb 1. to help some-<lb/>thing to happen 2. to help someone</sense>
</entry>
```

Figure 6: Structured output of the "Lexical Entry" model's adjusted version

```
<entry>
    <form>aid /e□d/ noun</form>
    <sense>
        <sense>1. help, especially money, <lb/>food or other gifts given to people living <lb/>
            in difficult conditions aid to the earth-<lb/>quake zone an aid worker (NOTE: This
            <lb/>meaning of aid has no plural.) in aid <lb/>of in order to help We give money in <lb/>
            aid of the Red Cross. They are collect-<lb/>ing money in aid of refugees</sense><pc>. </pc>
        <sense>2. some-<lb/>thing which helps you to do something <lb/>kitchen aids</sense> í
    </sense>
    <sense>
        <gramGrp>verb</gramGrp>
        <sense>1. to help some-<lb/>thing to happen</sense>
        <sense>2. to help someone</sense>
    </sense>
</entry>
```

Figure 7: Structured output of the "Sense" model

line based features used in GROBID's first model.[3] Our choice was based simply on the assumption of the general nature of such features. Moreover, the experiments on several samples showed a high and fast performance.

As explained in our approach, we tried to rely on a restricted list of features for the rest of the models, where we drop the ones that are most likely to produce bias. We chose to use features on the token level to structure the lexical information. For the first version of our system, we are experimenting the use of one list with 16 features:[4] 8 based on the text and the rest carrying the layout aspects of each token, such as the change of font or line breaks.

4. Experiments

4.1 Models

The resulting models and their corresponding labels are the following:

- **Dictionary Segmentation:** This is the first model and has as its goal the segmentation of each dictionary page into three main blocks, where each block corresponds to a TEI label: <fw type="header"> for information in the header, <ab type="page"> for all the text in the body of a page and <fw type="footer"> for footer information.

[3] https://github.com/kermitt2/grobid/blob/master/grobid-core/src/main/java/org/grobid/core/features/FeaturesVectorSegmentation.java

[4] https://github.com/MedKhem/grobid-dictionaries/blob/master/src/main/java/org/grobid/core/features/FeatureVectorLexicalEntry.java

```
<entry>
   <form type="lemma">
      <orth>aid</orth>
      <pron>/eɪd/</pron>
   </form>
   <gramGrp>
      <pos>noun</pos>
   </gramGrp>
   <sense>
      <sense>1. help, especially money, <lb/>food or other gifts given to people living <lb/>
         in difficult conditions aid to the earth-<lb/>quake zone an aid worker (NOTE: This
         <lb/>meaning of aid has no plural.) in aid <lb/>of in order to help We give money in <lb/>
         aid of the Red Cross. They are collect-<lb/>ing money in aid of refugees</sense><pc>. </pc>
      <sense>2. some-<lb/>thing which helps you to do something <lb/>kitchen aids</sense> í
   </sense>
   <sense>
      <gramGrp>
         <pos>verb</pos>
      </gramGrp>
      <sense>1. to help some-<lb/>thing to happen</sense>
      <sense>2. to help someone</sense>
   </sense>
</entry>
```

Figure 8: Final output of all the models

For the sake of simplicity, for training the models (see Section 4.3) we use: <headnote> to refer to <fw type="header">, <body> referring to <ab type="page"> and <footer> to refer to <fw type="footer">. But we respect the original labels for the final TEI output.

- **Dictionary Body Segmentation:** The second model gets the page body, recognized by the first model, and processes it to recognize the boundaries of each lexical entry by labelling each sequence with <entry> label.

- **Lexical Entry:** The third model parses each lexical entry, recognized by the second model, to segment it into four main blocks: <form> for morphological and grammatical information, <etym> for etymology, <sense> for all sense information, <re> for related entries.

- **Form:** This model analyses the <form> block, generated by the Lexical Entry model, and segments its contained information. We have for the moment three labels for this model: <orth> for the lemma, <pron> for pronounciation and <gramGrp> for grammatical information, such as part of speech, gender, number, etc.

- **Sense:** The Sense model has two goals. First, to extract the grammatical information <gramGrp>, that could exist. Second, to segment the first level senses, by structuring them in <sense> sequences.

- **Grammatical group:** The last model in our temporary hierarchy has the of segmenting the grammatical information <gramGrp>, extracted by previous models

For each model, we reserved two extra labels: <pc> for punctuation such as separators between text information or any markup text. A second label, <dictScrap>, is used to contain any information that couldn't be classified in one of the main labels of the model.

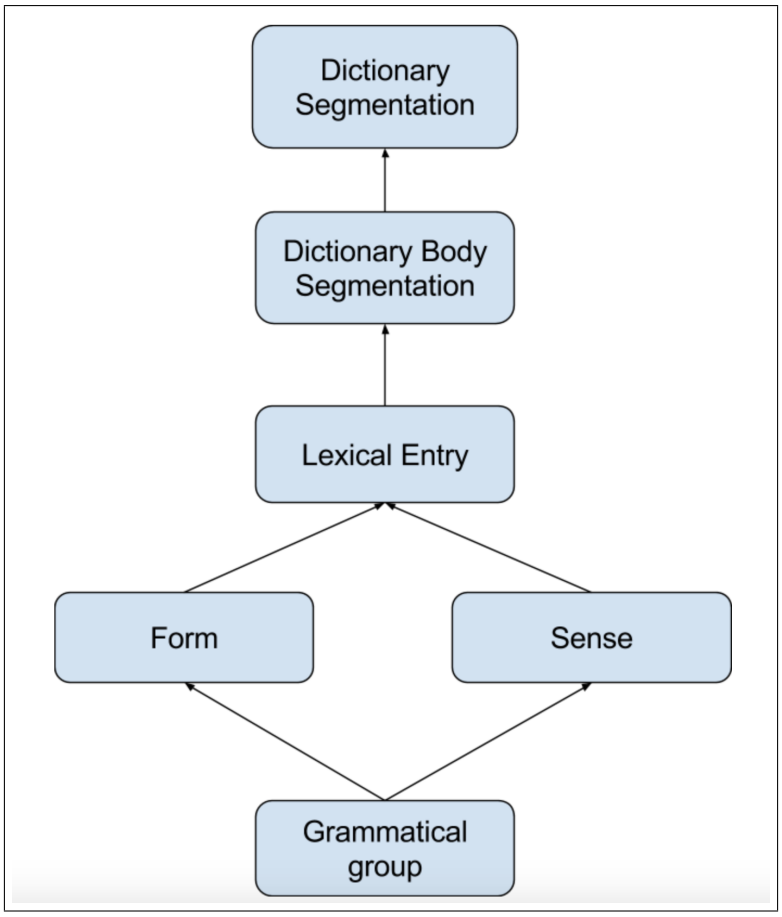


Figure 9: Selected models

4.2 Lexical Samples

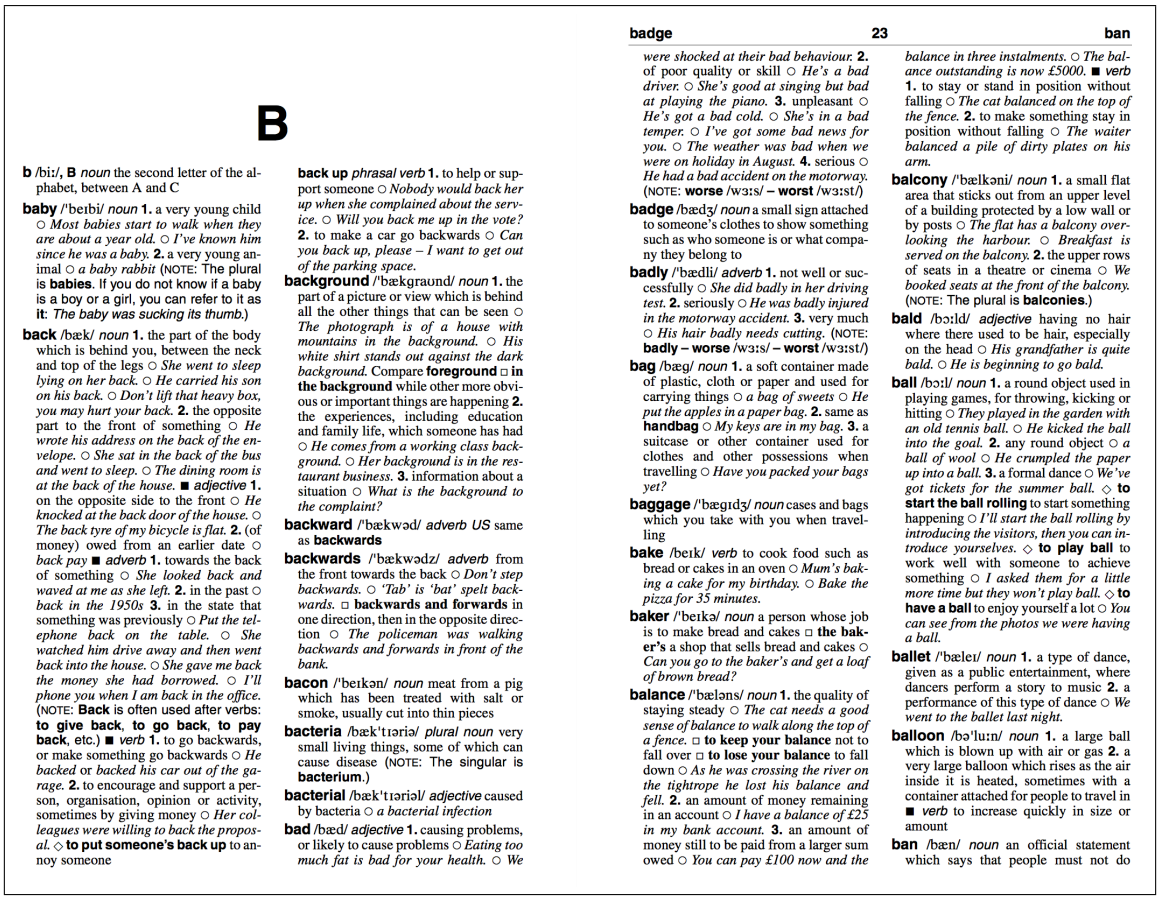
We carried out our experiments by applying our models to several dictionaries and given the inconsistency that some presented, mainly due to digitization issues, we selected two resources that represent several differences on many levels.

4.2.1 Digital dictionary

"Easier English Basic Dictionary" (EEBD, 2009) is a monolingual dictionary for English which contains over 5,000 entries, published in 2009. For our experiments, we used the 370 pages containing the body of the dictionary. The version which we used, is a digitally born one. In other words, no OCR processing has been performed to generate the resource in its electronic format. As Figure 10 illustrates, the dictionary has a very modern and basic layout and its markup system is spread over the entries to mark the transition of the lexical information presented. We chose this digital sample to be our baseline, since it contains very clean text and clear lexical information modelling.

4.2.2 Digitized dictionary

To take the experiments to the next level, we chose a dictionary that has been OCRized and that encloses totally different lexical information. The dictionary was published in



B

b /biː/, **B** *noun* the second letter of the alphabet, between A and C

baby /ˈbeɪbi/ *noun* **1.** a very young child ○ *Most babies start to walk when they are about a year old.* ○ *I've known him since he was a baby.* **2.** a very young animal ○ *a baby rabbit* (NOTE: The plural is **babies**. If you do not know if a baby is a boy or a girl, you can refer to it as **it**: *The baby was sucking its thumb.*)

back /bæk/ *noun* **1.** the part of the body which is behind you, between the neck and top of the legs ○ *She went to sleep lying on her back.* ○ *He carried his son on his back.* ○ *Don't lift that heavy box, you may hurt your back.* **2.** the opposite part to the front of something ○ *He wrote his address on the back of the envelope.* ○ *She sat in the back of the bus and went to sleep.* ○ *The dining room is at the back of the house.* ■ *adjective* **1.** on the opposite side to the front ○ *He knocked at the back door of the house.* ○ *The back tyre of my bicycle is flat.* **2.** (of money) owed from an earlier date ○ *back pay* ■ *adverb* **1.** towards the back of something ○ *She looked back and waved at me as she left.* **2.** in the past ○ *back in the 1950s* **3.** in the state that something was previously ○ *Put the telephone back on the table.* ○ *She watched him drive away and then went back into the house.* ○ *She gave me back the money she had borrowed.* ○ *I'll phone you when I am back in the office.* (NOTE: **Back** is often used after verbs: **to give back**, **to go back**, **to pay back**, etc.) ■ *verb* **1.** to go backwards, or make something go backwards ○ *He backed* or *backed his car out of the garage.* **2.** to encourage and support a person, organisation, opinion or activity, sometimes by giving money ○ *Her colleagues were willing to back the proposal.* ◇ **to put someone's back up** to annoy someone

back up *phrasal verb* **1.** to help or support someone ○ *Nobody would back her up when she complained about the service.* ○ *Will you back me up in the vote?* **2.** to make a car go backwards ○ *Can you back up, please – I want to get out of the parking space.*

background /ˈbækgraʊnd/ *noun* **1.** the part of a picture or view which is behind all the other things that can be seen ○ *The photograph is of a house with mountains in the background.* ○ *His white shirt stands out against the dark background.* Compare **foreground** □ **in the background** while other more obvious or important things are happening **2.** the experiences, including education and family life, which someone has had ○ *He comes from a working class background.* ○ *Her background is in the restaurant business.* **3.** information about a situation ○ *What is the background to the complaint?*

backward /ˈbækwəd/ *adverb US* same as **backwards**

backwards /ˈbækwədz/ *adverb* from the front towards the back ○ *Don't step backwards.* ○ *'Tab' is 'bat' spelt backwards.* □ **backwards and forwards** in one direction, then in the opposite direction ○ *The policeman was walking backwards and forwards in front of the bank.*

bacon /ˈbeɪkən/ *noun* meat from a pig which has been treated with salt or smoke, usually cut into thin pieces

bacteria /bækˈtɪərɪə/ *plural noun* very small living things, some of which can cause disease (NOTE: The singular is **bacterium**.)

bacterial /bækˈtɪərɪəl/ *adjective* caused by bacteria ○ *a bacterial infection*

bad /bæd/ *adjective* **1.** causing problems, or likely to cause problems ○ *Eating too much fat is bad for your health.* ○ *We were shocked at their bad behaviour.* **2.** of poor quality or skill ○ *He's a bad driver.* ○ *She's good at singing but bad at playing the piano.* **3.** unpleasant ○ *He's got a bad cold.* ○ *She's in a bad temper.* ○ *I've got some bad news for you.* ○ *The weather was bad when we were on holiday in August.* **4.** serious ○ *He had a bad accident on the motorway.* (NOTE: **worse** /wɜːs/ **– worst** /wɜːst/)

badge /bædʒ/ *noun* a small sign attached to someone's clothes to show something such as who someone is or what company they belong to

badly /ˈbædli/ *adverb* **1.** not well or successfully ○ *She did badly in her driving test.* **2.** seriously ○ *He was badly injured in the motorway accident.* **3.** very much ○ *His hair badly needs cutting.* (NOTE: **badly – worse** /wɜːs/ **– worst** /wɜːst/)

bag /bæg/ *noun* **1.** a soft container made of plastic, cloth or paper and used for carrying things ○ *a bag of sweets* ○ *He put the apples in a paper bag.* **2.** same as **handbag** ○ *My keys are in my bag.* **3.** a suitcase or other container used for clothes and other possessions when travelling ○ *Have you packed your bags yet?*

baggage /ˈbægɪdʒ/ *noun* cases and bags which you take with you when travelling

bake /beɪk/ *verb* to cook food such as bread or cakes in an oven ○ *Mum's baking a cake for my birthday.* ○ *Bake the pizza for 35 minutes.*

baker /ˈbeɪkə/ *noun* a person whose job is to make bread and cakes □ **the baker's** a shop that sells bread and cakes ○ *Can you go to the baker's and get a loaf of brown bread?*

balance /ˈbæləns/ *noun* **1.** the quality of staying steady ○ *The cat needs a good sense of balance to walk along the top of a fence.* □ **to keep your balance** not to fall over □ **to lose your balance** to fall down ○ *As he was crossing the river on the tightrope he lost his balance and fell.* **2.** an amount of money remaining in an account ○ *I have a balance of £25 in my bank account.* **3.** an amount of money still to be paid from a larger sum owed ○ *You can pay £100 now and the balance in three instalments.* ○ *The balance outstanding is now £5000.* ■ *verb* **1.** to stay or stand in position without falling ○ *The cat balanced on the top of the fence.* **2.** to make something stay in position without falling ○ *The waiter balanced a pile of dirty plates on his arm.*

balcony /ˈbælkəni/ *noun* **1.** a small flat area that sticks out from an upper level of a building protected by a low wall or by posts ○ *The flat has a balcony overlooking the harbour.* ○ *Breakfast is served on the balcony.* **2.** the upper rows of seats in a theatre or cinema ○ *We booked seats at the front of the balcony.* (NOTE: The plural is **balconies**.)

bald /bɔːld/ *adjective* having no hair where there used to be hair, especially on the head ○ *His grandfather is quite bald.* ○ *He is beginning to go bald.*

ball /bɔːl/ *noun* **1.** a round object used in playing games, for throwing, kicking or hitting ○ *They played in the garden with an old tennis ball.* ○ *He kicked the ball into the goal.* **2.** any round object ○ *a ball of wool* ○ *He crumpled the paper up into a ball.* **3.** a formal dance ○ *We've got tickets for the summer ball.* ◇ **to start the ball rolling** to start something happening ○ *I'll start the ball rolling by introducing the visitors, then you can introduce yourselves.* ◇ **to play ball** to work well with someone to achieve something ○ *I asked them for a little more time but they won't play ball.* ◇ **to have a ball** to enjoy yourself a lot ○ *You can see from the photos we were having a ball.*

ballet /ˈbæleɪ/ *noun* **1.** a type of dance, given as a public entertainment, where dancers perform a story to music **2.** a performance of this type of dance ○ *We went to the ballet last night.*

balloon /bəˈluːn/ *noun* **1.** a large ball which is blown up with air or gas **2.** a very large balloon which rises as the air inside it is heated, sometimes with a container attached for people to travel in ■ *verb* to increase quickly in size or amount

ban /bæn/ *noun* an official statement which says that people must not do

Figure 10: Two pages from EEBD side by side

1964 but later digitized. The version we have is of relatively good quality but still presents some anomalies, where some text blocks are unextractable from the PDF.

The Fang-French & French-Fang dictionary (Galley, 1964) is a bilingual dictionary having over 500 pages of lexical entries split into two parts. As Figure 11 shows, the markup system is totally different from the EEBD, where field transition is mostly marked with a change of font rather than with specific markers. For our experiments, we worked on the first part, Fang-French Dictionary (FFD), containing over 390 pages.

4.3 Results

For the sake of conciseness, in this paper we present an evaluation of just four selected models out of six implemented, for each dictionary. We used the benchmark module provided by GROBID to measure the precision, recall and F1 scores.

In the following tables, token level gathers the measures for each different token, field level is for each continuous sequence of the same label (so a field, a sequence of several tokens which all belongs to the same labelled chunk, e.g. a lexical entry).

4.3.1 Dictionary Segmentation

For both dictionaries, we annotated seven pages, which we split into four for training and three for evaluation.

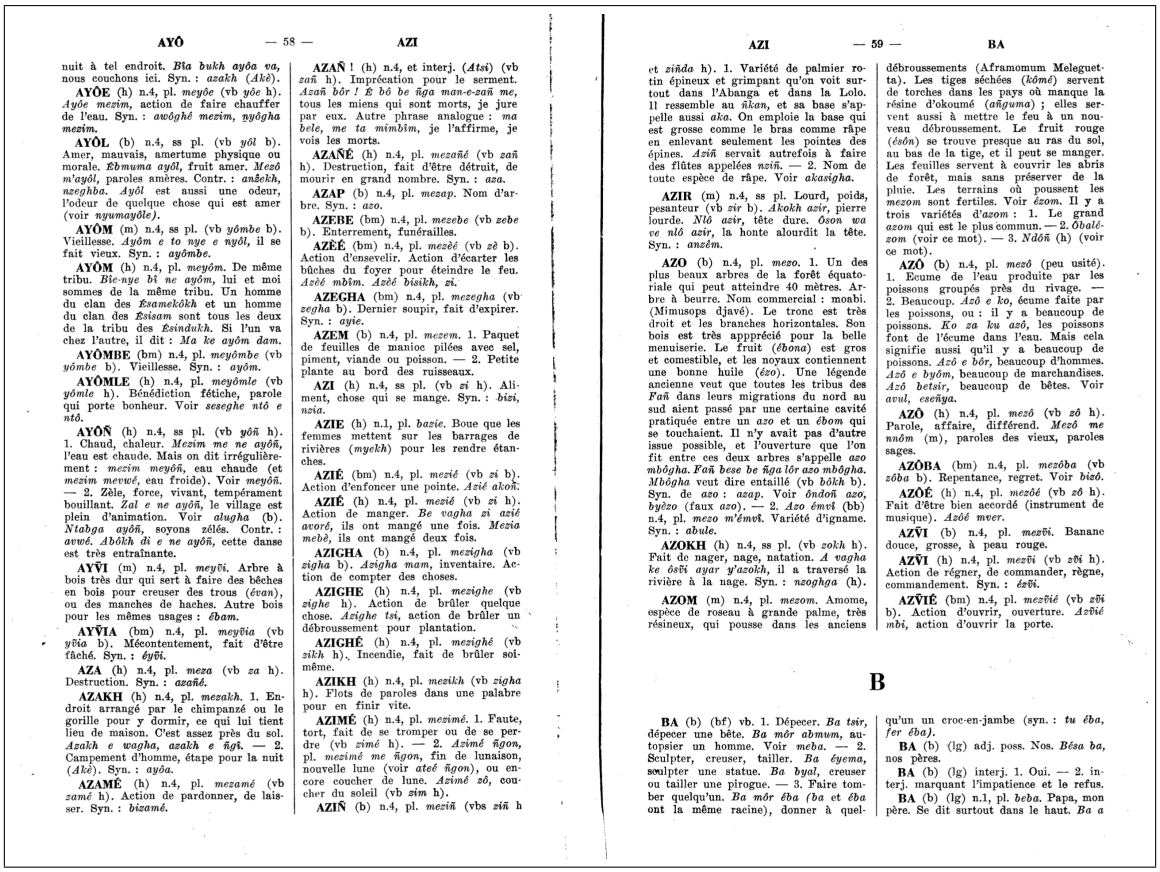


Figure 11: Two pages from FFD side by side

4.3.2 Dictionary Body Segmentation

For EEBD, we annotated five pages, which we split into 50 lexical entries for training and 27 for evaluation.

For FFD, we annotated seven pages with 91 lexical entries for training and 45 for evaluation.

4.3.3 Lexical Entry

For EEBD, we annotated eight pages, which we split into 76 entries for training and 24 for evaluation.

For FFD, we annotated three pages, which we split into 47 for training and 24 for evaluation.

4.3.4 Sense

For EEBD, we annotated six pages, which we split into 15 sense blocks for training and 15 for evaluation.

For FFD, we annotated four pages, which we split into 71 sense blocks for training and 19 for evaluation.

```
===== Token-level results =====


label                accuracy     precision    recall       f1

<body>               100          100          100          100
<headnote>           100          100          100          100
<dictScrap>          100          100          100          100

===== Field-level results =====

label                accuracy     precision    recall       f1

<body>               100          100          100          100
<headnote>           100          100          100          100
<dictScrap>          100          100          100          100
```

Table 1: Evaluation of "Dictionary Segmentation" model on EEBD

```
===== Token-level results =====


label                accuracy     precision    recall       f1

<body>               99.23        99.21        100          99.6
<headnote>           99.23        100          66.67        80

===== Field-level results =====

label                accuracy     precision    recall       f1

<body>               57.14        50           33.33        40
<headnote>           85.71        100          66.67        80
```

Table 2: Evaluation of the "Dictionary Segmentation" model on FFD

4.4 Discussion

The evaluation on both dictionaries shows a high performance by the first and second models to detect, respectively, the body part of a page and the boundaries of lexical entries. The header and punctuation predictions for the first two models are however low for the digitized sample. This could be explained by the quality of the text which sometimes led to the generation of feature values that bias the learning.

For the "Lexical Entry" model, the performance of the system remains high for the extraction of grammatical and morphological information on the English dictionary but with low precision on the Fang-French sample. The detection of related entries, which are contained only in the English dictionary, shows the limitation of our model to extract these constructs with the actual setup. We hypothesize that it is related, firstly, to a lack of annotated data and, secondly, to a lack of discriminative features. Nonetheless, the model performs relatively well for sense block detection on the English dictionary and slightly worse on the bilingual dictionary. The detection of the punctuation, representing the transition between the main fields of the model, is also limited in this model.

```
===== Token-level results =====

label                accuracy     precision    recall       f1

<entry>              100          100          100          100
<pc>                 100          100          100          100

===== Field-level results =====

label                accuracy     precision    recall       f1

<entry>              100          100          100          100
<pc>                 100          100          100          100
```

Table 3: Evaluation of the "Dictionary Body Segmentation" model on EEBD

```
===== Token-level results =====

label                accuracy     precision    recall       f1

<entry>              99.6         100          99.6         99.8
<pc>                 99.6         75           100          85.71

===== Field-level results =====

label                accuracy     precision    recall       f1

<entry>              75           61.02        80           69.23
<pc>                 88.28        75           100          85.71
```

Table 4: Evaluation of the "Dictionary Body Segmentation" model on FFD

The results of the final model reflect the reliability of our features to structure the sense information, when it has to focus on the boundaries of senses. But for the case of the senses aggregated by POS, more discriminative features should be added.

5. Related Works

This work takes place within the context of studies on lexicography and digital humanities fields, targeting the exploitation of digitized dictionaries. Most previous research (Khemakhem et al., 2009; Fayed et al., 2014; Mykowiecka et al., 2012) remained limited to the costly manual elaboration of lexical patterns, based on observing the organisation of the lexical information in a specific sample.

There have been, however, strong pointers to the usefulness of machine learning techniques, CRF in particular, to address the issue of decoding the complexity of lexical resources. Crist presented experiments for processing and automatically tagging linear text of two bilingual dictionaries, using CRF models. The goal has been purely experimental, proving the appropriateness of CRF for tagging tokens in digitized dictionaries. His exhaustive study also stressed the other processing issues, which are very important to the effectiveness and the evaluation of any parsing technique. Another recent study (Bago & Ljubešić, 2015), has addressed the issue of using CRF models to perform automatic

```
===== Token-level results =====

label                accuracy     precision    recall       f1

<form>               99.59        99.26        97.12        98.18
<pc>                 99.59        100          83.33        90.91
<re>                 89.62        0            0            0
<sense>              88.97        86.75        98.26        92.15

===== Field-level results =====

label                accuracy     precision    recall       f1

<form>               90.09        73.08        82.61        77.55
<pc>                 95.5         100          82.76        90.57
<re>                 90.99        0            0            0
<sense>              79.28        54.29        73.08        62.3
```

Table 5: Evaluation of the "Lexical Entry" model on EEBD

```
===== Token-level results =====

label                accuracy     precision    recall       f1

<form>               90.77        57.94        75.26        65.47
<pc>                 97.6         28.57        11.76        16.67
<sense>              92.45        96.46        93.59        95

===== Field-level results =====

label                accuracy     precision    recall       f1

<form>               59.12        2.94         4.17         3.45
<pc>                 85.4         28.57        11.76        16.67
<sense>              57.66        0            0            0
```

Table 6: Evaluation of the "Lexical Entry" model on FFD

language and structure annotation in a multilingual dictionary. The technique again has a very high accuracy in much less time than would be required for manual annotation.

Both of the mentioned machine learning approaches apply one CRF model to label the all the tokens of a dictionary. In such a bottom-up technique, the learner is overwhelmed by the number of labels to choose from at once, which increases the number of prediction errors. A huge amount of training data is also required per model to cover middle and high complexity dictionaries.

The novelty in our approach is that we reduce the scope of each bottom-up model by splitting the task over different models that process the lexical information in a top down fashion. Moreover, our system does not stop at the level of tagging the tokens, but enables the construction of blocks of lexical information in a format that facilitates the processing as well as the exchange of the output.

```
===== Token-level results =====

label                accuracy     precision    recall       f1

<gramGrp>            99.12        100          50           66.67
<sense>              99.12        99.1         100          99.55

===== Field-level results =====

label                accuracy     precision    recall       f1

<gramGrp>            88.89        100          50           66.67
<sense>              77.78        83.33        83.33        83.33
```

Table 7: Evaluation of the "Sense" model on EEBD

```
===== Token-level results =====

label                accuracy     precision    recall       f1

<sense>              100          100          100          100

===== Field-level results =====

label                accuracy     precision    recall       f1

<sense>              28.57        44.44        44.44        44.44
```

Table 8: Evaluation of the "Sense" model on FFD

6. Conclusion and Future Work

GROBID-Dictionaries in its first version has shown the promise of CRF cascading models to structure digitally born and digitized dictionaries, independently of the language and lexicographic style. Our experiments had the goal of, firstly, verifying our assumptions and, secondly, highlighting the strengths and the limitations of the implemented models. It is obvious that more focus should be given to the feature selection process, in order to reinforce the prediction of the models for certain labels and fields. Feature tuning should also be applied on larger annotated data with more varied instances. Therefore, we are planning to build a smart annotation tool with strong guidelines, to simplify the annotation process.

Our open source system could be used, after more tuning, to radically speed up the structuring of many digitized dictionaries in a unified scheme or to measure the structurability of OCRized lexical resources.

7. Acknowledgements

This work was supported by PARTHENOS. We would like to thank Patrice Lopez, the main designer of GROBID, for his continuous support and valuable advice.

8. References

Bago, P. & Ljubešić, N. (2015). Using machine learning for language and structure annotation in an 18th century dictionary. In *Electronic lexicography in the 21st century: linking lexical data in the digital age*. pp. 427–442.

Budin, G., Majewski, S. & Mörth, K. (2012). Creating Lexical Resources in TEI P5. A Schema for Multi-purpose Digital Dictionaries. *Journal of the Text Encoding Initiative*, 4(3), pp. 1–26.

Crist, S. (2011). Processing the Text of Bilingual Print Dictionaries. URL http://www.sean-crist.com/all/crist_dictionaries_20111210.pdf.

EEBD (2009). *Easier English Basic Dictionary: Pre-Intermediate Level. Over 11,000 terms clearly defined*. Easier English. Bloomsbury Publishing. URL https://books.google.de/books?id=nwVCBAAAQBAJ.

Fayed, D.M., Fahmy, A.A., Rashwan, M.A. & Kamel Fayed, W. (2014). Towards Structuring an Arabic-English Machine-Readable Dictionary Using Parsing Expression Grammars. *International Journal of Computational Linguistics Research*, 5(1).

Galley, S. (1964). *Dictionnaire fang-francais et francais-fang; suivi d'une grammaire fang par Samuel Galley. Avec une pref. de M.L. Durand-Reville*. H. Messeiller. URL https://books.google.de/books?id=Y-7zPgAACAAJ.

Khemakhem, A., Elleuch, I., Gargouri, B. & Hamadou, A.B. (2009). Towards an automatic conversion approach of editorial Arabic dictionaries into LMF-ISO 24613 standardized model. In *Proceedings of the Second International Conference on Arabic Language Resources and Tools*. pp. 260–265.

Lavergne, T., Cappé, O. & Yvon, F. (2010). Practical very large scale CRFs. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 504–513.

Lopez, P. & Romary, L. (2015). GROBID - Information Extraction from Scientific Publications. *ERCIM News*.

Mykowiecka, A., Rychlik, P. & Waszczuk, J. (2012). Building an electronic dictionary of old polish on the base of the paper resource. In *Proceedings of the Workshop on Adaptation of Language Resources and Tools for Processing Cultural Heritage at LREC*. pp. 16–21.

Adapting the M-ATOLL Methodology for the Generation of Ontology Lexicons to Non-Indo-European Languages: The Case of Japanese

Bettina Lanser, Philipp Cimiano

Semantic Computing Group, CITEC, Bielefeld University, Inspiration 1, 33619 Bielefeld, Germany
E-mail: blanser@cit-ec.uni-bielefeld.de, cimiano@cit-ec.uni-bielefeld.de

Abstract

In order to make the growing amount of conceptual knowledge available through ontologies and datasets accessible to humans, NLP applications need access to information on how this knowledge can be verbalized in natural language. One way to provide this kind of information are ontology lexicons, which apart from the actual verbalizations in a given target language can provide further, rich linguistic information about them. Compiling such lexicons manually is a very time-consuming task and requires expertise both in Semantic Web technologies and lexicon engineering, as well as a very good knowledge of the target language at hand. In this paper we present an alternative approach to generating ontology lexicons by means of the framework M-ATOLL. So far, M-ATOLL has been used with Indo-European languages that share a large set of common characteristics. We explore if M-ATOLL can also be used fruitfully with Non-Indo-European languages; for this purpose, we use M-ATOLL to generate a Japanese ontology lexicon for DBpedia.

Keywords: Ontology lexicalization; M-ATOLL; DBpedia

1. Introduction

As the amount of formalized conceptual knowledge available through datasets and ontologies grows, there is an increasing need to make this knowledge accessible to humans in an easy and intuitive way. One way to accomplish this is by means of language technology, e.g. in the form of question answering systems, that allows users to query repositories of conceptual knowledge through natural language. Of course, in order to e.g. map the natural language input onto the elements of the conceptual knowledge repository at hand, language technology systems that build upon repositories of conceptual knowledge need access to information on how the elements of the repository at hand can be verbalized in a given language. Ontology languages support the inclusion of such information to a certain extent, e.g. by means of `rdfs:label` or SKOS properties. However, these ontology-internal mechanisms usually do not provide further information about the labels' linguistic behavior, such as their part-of-speech or irregular inflectional forms they may take. In addition, labels only capture one canonical way of verbalizing an ontology element, but do not provide lexical variants.

As a result, in many scenarios external resources of linguistic information will be preferable in order to make resources of conceptual information accessible to language technology systems. Wiktionary[1] or WordNet(Miller, 1995), while providing rich linguistic information and lexical variants, do not contain any anchors between verbalizations and elements of a specific ontology.

One possible type of lexical resource are ontology lexicons (Prévot et al., 2010; McCrae et al., 2011b), which were specifically designed for the task of linking ontology elements to possible verbalizations in a given language enriched with various kinds of linguistic information. Conventionally, such ontology lexicons are generated manually, which is a very time-consuming task that requires expertise in Semantic Web technologies and lexicon engineering, as well as knowledge about the domain of the ontology. Furthermore,

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

in order to decide which verbalizations are appropriate for a given ontology element, in many cases one either needs to have a very good command of the target language at hand oneself, or one should at least be able to consult with native speakers, which in case of smaller target languages may pose a problem. While the latter problem may in principle be solved by translating an already existing ontology lexicon (McCrae et al., 2011a; Arcan & Buitelaar, 2013), corresponding systems have not yet reached an accuracy sufficient to produce high-quality lexicons off the shelf.

Therefore, this paper will deal with an alternative approach to ontology lexicalization that requires less manual effort: We will look at M-ATOLL (Multilingual, Automatic inducTion of OntoLogy Lexica; Walter, 2017), a framework for the (semi-)automatic generation of ontology lexicons in the RDF-based lemon format (McCrae et al., 2011b). Another main topic of this paper is ontology lexicalization specifically for Non-Indo-European languages: So far, M-ATOLL has been used with a number of Indo-European languages — English, German and Spanish — that share a rather large set of common characteristics. We investigated whether M-ATOLL can also be used fruitfully with Non-Indo-European languages and what kinds of adaptations to the framework would be necessary in order to make that work. Finally, we investigated whether lemon, the format for the specification of ontology lexicons used by M-ATOLL, in itself is flexible enough to support ontology lexicons in Non-Indo-European languages.

In order to investigate these topics we used M-ATOLL to generate a Japanese ontology lexicon for excerpts from DBpedia's ontology. We chose Japanese as our example language as it is one of the few Non-Indo-European languages for which a comparably large amount of NLP-related tools and resources as required by M-ATOLL is available. While working with a more underresourced language and seeing how the problems emerging from data sparseness in this case may be solved would definitely be worthwhile, in the context of this paper we wanted to focus on problems with language portation that are more directly related to the structure of M-ATOLL and lemon, respectively.

2. M-ATOLL

M-ATOLL (Walter, 2017) is a framework for the automatic induction of ontology lexicons in multiple languages. In general, the framework takes as its input at least an ontology, together with a corresponding knowledge base, and produces as its output a lexicon serialized in the lemon format that lexicalizes the input ontology. M-ATOLL is a combination of different approaches.

The corpus-based approach, which is M-ATOLL's main approach, lexicalizes only properties and is based on a dependency-parsed text corpus whose sentences M-ATOLL tries to match to predefined, language-specific dependency patterns. It consists of two main steps: First, M-ATOLL tries to extract relevant sentences from the corpus that may express a given property `p`, and preprocesses the sentences retrieved this way, as follows: First of all, for the given property `p` all triples are extracted from the knowledge base that contain this property as their predicate, i.e. which have the form

```
s p o .
```

Then, for each subject/object pair `s,o` retrieved this way, one selects all those sentences from the corpus for further processing that contain labels of the subject and object, i.e. which contain strings `s',o'` such that

```
s rdfs:label s' .
```

and

```
o rdfs:label o' .
```

are contained in the knowledge base. The dependency parses of the sentences which have been selected this way are then converted into RDF, and the nodes in the dependency tree that correspond to the subject and object labels based on which the sentence was selected are marked. In the second step of the corpus-based approach, the actual candidate lexicalizations are extracted from the sentences which were selected and turned into RDF in the preceding step: Each selected sentence is matched against a set of handcrafted, language-specific dependency patterns specified as SPARQL queries. Since the sentences are given in RDF, this amounts to a simple query operation. If there is a match between a sentence and a dependency pattern, a lexical entry is created. To do so, the output of the SPARQL query is matched onto one of several lemon-based templates. Finally, the candidate lexical entries retrieved this way are filtered, e.g. based on the number of sentences they were encountered in, in order to reduce noise in the final lexicon, and the actual lexicon is serialized as lemon RDF.

So far, all approaches covered by M-ATOLL support English, while the corpus-based approach also supports German and Spanish. Similarly, since it is the core approach of M-ATOLL, this paper will deal with adapting M-ATOLL's corpus-based approach to Japanese.

3. Adapting M-ATOLL to Japanese

3.1 Input Format

Since we want to port the corpus-based approach to the Japanese Wikipedia, the source data for generating the input for M-ATOLL are the texts from the Japanese Wikipedia in XML format, which can be downloaded from the site of the Wikimedia Foundation.[2] We extract the sentences from the XML file with an already existing script.[3] We then run the morphological analyzer MeCab[4] on the sentences, which splits them into their single tokens and provides further information about the tokens such as their part-of-speech. The result of MeCab is again used as the input to the dependency parser J.DepP;[5] its output for the following example is shown in Figure 1.

(1) *1943*年、 ロスアラモス国立研究所-を 建設した。
1943.year Los.Alamos.National.Laboratory-DOBJ constructed
In 1943, [someone] constructed the Los Alamos National Laboratory.

In contrast to most parsers for Indo-European languages, Japanese dependency parsers generate dependency structures that do not hold between single tokens, but between multi-word units called *bunsetsus*. For example, in Figure 1 multi-word unit 0, which contains

[2] https://dumps.wikimedia.org/jawiki/
[3] http://medialab.di.unipi.it/wiki/Wikipedia_Extractor
[4] http://taku910.github.io/mecab/
[5] http://www.tkl.iis.u-tokyo.ac.jp/~ynaga/jdepp/

```
# S-ID: 656657; J.DepP
* 0 2D
1943     名詞,数,*,*,*,*,*
年          名詞,接尾,助数詞,*,*,*,年,ネン,ネン
、          記号,読点,*,*,*,*,、,、,、
* 1 2D
ロスアラモス          名詞,一般,*,*,*,*,*
国立          名詞,一般,*,*,*,*,国立,コクリツ,コクリツ
研究所
名詞,一般,*,*,*,*,研究所,ケンキュウジョ,ケンキュージョ
を          助詞,格助詞,一般,*,*,*,を,ヲ,ヲ
* 2 -1D
建設          名詞,サ変接続,*,*,*,*,建設,ケンセツ,ケンセツ
し          動詞,自立,*,*,サ変・スル,連用形,する,シ,シ
た          助動詞,*,*,*,特殊・タ,基本形,た,タ,タ
。          記号,句点,*,*,*,*,。,。,。
EOS
```

Figure 1: Output of dependency parser J.DepP for example sentence 1

the tokens 1943, 年 and a comma, depends upon multi-word unit 2, which consists of the tokens 建設, し, た and a full stop. The grammatical information provided for each token comprises up to four part-of-speech tags of differing granularity, the inflection class and the given inflection form in case of verbs and adjectives, the base form of the token, its reading, and its pronunciation. Table 1 shows those part-of-speech and inflection type tags that were used in the formulation of the SPARQL queries later on. In the next step, we remove all punctuation marks from the parsed sentences, which facilitated writing the SPARQL queries in a subsequent step.

part-of-speech	part-of-speech subcategory 1	inflection type
名詞 noun	サ変接続 verbal (nouns that can form verbs by being followed by する or related verbs)	
動詞 verb	自立 main (i.e. non-auxiliary)	
		特殊・デス copula verb です
		特殊・ダ copula verb だ
助詞 particle	係助詞 dependency (comprises topic marker は)	
	連体化 adnominalizer (non-possessive の that joins nouns together)	

Table 1: Part-of-speech and inflection type tags used in the SPARQL queries

One of the tasks of M-ATOLL's sentence preprocessing component is to turn the input sentences into RDF. Since the Malt parser,[6] which had been used for dependency parsing the English and Spanish input to M-ATOLL (Walter, 2017: p. 144), uses the CoNLL format[7] as its output format, the sentence preprocessing component was already able

[6] http://www.maltparser.org/userguide.html
[7] http://ilk.uvt.nl/conll/

to deal with this format and turn it into RDF. Furthermore, a token-based dependency structure allows one to use both dependency relations among bunsetsus and among tokens in the specification of one's SPARQL queries, and in order to keep both options available, we wanted to transform the bunsetsu-based dependency structure into a token-based one, which could be better represented in the CoNLL format. Hence, we decided to transform the original output format of J.DepP into a modified version of the CoNLL format for further processing.

The dependency parse of example sentence 1 is again shown in Figure 2, this time in the CoNLL format. Table 2 shows a comparison between the features occurring in J.DepP's output and those employed in the CoNLL format. One of the main differences between the two formats is that in the CoNLL format instead of multi-word units each single token is assigned an index, and dependency relations hold between tokens. When transferring the J.DepP format into the CoNLL format we had to generate the token indices (`ID`) from the tokenization provided by MeCab. In contrast, `FORM` and `LEMMA` could be directly mapped from the respective columns in the J.DepP format. For the `CPOSTAG` and `POSTAG` columns we used the main part-of-speech tag column from the J.DepP format (column 2) and the first sub-part-of-speech tag column (3), respectively; hence, the information about the other two part-of-speech subtypes was lost in the transformation, which we considered not that problematic since most of the time those two columns are empty anyway. The information about inflection classes and forms (columns 6 and 7) was merged into the `FEATS` column in the CoNLL format separated by a vertical bar. In order to generate the correct values for the `HEAD` column, the following rules were used in order to transform the original bunsetsu-based dependency structure into a token-based one:

- If in the original dependency-based structure bunsetsu b_1 depends upon bunsetsu b_2, then in the token-based structure the last token of b_1 depends upon the last token of b_2.
- If a token belongs to a bunsetsu b and is not the last token within that bunsetsu, it depends upon the token that directly follows it within b.

As an example, Figure 3 shows how the bunsetsu-based dependency structure of sentence 1 would be transformed into a token-based structure. The remaining columns of the CoNLL format were left empty: Most Japanese dependency parsers such as J.DepP do not assign labels to the dependencies, hence no information was available for the `DEPREL` column. The remaining two columns seem to serve no real purpose in the context of M-ATOLL; at least they are referenced nowhere in the SPARQL queries for the Indo-European languages. The last two columns of the J.DepP format, which contain information about the reading and the pronunciation of the token at hand, were discarded in the transformation process, as this kind of information did not seem very relevant to the purpose of an ontology lexicon.

In order to keep the information about which tokens belong to which bunsetsu, we adopted the representation of multi-word units used in the CoNLL-U format,[8] which is a revised version of CoNLL aimed at being able to represent a larger variety of different languages:[9] Multi-word units are given in addition to the tokens they are comprised of, and instead of a single index they are assigned a range of indices, as shown in Figure 2. The remaining features are not specified for multi-word units.

[8] http://universaldependencies.org/format.html
[9] http://universaldependencies.org/introduction.html

J.DepP		CoNLL	
field number	field name/descr.	field number	field name/descr.
1	surface form	1	ID (token counter, starting at 1 for each new sentence)
2	part-of-speech	2	FORM (word form/punctuation symbol)
3	part-of-speech, subtype 1	3	LEMMA (lemma or stem; underscore if not available)
4	part-of-speech, subtype 2	4	CPOSTAG (coarse-grained part-of-speech tag)
5	part-of-speech, subtype 3	5	POSTAG (fine-grained part-of-speech tag)
6	inflection class (for verbs and adjectives)	6	FEATS (set of morphological and/or syntactic features, separated by |, underscore if not available)
7	inflection form (for verbs and adjectives)	7	HEAD (head of the current token; either a value of ID or zero)
8	lemma	8	DEPREL (type of the dependency relation to the head)
9	reading	9	PHEAD (projective head of the current token; either a value of ID, zero, or underscore if not available)
10	pronunciation	10	PDEPREL (type of the dependency relation to the projective head; underscore if not available)

Table 2: Types of information present for each token in the output format of MeCab/J.DepP (http://taku910.github.io/mecab/) and in the CoNLL format (Walter, 2017: 29)

```
1     1943          _          名詞          数          _|_      2          _
2     年    年          名詞          接尾          _|_      9          _
3     ロスアラモス    _     名詞     一般     _|_      4     _
4     国立    国立          名詞          一般          _|_      5          _
5     研究所    研究所          名詞          一般          _|_      6          _
6     を    を          助詞          格助詞          _|_      9          _
7     建設    建設          名詞          サ変接続          _|_      8          _
8     し    する     動詞     自立     サ変・スル|連用形     9          _
9     た    た     助動詞     _     特殊・タ|基本形     0     _
1-3     米国政府は _          _          _          _          _          _
4-5     1943年 _          _          _          _          _          _
6-11     第二次世界大戦の          _          _          _          _          _          _
12-13     最中に          _          _          _          _          _          _
14-17     ロスアラモス国立研究所を          _          _          _          _          _          _
18-20     建設した          _          _          _          _          _          _
```

Figure 2: Output of dependency parser J.DepP for example sentence 1, turned into CoNLL format

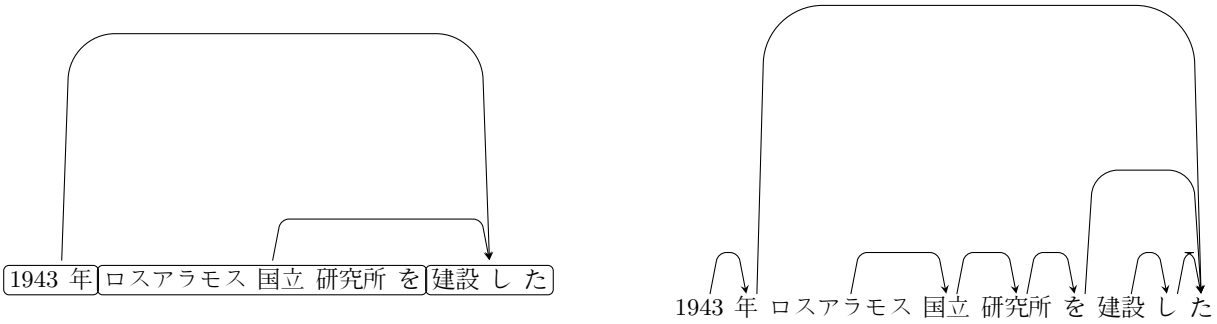


Figure 3: Exemplary transformation of bunsetsu-based into token-based dependency structure. The boxes indicate bunsetsu boundaries.

3.2 Dependency Patterns

3.2.1 Overview

We first manually defined eleven dependency patterns for Japanese in terms of SPARQL queries. Six of these patterns serve to retrieve noun lemmas, while the remaining five match verbs. We have not yet dealt with adjective lemmas and the respective SPARQL queries.

The patterns were retrieved based on five example properties from DBpedia's ontology, `parent`, `occupation`, `yearOfConstruction`, `crosses` and `nationality`. The approach to generating the patterns was similar to that described in Walter (2017: 55):

1. For a given property, we extracted all sentences from the Japanese Wikipedia that contain labels of entity pairs which are linked by the respective property in DBpedia's triple set.
2. Furthermore, we manually compiled a set of gold verbalizations our SPARQL patterns should be able to find. In part, we used verbalizations found through crowdsourcing as described in Lanser et al. (2016) for this.
3. We then searched the sentences from 1) for occurrences of these gold verbalizations. We looked at the dependency constructions they were embedded in, and watched out for frequently occurring patterns.

For example, we first looked at all sentences that contain on the one hand labels of entities that are linked by the property `parent` in DBpedia's triple store and on the other hand one of the verbalizations we had received through crowdsourcing for that property. This way, we found a number of sentences in which the entity label pairs and the verbalizations occur in the same kind of construction, such as the following:

(2) ヘンリー2世-の 母親 である 皇后 マティルダ-は これ-に
Henry.II-POSS mother COP empress Mathilda-TOP this-IOBJ
反対した
opposed
Empress Mathilda, who is the mother of Henry II, opposed this

(3) 宮崎吾朗-の 父親 である 宮崎駿-は 『ゲド戦記』-の
Goro.Miyazaki-POSS father COP Hayao.Miyazaki-TOP "Earthsea"-POSS
古く-から-の ファン であり
ancient-from-POSS fan COP

Hayao Miyazaki, who is the father of Goro Miyazaki, is an old fan of "Earthsea"

This structure also reoccurred for other properties, such as for `crosses` in the following example, which gave us confidence that it is indeed a general, not property-specific construction that should be incorporated in the set of dependency patterns M-ATOLL uses for Japanese. In general, when a given structure could only be found in sentences for one particular property, we decided based on intuition whether it may be a general or a property-specific structure.

(4) 木曽川-の 橋 である 愛岐大橋-は 慢性的な 渋滞-が
Kiso.river-POSS bridge COP Aichi.Bridge-TOP frequent congestion-SUBJ
発生している
was.happening

on Aichi Bridge, which is a bridge of the Kiso river, frequent congestions were happening

As a result, the respective pattern was added to the set of SPARQL queries.

3.2.2 Noun Patterns

Similarly to that which has been described for English, German and Spanish in Walter (2017: 55), most patterns we were able to identify for noun lemmas correspond either to an appositive or a copula construction. With respect to copula constructions we defined two SPARQL queries corresponding in English to the constructions *[e1] is the [lemma] of [e2]* (e.g. *Lydia Hearst is the child of Patty Hearst*) and *The [lemma] of [e2] is [e1]* (e.g. ボブ・ショーの職業はジャーナリスト *The occupation of Bob Shaw [is] journalist*). While in Japanese a number of different constructions may be considered appositions (Heringa, 2012), we only came along one of these construction types, where anchor and apposition are placed directly alongside. We generated two different patterns (*[e1]'s [lemma][e2]* and *[e1] is a NN of [e2][lemma]*) in which the apposition is embedded into one of its two most commonly occurring syntactic contexts, respectively, since the single, very general apposition pattern we used at first produced a lot of noise. We only found one further pattern that did not belong to either of these two groups, in which the lemma occurs as a direct object of a relative clause that contains the first entity as a further participant and has the second entity as its head.

3.2.3 Verb Patterns

For English, German and Spanish, there are separate patterns for transitive and intransitive verbs, as well as for verb occurrences in the active and passive voice, respectively (Walter, 2017: 131–136). For Japanese, in contrast, due to the use of particles to mark all

grammatical functions alike, and the way the passive voice gets marked simply through an auxiliary, one does not necessarily need to differentiate between patterns for transitive and intransitive verbs, and patterns for verbs in the active and passive voice, respectively. Hence, for Japanese one would actually only need one pattern for verbs in main clauses, and one pattern for verbs in relative clauses, respectively.

At first we allowed both entity labels to have arbitrary grammatical functions in our verb patterns; in particular, we did not require one of them to be in subject position. The reasoning behind this was that since Japanese is a pro-drop language and may omit any verb argument — including the subject — in principle also lexicon entries for verbs in which both entities occupy non-subject positions may be turned into well-formed sentences. However, when looking at the entries generated by this first version of the patterns it turned out that gold lemmas occurred only in clauses where one of the entity labels occupies the subject position, and that other kinds of clauses most of the time do not express the desired relationship between the two entities, as illustrated by examples 5 and 6 below. Hence, in order to reduce noise at current all verb patterns only match clauses where the label of one of the entities from the triple store is most probably in subject position, i.e. where it is either marked by the subject particle が or the topic particle は without any preceding particles, which most of the time indicates that it is a substitute for the subject particle.

In contrast to the English, German and Spanish patterns for verb occurrences in the passive voice (Walter, 2017: 131–136) the Japanese verb patterns also match clauses in the passive voice without an overt agent, i.e. clauses which when transferred into active voice would not have an overt subject, as exemplified by sentences 7 to 10 below: It turned out that such clauses regularly contained gold lemmas (7) or expressed the desired relation between the entities from the triple store by some other matching verbalization (9).

(5) *property:* ***parent***

安楽公主-と共に　中宗-を　毒殺した
Yasushi.Kura-along.with　Nakasune-DOBJ　poisoned
[someone] poisoned Nakasune along with Yasushi.Kura

(6) *property:* ***occupation***

会長職-を　李健熙-に　返上した
chairman.position-DOBJ　Lee.Kun.he-IOBJ　gave.up
[someone] gave up the chairman position to Lee Kun-he

(7) *property:* ***yearOfConstruction***

グラニット鉄道-は、　*1826*年*4*月*1*日-に　着工された
Granite.railway-TOP　1826.year.4.month.1.day-on　was.started
[the construction of] the Granite Railway was started on April 1, 1826

(8) グラニット鉄道-を　*1826*年*4*月*1*日-に　着工した
Granite.railway-DOBJ　1826.year.4.month.1.day-on　started
[someone] started [the construction of] the Granite Railway on April 1, 1826

(9) *property:* `occupation`

声優-として 植田佳奈-が 採用された
voice.actor-as Kana.Ueda-SUBJ was.employed
Kana Ueda was employed as a voice actor

(10) 声優-として 植田佳奈-を 採用した
voice.actor-as Kana.Ueda-DOBJ employed
[someone] employed Kana Ueda as a voice actor

3.3 Lexicon Entry Generation

When a sentence matches one of the SPARQL queries, in order to create the actual lexicon entry M-ATOLL matches the output of the SPARQL query to one of several templates, which roughly correspond to the different (sub-)parts-of-speech a candidate verbalization may belong to and generate a lemon-based lexicon entry for the candidate verbalization at hand. The syntactic behavior of the candidate verbalization is defined in terms of one of the subcategorization frames specified in the linguistic ontology LexInfo (Cimiano et al., 2011), which describe the syntactic argument structure of candidate verbalizations.

As mentioned before, the noun patterns for Japanese are very similar to those defined for English, German and Spanish, and accordingly the already existing template `NounWithPrep` would in principle have been a rather good match for generating lexicon entries for Japanese candidate noun verbalizations. However, when this template is used, throughout the resulting lexicon entry the term *preposition* is used, as shown by the following English entry:

- canonical form: *discoverer*
- part-of-speech: common noun
- subcategorization frame: noun PP frame
  arguments:
  - copulative argument e_1
  - prepositional object e_2 with preposition *of*
- semantic reference: `discoverer`
  arguments:
  - subject e_1
  - object e_2

Since Japanese particles are not pre- but postpositions, this terminology would be unfavorable in Japanese lexicon entries. Hence, we defined a kind of more general template `NounWithAdpos` that only differs from `NounWithPrep` in that it references adpositions instead of prepositions:

- canonical form: 発見者
- part-of-speech: common noun
- subcategorization frame: noun AdP frame
  arguments:

  - copulative argument e_1
  - adpositional object e_2 with adposition の
- semantic reference: `discoverer`
  arguments:
  - subject e_1
  - object e_2

Since in Japanese all verb arguments are marked the same way, in contrast to English, German and Spanish we do not differentiate between templates for transitive and intransitive verbs with an adpositional argument. Rather, we make use of two new templates, `ActiveVerb` and `PassiveVerb`, that each create lexicon entries which reference a subcategorization frame with a subject and an adpositional object. For example, for sentence 7 the `PassiveVerb` template would be invoked and the following entry would be created:

- canonical form: 採用される
- part-of-speech: verb
- subcategorization frame: passive AdP frame
  arguments:
  - subject e_1
  - adpositional object e_2 with adposition として
- semantic reference: `occupation`
  arguments:
  - subject e_1
  - object e_2

Here, transitive and intransitive verbs only differ in that for transitive verbs the marker of the adpositional object is always を, while for intransitive verbs it is any other marker. While it would be possible to use only one single template for all Japanese verb lemmas by turning the passive verbs into their active form, in cases such as example 7 or 9 this would lead to entries without a subject.

4. Evaluation

4.1 SPARQL Queries

In order to check how comprehensive our SPARQL queries for dependency patterns are, we took the gold lexicon for the properties we used to generate the SPARQL queries and looked at how many instances of the lemmas from this gold lexicon are found by our queries, and how many gold instances are occurring overall in positions where they may in principle express a relationship between subjects and objects from DBpedia's triple store. In order to determine the latter value, for each example property we retrieved all sentences containing labels of elements linked in DBpedia's triple store by the respective property, and counted how often the gold lemmas for the property at hand occurred between or behind these triple subjects and objects. Since Japanese is a strongly head-final language, this should cover all instances of the gold lemmas that may potentially express a relation between triple subjects and objects. The results are shown in Table 3, together with counts of how often each gold lemma was actually found by our SPARQL queries.

Out of the 29 lemmas in the gold lexicon, seven were not found at all by the SPARQL queries, which corresponds to a recall of 0.76. In only one case this is due to the lemma not occurring between or behind triple subjects and objects at all. Furthermore, the overall coverage of instances of the gold lemmas by the SPARQL queries was rather low: for example, for a lemma such as 務める (to serve) only 16 instances are found by SPARQL queries, while over 300 occur between or behind triple subjects and objects. As the number of instances found for a given lemma may serve as an important parameter when deciding which generated entries to keep in the lexicon and which to discard, we first looked into how to improve the overall coverage of our SPARQL queries in terms of found instances, and whether in the process the recall, i.e. the coverage in terms of found lemmas, may improve as well.

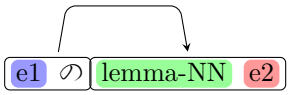
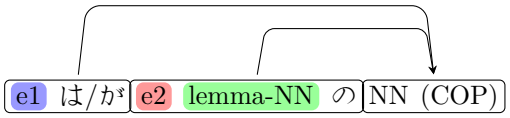
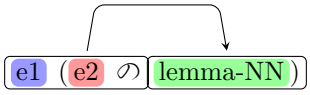
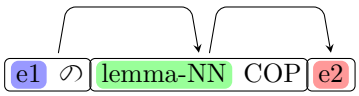
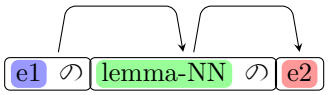
For each instance of a gold lemma found between or behind a triple subject and object, we constructed the minimal path between these three elements within the sentence's dependency tree, and grouped together all instances that share the same path structure.

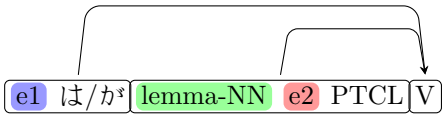
This analysis showed that the vast majority of dependency patterns occurs only once. Basing new SPARQL queries or modifications to existing queries on patterns that only occur a few times overall would probably not be very worthwhile.

Therefore, we looked at the 10 dependency patterns occurring most frequently in more detail, checking whether they were already covered by our SPARQL queries, and if not, whether it would make sense to build new SPARQL queries based on them, the results of which are shown in Table 4.

property	# sent.s	verbalization	# instances of lemma between/after triple subject and object	# sent.s found through SPARQL queries	% coverage
`crosses`	241	跨ぐ (to step over, to bridge)	5	2	40
		架かる (to span, to cross)	91	12	13.19
		かかる (writing variant of 架かる)	17	2	11.76
		またがる (to extend over)	1	1	100
		渡る (to cross over)	20	2	10
`nationality`	4660	国籍 (nationality)	9	2	22.22
		出身 (person's origin)	283	38	13.43
		生まれ (birthplace)	33	3	9.09
`occupation`	9531	仕事 (work)	58	0	0
		職業 (occupation)	10	0	0
		職 (job, position)	16	0	0
		生業 (job)	0	0	-
		勤める to work (for)	9	1	11.11
		務める to serve (as)	342	16	4.68
		活動 (activity)	311	5	1.61
		働く (to work)	9	0	0
`yearOfConstruction`	281	完成 (completion)	11	4	36.36
		竣工 (completion of construction)	2	0	0
		建設 (construction)	16	3	18.75
		建てる (to build)	5	3	60
`parent`	11,831	子供 (child)	104	2	1.92
		子 (child [of someone])	948	31	3.27
		父親 (father [of someone])	54	3	5.56
		父 (father)	651	69	10.60
		娘 (daughter)	928	65	7.00
		息子 (son)	1457	179	12.29
		親 (parent)	9	0	0
		母 (mother)	446	17	3.81
		母親 (mother [of someone])	49	1	2.04

Table 3: Number of instances of gold lemmas found between or after triple subjects and objects, and number of instances that are actually found by our SPARQL queries

Pattern	example	# sent.s	SPARQL pattern?
	フリードリヒ*4*世-の 息子 Friedrich.IV-POSS son フリードリヒ*5*世 Friedrich.V Friedrich IV's son Friedrich V	387	yes
	アン・ヒューズ-は Ann.Hughes-TOP イギリス 出身-の England origin-POSS 柔道選手。 judo.player Ann Hughes is a judo player of English origin.	128	yes
	スレイマン*1*世 Suleiman.the.Magnificent (セリム*1*世の 子) Selim.I-POSS child Suleiman the Magnificent (Selim I's child)	124	no (new pattern created)
	リチャード*1*世-の 父親 Richard.I-POSS father である ヘンリー*2*世 COP Henry.II Henry II, who is the father of Richard I	95	yes
	リュクルゴス-の 子-の Lykurgos-POSS child-ADN ペロプス Pelops Lykurgo's child Pelops	67	no (new pattern created)

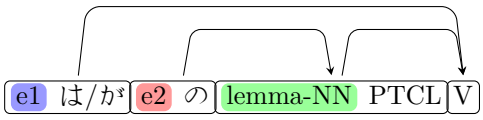


ウァレリアヌス-は 息子
Valerian-TOP son
ガッリエヌス-を
Galienus-DOBJ
ローマ帝国-の
Roman.Empire-POSS
西半分-を 任せた
western.half-DOBJ left

Valerian left the western half of the Roman Empire to [his] son Galienus.

60

no (too specific to `parent`?)

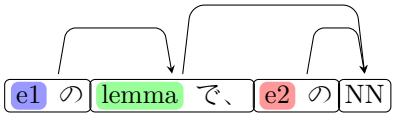


ラージャーラーム-は
Rajaram-TOP
シヴァージー-の
Shivay-POSS
息子-として 生まれた
son-as was.born

Rajaram was born as the son of Shivay

48

no (in ca. 50% of cases no relationship between e1 and e2 is expressed)

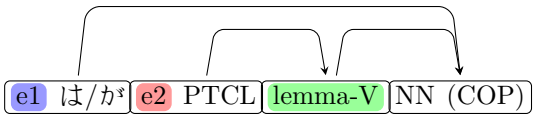


ロマノス2世-の 娘
Romanos.II-POSS daughter
で、 バシレイオス2世-の
COP Basilius.II-POSS
妹。
sister

[She] is the daughter of Romanos II and the sister of Basilius II.

47

no (expresses no direct relationship between e1 and e2)

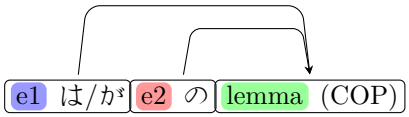


シャンジュ橋-は、
Change.bridge-TOP
セーヌ川-に 架かる
Seine-IOBJ to.cross
橋 である。
bridge COP

Pont au Change is a bridge that crosses the Seine.

38

no (new pattern created)



クレオメネス3世-は Cleomenes.III-TOP レオニダス2世-の 息子 Leonidas.II-POSS son である。 COP Cleomenes III is the son of Leonidas II.	37	yes	

Table 4: Most frequently occurring patterns over all lemmas from gold lexicon

Four out of these 10 dependency patterns were already covered by SPARQL queries. In addition, we wrote three more queries for patterns from the list that on the one hand seemed not too specific to a certain property and in which on the other hand the lemma seems to actually express a relationship between the triple subject and object in the majority of cases; the latter was decided based on a sample of 10 random instances of the respective pattern. The remaining three patterns were considered unsuitable for being turned into SPARQL queries: In one pattern the lemma does not express a relationship between the triple subject and object, but between the triple subject and another noun; in a further case, the pattern seems to convey a relationship between subject and object in only around half of all instances, and incorporating this pattern as a SPARQL query would hence most likely result in lots of incorrect entries. Finally, in the third case we suspected the pattern may be very specific to the property `parent`, and may produce lots of erroneous data for other properties. It should be noted that in addition to the patterns occurring most frequently in total, we also looked at the most frequent dependency patterns over those sentences that M-ATOLL did not cover yet. This way, it turned out that a number of sentences that the already existing SPARQL queries were supposed to match were not found yet, and according modifications were applied to the queries to improve their coverage.

Table 5 in the Appendix shows how these modifications and the introduction of the three new SPARQL queries influence the number of lemma instances found by M-ATOLL.

Overall, at least for some lemmas significantly more instances are found now; however, the recall has improved only very slightly: Only one further lemma is found, in only one sentence. In general, it seems that the applied changes lead to lemmas which were previously found frequently to be found even more often, while for lemmas which were found only a few times — or not at all — the numbers did not change much. In order to check if we could also improve coverage and recall for the less frequently found lemmas, we again looked at a list of most frequently occurring dependency patterns, this time based only on sentences not found by M-ATOLL yet and and a reduced set of gold lemmas with the five most frequently found ones being removed. This time the found patterns were of significantly lower quality: In most cases none of the sentences belonging to a given pattern express a direct relationship between triple subject and object — at least not by means of the lemma at hand — and one further pattern which already occurred in Table 4 seems too specific to the property `parent`. Since any further dependency pattern occurring in the data would at most match three instances of less frequently covered lemmas, we

decided that looking at further patterns would probably not be worthwhile and did not apply any further changes to our set of SPARQL queries.

4.2 Verbalizations Retrieved by M-ATOLL

In order to test how well the SPARQL patterns generalize, we computed precision, recall and f-measure for the ontology lexicon generated by M-ATOLL on the properties used already in the preceding section, and compared these values to those of another M-ATOLL lexicon generated for five new properties, `author`, `bandMember`, `foundingYear`, `languageFamily`, and `locationCity`. The results for the old set of example properties are shown in Figure 4, while the results for the five new example properties are depicted in Figure 5. As mentioned before, the entries created by M-ATOLL should be filtered in some way; we looked at a filtering strategy filtering strategies, both based on the number of times a given lemma has been found in the corpus: we sorted the entries according to the number of occurrences of their lemmas in descending order, and included only the first x entries from this list in the final lexicon.

So for example, the lexicon whose values are given at point four of the x-axis would contain the entries for the four most frequently occurring lemmas. One should note that since multiple lemmas may occur the same amount of times, the sorting of the entries may not be definite, and different entries may show up in the final lexicon if the filtering process is repeated. In contrast to other filtering strategies where e.g. only lemmas are included in the final lexicon that occur a certain number of times, this filtering strategy may be of advantage if one always wants to have a certain, fixed number of entries in one's final lexicon. Furthermore, it may be preferable when the number of entries M-ATOLL is able to extract differs significantly among different properties: For a property for which only a few entries are extracted, lemmas with only one or two occurrences may already be good verbalizations, while for a property with hundreds or thousands of generated entries such lemmas would most probably not be suitable.

As can be seen from Figure 4 and 5, the measures are roughly comparable among both lexicons. For smaller lexicons precision is higher for the lexicons based on the old properties, while for larger lexicons recall is slightly higher for the lexicons based on the new properties. However, overall the numbers seem to suggest that the SPARQL queries used for Japanese by M-ATOLL are not overfitted to the properties we used in Section 4.1.

While the recall of both the lexicon for the old and new properties can be brought to a halfway acceptable level with the right filtering strategy, precision is overall very low, i.e. only few of the entries in the M-ATOLL lexicons correspond to entries from the manually created gold lexicons. Therefore, for the M-ATOLL lexicon covering the new properties we looked at the top 20 entries received by the second filtering strategy described above, and checked whether these entries are really of low quality for the most part, or whether there may be some problem with the gold lexicon in terms of coverage instead. The original entries of the gold lexicon, plus additional entries from the top 20 lexicon we considered appropriate for verbalizing the respective property, are shown in Table 6.

For every property there were at least five such additional entries. For the most part they were not included in the gold lexicon due to a mismatch in semantic granularity: Some of these verbalizations are more specific than the property at hand, such as ボーカリス

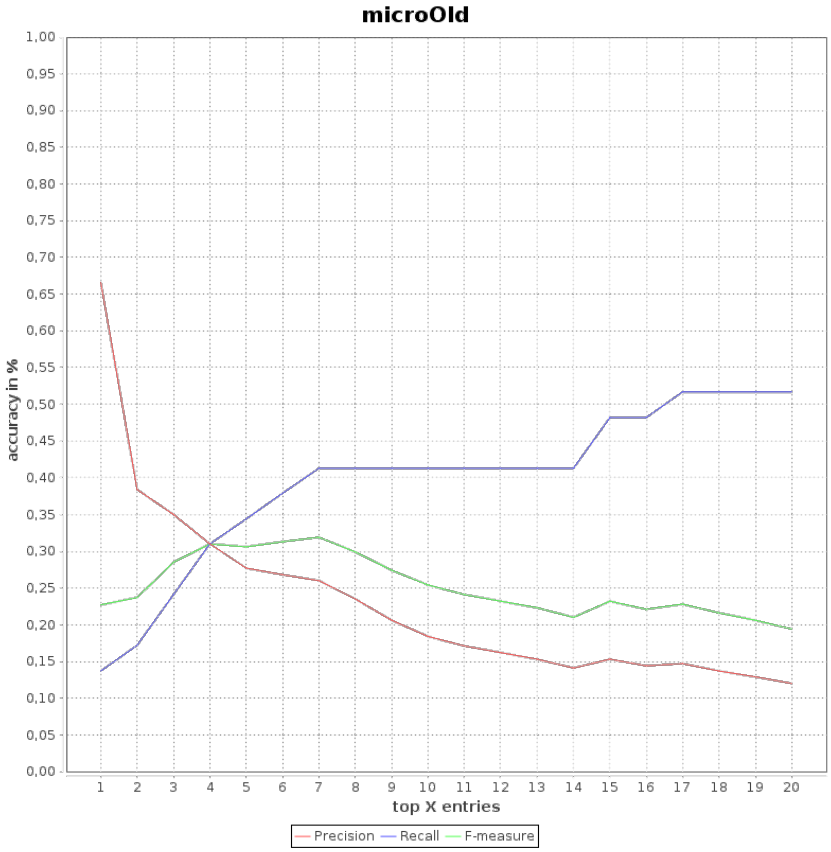


Figure 4: Precision, recall and f-measure for the lexicons generated by M-ATOLL for the old example properties; entries are filtered based on where in a list sorted according to the number of lemma occurrences in which they occur

ト (*vocalist*) as a verbalization of `bandMember`, while in other cases they are considerably more general, such as 一つ (*one [of several]*) as a verbalization of `languageFamily` or `locationCity`. Whether or not one would consider such verbalizations appropriate for being included in the final lexicon decidedly depends upon the application area at hand: In case of natural language generation, when the system needs to know which verbalizations it can use in its output, verbalizations more specific than the property at hand may lead to erroneous output, such as *Don Quixote is a manga by Cervantes*, at least if no further information about the semantics of those lemmas is provided in their respective entry. In contrast, more general terms, such as 一つ (*one [of several]*) in スペイン語はロマンス諸語の一つです *Spanish is one of the Romance languages*, would work for this application area. Conversely, in case of natural language understanding, where the system needs to figure out if a given natural language input contains a reference to a given property, more specific verbalizations would be acceptable. For example, if the system received an input of the form *Don Quixote is a novel by Cervantes*, and no other properties apart from `author` are linked to that verbalization in the lexicon, it could be sure that the `author` property is expressed in that sentence. However, very general terms such as 一つ (*one [of several]*), which would tend to be linked to a larger number of different properties, may lead to the system choosing the wrong property.

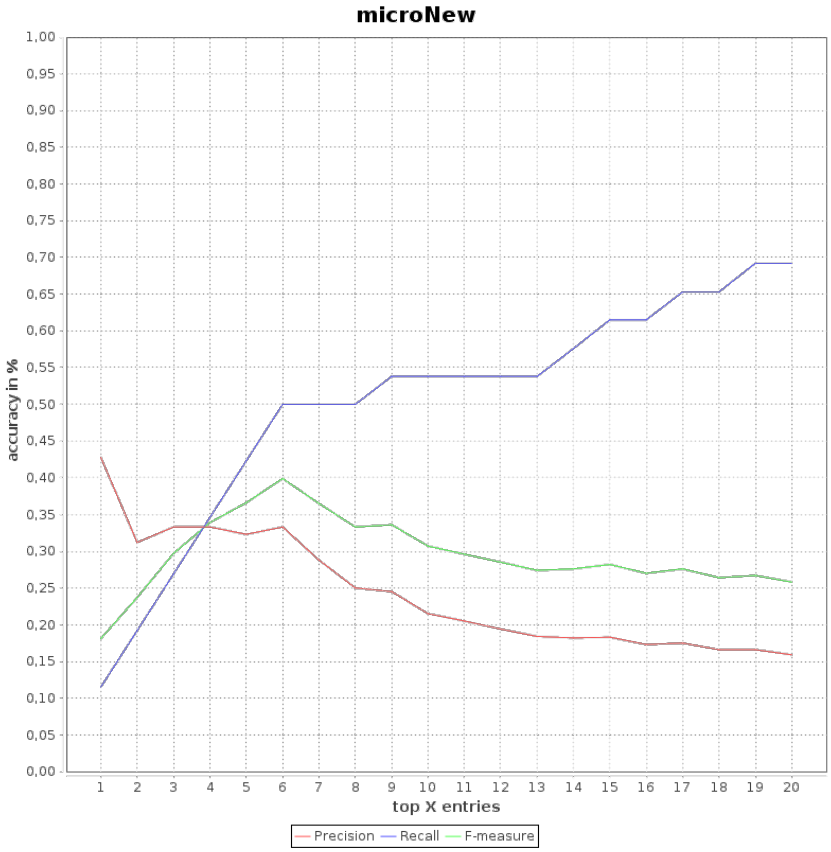


Figure 5: Precision, recall and f-measure for the lexicons generated by M-ATOLL for five new example properties; entries are filtered based on where in a list sorted according to the number of lemma occurrences in which they occur

As Figure 6 shows, if the additional verbalizations from table 6 are added to the gold lexicon, precision for the lexicon based on the new properties significantly increases for both filtering strategies. Which number of lemma occurrences or number of entries one should use as one's threshold, i.e. whether one should prefer higher precision or higher recall, again depends on the application area at hand: In case of natural language understanding one would want access to as many different potential verbalizations as possible, hence recall would be more relevant, while in case of natural language understanding one would want to make sure that no incorrect verbalizations are used in the output, which would make precision more important.

Alternative filtering mechanisms, such as those discussed in Walter (2017), may help to further improve precision.

5. Conclusion

In this paper we explored how M-ATOLL can be used to generate ontology lexicons for Non-Indo-European languages. For this purpose we used M-ATOLL to generate a Japanese ontology lexicon for DBpedia. The three main aspects that required manual work were the adaptation of the output format of the Japanese dependency parser to the input format expected by M-ATOLL, the generation of the language-specific dependency

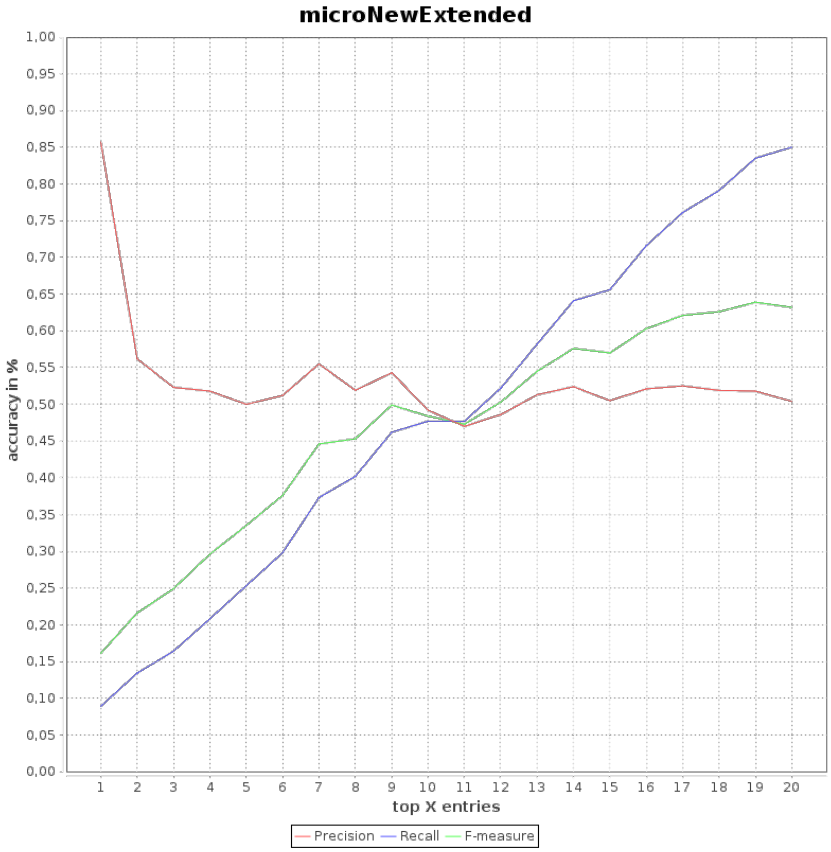


Figure 6: Precision, recall and f-measure for lexicon generated by M-ATOLL for new properties, with additional lemmas from Table 6 in gold lexicon

patterns required by M-ATOLL's corpus-based approach, and the specification of new lexicon entry templates. We showed how the most laborious of these three tasks, the generation of dependency patterns, can be partly automatized in order to reduce the temporal effort. We could show that M-ATOLL is a viable approach to the generation of ontology lexicons also for Non-Indo-European languages. Furthermore, due to it not being reliant on some specific grammatical framework or inventory of linguistic categories, lemon, the format lexicons generated by M-ATOLL are specified in, turned out to be very suitable for being used with Japanese, as can be seen e.g. from the ease with which we could generate new lexicon entry templates.

6. Acknowledgements

This work was supported by the Cluster of Excellence Cognitive Interaction Technology CITEC (EXC 277) at Bielefeld University, which is funded by the German Research Foundation (DFG).

7. References

Arcan, M. & Buitelaar, P. (2013). Ontology Label Translation. In L. Vanderwende, H.D. III & K. Kirchhoff (eds.) *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings,*

June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA. The Association for Computational Linguistics, pp. 40–46. URL http://aclweb.org/anthology/N/N13/N13-2006.pdf.

Cimiano, P., Buitelaar, P., McCrae, J. & Sintek, M. (2011). LexInfo: A Declarative Model for the Lexicon-Ontology Interface. *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(1). URL http://www.websemanticsjournal.org/index.php/ps/article/view/182.

Heringa, H. (2012). *Appositional Constructions*. Ph.D. thesis, Rijksuniversiteit Groningen.

Lanser, B., Unger, C. & Cimiano, P. (2016). Crowdsourcing Ontology Lexicons. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk & S. Piperidis (eds.) *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*. European Language Resources Association (ELRA), pp. 3477–3484. URL http://www.lrec-conf.org/proceedings/lrec2016/summaries/217.html.

McCrae, J., Espinoza, M., Montiel-Ponsoda, E., Aguado-de Cea, G. & Cimiano, P. (2011a). Combining Statistical and Semantic Approaches to the Translation of Ontologies and Taxonomies. In *Proceedings of the Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation*, SSST-5. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 116–125. URL http://dl.acm.org/citation.cfm?id=2024261.2024274.

McCrae, J., Spohr, D. & Cimiano, P. (2011b). Linking Lexical Resources and Ontologies on the Semantic Web with lemon. In *Proceedings of the 8th Extended Semantic Web Conference (ESWC)*. pp. 245–259.

Miller, G.A. (1995). WordNet: A Lexical Database for English. *Commun. ACM*, 38(11), pp. 39–41. URL http://doi.acm.org/10.1145/219717.219748.

Prévot, C.R.H.L., Calzolari, N., Gangemi, A., Lenci, A. & Oltramari, A. (2010). *Ontology and the lexicon: a multi-disciplinary perspective (introduction)*. Studies in Natural Language Processing. Cambridge University Press, pp. 3–28.

Walter, S. (2017). *Generation of multilingual ontology lexica with M-ATOLL : a corpus-based approach for the induction of ontology lexica*. Ph.D. thesis, Universität Bielefeld.

property	verbalization	# instances of lemma between/after triple subject and object	# sent.s found through SPARQL queries		% coverage	
			before analysis	after analysis	before	after
crosses	跨ぐ (to step over, to bridge)	5	2	2	40	40
	架かる (to span, to cross)	91	12	47	13.19	51.65
	かかる (writing variant of 架かる)	17	2	4	11.76	23.53
	またがる (to extend over)	1	1	1	100	100
	渡る (to cross over)	20	2	3	10	15
nationality	国籍 (nationality)	9	2	2	22.22	22.22
	出身 (person's origin)	283	38	119	13.43	42.05
	生まれ (birthplace)	33	3	6	9.09	18.18
occupation	仕事 (work)	58	0	0	0	0
	職業 (occupation)	10	0	0	0	0
	職 (job, position)	16	0	0	0	0
	生業 (job)	0	0	0	-	-
	勤める to work (for)	9	1	1	11.11	11.11
	務める to serve (as)	342	16	40	4.68	11.70
	活動 (activity)	311	5	6	1.61	1.93
	働く (to work)	9	0	0	0	0
yearOfConstruction	完成 (completion)	11	4	5	36.36	45.45
	竣工 (completion of construction)	2	0	1	0	50
	建設 (construction)	16	3	3	18.75	18.75
	建てる (to build)	5	3	3	60	60
parent	子供 (child)	104	2	36	1.92	34.62
	子 (child [of someone])	948	31	102	3.27	10.76
	父親 (father [of someone])	54	3	8	5.56	14.81
	父 (father)	651	69	106	10.60	16.28
	娘 (daughter)	928	65	102	7.00	10.99
	息子 (son)	1457	179	337	12.29	23.13
	親 (parent)	9	0	0	0	0
	母 (mother)	446	17	24	3.81	5.38
	母親 (mother [of someone])	49	1	2	2.04	4.08

Table 5: Number of instances of gold lemmas found between or after triple subjects and objects, and number of instances that are actually found by our SPARQL queries, after new patterns found through analysis of minimal dependency paths have been added

property	*original lemmas*	*additional lemmas found in top 20 entries of lexicon generated by M-ATOLL*
`author`	著者 (writer)	著す (to write [a book])
	書く (to write)	漫画 (manga)
	作家 (novelist)	小説 (novel)
	著作家 (author)	執筆 (writing [as a profession])
	作品 (work, opus)	原作者 (original author)
	作 (work [of art])	
	作者 (author)	
`bandMember`	バンドメンバー (band member)	ギタリスト (guitarist)
	メンバー (member)	ボーカリスト (vocalist)
	所属 (to belong to [used for humans])	ボーカル (abbrev. of vocalist)
		リーダー (leader)
		結成 (to form [a group of people, e.g. band, team])
		ヴォーカル (altern. writing form of abbrev. of vocalist)
		音楽ユニット ("music unit"; certain type of J-Pop band)
		ベーシスト (bassist)
		ユニット ("unit")
		音楽ユニットギタリスト (music unit guitarist)
		ロックバンド (rock band)
`foundingYear`	設立 (founding)	組織 (organization, construction)
	創立 (establishment)	発足 (start)
	創設 (founding)	建国 (founding of a nation)
	創始 (creation)	創業 (establishment [of a business])
	成立 (coming into existence)	独立 (becoming independent)
		始まる (to start)
`languageFamily`	属す (to belong to)	一種 (one kind, variety)
	言語 (language)	一つ (one [of several])
	含む (to include)	分類 (classification)
		ひとつ (*writing variant of* one)
		種 (kind, variety)
`locationCity`	所在する (to be located)	置く (to put, to place)
	都市 (city)	本社 (head office)
	場所 (location)	存在する (to exist)
		設立 (founding)
		会社 (company, corporation)
		本拠地 (headquarters)
		行う (to perform, to take place)
		創業 (establishment [of a business])
		構える (to set up)
		一つ (one [of several])

Table 6: Entries of gold lexicon for new properties

The Orkney Dictionary: Creating an Online Dictionary Efficiently from a Printed Book

Thomas Widmann, Phyllis Buchanan

Complexli Limited, 27 Kinloch Road, Newton Mearns, Glasgow G77 6LY, Scotland
E-mail: thomas@complexli.com, phyllis@complexli.com

Abstract

A great number of older dictionaries were compiled before the world of lexicography moved into the digital era. The result is that many older texts exist only in book format even though they contain a wealth of information that could still be extremely relevant today. A great deal of work went into these historical texts and some smaller languages and dialects are represented only in this format. Losing this information simply because the cost involved in digitising such resources is prohibitive would represent a wasted opportunity. In this paper we will demonstrate an efficient and cost-effective solution for converting these paper products into online resources. We will base the paper on the conversion of The Orkney Dictionary which we undertook in 2016.
In our approach, the book goes through the following phases: we began with the paper book, moved onto visual markup (HTML), this was converted to a simple tagging structure which formed the basis for the XML and then HTML. Finally the text was put into WordPress. Despite the numerous steps involved, many of them are standard components that can be reused, which is why it constitutes an efficient, low-cost way forward for retrodigitisation.

Keywords: XML conversion; online dictionaries; retrodigitisation

1. Introduction

A lot of dictionaries were compiled before the advent of digital lexicography, and today all that exists is a printed book. Many of these dictionaries are still of interest, for instance because they describe small or historical languages or dialects for which no digital resources exist.

However, the money available for digitising such books is often scarce, and thus many of these dictionaries never get the chance to move into the digital world.

In this paper, we shall demonstrate an efficient approach for converting a paper dictionary into an online product, based on the conversion of the Orkney Dictionary from a book (Flaws & Lamb, 1996) to a website (www.orkneydictionary.scot) which we undertook in 2016. In our approach, the book goes through the following phases:

1. Paper book
2. Visual markup (HTML)
3. Simple tagging
4. Fully nested XML tagging
5. HTML
6. WordPress

Although this process seems to contain many steps, many of them are standard components that can be reused, which is why this is an efficient, low-cost way to digitise an old dictionary.

2. The project

We were contacted in early 2016 by Simon W. Hall, who at the time was Education Scotland's Scots Language Coordinator in Orkney, with a view to turning Margaret Flaws and Gregor Lamb's seminal *Orkney Dictionary* from 1996 into an online dictionary.

This dictionary describes the Orkney (or Orcadian) dialect of Scots (the language derived from Old English that is spoken in the Scottish Lowlands and the Northern Isles, as well as parts of Ulster). The Orkney dialect is still widely spoken, for instance in local radio.

The project was funded by the Orkney Heritage Society, and the budget was quite limited compared to similar projects that we have undertaken for commercial dictionary publishers.

The Orkney Dictionary is not a large dictionary by any means, containing just over 2000 headwords on the Orkney-English side, and just shy of 1500 headwords on the other, as well as a few chapters describing the grammar and general orthographical principles.[1]

That said, it is the main dictionary describing the Orkney dialect, and copies of it are found in homes and schools everywhere on the islands. Making it available online for free was therefore of immense benefit to many people in Orkney.

We could have converted the dictionary to HTML directly from PDF without the detour via XML, but that would have made it harder to implement a decent search interface and live cross-references. We also thought that it would be a good opportunity to create a modern look, making good use of whitespace and colour. We therefore decided to undertake a proper conversion from PDF to XML instead, although this was going to make it challenging to stay within the budget.

The way we squared the circle was by reusing different components that we had created over a number of years.

3. The steps

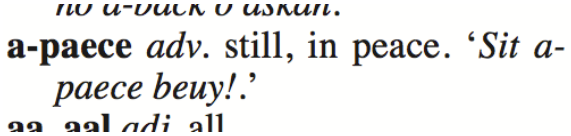
In the following we will look at the individual steps making up the conversion process.

3.1 Paper book to HTML

This first step was to convert the PDF file to a simple HTML file containing only visual mark-up.

In our case we were lucky enough to receive a PDF of the published book, but the process would have been similar if we had worked from a paper book, in which case we would have had to get it scanned or double-keyed instead.

A typical entry in *The Orkney Dictionary* looks like this:



It should be clear from a quick glance that this is not an straightforward format to identify the structure from. For instance, italics are used for both part-of-speech labels and examples, and full stops are everywhere.

[1] These chapters were converted separately to WordPress pages, but this is not of any interest here, given the lack of lexicographic content.

In this case, we used an on-line service to convert the PDF to Word format, and then an OpenOffice plug-in to create XHTML. This was based on trial and error, and different PDF files might have been easier to convert using different tools.

We now had a HTML file containing entries such as this:

```
<p class="Textbody">
  <span style="font-family:Times;font-weight:bold;font-size:14.666667px"
    >a-paece </span>
  <span style="font-family:Times;font-style:italic;font-size:14.666667px"
    >adv. </span>
  <span style="font-family:Times;font-size:14.666667px"
    >still, in peace. ‘</span>
  <span style="font-family:Times;font-style:italic;font-size:14.666667px"
    >Sit a- paece beuy!</span>
  <span style="font-family:Times;font-size:14.666667px">.’ </span>
</p>
```

This is a rather neat example. Many entries were interrupted by `<p>` tags, and in a few cases the text was not even sequential, but skipped back and forwards between the two columns. In general it was a decent file to base the next step on, though.

3.2 HTML to simple tagging

We now needed to convert the HTML to our own simple tagging format. This is a way to identify logical elements such as headwords, translations and examples without worrying about creating any explicit structure yet. Of course there might be tags indicating the beginning of grammatical categories or of new senses, but nothing is nested at this stage.

This mark-up is optimised for manual editing in Emacs so that a lexicographer is able to correct any conversion errors efficiently. We also developed a few Emacs macros to make it easy to undertake common editing operations, such as splitting up a translation containing a comma, or upgrading a phrase (together with its translation and other associated information) to a full entry.

Apart from inline tags (such as `<b>...</b>`, which are left as-is, every tag starts on a new line, and a TROFF-like notation (i.e., `.tag` without an end tag) is used instead of XML syntax (`<tag>...</tag>`) to minimise typing.

In this case, the program would convert the above example to this:

```
.hw a-paece
.ps adv
.tr still, in peace
.qu Sit a-paece beuy!
```

The advantage of this tagging system is that any conversion errors will jump out immediately. For instance, if the beginning quotation mark after “peace” had not been correctly identified, we might have ended up with something like this:

```
.hw a-paece
```

```
.ps adv
.tr still, in peace. ‘
.ph Sit a-paece beuy!.’
```

It is very easy to see that something is wrong here – much easier than spotting an error in HTML, and much easier to fix than doing it in XML (where a change to the structure might be necessary).

In this case, the conversion program did a very good job – the only correction needed was to split up the two translations that had been separated by a comma, resulting in this:

```
.hw a-paece
.ps adv
.tr still
.tr in peace
.qu Sit a-paece beuy!
```

Another common issue was that many headwords had become phrases (especially when they were derivations), e.g.:

```
.hw blether
.ps v
.tr talk nonsense
.ps n
.tr chatterbox
.ph bletherskate
.ps n
.tr someone who talks nonsense
```

All we needed to do in order to change the structure here was to replace `.ph` with `.hw`:

```
.hw blether
.ps v
.tr talk nonsense
.ps n
.tr chatterbox

.hw bletherskate
.ps n
.tr someone who talks nonsense
```

The equivalent change in XML would have required much more typing (or complex macros).

To convert the HTML file to this simple tagging format, we wrote a Perl program making heavy use of regular expressions. It would first replace the `<span style="...">` tags with more intuitive tags (such as `<b>` and `<i>`), and then replace them with our simple tagging tags based on the context.

For instance, a bold chunk of text at the beginning of a paragraph would become a headword (`.hw`), anything following this (separated by a comma) would become an alternative

form of the headword (`.ha`), and any other chunk of bold text would become a phrase (`.ph`). Any phrase enclosed in quotation marks would finally be turned into a quotation (`.qu`).

The program got many things right, and when it made an error, it was often very obvious and easy to fix, as described above.

3.3 Simple tagging to rich XML

At this point, we needed to design a DTD describing the resulting XML.

We considered using a standard TEI Dictionary structure (Text Encoding Initiative, 2016), but we thought that it was too verbose in places, made unnecessary distinctions for our purposes and yet conflated distinctions made in our source, so we created a custom DTD that mimicked the implicit structure adopted by the authors of *The Orkney Dictionary*.

This decision was aided by the fact that it seemed unlikely the data from the project would be reused elsewhere, given the low number of Orkney dialect speakers. If data sharing becomes important at a later date, it should be eminently possible to convert the data to a standard TEI structure.

It is important to bear in mind at this point that this conversion was being done on a very small project, so a decision had to be made quickly and pragmatically. In an ideal world, we would have spent some time exploring the XML structures used by similar projects and liaising with other experts in the field, but that would have left us with practically no time to undertake the actual conversion.

The relevant subset of our DTD relating to the entry we have been examining above looks as follows:

```
<!ELEMENT entry (hw, gram+)>
<!ELEMENT hw (#PCDATA)>
<!ELEMENT gram (pos, sense+)>
<!ELEMENT pos (#PCDATA)>
<!ELEMENT sense (tran+, quotes*)>
<!ELEMENT tran (#PCDATA)>
<!ELEMENT quotes (quote)>
<!ELEMENT quote (#PCDATA)>
```

In order to convert our simple tagging format to XML conforming to this DTD, we used our own conversion program (written in Perl, C, Flex and Bison) to convert it to highly structured XML (see Section 4 for more details on this program).

This program reads a description of the resulting XML file that is similar to the DTD, so if this has been done correctly, it should in theory always produce valid XML (apart from attribute values and inline tags and a few other things that are external to the program[2]).

[2] It would be relatively simple to extend the program with functionality to check that inline tags only get used in the correct locations, and that attribute values always are taken from a closed set, but we have found it is just as easy simply to validate the resulting files against the DTD afterwards.

It is completely reusable, and we have used it for many XML conversion projects over the years.

The relevant subset of the grammar that the conversion program reads looks as follows:

```
entry;hw gram+
hw;.hw
gram;pos sense+
pos;.ps
sense;tran+ quotes*
tran;.tr
quotes;quote
quote;.qu
```

It should be clear that this corresponds very closely to the DTD. Apart from the syntax, the biggest difference is that the `#PCDATA` bits have been replaced with the relevant tags used in our simple tagging system.

If the simple tagging input does not conform to this structure, the conversion program will produce an error message when it encounters the first offending line. For instance, if the `.ps` had been omitted, it would complain when it saw the `.tr` tag. Because of this, the conversion from our simple tagging format to XML is very safe.[3] However, running it can be quite an iterative process, requiring the lexicographer to correct the simple tags in the text when it cannot be converted.

As an example, imagine that the following entry were encountered:

```
.hw a-paece
.tr still
.tr in peace
.qu Sit a-paece beuy!
```

One would have three alternatives here: (1) To insert the missing `.ps` tag; (2) to amend the `gram` rule from `gram;pos sense+` to `gram;pos? sense+`; or (3) to convert "missing" `.ps` tags to a `<pos>` with a specific value that can then be corrected later. In this case, option (1) would clearly be best, but there are other cases where the other options might be preferable, for instance if the client wants to make any corrections themselves after the conversion has been completed.

In this case, our entry was transformed to the following bit of XML:

```
<entry>
  <hw>a-paece</hw>
  <gram>
    <pos>adverb</pos>
    <sense>
      <tran>still</tran>
      <tran>in peace</tran>
```

[3] It will always create XML that validates against the DTD if the grammar rules have been written correctly. However, there is no guarantee that the XML tags will have been used in a semantically correct fashion, for instance if the simple tags were wrong to start with.

```
        <quotes>
          <quote>Sit a-paece beuy!</quote>
        </quotes>
      </sense>
    </gram>
  </entry>
```

At the end of this stage, the entire dictionary had been converted to XML, which could be edited using any XML editor or a proper dictionary editing system such as IDM's DPS. In the case of The Orkney Dictionary, we preferred to implement all corrections in the simple tagging format and then reconvert it, though, simply because our simple tagging system is easier to work with than the resulting XML structure.

3.4 XML to HTML

We now needed to convert the XML to HTML. For this purpose, we wrote a simple XSLT program.[4] The resulting HTML consisted mainly of `<div>`s and `<span>`s:

```
<div class="entry">
  <span class="hw">a-paece </span>
  <span class="pos">adverb </span>
  <span class="tran">still</span>
  <span class="punct"> &#8226; </span>
  <span class="tran">in peace</span>
  <span class="punct">  &#9758; </span>
  <span class="quote">‘‘Sit a-paece beuy!’’</span>
</div>
```

We did not have a specific design brief, but basically tried to find a modern dictionary design that the client would be happy with. However, given the simplicity of the XML structure used here, we do not find it likely that any design proposals would have been too hard to implement had the client so desired.

We also developed a previewer based on the formatting program. This was optimised to highlight any conversion errors by ensuring that all tags would output in a distinctive fashion. This was not designed to be pretty, but it was a great way to find the last remaining conversion errors.

We also created a CSS file to display the HTML, and the result was the same that can be seen on the website today.

3.5 HTML to WordPress

The resulting HTML was then stored in a few MySQL tables and uploaded to a standard WordPress installation with some added search functionality. We also quickly converted the chapters describing the grammar and orthography of the Orkney dialect of Scots to HTML and made them available as WordPress pages.

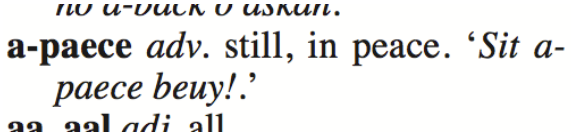
[4] There is probably no point in describing the XSLT program in any detail, given that it exhibits no features of great interest.

The WordPress theme is a child theme[5] of the *Twenty Sixteen* theme (WordPress.org, 2016), which implements the search functionality (including fuzzy matching) in PHP and incorporates the CSS code necessary to make the entries display correctly. Most of it could be reused with very few changes to create other online dictionaries.

At this point, the entry we have been looking at above now looks like this:

a-paece *adverb* still • in peace ☞ *"Sit a-paece beuy!"*

This is quite a difference from our starting point:



The XML can of course also be used for other purposes. It would for instance be relatively easy to create ICML from it in order to typeset the dictionary in InDesign (which is something we have done for another client), and the HTML could also be used to create a smartphone app. The cost of developing a dictionary app for The Orkney Dictionary would probably be prohibitive, but there is no reason why the costs could not be shared by a number of similar projects, and this is something we are currently looking into.

4. Our conversion program

In this section we shall describe our own conversion program (written in Perl, plain C, Flex and Bison) that we used to convert our simple tagging format to fully nested XML.

This program is extremely flexible and allows the conversion of many types of non-nested data to different XML structures. We have only used it for dictionary conversions, but there is no reason why it could not be used for other purposes as well.

It consists of two parts: A parser written in C, Flex and Bison, and a grammar preprocessor written in Perl that transforms the grammar (corresponding to the DTD) into Bison code.

Flex and Bison (open-source alternatives to Lex and Yacc) are standard programming tools used for writing parsers, e.g., for programming languages. Flex tokenises the input, and Bison takes these tokens and uses them to build a syntax tree. (Bison can only write parsers for context-free grammars.[6]) Variants for these tools exist for several different programming languages, but we used the one that produces C code, for the simple reason

[5] A WordPress theme is a collection of files that work together to produce a graphical interface with an underlying unifying design for a blog. A child theme is a theme that inherits the functionality and styling of another theme, called the parent theme. Child themes are the recommended way of modifying an existing theme.

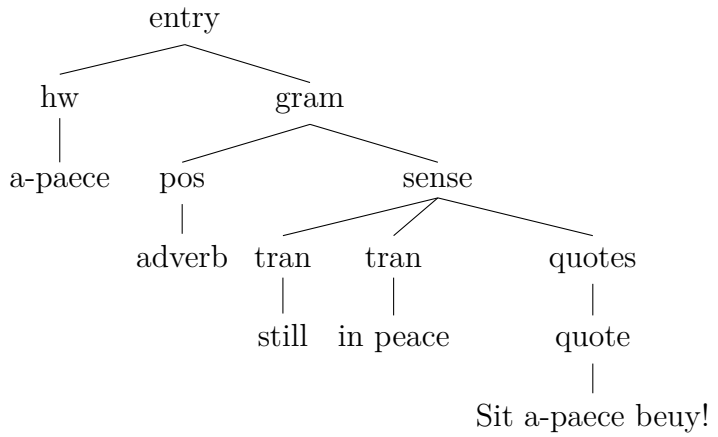
[6] A context-free grammar is a set of recursive rewriting rules that all have a single non-terminal on their left hand side.

that this was the one we were already familiar with.[7] The fact that the parser is written in plain C means that it is lightning fast. If no preprocessors or postprocessors are used, our parser can convert a large dictionary to XML in just a few seconds.

To understand how these tools can be used for producing XML from a flat structure, we need to realise that an arbitrary chunk of XML can be visualised as a tree. For instance, let us have a look at this entry:

```
<entry>
  <hw>a-paece</hw>
  <gram>
    <pos>adverb</pos>
    <sense>
      <tran>still</tran>
      <tran>in peace</tran>
      <quotes>
        <quote>Sit a-paece beuy!</quote>
      </quotes>
    </sense>
  </gram>
</entry>
```

The equivalent tree notation would look like this:



Bison is great at building trees like this. The syntax looks like this (`Dhw` and `CONTENTS` are two of the tokens produced by Flex, corresponding to `.hw` and the following text):

```
entry:
  hw gram_plus      { /* C code to add this to the tree */ }
| hw pron gram_plus { /* C code to add this to the tree */ }
/* ... */
;

hw:
  Dhw CONTENTS      { /* C code to add this to the tree */ }
```

[7] In fact, when we wrote the parser, we had to dust off our old copy of Aho et al. (1986), which was a very nostalgic experience.

```
;

gram_plus:
  gram              { /* C code to add this to the tree */ }
| gram_plus gram    { /* C code to add this to the tree */ }
;

/* many more rules here */
```

Bison will then write code that will select the correct rule based on the input.

It would be quite possible to write these Bison rules manually, but it would get rather tedious. Because of this, we have written a Perl program that makes it possible to specify the syntax in a much more compact way that almost mimics a DTD.

As a very simple example, let us imagine the input format only contains the tags `.hw` and `.tr`:

```
.hw a-paece
.tr still
```

Let us assume that this minimal entry should end up looking like this:

```
<entry>
  <hw>a-paece</hw>
  <tr>still</tr>
</entry>
```

This can be achieved with the following grammar rules:

```
entry;hw tr
hw;.hw
tr;.tr
```

If we want to allow sequences of headwords and translations, we can achieve this by adding `+` after the relevant tag, just like one would do in a DTD:

```
entry;hw+ tr+
hw;.hw
tr;.tr
```

In the same way, we can use `?` and `*` (again with the same semantics as in DTDs), e.g.:

```
entry;hw+ pos? tr+ quotation*
hw;.hw
tr;.tr
pos;.ps
quotation;.qu
```

It is also possible to add extra grouping tags. For instance, if every `.ps` tag starts a new grammatical category, we could write it like this:

```
entry;hw+ gram+
```

```
gram;pos tr* quotation*
hw;.hw
tr;.tr
pos;.ps
quotation;.qu
```

Although brackets cannot be used (yet), they can be emulated by using non-outputting grouping tags (starting with an underscore). For instance, if a `<gram>` can consist of a `<pos>` and a sequence of either `<tr>`s or `<def>`s, it could be expressed as follows:

```
gram;pos _trs_or_defs
_trs_or_defs
 ;tr+
 ;def+
```

We might add bracket notation in a future version of the parser to make such expressions follow the DTD syntax more closely.

Another feature is what we call named rules: These contain an underscore after the first word, such as `sense_first`. The tag they generate contains only the first word (`<sense>` in this case), but it increases readability and makes it possible to have many rules generating the same tag.

The tree that is built can be highly nested. For instance, consider the following grammar:

```
tag0;tag1
tag1;tag2
tag2;tag3
tag3;tag4
tag4;.tag
```

Based on this, the parser would turn the single line `.tag Hello world` into the following chunk of XML:

```
<tag0>
  <tag1>
    <tag2>
      <tag3>
        <tag4>Hello world!</tag4>
      </tag3>
    </tag2>
  </tag1>
</tag0>
```

The only real shortcoming of our parser is that it cannot look ahead in order to choose between two options – it knows the preceding lines and the current one, but it has no idea about what it will encounter later. For instance, imagine a situation where `.tr` tags have to be incapsulated in `<sense>` tags if and only if any `.tr` tag within the same `<gram>` structure is preceded by an `<lb>` tag. It would be logical to write rules such as these:

```
gram
```

```
 ;pos tr+
 ;pos _senses
_senses;sense_first? sense_extra+
sense_first;tr
sense_extra;lb+ tr
```

However, this would not work on the following structure:

```
.hw a-paece
.ps adverb
.tr still
.lb lit
.tr in peace
```

The parser would enter the first `gram` rule and then get stuck when it found the first label.

To deal with such situations, it is necessary to write preprocessors (we use Perl) to make the structure easier to parse by inserting extra tags. To resolve the conflict described above, the preprocessor might insert `.sense` tags like this:

```
.hw a-paece
.ps adverb
.sense
.tr still
.sense
.lb lit
.tr in peace
```

In this way, it is possible to generate XML for even highly complex and ambiguous structures.

In theory, it should also be possible to use Bison's *GLR* mode (this is an extension to handle nondeterministic and ambiguous grammars) instead of writing these preprocessors. GLR parsers handle Bison grammars that contain no unresolved conflicts in the same way as deterministic parsers. However, when there are unresolved shift/reduce and reduce/reduce conflicts, GLR parsers use the simple expedient of doing both, effectively cloning the parser to follow both possibilities. Each of the resulting parsers can again split, so that at any given time, there can be any number of possible parses being explored. Each of the cloned parsers eventually meets one of two possible fates: either it runs into a parsing error, in which case it simply vanishes, or it merges with another parser, because the two of them have reduced the input to an identical set of symbols (Free Software Foundation, 2015).

We have not explored this, but it would be worthwhile looking into it in the future.

We believe that our context-free parser is a great tool for converting a flat structure to XML. It is sadly not possible to make it freely available at the moment, as it would require a great deal of work to extend and document it before this could happen. It would not be particularly hard for other programmers familiar with Lex/Flex and Yacc/Bison to write something similar.

5. Discussion

It would have been much easier to take the HTML and put it on the web directly. However, going through the steps we did conferred many advantages:

1. It allowed us to present the dictionary in a modern way rather than being tied down to the old format.
2. Converting the data enabled us to catch many typos and inconsistencies that had been overlooked in the book.
3. Having proper XML made it easy to implement the search functionality (because it was clear what needed to be indexed), and it also made it simple to create live cross-references on the website.
4. If the authors decide to make any additions or alterations to the data in the future, they can do so using a modern dictionary editing system instead of a word processor.
5. It becomes possible to typeset the book in InDesign – this is unlikely to happen soon, however, as the book has recently been reprinted.
6. It becomes possible to create a smartphone app. This is likely to be the next step in this project.

It was only possible to deliver this project on budget because we already had our conversion program, which we had developed for other XML conversion projects. Most of the time was spent on converting the HTML into our simple tagging format, and on correcting remaining errors manually in this format.

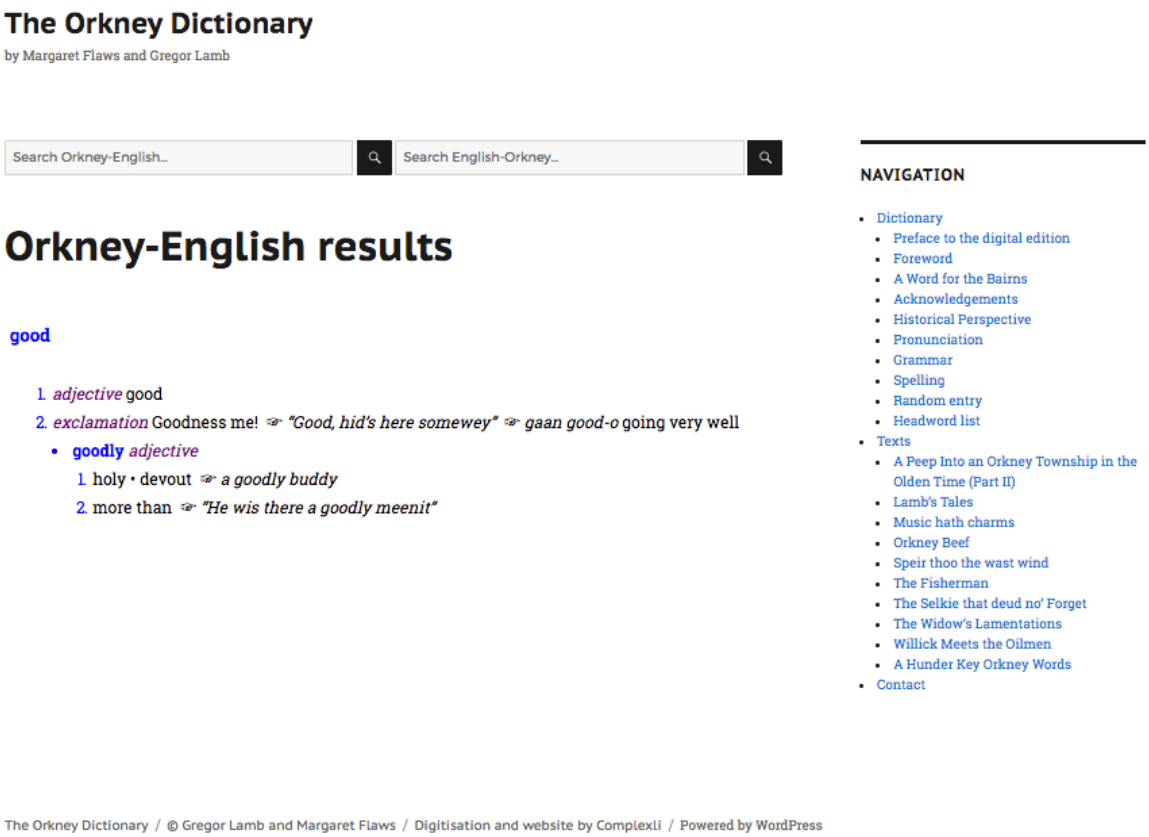


Figure 1: A screenshot from orkneydictionary.scot showing a sample entry

The online version of the dictionary (see the screenshot in Figure 1) has been very well received in Orkney. The past year has seen slightly fewer than 4,000 unique visitors, which might seem like a low figure, but it should be remembered that Orkney's total population is not much more than 20,000.

6. Conclusion

We have demonstrated that it is possible to digitise a paper dictionary and to create a website for it on a small budget without sacrificing quality.

We believe this is good way to convert legacy dictionaries into XML and onto the web. The key is to use standard components that can be reused in other projects, and to have simple data formats that are easy to edit using free tools.

7. References

Aho, A.V., Sethi, R. & Ullman, J.D. (1986). *Compilers: Principles, Techniques, and Tools.* Addison-Wesley.

Flaws, M. & Lamb, G. (1996). *The Orkney Dictionary.* Orkney Language and Culture Group.

Free Software Foundation (2015). GNU Bison – The Yacc-compatible Parser Generator. http://www.gnu.org/software/bison/manual/.

Text Encoding Initiative (2016). P5: Guidelines for Electronic Text Encoding and Interchange: Dictionaries. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/DI.html.

WordPress.org (2016). Twenty Sixteen. https://en-gb.wordpress.org/themes/twentysixteen/.

Good Examples for Terminology Databases in Translation Industry

Andraž Repar[1,2], Senja Pollak[3]

[1]Jozef Stefan International Postgraduate School, Jamova 39, 1000 Ljubljana, Slovenia
[2]Iolar d.o.o, Parmova 51, 1000 Ljubljana, Slovenia
[3]Jozef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
E-mail: repar.andraz@gmail.com, senja.pollak@ijs.si

Abstract

This paper deals with finding good examples for terminology database entries in the translation industry. When extracting terms from bilingual translation memory exchange files, it is very easy to also extract example sentences to showcase the use of the term in practice. However, there are usually a lot of sentences containing the term and selecting an appropriate example is not a straightforward task. In this paper, we explore the use of data mining techniques to find good term examples. After constructing the corpus from a large English-Slovenian bilingual file from a financial domain, we extract linguistic features and load them into the Weka data mining environment to analyze the performance of various classifiers, resulting in 0.8 precision for positive class (good examples) and 0.85 overall accuracy. While the model was tested only on one language combination, the nature of most features is language-independent which suggests that the model could be used successfully for other language combinations.

Keywords: terminology; good example; data mining; classification

1. Introduction

When building bilingual terminology databases with automatic term extraction from translation memory exchange files (TMX[1]), it often makes sense to include an example sentence[2] to see how the term in question behaves in context. But adding just any random sentence is hardly a good strategy – it is imperative that the sentence be as illustrative as possible. However, that is easier said than done. What at first appears to be a relatively straightforward task turns out to be anything but and a more systematic approach has to be taken. According to Kilgarriff et al. (2008), a good dictionary example must be:

- typical, exhibiting frequent and well-dispersed patterns of usage
- informative, helping to elucidate the definition
- intelligible to learners, avoiding gratuitously difficult lexis and structures, puzzling or distracting names, anaphoric references or other deictics which cannot be understood without access to the wider context.

Regarding the extraction of term examples, we can identify two lines of research. While on the one hand, definitions can be considered as optimal examples – with automated methods developed for several languages including English (Navigli & Velardi, 2010), Dutch (Westerhout, 2010), French (Malaisé et al., 2004), German (Storrer & Wellinghoff, 2006), Portuguese (Del Gaudio et al., 2014), and Slovene (Pollak et al., 2012) – other authors focus on extraction of good examples. Our work focuses on good examples, since in translation memories, the definitions are very rare, whereas including (good) examples is a feasible task.

[1] http://www.ttt.org/oscarStandards/tmx/tmx14-20020710.htm
[2] While it is certainly possible to include more than one example, additional examples have only marginal value. At the end of the day, these examples serve only as a supplement to the main part of terminology databases, such as terms and their definitions.

There are several existing related approaches available for good term example extraction. GDEX, a system described in Kilgarriff et al. (2008), lets the user define criteria for good dictionary examples and was designed to help lexicographers with identifying dictionary examples by ranking sentences according to how likely they are to be good candidates. It served as the basis for GDEX for Slovene (Kosem et al., 2011) whose approach was based on experience and arose from the assumption that experienced lexicographers can provide a useful set of heuristics based on their intuition and skills. In order to do so, they have come up with various criteria and then used them to filter out the unsuitable examples. Finally, Ljubešić & Peronja (2015) use a supervised learning approach to finding good dictionary examples in Croatian. They manually rank a set of monolingual example sentences into four categories and then use a regression algorithm. They obtain a precision of around 80 percent on the 10 top-ranked examples.

We take a similar approach but treat this issue as a binary classification problem. First, two domain experts annotated their own part of a large set of sentence pairs as either good or bad examples of the source/target sentence pair where only segments consisting of a sentence annotated as a good example for both languages are considered as positive examples (examples can be seen in Table 1). A set of linguistic features was then extracted to be used in the data mining phase. The features were extracted with Python scripts and the data were then loaded into Weka programming toolkit (Hall et al., 2009) to build a suitable classification model. The goal is to test the performance of various classifiers to try to find the most suitable one for good example selection.

Besides definitions and term examples, (semi-)automatic extraction of other types of knowledge-rich contexts (Meyer, 2001) is of great importance, especially for terminographic purposes. While one would normally look for only one, or at most a few, good term examples, researchers of knowledge-rich concepts are focusing on a larger subset of a corpus containing information that would be valuable to a human for the construction of a knowledge base (Barrière, 2004). Finding good term examples could thus be considered a sub-field of knowledge-rich context discovery.

This paper is structured as follows: Section 2 describes the data and the linguistic features, Section 3 describes the experimental setup, Section 4 describes the results and Section 5 contains the discussion of results, conclusion and plans for future work.

2. Data preparation

The examples in the dataset are from the domain of banking and finance. The data comes from a TMX file which is used by most translation applications to store completed translations – this means that the text is sentence aligned. It contains the source (English) and target (Slovenian) segments along with some metadata (date, translator name, project etc.). As a preliminary step, a monolingual terminology extraction process (adapted from Pollak et al. (2012)) was run on both sides of the TMX file and a subset of the extracted source and target terms was manually aligned.

Both sets of sentences – the source and target sets – were cleaned of various TMX tags, tokenized and POS-tagged (NLTK's Penn Treebank tokenizer and POS-tagger were used for English (Loper & Bird, 2002), whereas for Slovenian, the Penn Treebank tokenizer was again used for tokenization and the open-source Reldi tagger and lemmatizer was used for Slovenian (Ljubešić et al., 2016)). In addition, the Slovenian sentences were

English	Slovenian	
Allocation to (more) defensive stocks was the main detractor as high beta names rallied strongly amid the positive sentiments – though an overweight exposure to **equities** (versus bonds) has partially mitigated on the underperformance.	Razdelitev sredstev (bolj) obrambnim delnicam je najbolj zmanjšala donosnost, ker se je močno izboljšalo razpoloženje vlagateljev do imen z visokim koeficientom beta – čeprav je večja izpostavljenost **lastniškim vrednostnim papirjem** (v nasprotju z obveznicami) delno ublažila slabšo donosnost.	Bad
The resulting portfolio consisted essentially of financial stocks and **equities** from the energy, consumer goods and healthcare sectors.	Portfelj je vključeval predvsem finančne delnice ter **lastniške vrednostne papirje** energetskega, potrošniškega in zdravstvenega sektorja.	Good
d) In addition, deposits may be held and **money-market instruments** may be acquired; their value together with the value of the money-market funds held as defined in letter c), subject to the provisions of letter e), may total a maximum of 15 percent of Sub-Fund assets.	d) Poleg tega je dovoljeno imeti depozite in pridobiti **instrumente denarnega trga**. Njihova skupna vrednost skupaj z vrednostjo skladov denarnega trga v lasti, kot je določeno v točki c), lahko znaša največ 15 odstotkov sredstev podsklada v skladu z določili iz točke e).	Bad
If a Sub-Fund lends securities and **money-market instruments**, the borrower will normally either resell them quickly or has already done so.	Če podsklad posodi vrednostne papirje in **instrumente denarnega trga**, jih posojilojemalec hitro ponovno proda ali pa je to že naredil.	Good
As remuneration for administrative services rendered to the Company in its capacity as Management Company, BNP PAM Lux will receive a maximum annual fee of 0.15 percent calculated on the average of the **net asset values** of the assets of the various sub-funds of the Company for the period for which the fee is payable.	BNP PAM Lux prejme za administrativne storitve, ki jih v funkciji družbe za upravljanje opravlja za družbo, letno nadomestilo največ 0,15 odstotka, izračunano glede na povprečno **čisto vrednost sredstev** različnih podskladov družbe za obdobje, za obdobje, za katerega se plača nadomestilo.	Bad
Any subscription requests received before this closing time will be executed on the basis of the **net asset value** on the Valuation Day.	Zahtevki za vpis, prejeti v tem roku, bodo izvršeni na podlagi **čiste vrednosti sredstev** na obračunski dan.	Good

Table 1: Good/bad examples. The term in question is written in bold style

also lemmatized with the Reldi tagger and lemmatizer in order to facilitate searching for term positions in sentences. The sentences were transformed to the feature vector representation, where the target variable was a nominal variable with YES/NO classes corresponding to good term examples (positive class YES) and bad term examples (NO). For the manual annotation phase, 1,332 example bilingual sentence pairs for various terms were annotated (two professional translators each annotated one half of the examples). Because one sentence can contain multiple terms, individual sentences (sentence pairs) can be used multiple times for different terms. The dataset produced was somewhat imbalanced (962=NO, 370=YES).

The linguistic features extracted as attributes are listed in Table 2.

Altogether, there were five nominal (three of them had binary values, two had multiple nominal values) and 15 numeric attributes. While most of the features were designed to be language-independent, a few target language features were created with a specific characteristic of the target language in mind (e.g. target personal pronouns and target demonstrative pronouns are aimed at the propensity of the Slovenian language for using pronouns instead of repeating full words).

As mentioned above, the target variable was a nominal variable with YES/NO classes.

Short name	Description	Value
SLength	Source sentence length in characters	Numeric
TLength	Target sentence length in characters	Numeric
TLen by SLen	Target length divided by source length	Numeric
STermPos	Position of term in the source sentence	Numeric
TTermPos	Position of term in the target sentence	Numeric
SNoDig	Number of digits in the source sentence	Numeric
TNoDig	Number of digits in the target sentence	Numeric
SNoWeirdChar	Number of weird characters (brackets, asterisks, hyphens, dashes etc.) in the source sentence	Numeric
TNoWeirdChar	Number of weird characters (brackets, asterisks, hyphens, dashes etc.) in the target sentence	Numeric
TPPron	Number of personal pronouns in the target sentence	Numeric
TDPron	Number of demonstrative pronouns in the target sentence	Numeric
NoComma	Number of commas in the target sentence	Numeric
NoFullstop	Number of fullstops in the target sentence	Numeric
SNonInitCapWords	Number of capitalized words not in the initial position in the source sentence	Numeric
TNonInitCapWords	Number of capitalized words not in the initial position in the target sentence	Numeric
SCap_Punc	Checks whether the source sentence starts with a capitalized word and ends with a punctuation mark	Binary
TCap_Punc	Checks whether the target sentence starts with a capitalized word and ends with a punctuation mark	Binary
SPassV	Checks whether the source sentence contains a passive voice form	Binary
TargetCase	Checks the grammatical case of the target term	Nominal
InitWrdType	Checks the word type of the first word of the target sentence	Nominal

Table 2: Extracted features

3. Experimental setup

This section describes the selection of algorithms, feature transformation and feature selection, and presents the evaluation method.

3.1 Algorithms

We tested and compared the following algorithms implemented in the Weka data mining toolkit (Hall et al., 2009):

- Naïve Bayes is a simple probabilistic classifier.
- The J48 decision tree algorithm in Weka is an implementation of the C4.5 decision tree algorithm. It produces a pruned decision tree which offers good visualization of the data.
- The IBk classifier is Weka's implementation of the k-nearest neighbors approach to classification. Classification is performed on the basis of the majority class of k-nearest neighbors.
- JRip implements a propositional rule learner, Repeated Incremental Pruning to Produce Error Reduction (RIPPER). It generates a set of IF rules which provide an easily interpretable description of the data.
- SMO in Weka implements John Platt's sequential minimal optimization algorithm for training a support vector classifier.
- ZeroR is the majority class classifier used as a baseline.

3.2 Discretization

As the dataset contains a mix of numeric and nominal attributes, discretization could potentially prove beneficial. All classifiers (i.e. the implementation of the classifier in Weka) from Section 3.1 by default support numeric attributes, but Weka also offers a separate discretization functionality which was tested to see if it offers any improvements.

Supervised discretization was used. In order to avoid overfitting when using cross-validation (because supervised discretization takes into account class values), we have used the FilterClassifier meta classifier in Weka which allows you to specify a filter (i.e. supervised discretization) and apply it only on the training data leaving the test data untouched.

3.3 Feature selection

Kosem et al. (2011) discovered that some features are more significant than others. We wanted to test that using Weka's feature selection functionality. Specifically, we selected the AttributeSelectedClassifier in Weka which allows you to select the evaluator for feature selection before running the classifier itself. We chose the WrapperSubsetEval evaluator and selected the respective classifier to select the best possible features (e.g. for J48, first feature selection was performed with the J48 classifier, then a J48 model was built using the selected features).

3.4 Evaluation method

The performance of the classifiers was evaluated in the 10-fold cross-validation setting using the following basic measures: accuracy, precision, recall and F-score. Because we normally have several example sentences per term and we only really need one good example to be included in the termbase, the most important measure for our task is the precision of the positive class (i.e. true positive examples vs all classified positive examples).

$$precision = \frac{tp}{tp + fp} \tag{1}$$

4. Results

Since the dataset is imbalanced, it makes sense to compare the performance of the classifier with the ZeroR classifier which classifies all examples in the majority class (i.e. bad example). Apart from Naïve Bayes with the default configuration, all classifiers return an accuracy higher than the ZeroR baseline. The highest accuracy was recorded with the J48 classifier in combination with feature selection (85.21%). For detailed results, see Table 3.

Feature discretization: Naïve Bayes, J48, IBk and SMO have all exhibited improved precision after feature discretization, but this improvement came in the majority of cases at the expense of lower recall. However, we are primarily interested in precision meaning that discretization has a positive influence on the performance of these classifiers for this task. On the other hand, the performance of JRip slightly decreased with discretization.

Parameter fine-tuning: We have experimented with different parameter settings. For J48, different minimum numbers of objects were tested. Figure 1 plots precision as the parameter MinNumObj increases. Tests were performed with and without discretization. Without

	Precision (positive class)	**Recall (positive class)**	**F-score (positive class)**	**Accuracy**
ZeroR	0	0	0	0.7222
Naïve Bayes	0.440	**0.916**	0.595	0.653
cNaïve Bayes with discretization	0.554	0.819	0.661	0.766
Naïve Bayes with feat. selection	0.534	0.632	0.579	0.745
J48 (MinNumObj=2)	0.734	0.686	**0.709**	0.844
J48 (MinNumObj=9)	0.745	0.665	0.703	0.844
J48 with discretization (MinNumObj=2)	0.753	0.568	0.647	0.828
J48 with discretization (MinNumObj=22)	0.770	0.543	0.637	0.828
J48 with feat. selection (MinNumObj=2)	**0.801**	0.622	0.700	**0.852**
SMO	0.644	0.573	0.607	0.794
SMO with discretization	0.700	0.630	0.663	0.822
SMO with feat. selection	0.646	0.562	0.601	0.793
IBk (k=1)	0.635	0.673	0.654	0.802
IBk (k=7)	0.686	0.619	0.651	0.815
IBk with discretization (k=9)	0.732	0.635	0.680	0.834
IBk with feat. selection (k=9)	0.732	0.643	0.685	0.836
JRip	0.738	0.570	0.643	0.824
JRip with feat. discretization	0.735	0.616	0.671	0.832
JRip with feat. selection	0.763	0.576	0.656	0.833

Table 3: Classification results with different algorithms and parameter settings

discretization, the largest precision was achieved with MinNumObj set to 9 (0.745). This setting results in a tree with 29 leaves. With discretization, the best precision was achieved when MinNumObj was set to 22 (0.770). At this point the tree had nine leaves. Discretizing the data allows us to achieve better precision with the added bonus of having fewer leaves which makes the tree easier to interpret. As can be seen in Figure 2, discretization has a positive influence also on the precision of the IBk[3] classifier. The largest precision is achieved with k set to 9 (0.732). Without discretization, precision never breaks the 0.7 barrier. For Naïve Bayes, SMO and JRip, we have not been able to improve considerably the performance of these two classifiers by adjusting the respective parameters and have used the default parameters throughout the analysis.

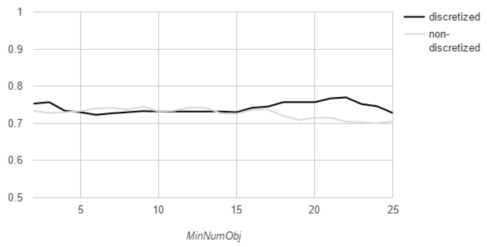


Figure 1: Precision of J48

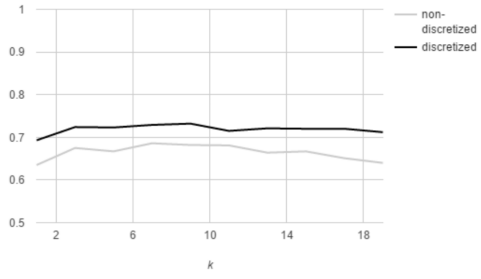


Figure 2: Precision of IBk

Feature selection: We tested the role of feature selection by applying the Weka's feature selection functionality, described in Section 3.3. In terms of precision, feature selection

[3] To avoid ties in a binary classification problem, we have only used odd values of k.

improved the precision of all classifiers, but this improvement came at the expense of recall (for details see Table 3). When applying the feature selection, the number of features has fallen considerably for all classifiers (e.g., from all the features seven were selected for J48, and six for JRip), except for SMO where the number of features fell only marginally to 17. In Table 4 the features resulting from the feature selection process are listed. It can be seen that the target term position, number of commas and the source initial capitalization/final punctuation are present in almost all the classification models (in four out of five models).

Classifier	**Selected features**
NaiveBayes	TLength, SCap_Punc, TargetCase
J48	TLength, STermPos, SNoDig, TNoDig, TTermPos, TCap_Punc, NoComma
SMO	SLength, TLength, STermPos, SNoDig, TNoDig, SNoWeirdChar, TNoWeirdChar, SCap_Punc, TtermPos, TCap_Punc, TargetCase, TDPron, NoComma, NoFullstop, InitWrdType, SNonInitCapWrds, TNonInitCapWrds
IBk	TNoWeirdChar, SCap_Punc, TTermPos, TargetCase, NoComma, NoFullstop
JRip	SCap_Punc, TTermPos, TCap_Punc, TargetCase, TPPron, NoComma

Table 4: The most informative features (feature selection results)

Model interpretation: For model interpretability the most interesting results were produced with the JRip classifier which produces a set of easily interpretable rules. In Figure 3 we present the JRip model with the feature selection step. For example, Rule 1 says that for a bilingual sentence pair to be a good term example, the target term has to be positioned within the first four words (the first position is 0) from the beginning of the sentence, there should be only one or no commas in the target sentence and the target term should be in the instrumental case. Interestingly, this rule contains no mention of any source features which could indicate that there is a strong relationship between the source and target sentences (i.e. if a sentence is a good example in one language, its corresponding pair will also be a good example in the other language). This rule has a very high precision (0.917), and covers 72 examples. Subsequently new rules are formed to cover other, still uncovered, instances.

```
1. If (TTermPos <= 3) and (NoComma <= 1) and (TargetCase = i) => Target Variable=YES (72.0/6.0)

2. (TTermPos <= 10) and (TargetCase = n) and (TTermPos <= 2) and
(NoComma <= 2) and (NoComma >= 1) and (TTermPos >= 2) => Target Variable=YES (51.0/8.0)

3. (TTermPos <= 10) and (TargetCase = n) and (TTermPos <= 1) and (NoComma <= 2) and
(TCap_Punc = TRUE) => Target Variable=YES (121.0/32.0)

4. (TTermPos <= 11) and (TargetCase = l) and (TTermPos <= 3) and (TTermPos >= 2)
=> Target Variable=YES (29.0/6.0)

5. (TTermPos <= 12) and (NoComma <= 1) and (NoComma >= 1) and (SCap_Punc = TRUE) and (TTermPos
<= 7) and (TTermPos >= 2) => Target Variable=YES (54.0/20.0)

6.  => Target Variable=NO (1005.0/115.0)
```

Figure 3: JRip rules on the dataset with feature selection.

We also analysed the results without feature selection and compared the results on all features and all features with discretization. We provide the first JRip rule for each. For the representation without discretization and feature selection the first rule (covering 134 instances, out of which 18 are misclassified, leading to the rule precision of 0.866) is the following:

1. (STermPos <= 6) and (TNoWeirdChar <= 0) and (TLength >= 13) and (TTermPos <= 3) and (SLength <= 28) => Target Variable=YES (134.0/18.0)

Again, the term position is important (the source and target term position), and other features are the length of the target and source sentences, and no weird characters in the target sentence.

On the discretized features the first rule (with precision of 0.795) is the following:

1. (TTermPos = '(0.5-3.5]') and (SLength = '(-inf-27.5]') and (SNonInitCapWrds = '(-inf-3.5]') and (SCap_Punc = TRUE) => Target Variable=YES (253.0/52.0)

As in the previous rule, the term position is important, (but here it is just the target term's position that was selected), followed by the requirement for a low number of non-initial capitalized words and the need for the first word in the source sentence to be capitalized and the source sentence to contain a final punctuation mark.

Comparison with GDEX: We can align our findings with the findings of characteristics of good examples for lexicography – the GDEX for Slovene by Kosem et al. (2011).While a direct comparison is not possible due to the different setup (e.g. GDEX only deals with monolingual data and because of the different features involved), there are nevertheless some comparisons to be made. In GDEX, the following features were found to be the most relevant: preferred sentence length, relative keyword position in the sentence, penalty for keyword repetition, penalty for words exceeding the prescribed maximum length, and penalty for sentences exceeding maximum length. As seen in Table 5, some of the most prominent features offered by the feature selection functionality in this paper are similar: source length and target term position are closely related with preferred sentence length and relative keyword position in GDEX. Looking at the values produced by the JRip classifier on the discretized data for these two features, we can observe similarities with GDEX results (see Table 5).

GDEX (Slovene1 configuration)	**Our approach**
Relative keyword position between 0-20% of the sentence	Target term position = '(0.5–3.5]'
Preferred sentence length min 8 and max 30 words	Source length (characters) = '(-inf–27.5]

Table 5: Comparison with GDEX on target term position and sentence length

5. Discussion and conclusions

We presented the data mining experiments on the task of finding good term examples for terminological databases, where the input files are parallel sentences from a translation memory.

Overall, all classifiers apart from Naïve Bayes with the default configuration have provided some level of improvement in accuracy over the ZeroR classifier which classifies all instances into the same class (bad examples) (see Table 3). In terms of precision of the positive class—which is also the most relevant measure for our goal—as well as overall

accuracy, the best classifier for this task seems to be J48 (with feature selection and minimum number of objects set to two) and the worst Naïve Bayes – the difference between the highest (J48, 0.801) and the lowest precision (Naïve Bayes, 0.440) is around 50%. Weka's implementations of k-nearest neighbours (IBk), support vector machine (SMO) and JRIP also perform quite well and could be good candidates for future research into good term example extraction. However, it is important to note that SMO is considerably more demanding in terms of processing power and takes much longer to complete, which can be a significant factor for practical applications; it also provides less interpretable results.

Fine-tuning parameters of the classifiers J48 and IBk provided some improvement in performance as well as reducing the leaf count in a J48 decision tree. We were unable to increase the performance of the other three classifiers by fine-tuning their respective parameters.

Supervised discretization has proved to be beneficial for the precision of classifiers with all classifiers (except for JRip) improving their results after discretization. The same holds true for feature selection. In general, the improvements due to feature selection were greater than the improvements due to discretization. Feature selection also considerably reduces the number of significant features with the number ranging from three to seven (out of the 20 available), except for SMO where the number of features remained relatively high even after feature selection.

Finally, the JRip classifier provides a set of easily interpretable rules. Some of these rules have even higher precisions (e.g. 0.917) than individual classifiers.

While the results are promising, there is certainly room for improvement. The obvious route to take would be to explore the combination of discretization and feature selection, because we have seen that both improve the precision of the classifiers. Moreover, having a larger dataset with more diverse data from different domains would most likely improve the ability to apply the model to any domain. We have not tested our classifier on language pairs other than English-Slovenian, but most of the extracted features are language independent which suggests that this classifier could also be used successfully for other language pairs. This is something we plan to test in the future.

Finally, the dataset is complex and treating this issue as a binary classification problem may be too simplistic to accurately reflect the differences between various sentences in the dataset. In the future, it would make sense to repeat the experiment with numeric scores (e.g. five being the best example, one being the worst) instead of YES/NO values which would allow us to test regression algorithms. Moreover, extracting word type sequences would allow us to discover the most typical sentence structures of good examples and including features describing word frequencies in reference and domain-specific corpora would unlock a completely new level of analysis.

This paper is part of a larger research into developing a comprehensive terminology extraction system for a translation service provider. In addition to extracting terms and good term examples, we will focus on other types of information that can be extracted from TMX files, such as definitions, collocations or domains. Having the ability to quickly and accurately extract good term examples would be of great benefit to this system.

6. Acknowledgements

This research was done as part of cooperation between the JSI Institute and Iolar d.o.o. in the scope of the TermIolar project.

7. References

Barrière, C. (2004). *Knowledge-Rich Contexts Discovery*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 187–201. URL http://dx.doi.org/10.1007/978-3-540-24840-8_14.

Del Gaudio, R., Batista, G. & Branco, A. (2014). Coping with highly imbalanced datasets: A case study with definition extraction in a multilingual setting. *Natural Language Engineering*, 20(3), p. 327–359.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. & Witten, I.H. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explor. Newsl.*, 11(1), pp. 10–18.

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M. & Rychlỳ, P. (2008). GDEX: Automatically Finding Good Dictionary Examples in a Corpus. In *Proceedings of the 13th EURALEX International Congress*. Barcelona, Spain: Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra, pp. 425–432. URL https://euralex.org/publications/gdex-automatically-finding-good-dictionary-examples-in-a-corpus/.

Kosem, I., Husak, M. & McCarthy, D. (2011). GDEX for Slovene. In *Electronic lexicography in the 21st century: new applications for new users: Proceedings of eLex 2011*. Bled, Slovenia: Trojina, Institute for Applied Slovene Studies, pp. 151–159. URL http://elex2011.trojina.si/Vsebine/proceedings/eLex2011-19.pdf.

Ljubešić, N., Klubička, F., Agić, Ž. & Jazbec, I.P. (2016). New inflectional lexicons and training corpora for improved morphosyntactic annotation of Croatian and Serbian. In *Tenth International Conference on Language Resources and Evaluation*. Portorož, Slovenia: European Language Resources Association, pp. 4264–4270. URL http://www.lrec-conf.org/proceedings/lrec2016/pdf/340_Paper.pdf.

Ljubešić, N. & Peronja, M. (2015). Predicting corpus example quality via supervised machine learning. In *Electronic Lexicography in the 21st Century: Linking Lexical Data in the Digital Age: Proceedings of eLex 2015*. pp. 477–485.

Loper, E. & Bird, S. (2002). NLTK: The Natural Language Toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 63–70. URL http://dx.doi.org/10.3115/1118108.1118117.

Malaisé, V., Zweigenbaum, P. & Bachimont, B. (2004). Detecting Semantic Relations between Terms in Definitions. In *COLING 2004 CompuTerm 2004: 3rd International Workshop on Computational Terminology*. Geneva, Switzerland: COLING, pp. 55–62.

Meyer, I. (2001). Extracting knowledge-rich contexts for terminography. *Recent advances in computational terminology*, 2, p. 279.

Navigli, R. & Velardi, P. (2010). Learning word-class lattices for definition and hypernym extraction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 1318–1327. URL http://dl.acm.org/citation.cfm?id=1858815.

Pollak, S., Vavpetic, A., Kranjc, J., Lavrac, N. & Vintar, S. (2012). NLP workflow for online definition extraction from English and Slovene text corpora. In *11th Conference on Natural Language Processing, KONVENS 2012, Empirical Methods in Natural Language Processing*. Vienna, Austria: ÖGAI, pp. 53–60. URL http://www.oegai.at/konvens2012/proceedings/10_pollak12o/.

Storrer, A. & Wellinghoff, S. (2006). Automated detection and annotation of term definitions in German text corpora. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC-2006)*. Genoa, Italy: European Language Resources Association (ELRA), pp. 2373–2376. URL http://www.lrec-conf.org/proceedings/lrec2006/pdf/128_pdf.pdf.

Westerhout, E. (2010). *Definition Extraction for Glossary Creation: A Study on Extracting Definitions for Semi-automatic Glossary Creation in Dutch*. International series. LOT.

Introducing Lexonomy: an Open-Source Dictionary Writing and Publishing System

Michal Měchura

Natural Language Processing Centre, Masaryk University, Brno, Czech Republic
E-mail: michmech@mail.muni.cz

Abstract

This demo introduces Lexonomy www.lexonomy.eu, a free, open-source, web-based dictionary writing and publishing system. In Lexonomy, users can take a dictionary project from initial set-up to final online publication in a completely self-service fashion, with no technical skills required and no financial cost.

Keywords: dictionary writing systems; online dictionaries; XML editing

1. Introduction

Lexonomy is a web-based platform for writing and publishing dictionaries. Its mission is to be an easy-to-use tool for small to medium-sized dictionary projects. In Lexonomy, individuals and teams can create a dictionary, design an arbitrary XML structure for the entries, edit entries, and eventually make the dictionary publicly available as a 'microsite' within the Lexonomy website. Lexonomy exists in order to lower the barriers of entry into modern born-digital lexicography.[1] Compared to other dictionary writing systems[2] it requires no installation or set-up, expects no knowledge of coding or programming,[3] and is free from financial cost. It is simply a website where lexicographers can sign up and start working.

Each Lexonomy user logs in with a user name and password. Users are allowed to create an unlimited number of dictionaries. The process of creating a dictionary consists of deciding what it should be called (this can be changed later) and what its URL should be, for example www.lexonomy.eu/mydictionary. This is the address at which the dictionary will eventually be publicly viewable, if and when its creators decide to make it public. By default, newly created dictionaries are not publicly viewable.

Once a dictionary has been created, the user who created it may add additional users and these can all start adding and editing entries. The rest of this introduction to Lexonomy will unfold in a logical sequence. Fist we will introduce features related to **dictionary planning**: specifying the structure of entries etc. Second, we will look at **dictionary editing** with Lexonomy's built-in XML editor. Third, we will show how Lexonomy can be used as a platform for online **dictionary publishing**.

2. Entry structure

Dictionary entries in Lexonomy are stored as XML documents and their structure is defined by a schema which is unique to each dictionary. Users can choose a predefined schema while creating a new dictionary (the options are *monolingual dictionary*, *bilingual*

[1] One implication of this is that Lexonomy is not a good match for retro-digitized dictionaries.

[2] The reader is kindly asked to read what follows as a mission statement rather than an empirically verified fact. A thorough comparison of existing dictionary writing systems is beyond the scope of this paper.

[3] Familiarity with XML is a plus but Lexonomy users are not expected to be able to hand-code XML.

dictionary and so on) and customize it later or, if they prefer, they can start from a completely blank schema.

A Lexonomy schema is similar to a DTD (Document Type Definition): it lists the XML elements which are allowed to appear in the entries and specifies how they may be nested, how many of them must or may be there, which attributes they may or must have, what their values may be and so on. In a conventional dictionary writing system the schema would typically be hand-coded by an IT specialist. Lexonomy, on the other hand, offers a visual schema editor where users can define the structure of their entries without having to hand-code anything.

The left-hand side of the screen (Figure 1) contains a list of XML elements and attributes. The tree structure indicates how they may be nested, such that the top-most element will be the root element of every entry. It is up to the user to decide what the elements and attributes are called and how they are nested. The right-hand side of the screen then contains detailed settings for the selected element or attribute: this is where the user specifies what child elements or attributes the element may contain, in what order, how many of each, and what content they are allowed to hold.

2.1 Element content

The content of each element can be constrained by making a choice from these options:

- *Child elements*: elements of this name will contain other elements.
- *Text*: elements of this name will contain plain text.
- *Text with markup*: elements of this name will have mixed content (= plain text interlaced with other XML elements).
- *Value from list*: elements of this name will contain a value from a predefined list.
- *Empty*: elements of this name will have no content.

Depending on the type of content chosen, the schema editor will offer different additional options. If the content is *Child elements* or *Text with markup*, we can specify the child elements as in Figure 1. The *min* and *max* numbers control how many instances of the child element must be present inside the parent element: *min* = 1 and no *max* means 'one or more', *max* = 1 and no *min* means 'none or one', no *min* and no *max* means 'none, one or more', and so on. We will see in a later section how Lexonomy imposes these constraints while the lexicographer is editing an entry.

If the element's content is set to *Value from list*, we can specify the values on that list, along with optional captions (Figure 2). We will see later how Lexonomy's XML editor makes use of this setting by giving the lexicographer a menu to choose from when inputting an attribute value. The captions are used instead of values for visualization to end-users.

2.2 Attributes

Besides child elements, XML elements in Lexonomy can have XML attributes. When specifying that an element can have an attribute, we can declare the attribute optional or obligatory, as in Figure 1. Further settings for attributes are a subset of those for elements: an attribute's content can be either *Text* or *Value from list* (Figure 3).

According to the XML standard,[4] the attributes of an element are considered unordered: the order in which they appear in the XML document is insignificant. But, as a convenience to human users, Lexonomy makes sure that attributes always appear in the order in which they are listed in the schema.

2.3 Element nesting

It is possible in Lexonomy for elements of a certain name to appear as children under parent elements of more than one type. For example, if your dictionary has separate elements for senses and subsenses, say `<sense>` and `<subsense>`, they can both have child elements called `<definition>`, `<example>` etc. (Figure 4). Element nesting can be recursive, too: it is possible to allow `<sense>` elements to appear inside `<sense>` elements (Figure 5).

2.4 Expressivity of the schema formalism

The schema formalism used internally by Lexonomy and exposed through its schema editor is approximately as expressive as a DTD (Document Type Definition). The only major point of difference is how child elements are ordered. In Lexonomy, child elements (under parents whose content is *Child elements*) must appear in exactly the same order in which they are given in the schema, while a DTD allows more flexibility in this regard.

3. Entry editing

Once the structure has been finalized lexicographers can start working on the actual entries. Lexonomy's entry browser and editor offers a familiar interface with an entry list on the left-hand side and entry details on the right-hand side (Figure 6). Clicking the *Edit* button opens the entry in Lexonomy's built-in XML editor (Figure 7).

The XML editor in Lexonomy[5] emulates the look and feel of a text editor with syntax highlighting, code folding and autocompletion. It is, however, not a text editor: lexicographers edit XML by clicking on things, selecting options from context menus, selecting attribute values from picklists, dragging and dropping elements around and so on. This serves the dual purpose of making the lexicographer aware that he or she is manipulating XML while simultaneously making it impossible for them to corrupt the entries by entering non-well-formed XML. In fact, no knowledge of XML syntax is needed for working with Lexonomy: the angle brackets and other formalities of XML syntax are merely a kind of 'decoration'. Users who are not comfortable with the XML notation can turn it off completely and switch Lexonomy into *laic mode* (Figure 8).

3.1 Knowing where to click

Almost everything in the XML editor is clickable:

- Click the name of an element (it its opening or closing tag) to get a menu with options for adding child elements, for adding optional attributes, and also for removing the element itself. The options offered are in accordance with the schema.

[4] https://www.w3.org/TR/REC-xml/#attdecls
[5] The XML editor is actually a separate software product called Xonomy: www.lexiconista.com/xonomy

- Click the name of an attribute to get a menu with an option to remove the attribute.
- Click the value of an attribute to get a pop-up box for editing the value. This will be either a text box or a menu to choose from a list, as per the schema.
- Click a text node (= a stretch of text between tags) to get a pop-up box for editing the text. Again, this will be either a text box or a menu to choose from a list, as per the schema.

To change the order of elements (for example to re-order senses) or to move an element to a different location inside the entry (for example to move an example sentence from one sense to another) you can use the 'drag handle' (six grey dots) beside the opening tag of each element. As you drag this with the mouse, Lexonomy will show you 'drop targets' (grey spots) in different places in the entry: these are locations where you can legally drop the element you are dragging ('legally' here means 'the schema allows it').

3.2 Keyboard navigation

A frequent complaint by users of web-based editing interfaces[6] is that the work is slow because there is 'too much clicking' involved. For increased productivity and ergonomics, Lexonomy makes it possible for lexicographers to perform the most repetitive tasks with the keyboard as well as the mouse. While editing an entry in the XML editor, the following keyboard shortcuts are available:

- The cursor keys `up` and `down`, `left` and `right` to navigate around the hierarchical structure of the entry, from tag to tag, from attribute to attribute, and so on.
- When an element has the plus sign next to its opening tag, `Ctrl + right` can be used to expand it and `Ctrl + left` to collapse it again.
- Press `Enter` to open the menu or pop-up editor associated with the currently highlighted element, attribute, attribute value or text node. Then press `Esc` to close it again.
- When a pop-up menu is open, use the cursor keys `up` and `down` to move up and down the menu, and `Enter` to select an item from the menu.
- If the entry is very long and has a scrollbar next to it, you can use `Ctrl + up` and `Ctrl + down` to scroll the entry up and down.

These keyboard shortcuts work when the entry editor is focused. If it is not focused (you will know because the keyboard shortcuts are not working) you can press `Alt + right` at any time to focus it. Similarly, you can press `Alt + left` at any time to focus the entry list on the left hand side of the screen. When the entry list is focused, the following keyboard shortcuts can be used:

- The cursor keys `up` and `down` to move up and down the list.
- `Enter` to the currently highlighted entry.
- `Ctrl + up` and `Ctrl + down` to scroll the entry list up and down.

Last but not least, the following keyboard shortcuts are available at any time, regardless of which side of the screen is focused:

[6] Based on the author's long career in building, and dealing with users of, such interfaces.

- When an entry is being displayed on the right-side of the screen, you can press `Ctrl + Shift + E` to open it for editing: this is the same as pressing the *Edit* button. Then press `Ctrl + Shift + E` again to cancel editing and switch back to viewing: this is the same as clicking the *Cancel* button.
- `Ctrl + Shift + S` to save the entry being edited: this is the same as clicking the *Save* button.
- `Ctrl + Shift + N` to start creating a new entry: this is the same as clicking the *New* button.
- `Ctrl + Shift + T` to move move the cursor into the search box in the top left-hand corner of the screen.

In all keyboard shortcuts mentioned here, Mac users can (but do not have to) substitute the `Cmd` key for the `Ctrl` key.

3.3 Editing inline markup

One area which tends to be particularly troublesome for XML editors is 'mixed content': situations in which an XML element contains a mixture of text and other XML elements. Here is how Lexonomy handles it. If the schema says that the content of an element is *Text with markup*, Lexonomy lets the lexicographer edit its text as if it were normal plain text: clicking it opens a pop-up text box. Additionally, a thin grey line appears underneath the text and the lexicographer can click on this to select stretches of text and annotate them with inline XML markup. When a stretch of text is selected, a menu will appear with options for 'wrapping' that selected text with XML elements (see Figure 9). The options on that menu come from the schema. Once markup has been inserted, it is again possible to click the inline element and a menu will appear with an option to remove ('unwrap') the element.

3.4 Entry validation

While working with an entry, the options that appear in menus and dialogs conform to the dictionary's schema: users are only allowed to add child elements to parents that may have them, and so on. When adding a new element into the entry, Lexonomy will automatically pre-populate the element with everything it needs to have, as per the schema: obligatory attributes, the correct number of child elements and so on. The same happens when creating a new entry: Lexonomy will automatically launch with a 'prefabricated' blank entry which conforms to the schema as much as possible: for example, if your schema says that every `<entry>` must have at least one `<headword>`, then every new `<entry>` will come with one (empty) `<headword>` already inserted.

As you make changes to the entry, Lexonomy is constantly validating it against the schema. If you make an edit which is not allowed by the schema, such as insert more child entries than the schema allows, Lexonomy will notify you with a small warning triangle next to the offending element or attribute (Figure 10). As a general rule, however, Lexonomy's approach to entry validation is permissive: it gives warnings but it will not prevent you from saving an invalid entry (= an entry that does not conform to the schema).

4. Advanced settings

Each dictionary hosted in Lexonomy comes with an extensive configuration screen (Figure 11). Many settings on this screen are of an advanced nature and we will explore some of those in this section.

4.1 Where is the headword?

In Lexonomy, dictionary authors themselves decide what names the XML elements and attributes in their entries will have. There is no requirement to use a standard vocabulary of names such as `<entry>` or `<headword>`, these can have any names at all, including names in other languages than English.[7] But, at the same time, Lexonomy needs to understand what (at least some of) those element names mean. For example, it needs to know where to find the headword in each entry.

The *Headwords* area on the configuration screen is where the dictionary administrator can make such information explicit (Figure 12). Lexonomy uses this information for various things, including listing the entries by headword in the entry list on the left-hand side of the editing screen. If you make no selection here, Lexonomy will try to guess where the headword is by simply taking the first non-empty text node it finds in each entry.

In many dictionaries, headwords are 'annotated' with additional elements such as homograph numbers and part-of-speech labels. These can be made to appear in the entry list by selecting them in the *Headword annotations* section. Headwords are displayed in bold fond and are searchable (more about searching later), while annotations are displayed in non-bold font and are not searchable, but are taken into consideration for alphabetical sorting.

4.2 Alphabetical order

When listing entries by headword, the question of alphabetical order unavoidably comes up. Alphabetical order depends not only on the alphabet used (Latin, Cyrillic etc.) but also on the language (e.g. *ä* is sorted right after *a* in German but at the end of the alphabet after *z* in Swedish) and, in extreme cases, even on personal preference. Lexonomy takes an agnostic view and allows dictionary authors to set up their own alphabetical order by simply inputting a linebreak-delimited sequence of characters into the *Headwords* area of the configuration screen (see Figure 12; characters that appear on the same line with a space between them are sorted as if they were the same). There is a default sort order which dictionary administrators can customize, for example by moving characters around or by adding characters for their language. Alphabetical sorting in Lexonomy is always case-insensitive.

The sorting algorithm supports digraphs, that is, sequences of characters which are sorted as if they were a single character, such as the Czech *ch* which sorts between *h* and *i* or the Welsh *ng* which sorts between *g* and *h*. All the dictionary administrator needs to do is include the digraph in the correct place in the alphabetical order, e.g. `ch` on a separate line between `h` and `i`.

[7] The names of XML elements and attributes in Lexonomy can even contain non-ASCII characters, such as extended Latin characters and characters from other alphabets.

4.3 Search

Another thing which is under the dictionary author's control is the extent to which the dictionary is searchable by typing some text into the search box (in the top left corner of the editing screen, and also on the dictionary's public home page if the dictionary is publicly viewable). By default, searching means searching for headwords, and typing anything into the search box will return a list of entries whose headwords contain that sequence of characters. But dictionary administrators can search-enable other XML elements too, and this is done in the *Search* area of the configuration screen (Figure 13). For example, if yours is a bilingual dictionary and if you would like reverse searches to be possible, you can search-enable the elements containing the translations. Then, when you search for a sequence of characters, Lexonomy will return a list of entries where either the headword or one of the translations match (Figure 14).

Search in Lexonomy is always based on simple substring matching: when you search for *go* you will get entries where this sequence of characters occurs in one of the search-enabled elements, regardless of where in the element it is: *gorge*, *mango*, *mongoose* as well as *go* itself. In other words, search in Lexonomy is not linguistically 'clever': it is aware of neither word boundaries nor word inflection (e.g. a search for *bring* does not match *brought*), as these features are language-dependent. One implication of this is that Lexonomy's search functionality is really only suitable for short strings of text (such as headwords and their translations) but will not perform as well as full-text search (e.g. for example sentences or for definitions).

5. Entry formatting

Beside the schema designer and the entry editor, a third crucial feature of Lexonomy is its formatting designer. This is where users can design the visual appearance of their entries. In a conventional dictionary writing system this task would typically be achieved by hand-coding an XSL and/or CSS stylesheet, and an IT specialist would be required for the job. In Lexonomy, users can design the look of their entries themselves, without knowledge of any stylesheet language.

Similarly to the schema editor, the user sees a hierarchical list of elements and attributes on the left-hand side of the screen, while the right-hand side is where he or she sets the formatting properties of the selected element or attribute (Figure 15). A randomly selected entry is shown on the right on which all formatting changes are previewed in real time to help lexicographers understand the visual impact of their choices.

Under *Visibility* you select whether the element or attribute is shown at all (the default is *Shown* for elements and *Hidden* for attributes), and under *Layout* you select whether the element is separated from other elements by line breaks or not. The rest of the screen is for setting individual formatting properties of the element or attribute:

- *Separation from other content*: the options are *whitespace* or *none*. For inline elements *whitespace* means that there is a space character between it and any elements that precede or follow it. For line-breaked elements *whitespace* means there is an additional amount of vertical space (approximately half the height of a line of text) above and below.

- *Indentation and bulleting*: the options include various kinds of bullets (round, square-shaped etc) and various sense numbering patterns. It goes without saying[8] that senses in Lexonomy are numbered automatically at display-time and that sense numbers should not be included in entries explicitly.
- *Box border*: the options are *dotted*, *thin* and *thick* for putting a visual border around the element.
- *Background colour*: the options are *none*, *yellow*, *blue* and *grey*.
- *Outer punctuation*: these indicate how the element should be separated from other elements by punctuation such as commas, semicolons or brackets.
- *Text colour*: the options are *none*, *red*, *blue*, *green* and *grey*.
- *Text slant*: the options are *none* and *italic*.
- *Text weight*: the options are *none* and *bold*.
- *Inner punctuation*: the options are the same as *outer punctuation* above. The difference is that inner punctuation is inserted in the same colour, slant and weight as the content while outer punctuation is not: it is 'outside' the scope of font formatting.

5.1 Expressivity of the formatting formalism

Depending on your perspective, the formatting properties available in Lexonomy may seem either carefully curated or inconveniently constrained. The truth is a bit of both. Lexonomy's formatting mechanism is certainly not nearly as expressive as stylesheet languages such as XSL and CSS. On the other hand, the full gamut of XSL and CSS would probably be too confusing for the average lexicographer and would likely lead to amateurish misuse. Lexonomy wants all dictionaries to look good in it, but also, it wants lexicographers themselves to be in control of the formatting of their dictionaries – this calls for simplification. Time will tell whether this level of simplification is the right one.

6. Online publishing

Finally, a dictionary can be made available to the public as a 'micro-site' within Lexonomy, e.g. `www.lexonomy.eu/mydictionary`. This does not require any complicated work, the user merely needs to change a few settings in the dictionary's configuration section (Figure 16).

When a dictionary is made public, Lexonomy gives it a simple user interface which allows the dictionary to be searched and browsed (Figure 17). The home page offers a random selection of headwords and a search box. Search here works exactly like it does in the edit screen. Each individual entry has its own page with its own URL, and the headword's alphabetical neighbourhood is displayed on the side (Figure 18). The interface is responsive (therefore mobile screen-friendly) and optimized for indexing by search engines.

When a dictionary has been made public, the public interface is of course viewable by anybody, regardless of whether they are currently logged into Lexonomy or not. If logged in, and if the user has editing access to the dictionary, an *Edit* link is shown beside the dictionary title. When a dictionary has not been made public yet, the dictionary's home page is essentially the same but has only the dictionary's name and optional description (both supplied by the dictionary author) and nothing else.

[8] But let us say it anyway.

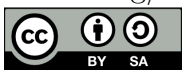
7. Conclusion

This concludes our brief introduction to Lexonomy. We have seen how Lexonomy can be used to develop a dictionary from initial set-up to final online publication. Hopefully the reader is now convinced that Lexonomy is a good home for small-to-medium sized dictionary projects. What remains is to mention a few administrative and house-keeping matters.

Lexonomy was originally created as a training tool for a lexicographic training event organized by the European Network of e-Lexicography[9] in May 2016 in Ljubljana, Slovenia. The version of Lexonomy presented here has been completely rewritten since then, contains several new or improved features, and the author believes it is fit for real-world applications.

Lexonomy is and will continue to be open-source software, licensed under the MIT Licence.[10] The source code is hosted in Lexonomy's GitHub repository.[11] Teams who do not want to use Lexonomy's 'home' installation at www.lexonomy.eu can download the source code, set up a local installation on their own server and customize it to their requirements. Lexonomy is written in Node.js,[12] a technology which makes it capable of running on both Linux and Windows servers.

Lexonomy will continue to be actively developed over the next number of years, thanks partly to financial support from Lexical Computing, the makers of Sketch Engine,[13] a popular corpus query system.



[9] http://www.elexicography.eu/
[10] https://opensource.org/licenses/MIT
[11] https://github.com/michmech/lexonomy
[12] https://nodejs.org/
[13] https://www.sketchengine.co.uk/

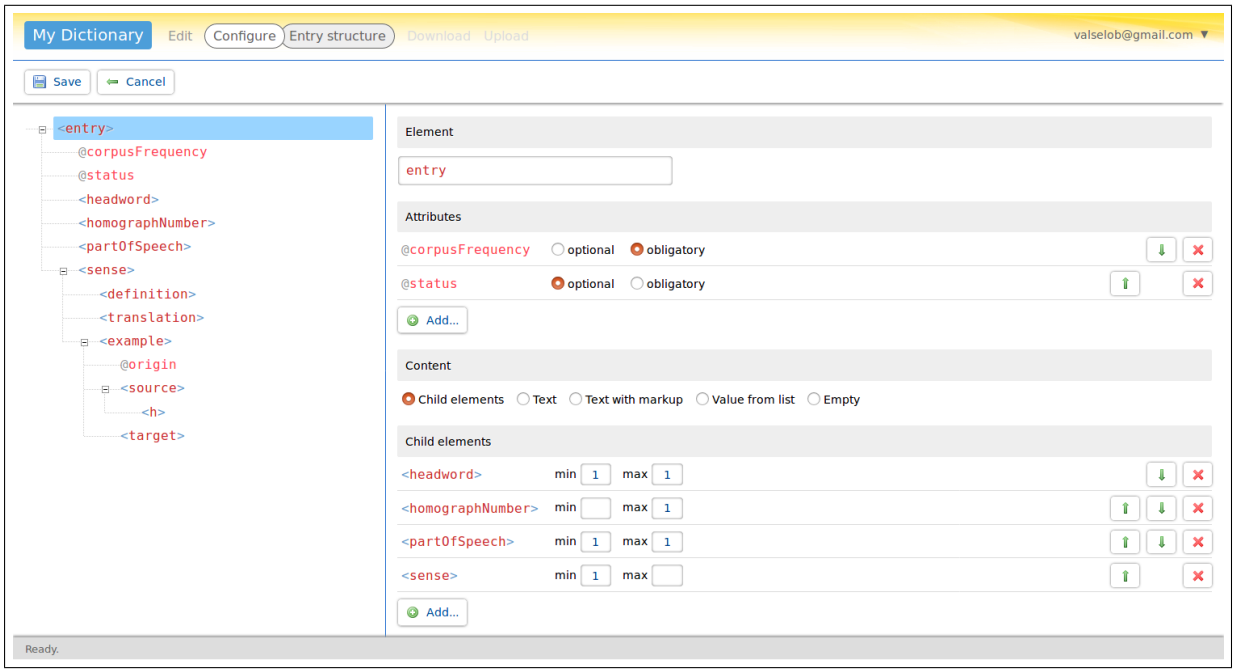


Figure 1: Entry schema editor

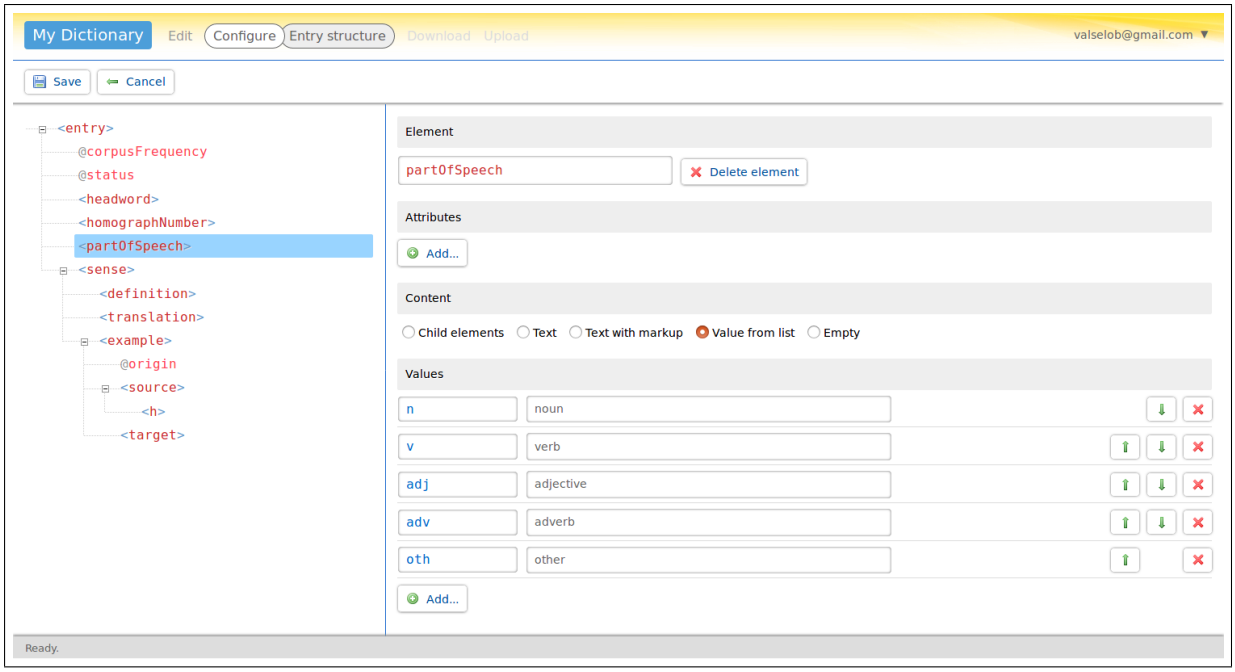


Figure 2: Specifying the values that can appear in an element

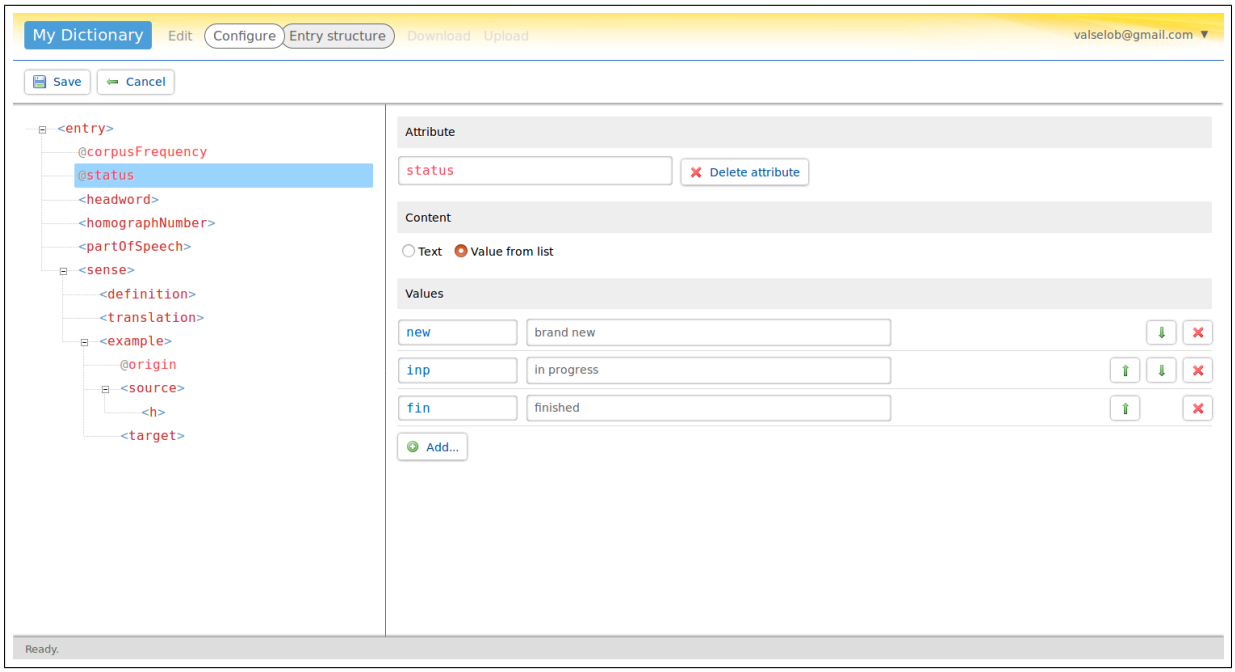


Figure 3: Specifying the content of an attribute

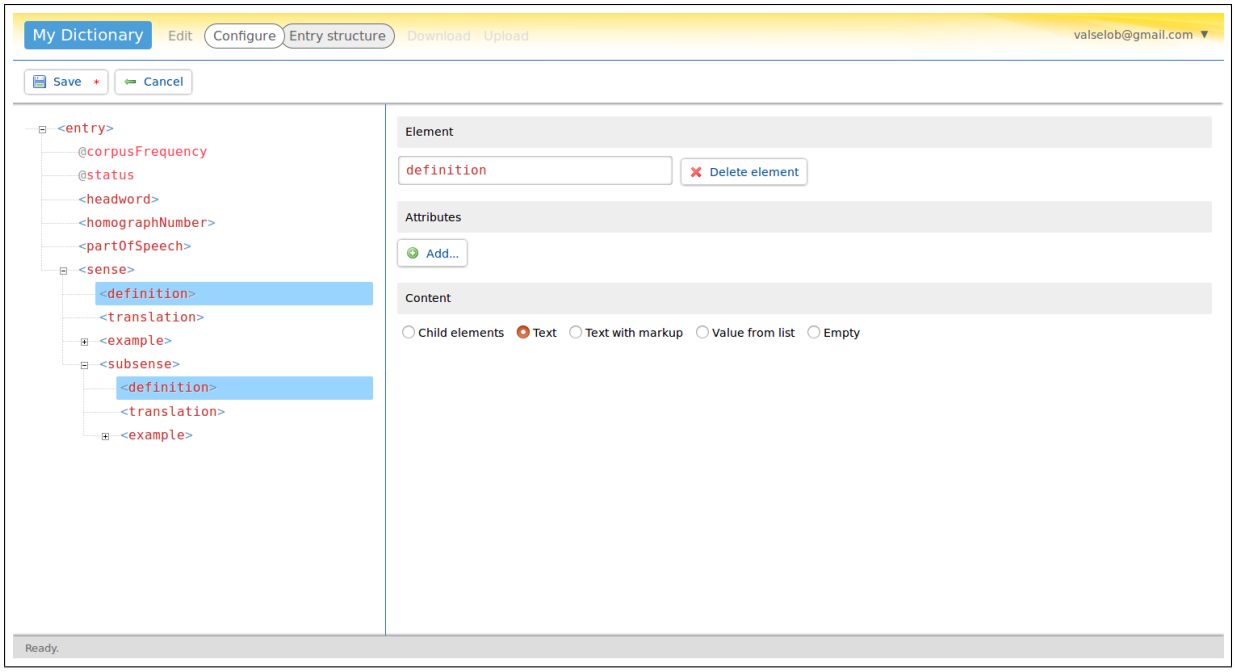


Figure 4: Allowing elements of a given name to appear under more than one type of parent

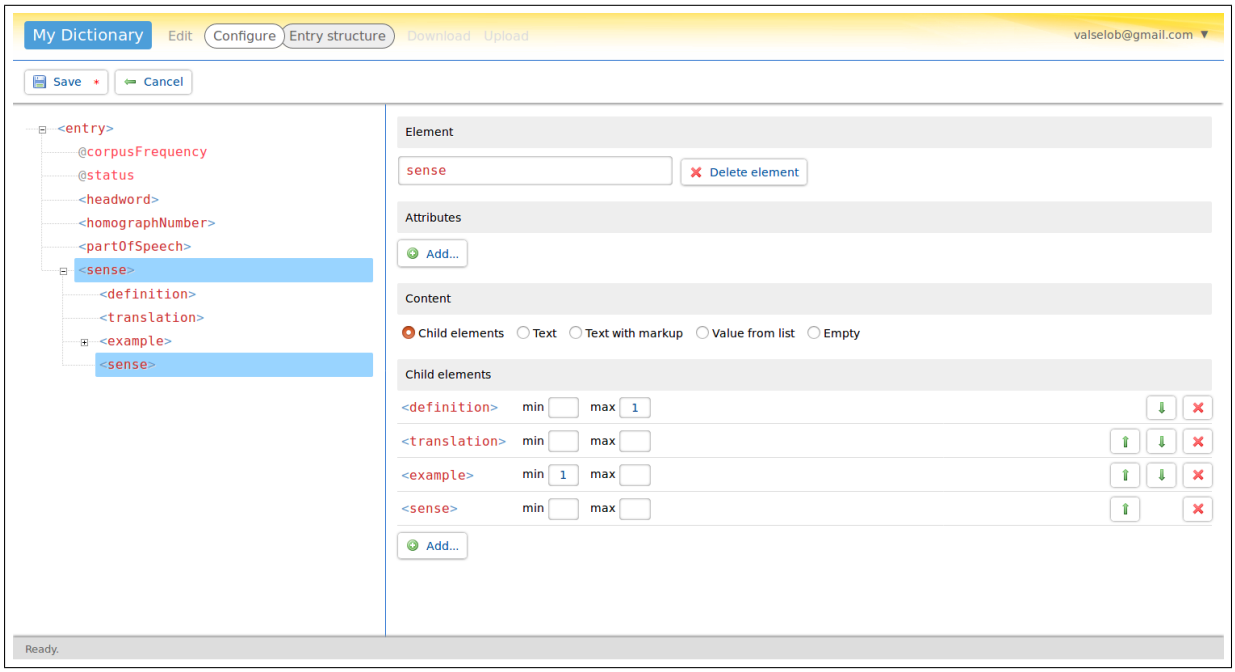


Figure 5: Recursive element nesting

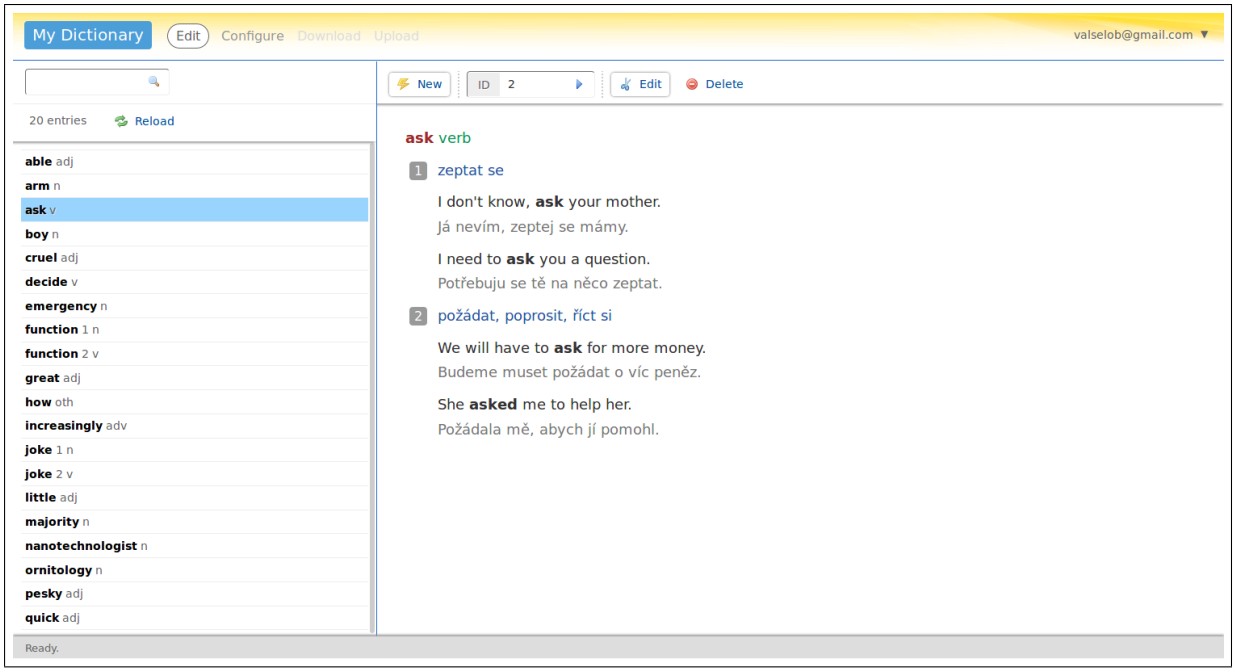


Figure 6: Browsing and viewing a dictionary

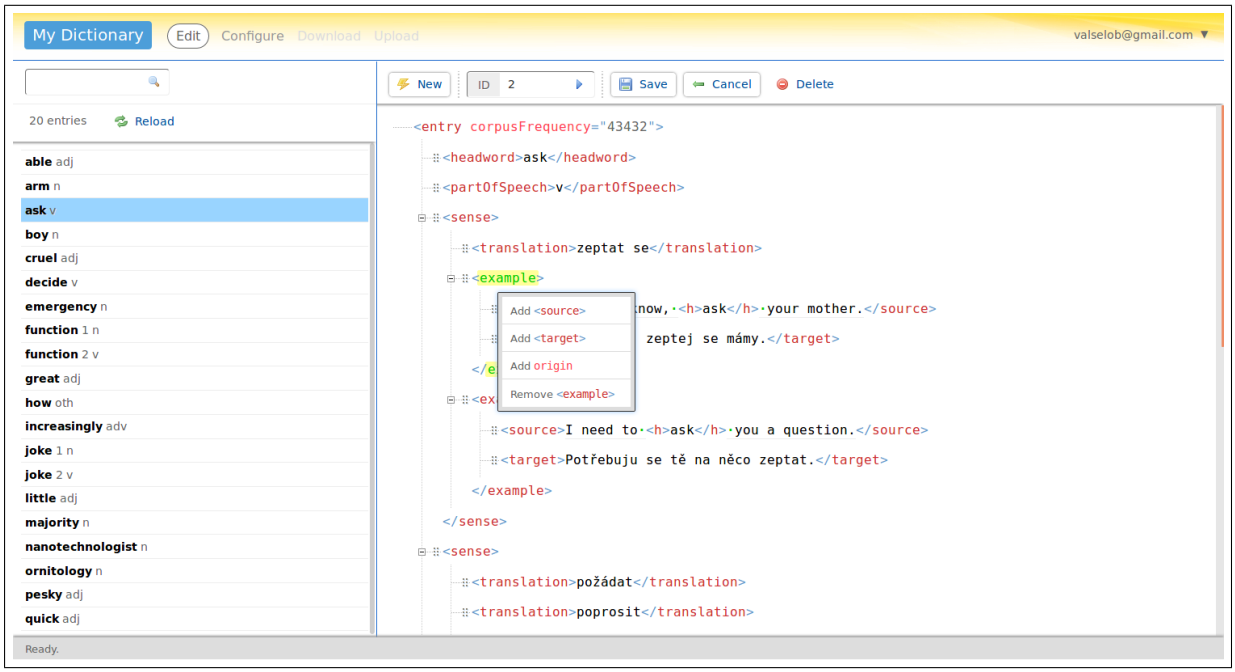


Figure 7: Entry editor

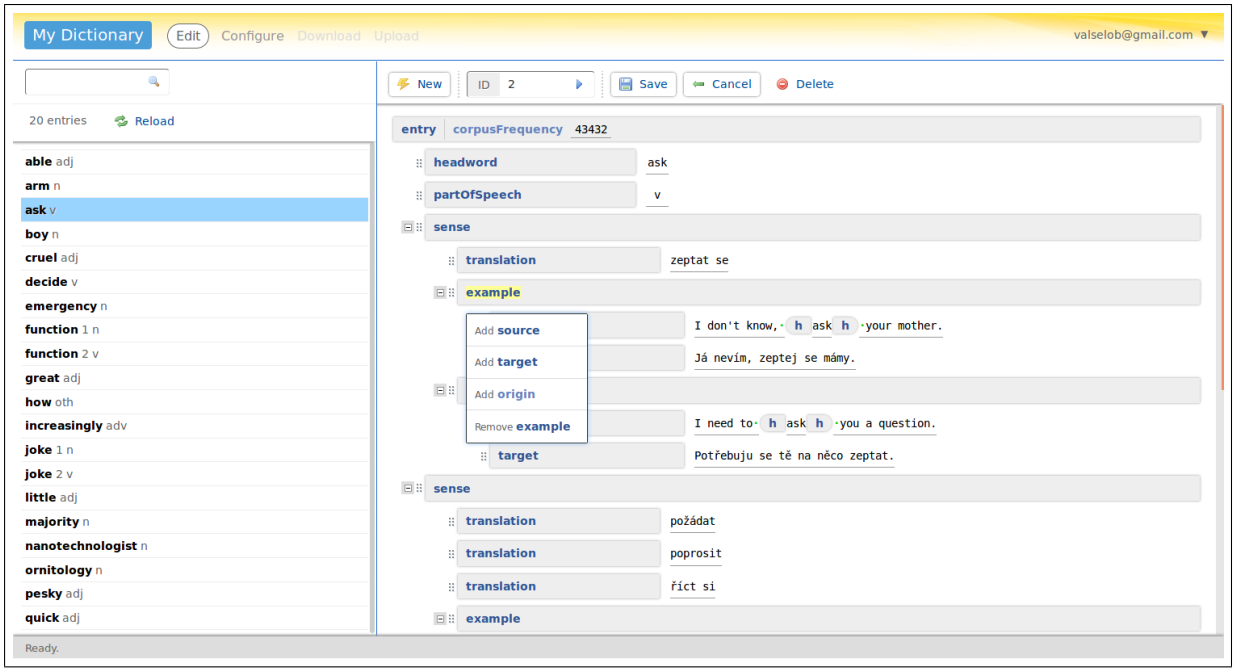


Figure 8: Entry editor in laic mode

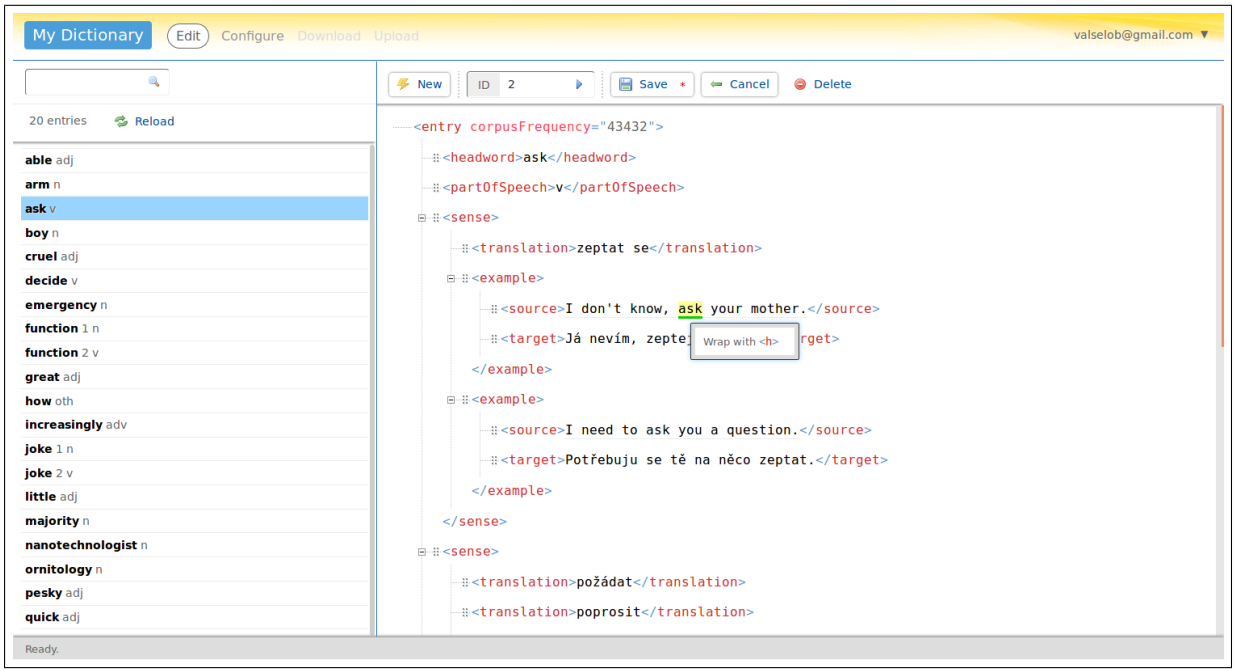


Figure 9: Inserting inline markup

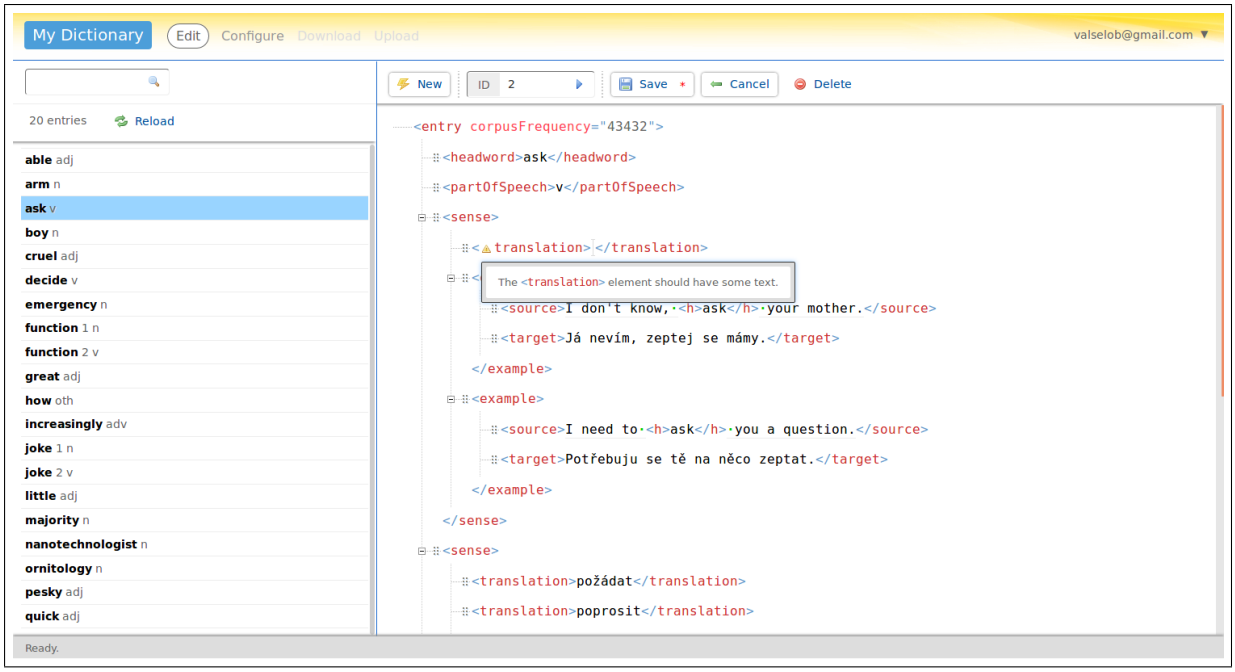


Figure 10: XML validation

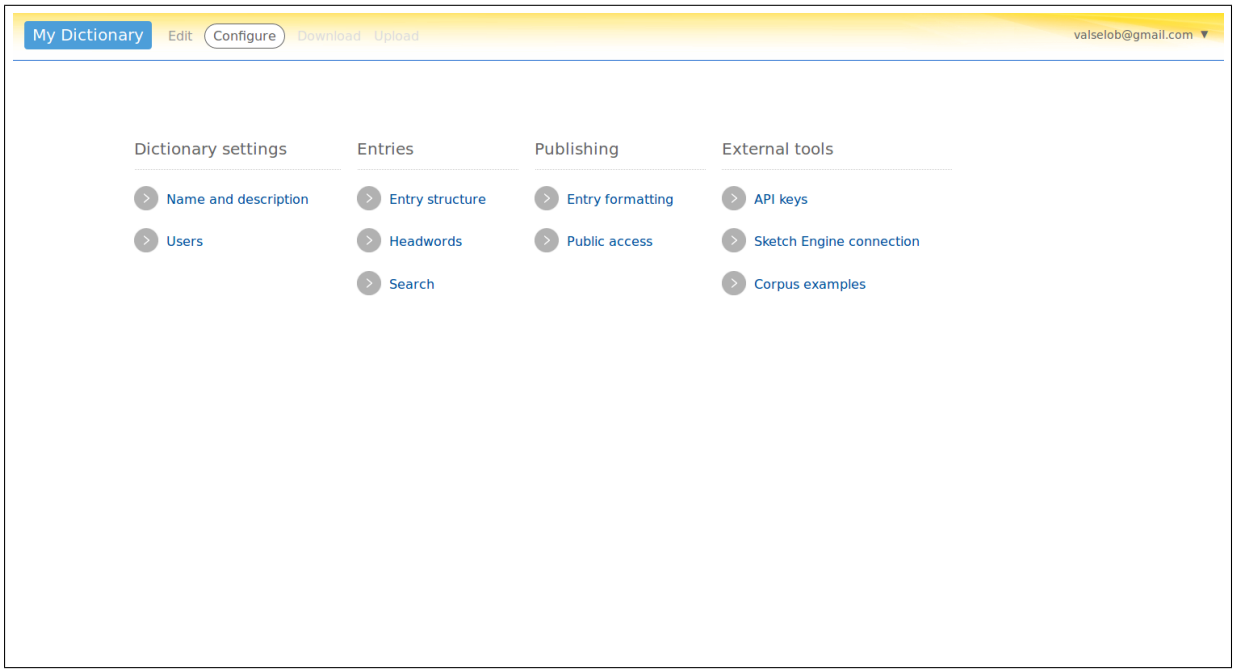


Figure 11: Configuration screen

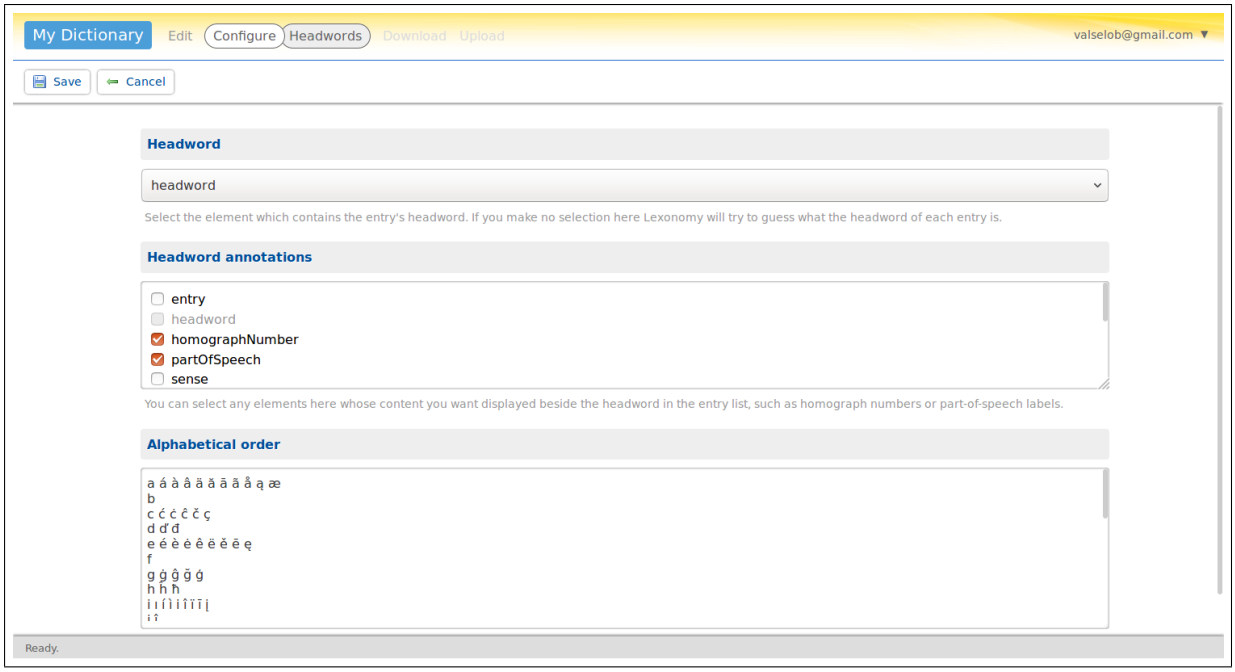


Figure 12: Headwords area of the configuration screen

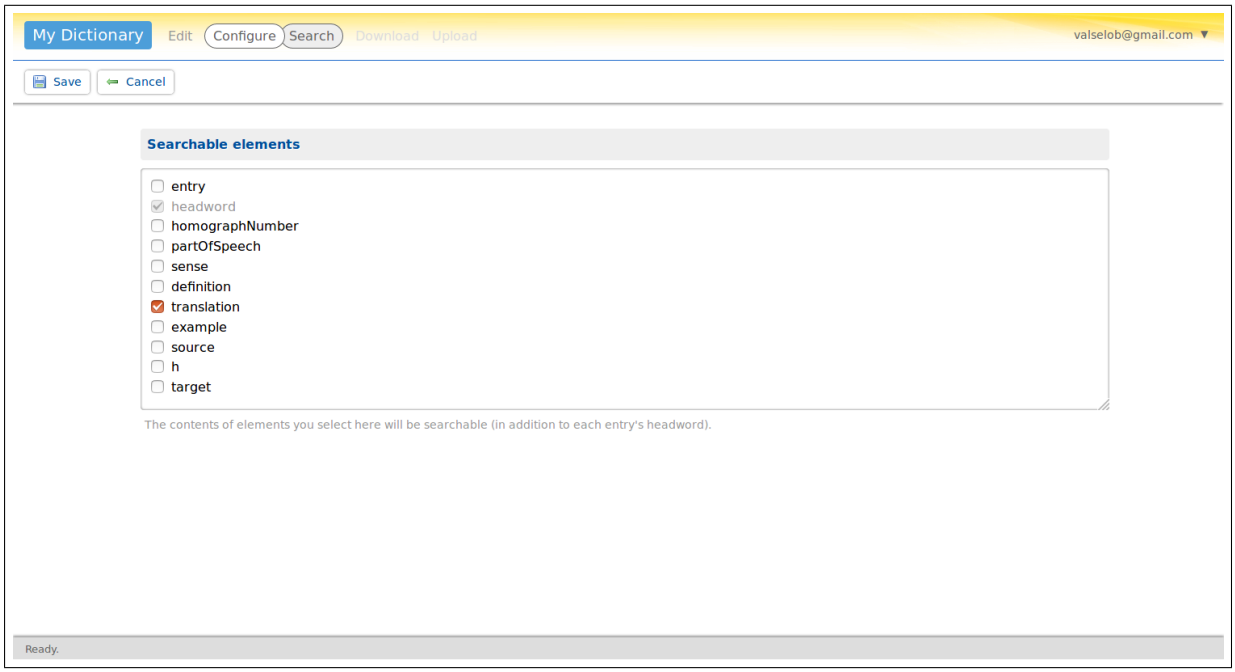


Figure 13: Search area of the configuration screen

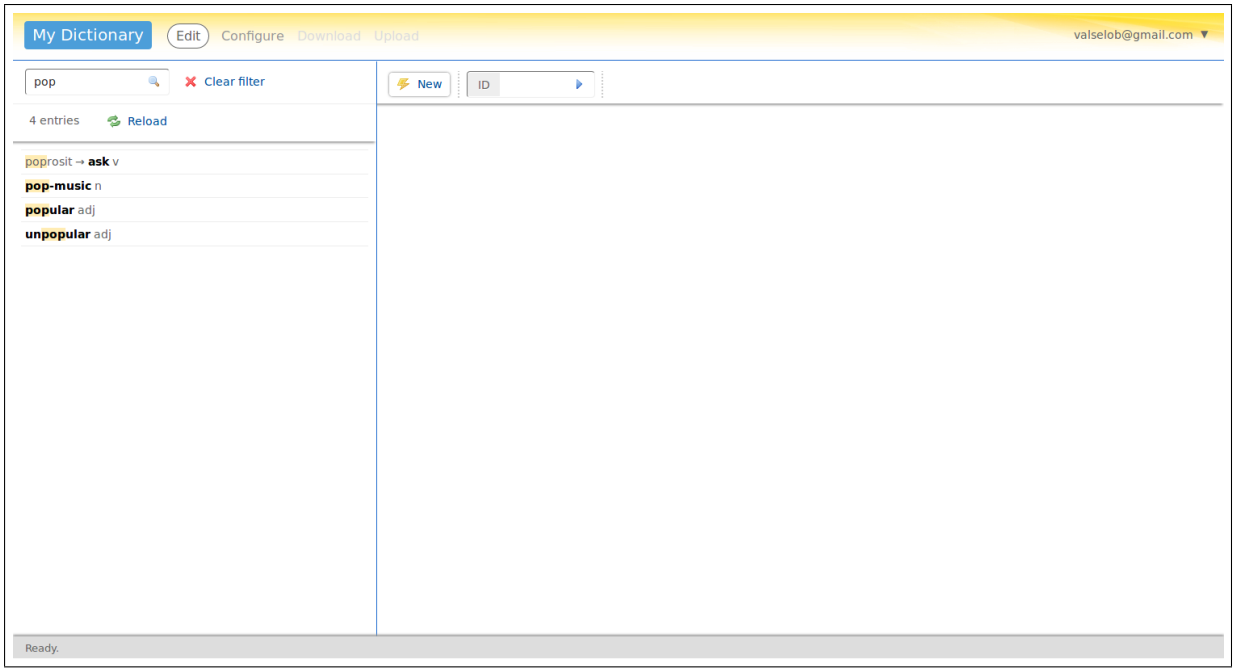


Figure 14: Example of search results

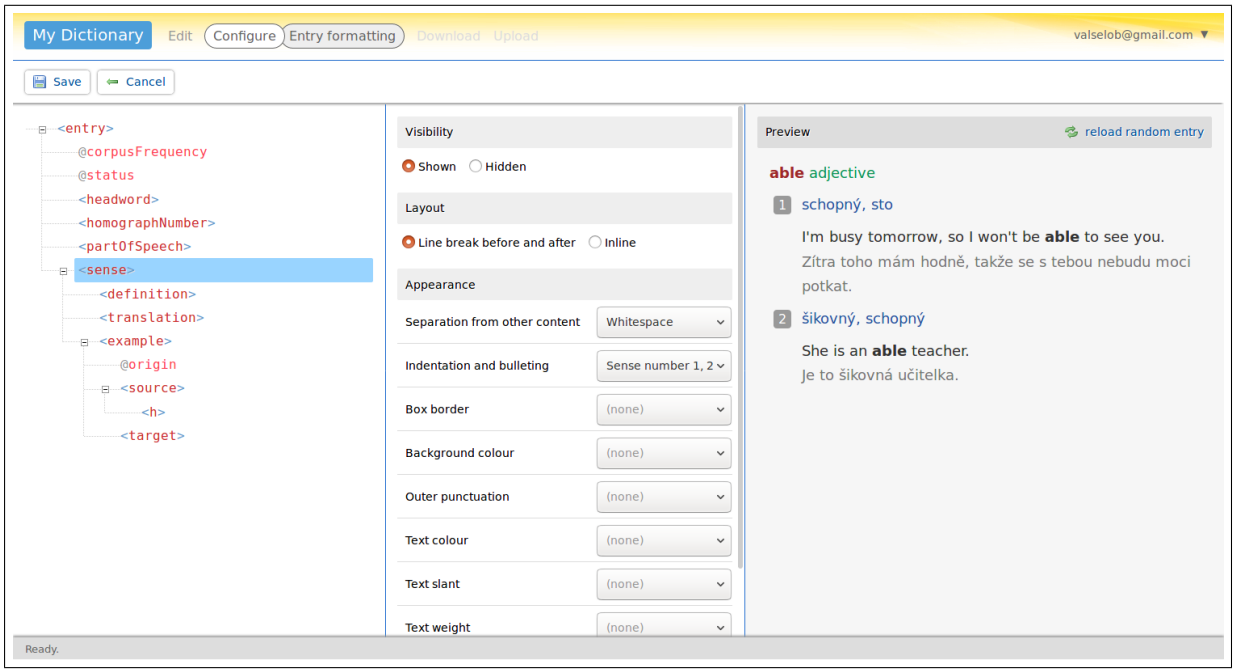


Figure 15: Formatting designer

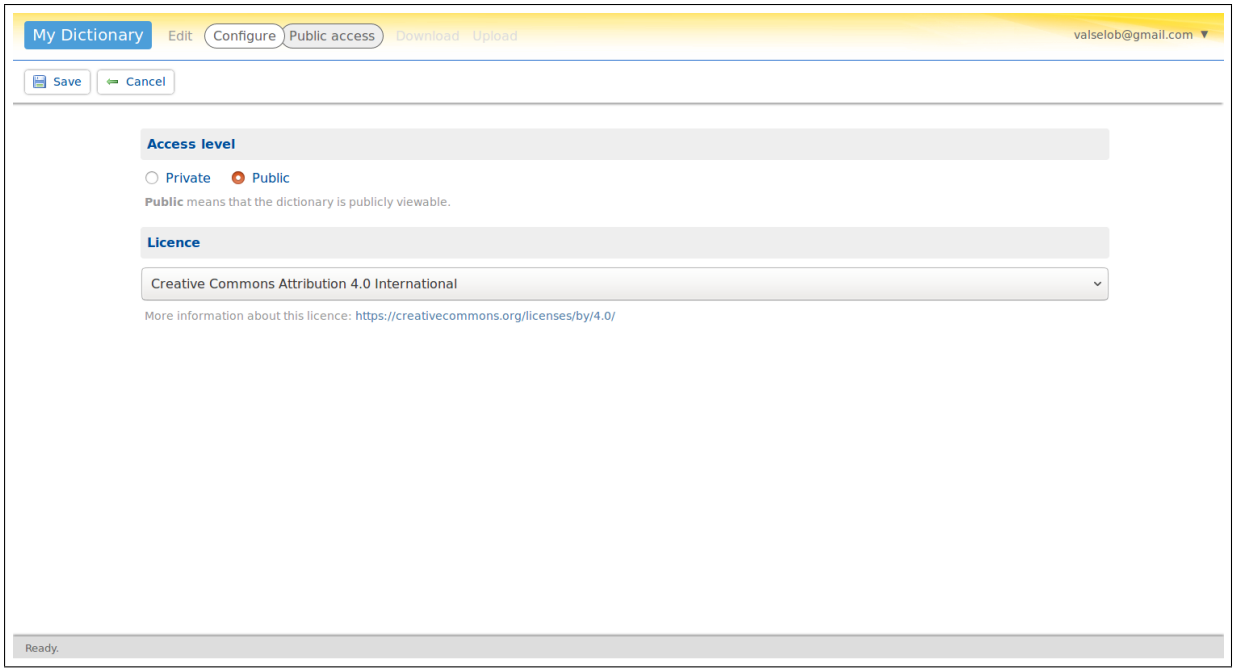


Figure 16: Public access settings

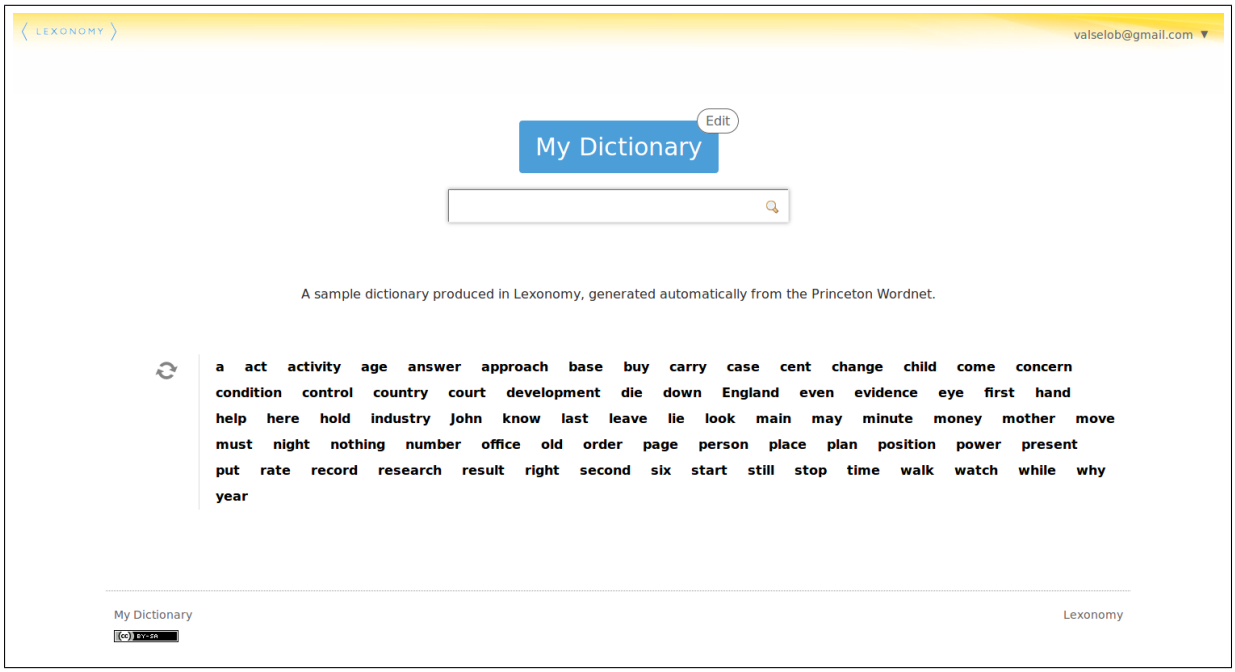


Figure 17: The public homepage of a dictionary

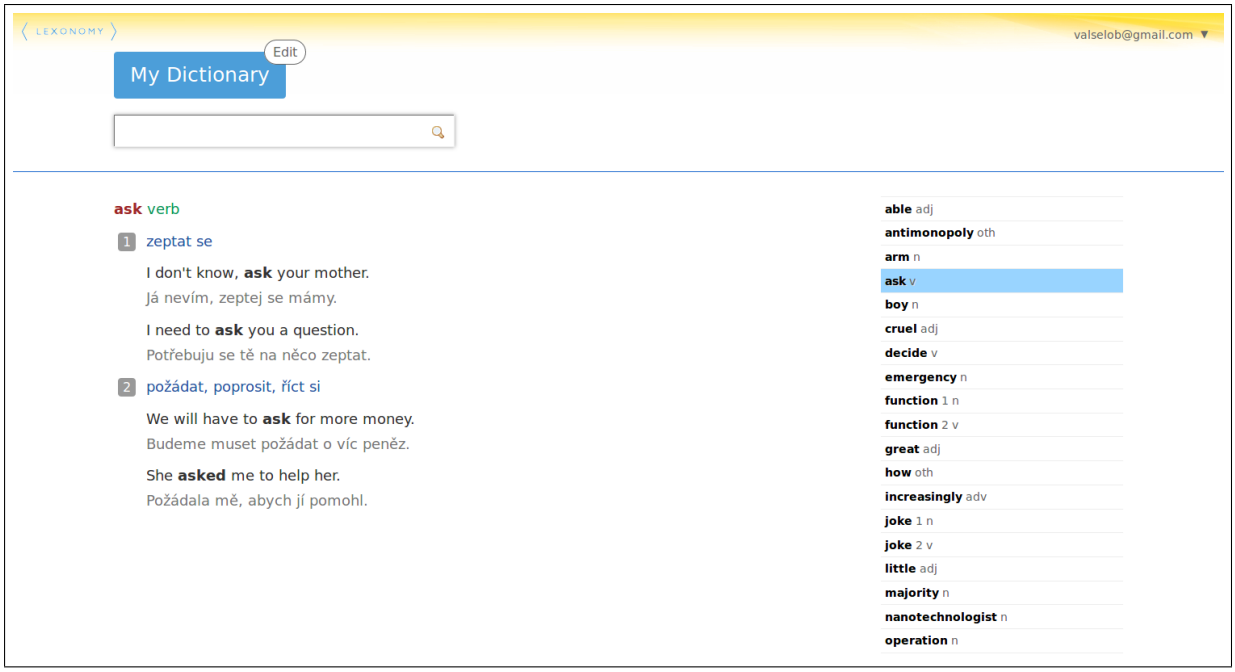


Figure 18: The public page of a dictionary entry

From Printed Materials to Electronic Demonstrative Dictionary – the Story of the National Photocorpus of Polish and its Korean and Vietnamese Descendants

Łukasz Borchmann, Daniel Dzienisiewicz, Piotr Wierzchoń

Institute of Linguistics
Adam Mickiewicz University
Poznań, Poland
E-mail: {borch, dzienis, wierzch} @amu.edu.pl

Abstract

The most popular form of lexicographic exemplification is plain-text transcript. Apart from the doubtless advantages of such a quotation method, it may be perceived as a kind of trade-off when considering readability, accessibility, simplicity, accuracy, and even the logistics of a documentation project. Another approach is to gather and present excerpts in the form in which they were originally published, that is, as the clippings from publications (this is referred to as *photodocumentation*).

The photodocumentary technique is a distinctive feature of both the National Photocorpus of Polish and its Korean and Vietnamese descendants. The main goal of the first of the above-mentioned projects was to describe around 250,000 lexical units, which would be enough to outperform all of the 20th-century dictionaries of Polish. Even more momentously, the process was entirely corpus-driven – that is, all of the principial lexicographic works preceding the project were intentionally ignored. As a result, the material contains largely the words of which linguists were unaware of or which were perceived as later neologisms under leading derivative models of Polish.

This article describes the projects from their early stages, namely the acquisition of printed materials, to the final level of development where an electronic lexicographic tool is made available to both amateur and professional users. Also described is the struggle to avoid unthinking imitation of p-lexicographic techniques. The methodology had to be adapted to meet modern web usability standards.

Keywords: e-lexicography; photodocumentation; corpus linguistics; computational linguistics; digitisation

1. Introduction

Lexicography, from a discipline built around traditional, deeply philological methods, has transformed into an interdisciplinary field involving both linguistics and computer science. This transformation is well reflected in many aspects of the National Photocorpus of Polish (NFJP) project and its Korean and Vietnamese descendants.

Three key ideas behind this lexicographic project are outlined in the following sections.

1.1 Photolexicography

Firstly, the project is based on photolexicography, a documented subdiscipline of applied linguistics in which every lexical unit is presented in exactly the same form as it appeared in print, along with its lexicographically relevant context (see Figure 1).

The method, which originated nearly a decade ago, is still progressing dynamically, not only contributing to the development of the basis for lexico-derivational models of 20th-century Polish, but also finding applications in a variety of new analyses, descriptions and glosses.

The advantage of the photodocumentary approach to quotation is that it prevents the risk of erroneous recreation or inaccurate recording of text, and, what is more, it presents maximally complete information, preserving both the textual contents and the original typographic layout (Małek, 2008; Wierzchoń, 2009).

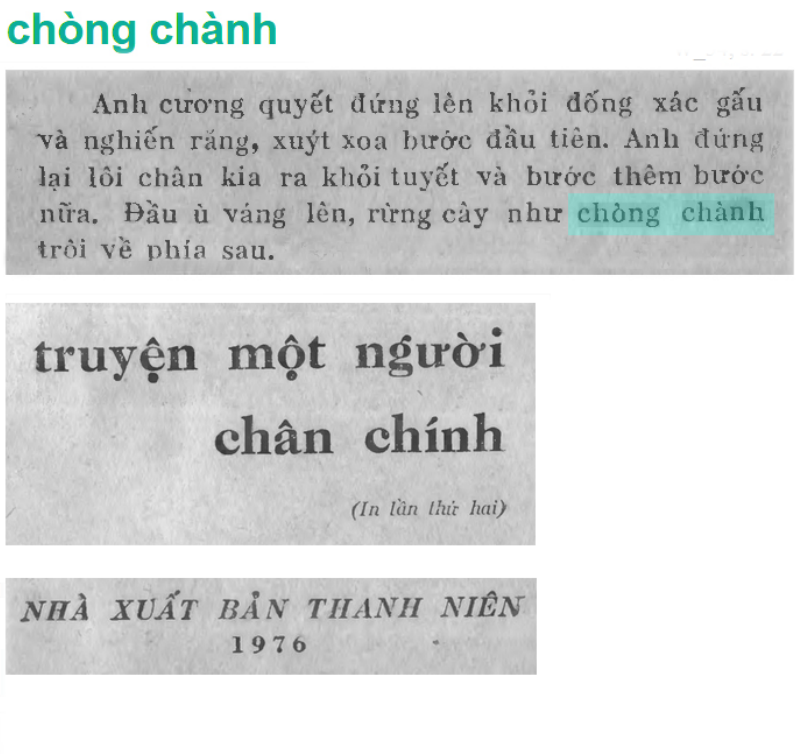


Figure 1: Vietnamese excerpt in the original form, that is, as a clipping from a publication (an example of *photodocumentation*)

1.2 Demonstrative dictionary

Оазис. *Образно.* ...Чтобы спасти *оазис* своей индивидуальности (А.И.Титаренко).

Обвод. По мысли Леонардо да Винчи, первый рисунок – это тень предмета, освещенного костром. Первобытный человек начинает рисовать, осваивая технику «*обвода*». Пещеры сохранили десятки таких примеров (В.Н.Дублянский).

Обгонять. Поменьше запретов – запреты уменьшают самостоятельность, ухудшают мышление, снижают ответственность. «Уверен – *обгоняй*» (Ю.Крелин).

Обезлюдеть. Дедушка рассказывал, что он отчетливо помнит два дня войны – начало и Победу. Он так описывал начало войны: «Из окна было видно море, легкий туман над ним. Солнце. Мы с ребятами проснулись рано, сидели за книжками. И вдруг за городом забухали пушки, зенитные разрывы в синеве неба, вой самолетов. Самым тягостным и тревожным было в тот день то, что Феодосия, казалось, враз *обезлюдела*. На улицах только военные» (ИМС, М.Полякова).

Figure 2: Extracted from *Словарь богатств русского языка*

Secondly, the NFJP project aims to create a demonstrative dictionary – a new type of work with its origins in Russian lexicography, as described in the 2003 work ***Словарь богатств русского языка*** (Figure 2; Kharchenko, 2003).

The authors of the original demonstrative dictionary aimed to present the wealth of the language and its curiosities of which people become unaware through everyday experience (Bobunova, 2013: 180). Aimed at the promotion of the lexical abundance of the Russian language, the project popularised, among others (Kharchenko, 2015):

- rare words discovered in texts and historical dictionaries, recorded with a view to reviving them;
- aphorisms that are not commonly known, mostly taken from the works of local writers from the 1970s, 1980s and 1990s;
- extracts from literary, popular-scientific and scientific texts where a given word was used in such a way that it deserved recognition and quotation;
- *biographemes* (*биографемы*), namely microdescriptions of family history and genealogical notes;
- attestations of the use of metaphors in the periods in which they were formed and when the motivational basis for formulating them was clear.

The above list does not exhaust the contents of the dictionary, but it enables us to comprehend the intentions of its authors of the enterprise. It also records idioms, sayings, proper names and lexical items used solely by particular authors.

There are numerous analogies between the premises of a photocorpus and the concept of a demonstrative dictionary, which lead us to consider NFJP a distinctive variety of the latter, referring to a related lexicographic tradition and a similar means of preservation and promotion of a national legacy.

Despite the fact that the two projects are closely related, one can distinguish methodological differences, which is evidenced by the fact that in its nature the demonstrative dictionary is a traditional work and the material contained in it is a result of decades of manual *gathering of words* (Kharchenko, 2015), as such an activity is described by (Małek, 2008).

1.3 Electronic lexicography

Thirdly, not only is NFJP a repository of lexical inventory, but it is also an e-lexicographic tool (for instance, involving such features as e.g. morphological tagging and searching with the use regular expressions – see Section 3.1).

Nowadays both the theory and practice of lexicography are deeply rooted in information science, which is reflected in the present work as well as in the NFJP project and methodology.

With the transformation of lexicography, the issue arose as to whether a theory setting a new direction for computational studies should be devised. Some claimed, however, that lexicographers should adhere to the concepts dating from the era of p-lexicography. A potential advantage of electronic dictionaries over traditional ones, as noted by (Nichols, 2010), is liberation from the limits set by the space taken by entries concerning their number and exemplifications as well as the length of the definition. Such limits are practically non-existent in the case of electronic dictionaries.

In the pre-electronic era the immediate elimination of errors was impossible – this difference is also indicated by (Nichols, 2010), who states that error correction can be performed online at any moment.

The above-mentioned possibilities can be recognised as reactions to problems of which traditional lexicographers are commonly aware. The advantage of e-lexicography is the

fact that a website constitutes a much more effective material than paper, due to its interactivity.

As a point of reference, one may consider a division of e-lexicographic tools into four categories (Tarp, 2011: 57–62):

1. digitised dictionaries, originally published in paper form;
2. dictionaries originally developed in a digital form, although with data structured as in traditional dictionaries – despite the more effective access (e.g. due to the headword search function) these are projects based on *utraditional models and concepts which have been taken over uncritically from the era of p-lexicography*;
3. tools with *dynamic contents and dynamically generated data*, crossing the borders of conventional lexicography, offering configurable functions enabling the dictionary to be adjusted to specific needs and expectations;
4. e-lexicographic tools, that are expected to be implemented in the future, which will enable one to combine the data from a previously prepared database with the data accessed online, so that it will be possible *de facto* to create and re-represent entries in real time.

One may familiarise oneself with real interactivity through two existing collections. These examples of projects from the third of the above categories are *Den Danske Ordbog* and the *Macmillan Dictionary and Thesaurus*.

Contrary to that which traditionally oriented scholars might claim, abandoning the idea of planning and developing a dictionary in its traditional form is a necessary step in order to access the broader perspective of contemporary lexicographic tools (Gouws, 2011).

Viewing online dictionaries as a search tool and abandoning the vision of a repository containing data or a conventional dictionary, allows their usability to be tested in a way which has been successfully applied to IT systems (see Heid, 2011).

1.4 Photographic quotation: a desirable practice or a foreign body in the world of e-lexicography?

The description contained in the preceding section may give the impression that a photographic quotation is in some ways incompatible with the idea of e-lexicography, and that NFJP might be considered an example of a project based on uncritically acquired models and concepts from the era of p-lexicography, as it was put by (Tarp, 2011).

The methods applied in the process of searching for textual attestations and edition of entries undoubtedly fit within the discipline of computational lexicography, and are far removed from the traditional conservative approach to lexicography (Piotrowski, 2001; Atkins & Zampolli, 1994; Boas, 2009). Is it not the case that a photographic quotation, being a digitised form of paper material, reintroduces old models and concepts into a world which has the aim of reforming them? A text presented in the form of raster graphics resembles the worst practices of website creation.

To avoid this situation, actions were taken to adapt the concept of photographic quotation developed for paper publications to the reality of modern lexicographic applications. While

photographic quotations were still demanded for each item, the contents of the exemplum were also required in the form of regular text. At the present stage of development of the project, this is text that is recognised automatically. In the future, manual verification will be made possible.

An exemplum obtained in such a way is used as the alternative text of a photographic quotation (for search engine robots and people with disabilities), but with the help of developed tools, phonetic transcription would be possible, for instance. In this way we attempt to combine the accuracy of documentation with the possibilities related to access to the content of the quotation.

Naturally, the above discussion does not exhaust the issue of the position of NFJP in the world of contemporary e-lexicography – this question, considered in more general terms, is addressed in the next section. The present study describes the projects from their early stages, namely the acquisition of printed materials, to the final level of development where an electronic lexicographic tool is made available to both amateur and professional users.

2. The process

Not to mention the problems of digitisation, difficulties abound even when the materials have already been scanned, analysed with OCR software and tokenised. Because of OCR errors, some kind of positive lookup is helpful in order to select promising lexical units for further analysis.

The following sections describe these difficulties, as well as the process of verification and editing of units by qualified annotators. Figure 3 is an illustration of the entire process of creating the NFJP resource described in this part of the article, and may be helpful in resolving any ambiguities.

2.1 Acquisition, preparation and preprocessing of the materials

At the current stage of the project's development, materials from in-house digitisation (referred to as the *non-electronic canon*) have been used in addition to materials from Polish digital libraries (the *electronic canon*). The non-electronic canon consists of approximately 4,000 books received free from non-electronic libraries which planned to recycle them, while 2,000 additional books from the electronic canon were selected to balance the corpora diachronically.

Information exchange at Polish digital libraries takes place using the OAI protocol. Most of the publications stored by *dLibra*[1] are from the pre-war period, up to 1939. The digital libraries also store various types of collections (printed matter, press cuttings, audiovisual materials). As a result, over a period of more than 10 years, a collection of over three million digitised library items has been built up. This material is described according to the Dublin Core scheme.

Unfortunately, the Polish digital library system does not offer normalised metadata, such as publication type or even year of publication, which are vital for many purposes. The

[1] A program used for the collection, editing and sharing of digital publications, developed at the *Poznań Supercomputing and Networking Centre*.

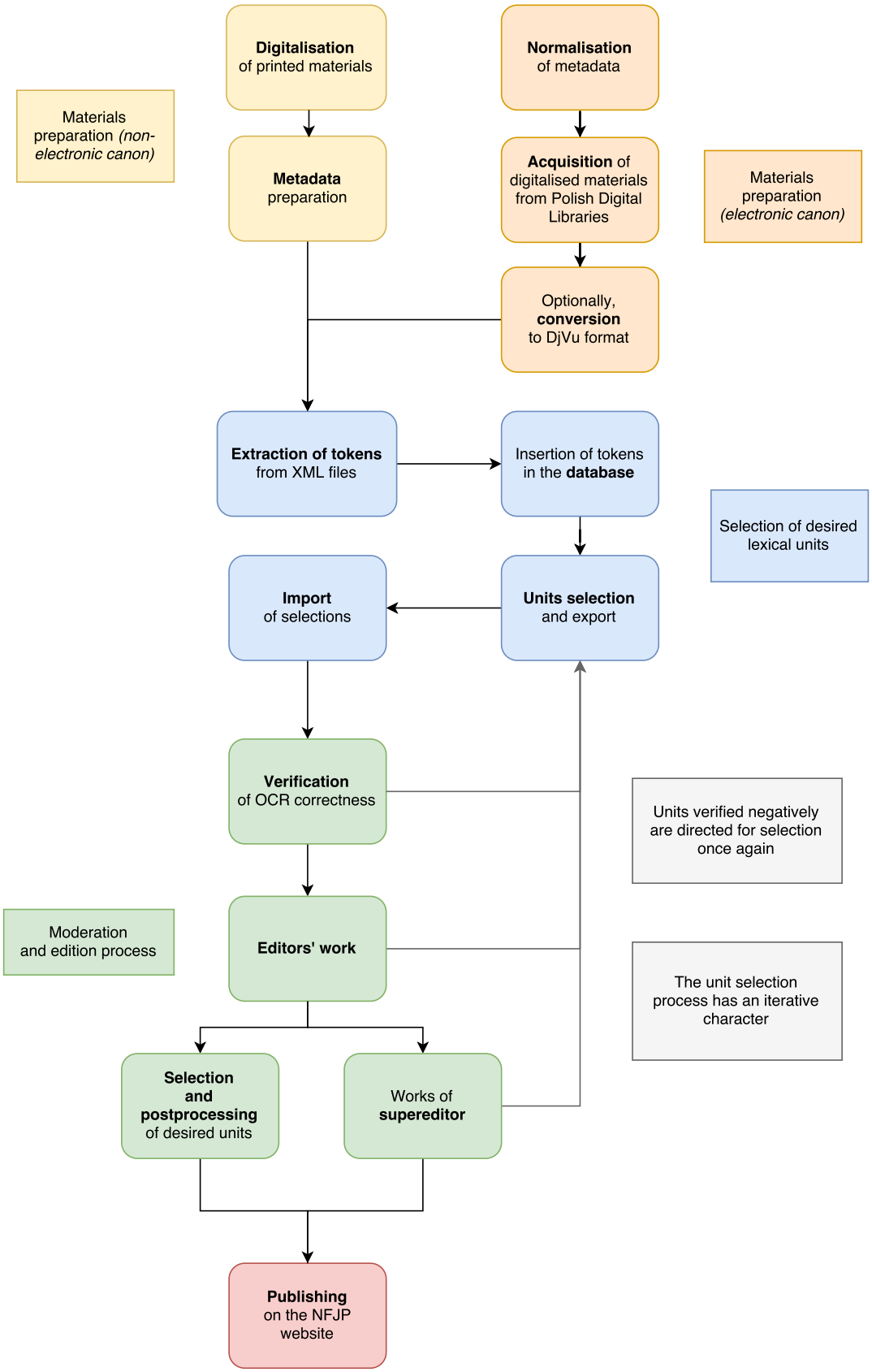


Figure 3: The process of creating the NFJP resource

structured data available via the OAI-PMH mechanism contain subject, type and date elements, but the practice of their use varies between and within libraries, so that automatic or semi-automatic normalisation had to be performed to convert this data to a form that would be easily usable by a computer program. Consider, for example, the following instances of text contained in the date field:

1884	rok obiegu 1940	1944 (Ausgabe Nr 1)
20 stycznia 2010	[ca 1914]	1850 ?
[post 1741]	b.d.	[ok. 1850]
[ok. 1930]	[192?]	[post 1658]
1920.03.27	1877	1800/1900
1936.11.18	22 II 1763	lata miedzywojenne
1785-1819	ante 1945	lata 30. XX w.
1983-	19w.	poczatek XIX w.
[XVIII/XIXw.]	12 III 1763	
mar-09	[1836]	
1852 November	27-lut-08	

Moreover, resources are available in different file types, so that within one digital library some publications may be published as multiple PDF files, and others as single or multiple DjVu files.

Before further processing, the materials obtained from these two heterogeneous sources were unified to single DjVu files, and for each of them XML files containing information about the text layer were created (with the use of the *djvutoxml* command from the *DjVuLibre* package). Years of publication from the electronic canon were normalised using a rule-based algorithm which selected the most pessimistic option, that is, the last year valid for a given textual date or period. Not only the date field was used, but also the title, which sometimes contains a more specific date (for example, there are cases where the date field contains a period, while there is a four-digit year within that period available in the title field).

2.2 Selection of lexical units for further processing

The content of an XML *word* tag was treated as a token, normalised, and inserted into a relational database with the structure presented in Figure 4 (names of tables and fields are self-explanatory). Obviously, not all of the unique tokens are correct Polish words (in fact, only around 10–15% are). To ensure low editing costs, because of OCR errors some kind of positive lookup needed to be used to select only promising lexical units for further analysis.

The first method that comes to mind is the use of dictionaries, and naturally this was attempted. However, the intention was to apply also a more sophisticated solution involving the generation of *verba possibilia*.

This term was coined to describe artificially created words on the basis of how morphological derivation works in a particular language. These few examples shed light on the method:

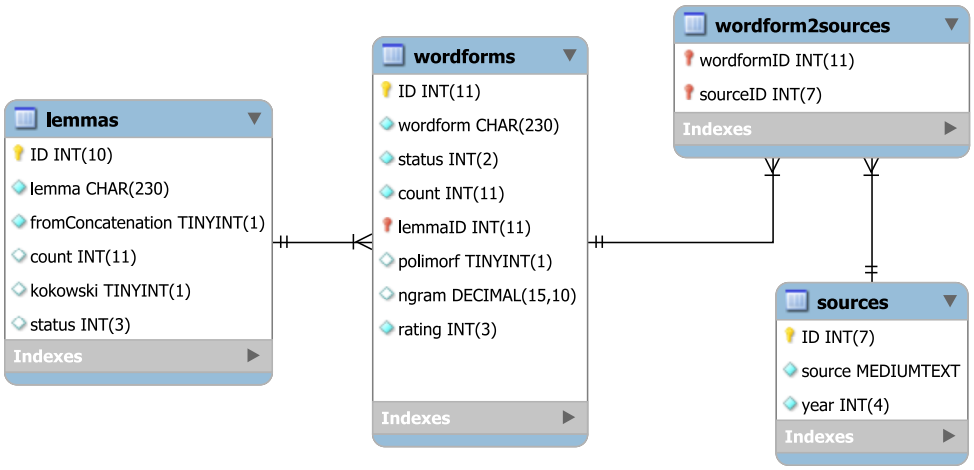


Figure 4: Schema of the database used in the process of selection of lexical units

- *naukowoczysty* 'scientifically clean' (concatenation of *naukowo* 'scientifically' and *czysty* 'clean');
- *panna-wdowa* 'spinster-widow';
- *samozaciemnienie* 'self-blackout' (concatenation of *samo* 'self' and *zaciemnienie* 'blackout').

One can also formulate rules to create unknown but probable words using the right-sided derivation, for example, using the equivalent of the English suffix *-zation/-sation* – Polish *-zacja*, Vietnamese *hóa* or Korean *화* (hwa):

- bình thường hóa 'normalisation'
- cách mạng hóa 'revolutionisation'
- chính thức hóa '*officialisation' (forms marked * probably do not exist within the English language, but the assumption that they will never be used in texts would be unreasonable)
- hoạt hóa '*activisation'
- hợp lý hóa 'organisation'
- 표준화 'standardisation'
- 세계화 'globalisation'
- 식민지화 'colonisation'

Many more unexpected findings can be obtained using two other methods applied within the NFJP project. The first of them is based on the assumption that unrecognised tokens that appear in a text in the context of known words are more likely to be correct Polish words than those which are never present in such a context. The second is the simple character-level n-gram word model (Jurafsky & Martin, 2000).

2.3 Verification and editing process

To verify the correctness of OCR and tokenisation, the panel shown in Figure 5 was prepared (the one shown was used during the preparation of the Great Photocorpus of

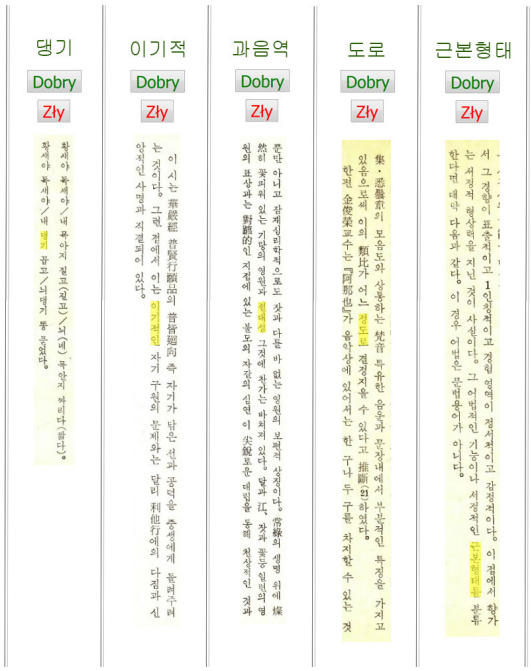


Figure 5: Initial verification of OCR for the purposes of the Great Photocorpus of Korean. The task of the reviewer was simply to check whether the highlighted word was equal to that recognised automatically

Figure 6: Editor's panel – part presenting the analysed unit

Korean, described more profoundly in Section 4; in case of other language variants it is analogous).

The approved units are then reviewed and annotated by editors with a strong linguistic background, who determine the lemma, the part of speech (in the case of phrases, instead of verb, for instance, verb phrase is presented as an option), and other grammatical categories (Figure 6). For the purposes of editing they are able to see the usage of the word in a broader context, up to the whole page.

During initial photodocumentation work, excerpts were cropped manually, as they were expected to meet certain rigorous conditions. Subsequently, as projects became more and more massive, steps were taken to make the cropping process fully automatic. Somewhat unexpectedly, the results of automatic methods proved to be indistinguishable from the manual ones, even without the use of machine-learning solutions. The currently utilised script uses heuristic methods based on recognised orthographic text (so as to take sentence beginnings and endings into account) and words' coordinates.

3. Functionality

The NFJP project is currently a fully functioning website, providing useful features for both amateur and professional users (see Figure 7 presenting entry structure). There are some new advanced features that will be released shortly; these will be discussed in a separate section below.

3.1 Publicly available

3.1.1 REGEX-based searching

The NFJP engine allows one to use Perl Compatible Regular Expressions while performing a search action. A systematic description of this formalism is not an aim of this work, thus we present only a few examples below.

The `$` character in REGEX syntax stands for an anchor to the end of the string. Thus the query `stylowy$` 'stylish, in style' would return results such as *ponadstylowy* 'abovestylish', *neostylowy* 'neostylish' and *emocjonalno-stylowy* 'emotionally-stylish'. Similarly, the `^` character matches the start of the string to which the regex pattern is applied; thus the query `^pseudo` would return such words as *pseudozdrajca* 'pseudotraitor', *pseudowynalazca* 'pseudoinventor', *pseudoszwabacha* 'pseudoschwabacher (a specific blackletter typeface)'.

A slightly more advanced example of a regular expression is `^.{4}$`, which returns words consisting of exactly four characters.

For more advanced examples of regular expressions usage see Friedl (2006), Good (2004) and Stubblebine (2003).

3.1.2 Search operators

Modern search engines provide a feature allowing one to make search results more precise using so-called search operators. A similar solution is implemented in the National Photocorpus.

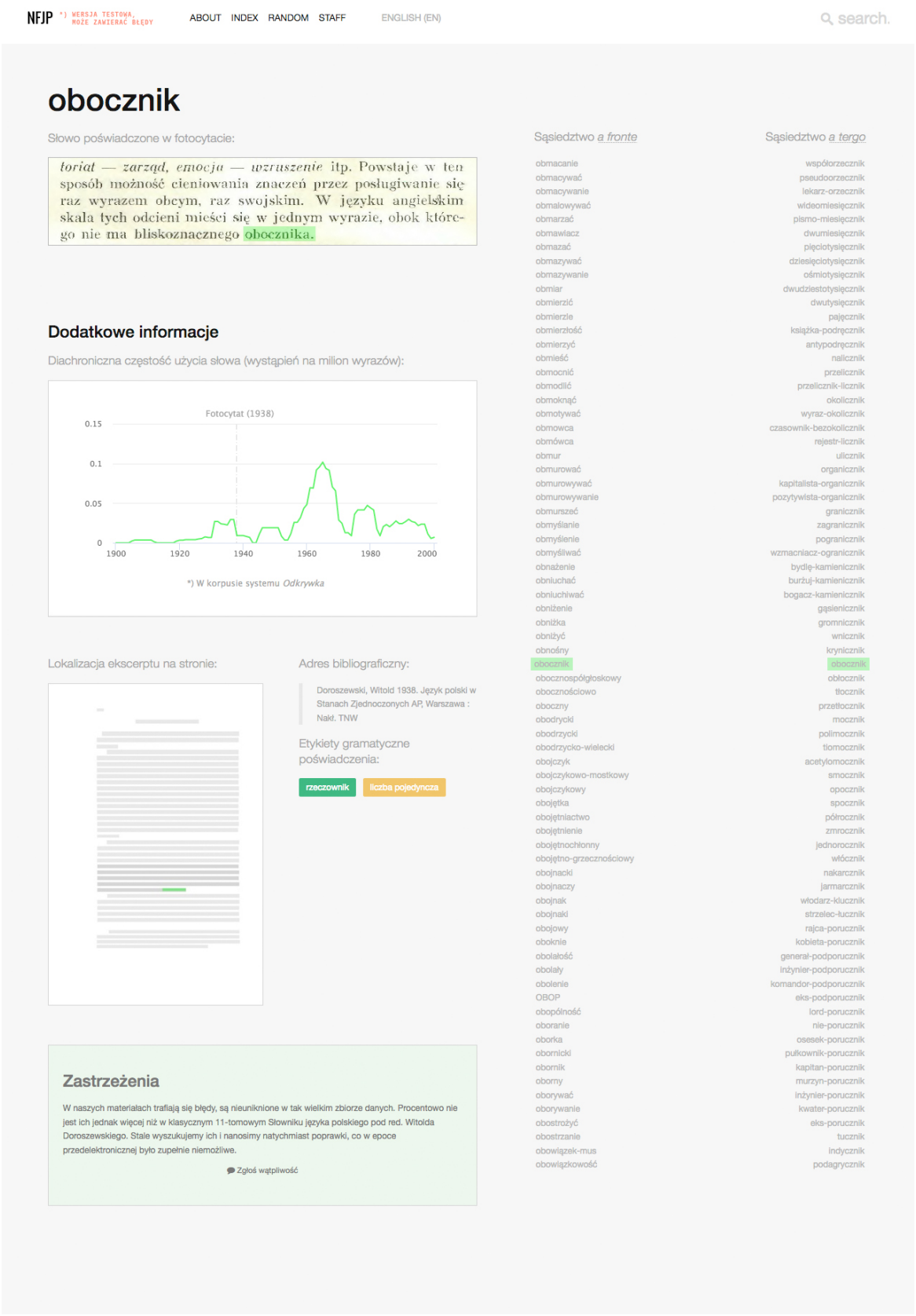


Figure 7: View of the entry for the word *obocznik*

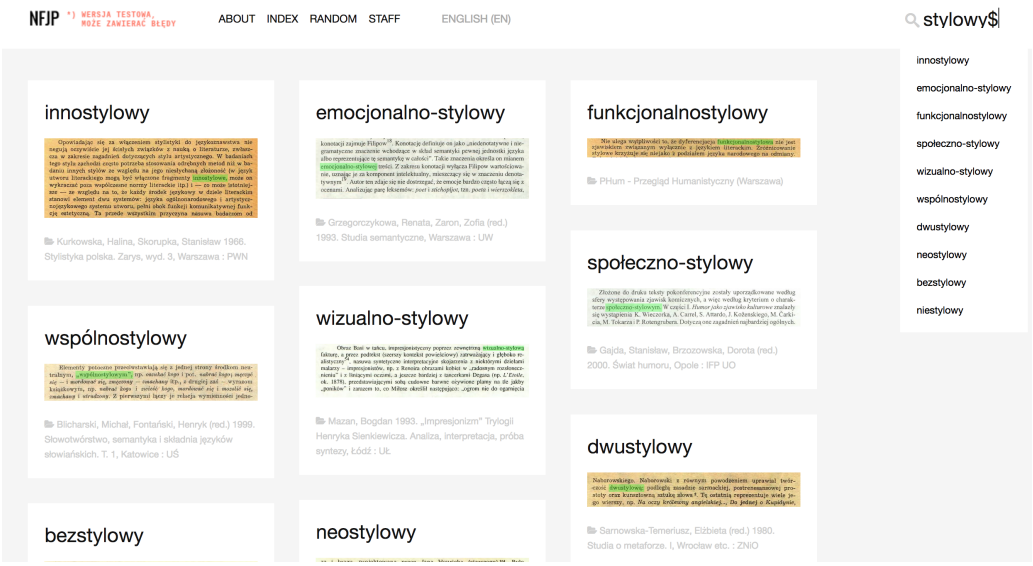


Figure 8: Results obtained with the use of a regular expression

Part of speech. Using the *pos* operator one can return results matching only the selected part of speech. Available values are: *verb, part, num, particle, pred, prep, adj, adv, subst, conj, interj, ppron, other*. For example, adding `pos:adj` to a query will cause it to return only adjectives.

Number. The string `number:pl` in a query will restrict the results to plurals only. Other available values are *sg, pt* (pluralia tantum) and *du* (dual).

Source. The string `source:IJ_698` in the search input will return only words found in the book *Encyklopedia techniki. Przemysł spożywczy* (Banecki et al., 1978), because *IJ_698* is its ID within the system.

Reflexive form. For the purposes of binary features, the feature operator was introduced. At present it allows one to restrict the results to reflexive verbs using `feature:reflexivum`.

Multiple search operators can be used in one query and they can be combined with regular expressions. For example `^s source:IJ_2788` will return words beginning with the letter *s* from the source with the selected ID.

3.1.3 *A fronte* and *a tergo* neighbourhood

On the details page of each entry, *a fronte* and *a tergo* neighbourhoods are presented. For example, for the entry *ślimaczenie sie* such a neighbourhood is:

ślicznіuchny	próżniaczenie
ślicznіutki	półmajaczenie
ślicznіutko	żydłaczenie
ślicznotka	rozkułaczenie
ślimaczenie sie	**(sie) ślimaczenie**
ślimaczo	przysmaczenie
ślimakowato	re-tłumaczenie
ślimakowo	przetłumaczenie
ślimakowo-wirnikowy	idiotłumaczenie

On the NFJP website 36 words above and below the displayed unit are visible (Figure 7), which is useful particularly in a research regarding word formation and inflection (Grzegorczykowa & Puzynina, 1973; Obrebska-Jabłońska et al., 1968).

3.1.4 Other features and materials

For each of the words relative usage frequency is shown, within the period 1900–2000 (count per million words in publications from each year). See Figure 9.

The website also contains materials in five languages (Polish, German, English, Russian and Japanese) describing the purpose of the project, its methodology and the significance of the results, as well as information regarding other projects focused on Polish vocabulary undertaken prior to NFJP, a bibliography, and a library containing information about all of the publications describing NFJP materials.

3.2 Case studies

3.2.1 Lexical inventions of Adolf Nowaczyński

The authors of the work *Archikastrat, emancypaństwo i krytykretyni...* analysed the linguistic creativity of Adolf Nowaczyński, a Polish writer, poet, playwright, critic, and social and political activist (Dzienisiewicz et al., 2017).

In the course of the analysis the authors distinguished five categories: words which had been commonly used before they first appeared in Nowaczyński's works (A), words which had occurred several times before they first appeared in Nowaczyński's works (B), words with single or several occurrences after they first appeared in Nowaczyński's works (C), words whose use might have originated within Nowaczyński's idiolect (D), and words discovered solely in Nowaczyński's writings (E).

To perform analyses of this type, one may utilise two functions available in NFJP: the diachronic frequency of a word, and the search operator source:, allowing one to select all of the units recorded for the first time in a given publication.

One of the publications included in the NFJP canon is *Góry z piasku* by Adolf Nowaczyński, where such units as *afiszowość, aluzjonizm, junaczość, omłacanie, katastrefa, powstydzenie, wyklecić, proteuszowo, regencki, renomista, zniewieścialec, lubownictwo, nieobmieciony, nawałesać sie, mieszczuszek, nieprzyłaczony, nierozpowity, nierozjatrzanie,*

niedźwigajacy, niekabłakowaty, nieświatowość, oblagowywanie and *złotorunny* were discovered.

Most of the presented words are especially interesting in terms of their word-formative features, e.g. *zniewiescialec* (a personal noun denoting 'an effeminate man'), *złotorunny* (an adjective derived from the phrase 'Golden Fleece'), *powstydzenie* (an unusual form of the word 'ashamedness' with the prefix *po-*; the common Polish form is *zawstydzenie*), *mieszczuszek* ('a little city slicker'; an original example of the use of the diminutive suffix *-ek*).

Some of the above units were included in the categories devised by the authors; however, some of them were not recorded by them, although they meet the criteria for category E, that is, words discovered only in Nowaczyński's writings. The corpus of the Discovermat system (which served as a point of reference for the authors) returns one result for the query *junaczość* from an article by Nowaczyński published in *Nowy Przeglad Literatury i Sztuki*.

3.2.2 NRF and RFN

In the period of the Polish People's Republic two names were used to denote Western Germany, namely, *Niemiecka Republika Federalna* (NRF) and *Republika Federalna Niemiec* (RFN). Both abbreviations are included in NFJP, thus their diachronic frequency of occurrence in texts can be traced (Figure 9; Dzienisiewicz, 2017).

3.3 Russian and Soviet lexical borrowings

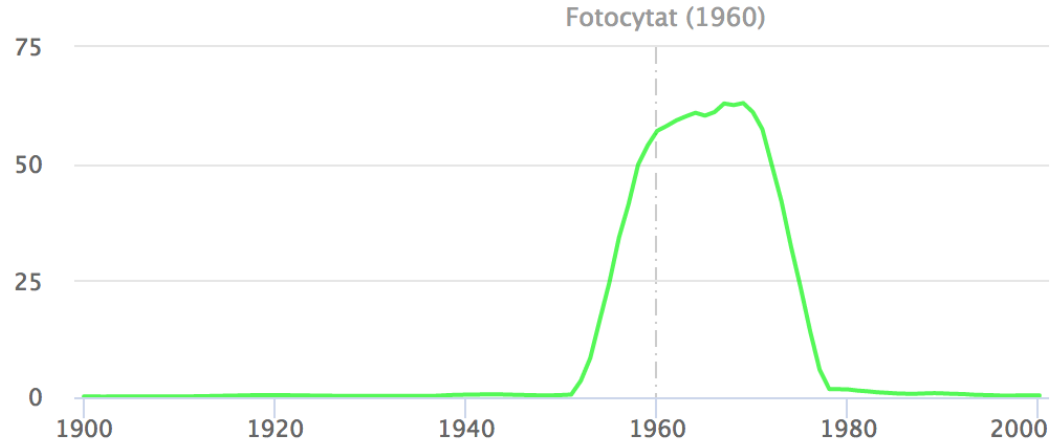
The list of publications available on the NFJP website enables one to distinguish several groups of sources which might include Russian and Soviet lexical borrowings, that is (Wawrzyńczyk, 2014):

- translations of Russian literary works (Chekhov, Dostoyevsky, Gogol, Lermontov, Pushkin, Solzhenitsyn, Tolstoy);
- translations of journalistic writings, diaries, letters and scholarly texts of, among others, Byelinsky, Herzen, Dostoyevsky, Zinovyev, Likhachov;
- diaries and correspondence of the Polish people who were sent to Russia and the USSR;
- works by Polish authors who lived in the Russian Partition.

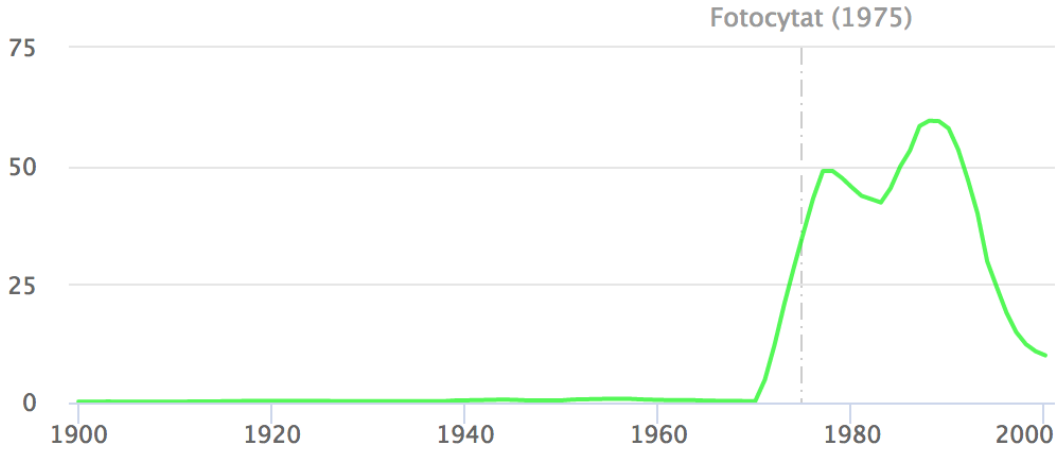
Using the source: operator one can obtain a list of words recorded for the first time in the above works. Even a cursory overview of the units brings to light some which might be of interest to scholars specialising in Russian borrowings, as it includes the following words: *niepuszkinowski, grażdański, sowchozowy, sofista-słowianofil, półimperiał*.

Even more interesting cases of words can be found in Pushkin's works (included in NFJP):

- the word *niedaleczko* discovered in the saying *Rzekłbym słóweczko, lecz wilk niedaleczko* (***Сказал бы словечко, да волк недалечко***, 'walls have ears');



— Chyba pomyłka, Panie Inżynierze — odpowiadał Doktor Piotrowski z maleńką nutką niepokoju w głosie. — To było specjalnie dla mnie sprowadzane z NRF-u. Zygmuś mi się zresztą z tym szybko uwinął. Choć co prawda słono extra kosztowało, ale miałem wóz na chodzie raz dwa. Ten chłopak po prostu cuda umie robić! Nie pytam już nawet jak...



takich zapisów nie spotyka się wcale. Oczywistą sprawą jest także to, że końcówki przypadkowe wyrażają skrótowce w wyższym stopniu zleksykalizowane, częściej używane lub należące do warstwy słownictwa potocznego, np. *piłkarze LZS-ów*, *RFN-u*, ale *ZSRR*.

Figure 9: Photographic quotations and diachronic frequency of occurrence of NRF (upper graph) and RFN (lower graph)

- the word *dych*, which appeared in the expression *ani słychu, ani dychu* (***Ни слуху ни духу***, 'there has been no news of somebody or something').

With the use of the described method a large-scale analysis of Russian borrowings can be conducted on the materials contained in NFJP.

3.4 Features to be released shortly

3.4.1 Morpheme segmentation

The automation of morpheme segmentation is not a trivial task and can be performed in various ways. Considering the fact that there are no large sets of annotated data for many languages and that creating them requires a huge amount of work, solutions based on unsupervised machine learning (Creutz & Lagus, 2007, 2005; Goldsmith, 2001) and minimally supervised machine learning techniques are popular. In the latter case models are learned from a small number of segmented words and a large number of unsegmented words (Ruokolainen et al., 2016). Fortunately, there are publications for Polish that make supervised machine learning techniques applicable without the need for additional annotating efforts, so that we can easily compare the performance of both approaches.

For the purposes of supervised machine learning two volumes of *The Dictionary of Derivational Nests of Modern Polish* were used (Jadacka & Bondkowska, 2002; Vogelgesang, 2001) with a total of 50,000 words. They required a pre-processing stage before performing supervised learning, because the format used was not segmented orthographic text. The only methodological difference between source segmentation and the one used in the described set is the abandoning of the null morpheme concept, which has no rational motivation in morpheme segmentation (nor in linguistics in general, cf. Mańczak, 1996: 11).

During the work the above set was split into random training and test subsets to perform cross-validation. The rule-based model was used as a baseline for machine learning techniques. It is similar to the one described by Yang (2007) but is simpler and based on a predefined list of morphemes.

In terms of supervised machine learning techniques, the problem of morpheme segmentation can be treated as a problem of binary classification, that is whether the morpheme boundary should or should not be placed between certain letters in a word (this approach is similar to the one described by Neubig et al., 2011 for Japanese). In order to determine the best classifier for this purpose, various methods available in the *scikit-learn* Python library were tested (Pedregosa et al., 2011). For each of the classifiers Confusion matrix was computed as well as other evaluation metrics, such as Accuracy, F1 score and Matthews correlation coefficient (MCC).

The optimal set of features seems to be similar to some of the features proposed for Arabic by (Monroe et al., 2014). In the case of the Polish language it consists of:

- a five-character window around the analysed character boundary;
- character n-grams made from the current character and up to the next four characters;

- character n-grams made from the current character and up to the previous four characters.

From the methods available within *scikit-learn*, only Decision Trees offers comparable results. Although the results of Decision Trees are weaker than those obtained using a linear Support Vector Classifier, its moderate effectiveness encourages us to check the results of combining both Decision Trees and SVC, using for instance a Voting Classifier. The idea is to combine different machine learning classifiers and use the average of the predicted probabilities offered by each of the combined methods. The method described, however, does not produce significantly better results.

A different approach to morpheme segmentation is to use a Conditional Random Fields statistical sequence modelling framework (Tseng et al., 2005). The problem is basically to predict a vector $y = \{y_0, y_1, \ldots, y_T\}$ of variables for a feature vector x. It can be solved by learning an independent per-position classifier that maps $x \mapsto y_s$ for each s, as was done in the above section, ignoring the sequential aspect of the data. By contrast, Conditional Random Fields refers to neighbouring samples and predicts a sequence of labels for a sequence of input sample (Sutton & McCallum, 2012).

For the purposes of this work, CRFsuite was used (Okazaki, 2007). This offers various training methods (such as Limited-memory BFGS, Orthant-Wise Limited-memory Quasi-Newton, Stochastic Gradient Descent, Averaged Perceptro, Passive Aggressive, Adaptive Regularization Of Weight Vector) and simple TSV input format.

The final CRF-based solution performed as efficiently as the best SVM-based solution in terms of evaluation metrics, even though it seems to outperform it when examining the results. It uses the Passive Aggressive training method (Crammer et al., 2006) and the following features (let c[t] be the current character in a word):

- a five-character window around the analysed character boundary (c[t- 2]|c[t-1]|c[t]|c[t+1]|c[t+2]);
- character n-grams made from the current character and up to four following characters (e.g. c[t]|c[t+1] for a bigram);
- character n-grams made from the current character and up to four previous characters (e.g. c[t-2]| c[t-1]|c[t] for a trigram);
- every single character within the word identified as e.g. c[t-4];
- c[t-2]|c[t-1] and c[t+1]|c[t+2];
- c[t-2]|c[t] and c[t]|c[t+2]=n|e.

Moreover, a family of methods for unsupervised learning of morphological segmentation was tested (e.g. one utilizing probabilistic generative models), as well as semi- (minimally) supervised machine learning (including a model trained on the full *National Corpus of Polish* skipping compounds with a random probability, this being expected to speed up the training considerably with only a minor loss in model performance; cf. Virpioja et al., 2013).

None of these attempts, however, resulted in a level of performance comparable to those obtained using the final SVM- and CRF-based models.

The features proposed in the literature for unrelated languages such as Chinese and Japanese are applicable to Polish with only minor modifications. The fact that the performance limit for three conceptually different methods stands at a similar level suggests that it is either a limit of machine learning methods (at least at this level of advancement) or a limit of training on the data set described in this paper. Observation of incorrect classifications reveals that they are sometimes related to the idea behind the *Dictionary of Derivational Nests of Modern Polish*, where some derivatives are presented without inherited morphological structure. This supports the second hypothesis.

Future work will focus on developing better training sets and on testing deep learning methods, as well as other ensemble combinations. Independently of this, the solution described in the present chapter is production-ready, and will be released shortly on the NFJP website.

3.5 Phonetic and phonematic transcription

Maria Steffen-Batóg proposed mechanisms of phonetic and phonematic transcription for Polish, based solely on the character context of a particular letter. The algorithm assumes iterative reading of input orthographic text (character by character) and matching of appropriate left and right context definitions from the tables of rules created by Steffen-Batogowa (1975) and Steffen-Batóg & Nowakowski (1997). In each of the tables the first row contains a formal definition of the right context, and the first column a definition of the left context. The proper transcription can be found at the intersection of the matching definitions.

The proposed formal definitions of left and right context (ca. 500 unique descriptions and many more combinations) were implemented using regular expressions. The correctness of the algorithm is currently being checked on the vast material of NFJP, and required fixes are continuously applied.

3.6 The formal definition of neologism

Matyka (2010) formulated three questions regarding neologisms:

- How can one objectively check whether a word is a new one?
- How one can determine its age?
- When should a lexicographer assume that a neologism is old enough to place it in his dictionary?

Answers to these and similar questions should consider that a word may be widespread within one group, but completely unknown within another.

For this purpose the Herfindahl–Hirschman Index was adapted. This is a measure of the size of companies in relation to their industry, widely applied in competition law as an indicator of the degree of competition (Calkins, 1983). It is expressed as the sum of squares of the shares:

$$HHI = \sum_{i=1}^{N} s_i^2$$

The HHI is the same as Simpson's index (Magurran, 1988: 39–40) used in ecology to measure the concentration of individuals classified into types (the two indices were proposed independently for analogous purposes). The HHI has also been used outside these fields, for instance to quantify level of political competition (Davidson et al., 2008).

In our case it reflects the concentration or dispersion of word usage among sources. A high value means that there are only a few sources to which the majority of word usage cases belong. The smaller the value, the greater the dispersion of the word among sources from a given year.

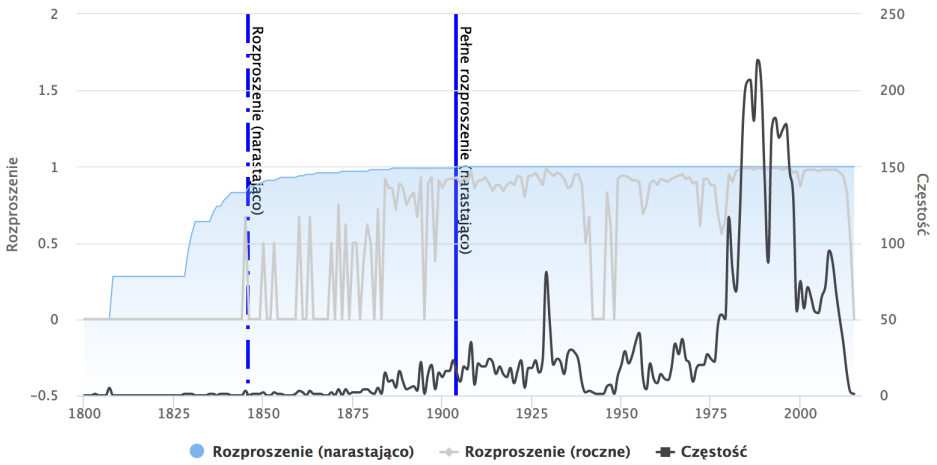


Figure 10: Word usage dispersion

In Figure 10 vertical lines denote some key moments, namely when HHI for the first time took a value smaller than 0.2 (interpreted in law and economics as unconcentrated industry) and the value 0 (highly competitive industry).

4. Discussion and perspectives

In the course of the development of NFJP, other e-lexicographic projects were derived from the original undertaking, namely the Great Photocorpus of 20th-Century Vietnamese and the Great Photocorpus of Korean. Created with the use of techniques developed while working on NFJP, the new enterprises provide us with some insights about the application of the original methodology to languages that are genetically unrelated to Polish.

Because in Vietnamese spaces are used not only to separate words, but also syllables (which may be words in themselves), from the perspective of photodocumentation procedures and software developed originally for Indo-European languages, such as Polish, an attempt to process Vietnamese words resembles in some way a multi-word expression analysis. Indeed, what we have done is treat Vietnamese words exactly as Polish multiword units within our system. The main difference relates to the above-mentioned problem; however, it is common to almost every natural language processing task involving Vietnamese, and thus has well-established solutions proposed in the literature. We decided to rely on the vnTokenizer, utilising the hybrid approach to word segmentation (Hông Phuong et al., 2008).

Figure 11: Newspaper from the 1970s with headlines written horizontally and article content vertically

In the Korean project a new problem arises, related solely to the automatic excerpt generation mechanism: text can be written either horizontally from left to right or vertically from top to bottom. What is more, both writing styles may be used on the same page, as shown in Figure 11.

The rest of the workflow, for both Korean and Vietnamese, remains almost entirely the same.

Despite the advancement of some features presented in this paper, plans are much more ambitious – for example, we intend to use methods generally not applied in the humanities, such as *word2vec* software, which can be used to determine semantic and syntactic relations between words (Mikolov et al., 2013c,a,b). These can be used in many ways – from simple visualisation of semantics to finding diachronic synonyms of a word and tracking changes of word meanings.

The future is near and will be even more e-.

5. Acknowledgements

Work supported by the Polish Ministry of Science and Higher Education under the National Programme for Development of the Humanities, 0014/NPRH3/H11/82/2014, *Narodowy Fotokorpus Jezyka Polskiego. Fotodokumentacja słownictwa XX w.* (National Photocorpus of the Polish Language).

6. References

Atkins, B. & Zampolli, A. (1994). *Computational approaches to the lexicon.* Oxford University Press.

Boas, H.C. (2009). *Multilingual FrameNets in Computational Lexicography: Methods and Applications.* Trends in Linguistics. Studies and Monographs 200. Mouton de Gruyter, 1 edition.

Bobunova, M. (2013). ***Русская лексикография** XXI века. **Учебное пособие***. Москва: **Флинта**.

Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S. & Singer, Y. (2006). Online Passive-Aggressive Algorithms. *The Journal of Machine Learning Research*, 7, pp. 551–585.

Creutz, M. & Lagus, K. (2005). Unsupervised Morpheme Segmentation and Morphology Induction from Text Corpora Using Morfessor 1.0. Technical Report A81, Publications in Computer and Information Science, Helsinki University of Technology.

Creutz, M. & Lagus, K. (2007). Unsupervised models for Morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing*, 4(1).

Dzienisiewicz, D. (2017). Na krzyż: NRF vs. RFN. http://re-research.pl/pl/post/2017-01-30-60105-na-krzyz-nrf-vs-rfn.html.

Dzienisiewicz, D., Graliński, F. & Wierzchoń, P. (2017). Archikastrat, emancypaństwo i krytykretyni – głos lingwochronologizatorów w sprawie kreatywności jezykowej Adolfa Nowaczyńskiego. In *Kreatywność jezykowa w przestrzeni publicznej*. In print.

Friedl, J. (2006). *Mastering Regular Expressions: Understand Your Data and Be More Productive*. O'Reilly Media.

Goldsmith, J. (2001). Unsupervised Learning of the Morphology of a Natural Language. *Comput. Linguist.*, 27(2), pp. 153–198. URL http://dx.doi.org/10.1162/089120101750300490.

Good, N. (2004). *Regular Expression Recipes: A Problem-Solution Approach*. Apresspod Series. Apress. URL https://books.google.pl/books?id=3ttQAAAAMAAJ.

Gouws, R. (2011). Learning, Unlearning and Innovation in the Planning of Electronic Dictionaries. In *E-Lexicography: The Internet, Digital Initiatives and Lexicography*. London: Bloomsbury Publishing, pp. 17–29.

Grzegorczykowa, R. & Puzynina, J. (1973). *Indeks a tergo do Słownika jezyka polskiego pod redakcja Witolda Doroszewskiego*. PWN.

Heid, U. (2011). Electronic Dictionaries as Tools: Toward an Assessment of Usability. In *E-Lexicography: The Internet, Digital Initiatives and Lexicography*. London: Bloomsbury Publishing, pp. 287–304.

Hông Phuong, L.ê., Thi Minh Huyên, N., Roussanaly, A. & Vinh, H.T. (2008). *A Hybrid Approach to Word Segmentation of Vietnamese Texts*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 240–249. URL http://dx.doi.org/10.1007/978-3-540-88282-4_23.

Jadacka, H. & Bondkowska, M. (2002). *Gniazda odrzeczownikowe*, volume 2 of *Słownik gniazd słowotwórczych współczesnego jezyka ogólnopolskiego*. Universitas.

Jurafsky, D. & Martin, J.H. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River, NJ, USA: Prentice Hall PTR, 1st edition.

Kharchenko, V. (2003). *Словарь богатств русского языка: редкие слова, метафоры, афоризмы, цитаты, биографемы*. Number t. 1-2 in Словарь богатств русского языка: редкие слова, метафоры, афоризмы, цитаты, биографемы. Изд-во Белгородского государственного университета.

Kharchenko, V. (2015). О демонстративном словаре русского языка. *Лексикография и коммуникация - 2015 : материалы I междунар. науч. конф.*, pp. 79–88.

Matyka, A. (2010). *Słowa – kładki, na których spotykaja sie ludzie różnych światów*, chapter O pojeciu neologizmu w jezykoznawstwie. Warszawa: Wydział Polonistyki UW, pp. 99–109.

Małek, E. (2008). *Ku fotoleksykografii*. Łódź: Instytut Rusycystyki Uniwersytetu Łódzkiego.

Mańczak, W. (1996). *Problemy jezykoznawstwa ogólnego*. Zakład narodowy im. Ossolińskich.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013a). Efficient Estimation of Word Representations in Vector Space. *CoRR*, abs/1301.3781. URL http://arxiv.org/abs/1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. & Dean, J. (2013b). Distributed Representations of Words and Phrases and their Compositionality. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani & K.Q. Weinberger (eds.) *Advances in Neural Information Processing Systems 26*. Curran Associates, Inc., pp. 3111–3119. URL http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf.

Mikolov, T., Yih, S.W.t. & Zweig, G. (2013c). Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT-2013)*. Association for Computational Linguistics, pp. 746–751. URL https://www.microsoft.com/en-us/research/publication/linguistic-regularities-in-continuous-space-word-representations/.

Monroe, W., Green, S. & Manning, C.D. (2014). Word Segmentation of Informal Arabic with Domain Adaptation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Baltimore, Maryland: Association for Computational Linguistics, pp. 206–211. URL http://www.aclweb.org/anthology/P14-2034.

Neubig, G., Nakata, Y. & Mori, S. (2011). Pointwise Prediction for Robust, Adaptable Japanese Morphological Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 529–533. URL http://dl.acm.org/citation.cfm?id=2002736.2002841.

Nichols, W. (2010). *English Learners' Dictionaries at the DSNA 2009*, chapter I've heard so much about you: Introducing the native-speaker lexicographer to the learner's dictionary. Tel Aviv: K Dictionaries, pp. 29–43.

Obrebska-Jabłońska, A., Dulewicz, I., Grek-Pabisowa, I. & I., M. (1968). *Indeks a tergo do Materiałów do słownika jezyka staroruskiego I.I. Srezniewskiego*. Państwowe Wydawnictwo Naukowe.

Okazaki, N. (2007). CRFsuite: a fast implementation of Conditional Random Fields (CRFs). http://www.chokkan.org/software/crfsuite/.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, pp. 2825–2830.

Piotrowski, T. (2001). *Zrozumieć leksykografie*. Wydawnictwo Naukowe PWN.

Ruokolainen, T., Kohonen, O., Sirts, K., Grönroos, S.A., Kurimo, M. & Virpioja, S. (2016). A Comparative Study of Minimally Supervised Morphological Segmentation. *Computational Linguistics*, 42(1), pp. 91–120.

Steffen-Batogowa, M. (1975). *Automatyzacja transkrypcji fonematycznej tekstow polskich*. Warszawa: PWN.

Steffen-Batóg, M. & Nowakowski, P. (1997). An algorithm for phonetic transcription of orthographic texts in Polish. In *Studies in phonetic algorithms*. Poznań: Soros, pp. 581–602.

Stubblebine, T. (2003). *Regular Expression Pocket Reference*. Sebastopol, CA, USA: O'Reilly & Associates, Inc., 1 edition.

Sutton, C. & McCallum, A. (2012). An Introduction to Conditional Random Fields. *Foundations and Trends® in Machine Learning*, 4(4), pp. 267–373. URL http://dx.doi.org/10.1561/2200000013.

Tarp, S. (2011). Lexicographical and Other e-Tools for Consultation Purposes: Towards the Individualization of Needs Satisfaction. In *E-Lexicography: The Internet, Digital Initiatives and Lexicography*. London: Bloomsbury Publishing, pp. 54–70.

Tseng, H., Chang, P., Andrew, G., Jurafsky, D. & Manning, C. (2005). A Conditional Random Field Word Segmenter for Sighan Bakeoff 2005. In *Fourth SIGHAN Workshop on Chinese Language Processing*. pp. 168–171.

Virpioja, S., Smit, P., Grönroos, S.A. & Kurimo, M. (2013). Morfessor 2.0: Python Implementation and Extensions for Morfessor Baseline. Technical Report 25/2013 in Aalto University publication series SCIENCE + TECHNOLOGY, Department of Signal Processing and Acoustics, Aalto University, Helsinki, Finland. URL https://aaltodoc.aalto.fi/handle/123456789/11836.

Vogelgesang, T. (2001). *Gniazda odprzymiotnikowe*, volume 1 of *Słownik gniazd słowotwórczych współczesnego jezyka ogólnopolskiego*. Universitas.

Wawrzyńczyk, J. (2014). *Jezyk, literatura i kultura rosyjska na stronie www.nfjp.pl*. Warszawa: Mila Hoshi.

Wierzchoń, P. (2009). Fotodokumentacja 3.0. *Jezyk. Komunikacja. Informacja*.

An Electronic Translation of the LIWC Dictionary into Dutch

Leon van Wissen[1], **Peter Boot**[2]

[1]Vrije Universiteit Amsterdam, The Netherlands
[2]Huygens ING, Amsterdam, The Netherlands
E-mail: l.van.wissen@vu.nl, peter.boot@huygens.knaw.nl

Abstract

LIWC (Linguistic Inquiry and Word Count) is a text analysis tool developed by social psychologists but now widely used outside of psychology. The tool counts words in certain categories, as defined in an accompanying (English-language) dictionary. The most recent version of the dictionary was published in 2015. We present a pipeline for the automatic translation of LIWC dictionaries into Dutch. We first make an automated translation of the LIWC 2007 version and compare it to the manually translated version of this dictionary. Then we use the pipeline to translate the LIWC 2015 dictionary. We also present the provisional Dutch LIWC 2015 dictionary that results from the pipeline. Although a number of categories require further work, the dictionary should be usable for most research purposes.

Keywords: Machine translation; Linguistic Inquiry and Word Count (LIWC); Google Translate

1. Introduction

LIWC (Linguistic Inquiry and Word Count, often pronounced 'Luke') is a lexical resource developed by social psychologist James Pennebaker and his team at the University of Texas (Pennebaker et al., 2001). Its lexical information is stored in a dictionary that groups English words into categories with psychological significance, such as emotions, cognitive processes, life concerns, social words and several categories of function words. This dictionary can be used in an application that processes a collection of texts and outputs the relative frequencies of words belonging to the categories in each of the texts. The distribution of those categories in the text can give insight into the psychological state of its author or can reflect an author's personal condition. The LIWC dictionary has been published in multiple versions (notably Pennebaker et al., 2001, 2007, 2015b) and the dictionary has been translated into many languages, mostly using the 2001 version as reference.

The 2015 version of LIWC introduces several new categories and sizable amounts of new words into existing categories, improving and fine graining the program's results. To use the capabilities of the LIWC 2015 program for Dutch text analysis, a Dutch version of the dictionary with the same structure and categories needs to be available. In this paper we therefore present an automated translation of the 2015 version of the LIWC dictionary into Dutch. The 2001 and 2007 versions were both manually translated into Dutch (Zijlstra et al., 2004; Boot et al., 2017). Since the process of manual translation is very labour-intensive, the experiment of trying an automated process is an obvious one. Our provisional translation is, as far as we know, the first LIWC translation based on the 2015 dictionary.

We show a method to automatically translate an English LIWC dictionary into Dutch, by using a pipeline of machine translation and combining part-of-speech tagging with different dictionary expansions through lexica. We first make an automated translation of the LIWC 2007 version and compare it to the manually translated version of this dictionary. The result of the procedure can then be used to evaluate the translation process and to translate the LIWC 2015 dictionary. We developed the pipeline by testing

it on the same corpus that was used in the evaluation of the manually translated version (Boot et al., 2017). Finally, we evaluate the method on the Dutch and English portion of the Dutch Parallel Corpus (Paulussen et al., 2013). For this, we use and extend the evaluation scripts and the Python LIWCtools script (Boot, 2016) that assisted the manual translation. The pipeline, as well as the lexical resources we use, are (in so far as the license allows for it) available in our GitHub repository.[1]

2. Background

2.1 LIWC dictionary

The LIWC program has been designed to work with multiple dictionaries, allowing users to input their own research- or language-specific data files. The program counts the occurrences of words in texts, based on the words contained in its dictionary. It does not take into account the words' context, nor does it do word sense disambiguation.[2] Usually, words will only be included in the dictionary under the category that is relevant for their most frequently used word sense. By the standards of computational linguistics, the program is very simple indeed. Still, it is a widely used research tool (see Tausczik & Pennebaker, 2010, for examples), also widely used outside its original field of psychology.

The LIWC dictionary (Pennebaker et al., 2015a) consists of a number of categories (identified by a number and label) and a number of words or terms, assigned to one or more of these categories. Terms are words or strings ending in the '*' wildcard. As the dictionary contains the term *administrat**, the LIWC program will count *administrator* and *administrative* in categories assigned to *administrat**. In Figure 1 an example of the dictionary layout is shown.

In the 2015 dictionary, there is a possibility to take into account multi-word expressions, though it is used only a few times. The LIWC categories are organised into partial hierarchies. The function word category contains the category of pronouns, which contains the category of personal pronouns, which contains the category of personal pronouns for the first person singular. There are also hierarchies for, among others, social words, for emotions, cognitions, biology, and, new in 2015, drives (a.o. achievement, risk, power).

The content and number of categories in the LIWC dictionaries has increased over the years. While the 2001 dictionary contained 2,319 words, the 2007 version contained 4,487 words and the 2015 version 6,549. The number of categories has been more or less stable (68 categories in 2001, 64 categories in 2007, 76 in 2015). However, both in 2007 and in 2015, a number of categories have disappeared and a number of new ones were created. New words have been added to existing categories, but words have also been removed from categories.

2.2 LIWC translation

The English LIWC dictionary has been translated into many languages, among others German (Wolf et al., 2008), French (Piolat et al., 2011), Spanish (Ramírez-Esparza et al.,

[1] https://github.com/LvanWissen/liwc-translation

[2] From the content of the 2015 English dictionary, it appears there might be a way of taking into account previous words' content or category. If this works, it would be an undocumented feature, and apparently only used to distinguish the various uses of (American) English *like*.

```
80          drives (Drives)
81          affiliation (Affiliation)
82          achieve (Achievement)
83          power (Power)
%
additional  21
address     112
adds        25  80 84  91
adequa*     80  82
adjust*     50  56
administr*  80  83 110
```

Figure 1: Example layout of a LIWC dictionary taken from the 2015 internal dictionary. The upper part of the excerpt shows categories (by number) and their definition. The lower part lists words and terms that are each assigned to one or multiple categories. The term *adequa** as well as all the words from a text starting with this string are for example assigned the 'drives (Drives)' and the 'achieve (Achievement)' categories.

2007) and Chinese (Gao et al., 2013). Translating a LIWC dictionary is not as straightforward as finding one or multiple equivalents for the English words. We mention three general complications. (i) Because words are assigned to multiple categories, the translator will have to check which equivalents fit into which categories. This led the creators of the Dutch 2007 translation to translate a word multiple times, for each of the categories in which it appeared. (ii) Another complication is presented by the wildcards: before an entry such as *manag** is translated, it has to be expanded into *manager*, *management*, *manageable*, *manage*, etc. (iii) Finally, in some cases, translating the dictionary requires finding corresponding words in a different culture. The Dutch 2007 translation for example includes names of Dutch labour unions in the category 'work', and Dutch beverages in the category 'leisure'.

Other problems are related to specific ways in which languages differ from English. In Romance languages, verbs are conjugated into many different forms. Do all of these forms have to be included in the dictionary? Because the subject of the sentence can often be deduced from the verb form, these languages use less personal pronouns than English does. To what extent does the translation need to take that into account? For Dutch a significant difference from English is its use of composite words: the English dictionary contains the entries *drug* and *addict**, but the Dutch equivalent of *drug addict* is a composite word *drugsverslaafde*, which would not necessarily appear in the dictionary when translating individual words.

Because of this, the translation of an LIWC represents a significant amount of work. The Dutch upgrade of the 2001 translation to 2007 took eight years. Yet, all translations known to us were compiled manually, except the translation into Catalan (Massó et al., 2013). Masso and his colleagues created a Catalan LIWC dictionary by automatically translating LIWC dictionaries from other Romance languages into Catalan. The main focus of their efforts is in assigning the words in the translation to the correct categories. They do not report an evaluation of their dictionary on a (parallel) corpus.

3. Translation procedure

We have developed a translation pipeline to translate an English LIWC dictionary into Dutch, which consists of the following steps:

3.1 Initalisation

The LIWC internal English dictionary is read and stored into a data structure that is listing words and their respective categories in a machine readable form. The categories from the source term are copied as is, with the exception of the function word categories (see below).

3.2 Wildcard expansion

Terms ending in an asterisk (*), which represent every word form in a text that starts with the preceding string, are resolved by looking for matching words in the Google n-gram corpus (Brants & Franz, 2006).[3] We use the frequency list of the unigram model. In order to remove noise, we only extract words that have a minimal frequency of 750,000 (which scales the corpus down to 46,717 tokens).

3.3 Translation

All words are sent for translation to the Google Translate interface[4] for a word to word translation. Since the online translations are bound to change due to improvements in the algorithm or user contributions and corrections, we store the translations to replicate and backtrack the procedure, if necessary.

3.4 Filtering

To prevent non-existing (malformed or not translated) Dutch words from entering the dictionary, words that are returned from the translation query are removed if they do not occur as token entry in the e-Lex corpus (NTU (Nederlandse Taalunie) [Dutch Language Union], 2006). We also discard any multiword expressions returned by the online translation.

3.5 Tagging

All translations are in this step tagged with part of speech information by TreeTagger (Schmid, 1994). The POS tags are converted to LIWC (function word) categories which are then added to the word's category information. We implement a conversion from POS tags to LIWC categories by using rules of the type shown in Figure 2.

3.6 Adding lemmas

In the same call, TreeTagger returns a lemmatised form of a word, which we recursively also tag, convert to LIWC functional categories using the same table and add to the dictionary as a separate entry.

[3] This corpus dates from 2006 and contains approximately 1 trillion words from the web from mostly English web pages. It is available online through the Linguistic Data Consortium (LDC).

[4] https://translate.google.com/. Although translating to Dutch was already possible for a long time, Google recently updated the system to include Dutch in its new Neural Machine Translation (Wu et al., 2016).

POS	description	LIWC-category	
adj	adjective	21	adjective
adv	adverb	13	adverb
conjcoord	coord. conjunction	14	conj
det_art	article	10	article
det_indef	indefinite pronoun	2,9	pron,ipron
det_poss	possessive pronoun	2	pron
int	interjection	125	filler

Figure 2: Example from a set of POS tags and their corresponding LIWC function word categories. We apply this mapping after tagging the words.

3.7 Adding other word forms

As a final step, we further extend the dictionary with word forms from a lemma list (NTU (Nederlandse Taalunie) [Dutch Language Union], 2015), which we again tag and add to the dictionary with both functional and content categories. If the word already exists, the category information is merged so that there exists only one entry in the resulting dictionary.

3.8 Handling function words

Since translating pronouns by a (statistical) machine translation system is known to be harder than translating content words due to differences in the way a language deals with pronouns (Guillou et al., 2016), we have chosen to exclude most function words from the translation process described above. We fill these categories based on the POS-tagging in the e-Lex lexicon[5] (NTU (Nederlandse Taalunie) [Dutch Language Union], 2006). We query the lexicon and ask it to return a list of all words meeting specified POS and category criteria. We retrieve for instance all first person singular pronouns by asking for all words that have POS equal to 'VNW' (voornaamwoord [=pronoun]) with categories '1' (first person) and 'ev' (enkelvoud [=singular]). The output is given in Figure 3. We add all those words to the dictionary, in the 'I'-category.[6]

mijzelf, m'n, mezelve, ik, mij, ikzelf, mijne me, eigen, mijn, waterdragen, 'k, mijns

Figure 3: List of first person singular pronouns from e-Lex for the 'I' category in LIWC.

3.9 Remove function words from content categories

We use similar lookups for words that we only allow in a certain category. Translation artifacts, faulty translations or inconsistencies in the lexicon can for example put a determiner inside one of the content categories, and its high frequency would have a large effect on the category scores. We specify for example that all determiners from e-Lex may only occur in the 'det' category of the LIWC dictionary.

[5] Formerly the TST-lexicon. The e-Lex lexicon is a Dutch lexicon (we use the one-word version) that contains over 600,000 word forms in ca. 200,000 entries with POS and category information (e.g. gender and number).

[6] The word 'waterdragen' (i.e. 'carry water', 'domestic service') is obviously an error. e-Lex is constructed from several other corpora that have been annotated semi-automatically and as such can contain errors. However, the problems that we found are minor.

3.10 Extending hierarchy

The LIWC dictionary has a hierarchical structure. As a final step in the translation pipeline we extend the scope of terms by also adding the parent category to its categories. This means that we also add a word that is part of the 'health' category (category id 72) to the parent 'bio' (category id 70) category. We use the completion function of LIWCtools (Boot, 2016) for this step, which takes the existing English dictionary as a model and projects its structure onto the newly translated Dutch one.

3.11 Wrap-up

When the translation is complete, the dictionary is stored in a format that can be used in the official LIWC program.

3.12 Manual correction

Although the dictionary that is created in the automatic procedure performs acceptably (see the sections below), errors are inevitable. The more frequent words among the errors have a measurable effect on the outcome. We decided to add a manual correction step to remove those from the dictionary. What we did was to compute, for each LIWC category and for both the Dutch and English dictionary, a list of the words that accounted for more than 1.5% of the hits in that category. For most categories, this produces a list of ca. 10 to 15 words. For the English words, we manually checked whether their main translation(s) occurred in the generated dictionary. If not, we added them. For the Dutch words, we checked whether these words belonged in the category. If not, we removed them. We also did a superficial inspection of the translated dictionary and corrected some of the more obvious errors.

4. Evaluation procedure

4.1 Corpus

The translation pipeline was designed, developed and tested on the same set of parallel Dutch and English texts that was used by Boot et al. (2017). The test corpus includes letters of Vincent Van Gogh, documents from the European parliament, TED-talk subtitles and Bible books. This corpus is also used to test the efficacy of the manual corrections to the dictionary.

In order to avoid the risk of overfitting to this development corpus, we use a separate corpus for the final evaluation of the dictionary. Here we use the Dutch Parallel Corpus (DPC, Paulussen et al., 2013).[7] From the test and evaluation corpora, we remove files with a low word count (<1,000) to prevent small files from influencing the results.

4.2 Calculations

We use the count functionality of LIWCtools to replicate the textual analysis function of the official LIWC software. Each Dutch text from the DPC is processed using the

[7] A corpus built from Dutch and English texts coming from a broad range of fields such as finance, science, culture and communication.

translated dictionary. Its English equivalent is processed by the English dictionary. The result is a table containing the coverage (expressed in relative frequency) per dictionary category (columns) for each individual processed file (rows). A sample is shown in Figure 4 below.

Filename	**function**	**pronoun**	**ppron**	**i**
education/dpc-vla-001191.txt	0.479	0.079	0.041	0.000
education/dpc-vla-001172.txt	0.482	0.05	0.029	0.000
education/dpc-mis-001909.txt	0.488	0.069	0.046	0.001
institutions/dpc-bal-001241.txt	0.54	0.142	0.088	0.011
institutions/dpc-gim-002525.txt	0.424	0.076	0.051	0.005

Figure 4: Example of the output that is created after processing text files from the parallel corpus. Shown is an excerpt of the data that shows five processed files (rows) and the share of several categories (columns) of the total amount of words of the text file. The format of the file is very close to the output of the official LIWC program.

We then calculate a correlation score and effect size (Cohen, 1992) for the corresponding columns (e.g. the function words in the Dutch texts with the function words in the English texts). Based on whether the data are normally distributed, either a Pearson or a Spearman correlation measure is used. For both English and Dutch we also compute the median, minimum and maximum frequencies.

The target values for our automatic translation are those of the Dutch manual (gold) translation of LIWC 2007. This translation achieved an average correlation of 0.77 with the English dictionary (effect size 0.39) on the DPC.[8]

5. Evaluation for the 2007 LIWC dictionary

We evaluate our automatic approach by comparing the correlation coefficient and effect sizes between the English 2007 dictionary and the manual translation with those for the English dictionary and the automatic translation.

As mentioned above, evaluating the manually translated 2007 dictionary on the DPC corpus results in an average correlation score of 0.77 (effect 0.39). Our automatically translated 2007 dictionary, without a manual correction step, scores a bit less with an average correlation coefficient of 0.72 (effect 0.72). Our translation does especially well for the function word categories with most correlations above 0.80. Only the impersonal pronouns category ('ipron') scores much lower compared to the manual translation. This is probably due to the word *niet* [=not] being included in the translation, which accounts for ca. 40% of the 'ipron' category. The adverb category is problematic too, as it has an effect size of 6.21. This is because a number of prepositions ended up in this category.

For the content word categories, some actually do better than the manual translation, e.g. 'home'. Given the large numbers of words in these categories, it is hard to say what is the cause of this improvement. The categories 'inclusive', 'body', 'ingest', 'time' and 'leisure' score lower on correlation. For the 'body' category, this is probably largely due

[8] This comparison and performance test was already done when the Dutch 2007 dictionary was presented (Boot et al., 2017). The translators then achieved a correlation of 0.80 (effect size: 0.35) on their test set. We did this comparison again on our own evaluation corpus.

to the ambiguous words *haar* [=hair, her] and *enkel* [=ankle, solely]. In other cases it is impossible to point to a few words to explain an unsatisfactory result. Some other categories do not score that well in the manual translation either (e.g. 'feel' and 'motion'). For the 'swear' category, this might be due to a lack of testing material in the corpus.

From preliminary testing, we know that a manual correction step can improve the result of the automatic 2007 translation with ca. 0.04 (correlation) and -.20 (effect size). That would bring us quite close to the results of the manual translation.

6. Evaluation for the 2015 LIWC dictionary

6.1 Procedure

For the automatic translation of the 2015 dictionary, we do not have the manual translation to compare the results. What we do have is the possibility to compare the results with that of the English dictionary on our test corpus. We first do an automatic translation and test the result against the test corpus, then add a manual correction and test again against the test corpus. Finally, we evaluate the end result against the evaluation corpus.

6.2 Results

Table 1 shows the average correlations and effect sizes for the different conditions. The initial automatic 2015 translation scores somewhat lower than the automatic 2007 translation. While most categories perform somewhere between acceptably and very well, the informal word categories perform very bad. The correction step does have a measurable effect, an effect that is largely retained when testing against the evaluation corpus.

Dictionary	**Corpus**	**Correlation d**	**Effect size r**
Automated 2015 translation	Test corpus	0.69	0.88
Automated 2015 translation with correction	Test corpus	0.73	0.52
Automated 2015 translation with correction	Evaluation corpus	0.73	0.59

r: correlation, d: effect size (Cohen's d).

Table 1: Average correlation coefficients and effect sizes for the Dutch LIWC 2015 dictionary.

6.3 Results by category

The numbers shown in Table 2 below give the results by category of the corrected dictionary on the evaluation corpus. The table should provide researchers with the information necessary to decide which LIWC 2015 categories should work the same in a Dutch-language context as in an English context.

By and large, the function word categories perform very well. Exceptions are the new categories 'adjectives', 'comparatives' ('greater', 'greatest', etc.) and 'interrogatives' ('where', 'how', etc.). For the adjectives, the explanation may be that the translation contains many more adjectives than the original; for the interrogatives, the explanation may be that in both languages these words can also occur as adverbs or pronouns. These categories clearly need more work, as does the category of quantitative words, which scores inexplicably low.

Some of the function word categories profited significantly from the manual correction, such as 'shehe' where we removed the male possessive pronoun *zijn*, as it is more frequently used as a verb (*to be*). For other categories we added words missing in the translation, such as the demonstrative pronouns that should have been in the impersonal pronouns category.

The psychological categories of emotion, social words and cognitive words again perform rather well. From the 'insight' category, maybe we should have removed *worden* [=become], which is translated correctly, but also serves as a passive auxiliary verb in Dutch. From 'friends', maybe we should have removed the word *kennis* which in Dutch is *acquaintance* as well as *knowledge*. The biological categories are less satisfactory, without clear culprits. In contrast, the new categories under 'drives' ('affiliation', 'power', 'reward', 'risk') perform generally well.

From the 'time orientation' group, 'focusfuture' could perform better. We might try to remove the verb *gaan* [=to go] which is often but certainly not always used to express a focus on the future. The categories from the 'personal concerns' group do generally well. But as noted, the informal categories perform very poorly. This was also true, though not quite to this extent, in the manual LIWC 2007 translation. The results are probably to some extent due to the test and evaluation corpora, that are heavily oriented to written language, and certainly do not contain terms from the netspeak category (a category where Dutch borrowed lots of terms from English). Another issue is probably that the translation engine will have been trained on written language. There are also some problems with the English categories: the 'nonfluencies' category for instance contains the word *well*, which is responsible for 85% of the category count, but of course has many other uses besides its use as a nonfluency. And, finally, in these categories cultural differences may play an important role. For example, Dutch often uses names of illnesses as swear words (Fletcher, 1996).

	Word counts						**Equivalence statistics**	
Category	**English**			**Dutch**			***r***	***d***
	Median	***Min***	***Max***	***Median***	***Min***	***Max***		
Word count	2,179	1,003	122,206	2,169	999	128,338	0.99*	0.00
Linguistic dimensions								
function words	46.60	27.96	60.67	51.33	32.37	62.62	0.94	0.94
pronoun	7.55	1.34	22.05	9.24	1.86	23.25	0.97	0.37
ppron	3.30	0.05	16.48	3.72	0.17	15.90	0.98	0.11
I	0.23	0.00	7.67	0.19	0.00	7.73	0.95*	0.03
we	0.56	0.00	3.99	0.52	0.00	4.24	0.97	0.07
you	0.18	0.00	2.40	0.26	0.00	2.33	0.91*	0.07
shehe	0.20	0.00	7.09	0.82	0.00	8.89	0.80*	0.30
they	0.51	0.00	3.31	0.75	0.00	5.85	0.79*	0.50
ipron	3.85	0.86	7.58	4.20	1.10	7.69	0.86	0.18
article	9.25	5.01	17.06	12.12	6.79	17.96	0.81	1.48
prep	15.29	9.83	20.39	16.48	12.49	21.51	0.74	0.78
auxverb	6.22	2.04	10.12	5.77	1.72	9.09	0.77	0.32
adverb	3.06	0.65	6.92	6.07	1.60	11.54	0.83	1.90

	Word counts						Equivalence statistics	
Category	**English**			**Dutch**			***r***	***d***
	Median	***Min***	***Max***	***Median***	***Min***	***Max***		
conj	5.24	2.18	8.22	6.03	2.87	9.46	0.85	0.74
negate	0.70	0.00	2.42	0.86	0.00	3.10	0.85	0.37
Other grammar								
verb	10.48	3.67	20.20	11.54	4.43	20.00	0.90	0.47
adj	4.43	1.76	7.83	6.28	3.19	10.63	0.52	1.84
compare	2.41	0.93	5.74	3.35	1.99	7.05	0.50	1.54
interrog	1.06	0.12	2.59	0.84	0.06	3.00	0.45	0.33
number	0.95	0.13	6.90	1.12	0.00	5.53	0.79*	0.16
quant	1.87	0.58	3.45	1.76	0.45	5.30	0.31	0.10
Psychological processes								
affect	4.12	0.97	9.50	2.62	0.79	7.09	0.81	1.13
posemo	2.80	0.44	7.35	1.70	0.47	4.94	0.77	1.07
negemo	1.03	0.00	7.02	0.83	0.00	5.30	0.85*	0.42
anx	0.21	0.00	2.59	0.18	0.00	1.27	0.71*	0.23
anger	0.19	0.00	3.81	0.15	0.00	2.30	0.82*	0.30
sad	0.19	0.00	1.16	0.19	0.00	0.84	0.57*	0.17
social	6.64	0.22	18.00	6.55	1.28	16.85	0.95	0.06
family	0.05	0.00	3.89	0.07	0.00	4.04	0.80*	0.16
friend	0.15	0.00	1.22	0.12	0.00	1.21	0.59*	0.20
female	0.07	0.00	7.21	0.48	0.00	7.92	0.67*	0.32
male	0.31	0.00	7.03	1.15	0.19	7.36	0.87*	0.55
cogproc	8.58	2.51	16.69	10.07	5.15	16.29	0.84	0.72
insight	1.78	0.39	3.98	2.41	0.94	4.99	0.56	1.04
cause	1.77	0.49	4.30	1.29	0.34	4.03	0.74	0.82
discrep	0.96	0.07	3.35	1.97	0.34	5.46	0.77	1.42
tentat	1.57	0.15	6.49	1.80	0.45	4.55	0.78*	0.30
certain	1.15	0.19	2.88	1.23	0.19	3.14	0.73	0.18
differ	2.08	0.09	5.76	2.44	0.52	5.45	0.88	0.33
percept	1.29	0.07	7.25	0.99	0.04	4.71	0.87*	0.39
see	0.54	0.00	5.88	0.43	0.00	3.39	0.74*	0.36
hear	0.30	0.00	3.45	0.23	0.00	2.77	0.89*	0.20
feel	0.24	0.00	2.05	0.22	0.00	2.48	0.61*	0.21
bio	0.82	0.00	7.06	0.51	0.05	4.71	0.75*	0.48
body	0.17	0.00	3.29	0.16	0.00	2.49	0.69	0.16
health	0.41	0.00	5.09	0.23	0.00	3.11	0.62*	0.48
sexual	0.00	0.00	2.63	0.00	0.00	1.60	0.60*	0.16
ingest	0.15	0.00	2.94	0.10	0.00	1.43	0.64*	0.32
drives	7.75	2.62	16.28	5.49	1.94	12.50	0.88	0.89
affiliation	1.86	0.00	7.28	1.47	0.08	6.11	0.90	0.26
achieve	1.76	0.15	4.75	1.40	0.19	3.49	0.80	0.61
power	3.12	1.20	9.75	2.09	0.61	7.50	0.76	0.86
reward	1.02	0.07	2.63	0.78	0.00	2.67	0.65	0.61

	Word counts						Equivalence statistics	
Category	**English**			**Dutch**			***r***	***d***
	Median	***Min***	***Max***	***Median***	***Min***	***Max***		
risk	0.58	0.00	4.31	0.44	0.00	2.60	0.71*	0.40
Time orientation								
focuspast	2.22	0.49	10.76	3.38	1.38	10.93	0.87*	0.57
focus present	6.52	1.96	11.71	9.39	3.37	15.64	0.66	1.30
focusfuture	0.97	0.15	4.04	1.85	0.52	4.24	0.63*	1.29
relativ	13.87	7.74	19.28	13.97	9.56	19.26	0.75	0.06
motion	1.62	0.32	4.60	1.42	0.29	2.97	0.55	0.53
space	7.89	3.43	13.71	7.70	4.45	12.18	0.76	0.15
time	4.24	1.63	7.75	5.12	2.50	7.44	0.71	0.82
Personal concerns								
work	4.82	0.53	14.60	2.77	0.46	9.22	0.85	0.95
leisure	0.43	0.00	4.48	0.26	0.00	3.36	0.80*	0.45
home	0.19	0.00	2.02	0.11	0.00	2.18	0.64*	0.37
money	1.14	0.00	9.22	0.71	0.00	5.95	0.92*	0.55
relig	0.09	0.00	6.02	0.03	0.00	3.54	0.68*	0.24
death	0.06	0.00	2.23	0.03	0.00	1.80	0.81*	0.20
informal	0.17	0.00	2.99	1.23	0.07	3.86	0.31*	2.22
swear	0.00	0.00	0.40	0.00	0.00	0.26	0.42*	0.22
netspeak	0.00	0.00	2.99	0.21	0.00	3.38	0.35*	0.80
assent	0.03	0.00	0.60	0.00	0.00	0.40	0.50*	0.43
nonflu	0.07	0.00	0.43	0.00	0.00	0.09	0.17*	1.51
filler	0.00	0.00	0.13	0.95	0.07	3.31	0.20*	2.71

r: correlation, d: effect size (Cohen's d).

*: Correlations with * were computed using Spearman's rank correlation coefficient.

Table 2: Results of equivalence test on translated Dutch and English dictionary.

7. Conclusion

We presented a pipeline for automatic translation of the LIWC dictionary from English into Dutch. The result of a comparison between an automatic translation of the 2007 and the manually translated version shows that the automatic translation is nearly as good as the manual one when looking at the correlation coefficients. When repeating this translation procedure for the new 2015 dictionary, we are able to produce a dictionary with an average correlation coefficient of 0.69 (effect 0.88) to the English dictionary. Manual correcting boosts these numbers to 0.73 (effect 0.59), a score that is again very close to the one reached by the manual (2007) translation.

We should note that the correlations for the informal word categories (netspeak, swear words, etc.) are considerably less satisfactory. There are a number of underperforming categories as well among the psychological processes and function words. Still, the au-

tomatic translation as a whole performs well. This is all the more remarkable as our automatic translation does not take into account some of the aspects of translation that we discussed in the Background section 2.2 and that a human translator will care about, such as the assignment of translated words to the fitting LIWC category or the use of composite words in Dutch.

Given the fact that a manual translation of an LIWC dictionary is a very time-consuming task, the automatic translation should therefore be considered a serious alternative, at least for those languages for which a sufficient number of linguistic resources is available. Further improvement (manual or automatic) is always possible.

As is unavoidable in any automatic treatment of language, the translated dictionary does contain errors. However, given the fact that the categories contain many words, most only responsible for a tiny fraction of the total of words in its category, errors are not necessarily problematic. It is also in the nature of a tool such as LIWC, that does not do word sense disambiguation, that words are occasionally misclassified. In spite of the errors, the resulting (provisional) dictionary should be usable for most research purposes. We invite researchers in psychology, digital humanities and other fields to validate its usability in the context of practical research.

8. Acknowledgements

We thank Isa Maks for the suggestion that led to this paper and we thank Sem Zweekhorst for his contributions to the manual corrections.

9. References

Boot, P. (2016). LIWCtools. Tools for working with LIWC dictionaries (Version 0.0.1). URL https://github.com/pboot/LIWCtools.

Boot, P., Zijlstra, H. & Geenen, R. (2017). The Dutch translation of the Linguistic Inguiry and Word Count (LIWC) 2007 dictionary. *Dutch Journal of Applied Linguistics*, 6(1).

Brants, T. & Franz, A. (2006). Web 1T 5-gram corpus version 1.1. *Google Inc.*

Cohen, J. (1992). A power primer. *Psychological bulletin*, 112(1), p. 155.

Fletcher, W.H. (1996). Come down with cholera: Disease names in Dutch strong language. *Canadian Journal of Netherlandic Studies*, 17, pp. 231–239.

Gao, R., Hao, B., Li, H., Gao, Y. & Zhu, T. (2013). Developing simplified Chinese psychological linguistic analysis dictionary for microblog. In *International Conference on Brain and Health Informatics*. Springer, pp. 359–368.

Guillou, L., Hardmeier, C., Nakov, P., Stymne, S., Tiedemann, J., Versley, Y., Cettolo, M., Webber, B. & Popescu-Belis, A. (2016). Findings of the 2016 WMT shared task on cross-lingual pronoun prediction. In *Proceedings of the First Conference on Machine Translation (WMT16), Berlin, Germany. Association for Computational Linguistics*. pp. 525–542.

Massó, G., Lambert, P., Penagos, C.R. & Saurí, R. (2013). Generating New LIWC Dictionaries by Triangulation. In *Asia Information Retrieval Symposium*. Springer, pp. 263–271.

NTU (Nederlandse Taalunie) [Dutch Language Union] (2006). e-Lex Version 1.1.1.

NTU (Nederlandse Taalunie) [Dutch Language Union] (2015). URL http://taalunieversum.org/sites/tuv/files/downloads/dutch_lemmas.txt.

Paulussen, H., Macken, L., Vandeweghe, W. & Desmet, P. (2013). Dutch parallel corpus: A balanced parallel corpus for Dutch-English and Dutch-French. In *Essential Speech and language technology for Dutch*. Springer, pp. 185–199.

Pennebaker, J.W., Booth, R.J., Boyd, R. & Francis, M.E. (2015a). Linguistic Inquiry and Word Count: LIWC2015 Operator's Manual. URL https://s3-us-west-2.amazonaws.com/downloads.liwc.net/LIWC2015_OperatorManual.pdf.

Pennebaker, J.W., Boyd, R.L., Jordan, K. & Blackburn, K. (2015b). The development and psychometric properties of LIWC2015. URL http://www.liwc.net.

Pennebaker, J.W., Chung, C.K., Irel, M., Gonzales, A., Booth, R.J. & Framework, T.L. (2007). The Development and Psychometric Properties of LIWC2007. URL http://www.liwc.net/LIWC2007LanguageManual.pdf.

Pennebaker, J.W., Francis, M.E. & Booth, R.J. (2001). Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, 71.

Piolat, A., Booth, R.J., Chung, C.K., Davids, M. & Pennebaker, J.W. (2011). La version française du dictionnaire pour le LIWC: modalités de construction et exemples d'utilisation. *Psychologie française*, 56(3), pp. 145–159.

Ramírez-Esparza, N., Pennebaker, J.W., García, F.A., Suriá Martínez, R. et al. (2007). La psicología del uso de las palabras: Un programa de computadora que analiza textos en español. *Revista Mexicana de Psicología*, 24(1), pp. 85–99.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*. p. 154.

Tausczik, Y.R. & Pennebaker, J.W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology*, 29(1), pp. 24–54.

Wolf, M., Horn, A.B., Mehl, M.R., Haug, S., Pennebaker, J.W. & Kordy, H. (2008). Computergestützte quantitative Textanalyse: Äquivalenz und Robustheit der deutschen Version des Linguistic Inquiry and Word Count. *Diagnostica*, 54(2), pp. 85–98.

Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K. et al. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv preprint arXiv:1609.08144*.

Zijlstra, H., Van Meerveld, T., Van Middendorp, H., Pennebaker, J.W. & Geenen, R. (2004). De Nederlandse versie van de 'linguistic inquiry and word count'(LIWC). *Gedrag Gezond*, 32, pp. 271–281.

Extracting an Etymological Database from Wiktionary

Benoît Sagot

Inria
2 rue Simone Iff, 75012 Paris, France
E-mail: benoit.sagot@inria.fr

Abstract

Electronic lexical resources almost never contain etymological information. The availability of such information, if properly formalised, could open up the possibility of developing automatic tools targeted towards historical and comparative linguistics, as well as significantly improving the automatic processing of ancient languages. We describe here the process we implemented for extracting etymological data from the etymological notices found in Wiktionary. We have produced a multilingual database of nearly one million lexemes and a database of more than half a million etymological relations between lexemes.

Keywords: Lexical resource development; etymology; Wiktionary

1. Introduction

Electronic lexical resources used in the fields of natural language processing and computational linguistics are almost exclusively synchronic resources; they mostly include information about inflectional, derivational, syntactic, semantic or even pragmatic properties of their entries. Because this information is formalised, it can be used by automatic tools.

Conversely, diachronic information such as etymology is virtually absent from electronic resources, only being present in printed or online dictionaries. The few exceptions, such as *The Tower of Babel* database[1] or the *PIElexicon* project,[2] often rely on comparative and etymological principles that are, at best, obsolete or non-consensual,[3] and, at worst, unanimously rejected by the scientific community.[4] Only EtymWordNet (de Melo, 2014), to which we shall come back below, is an outlier in this regard, although it has other severe limitations.

The availability of formalised, detailed and large-coverage etymological databases would make it possible to develop automatic tools targeted towards historical and comparative linguistics. Modelling language evolution and reconstructing proto-languages—ancestors of attested languages—rely on a very large amount of lexical information often covering dozens, if not hundreds of languages. For some language families, such as Indo-European or Semitic languages, almost two centuries of careful work has resulted in a fairly clear understanding of lexical diachrony. However, even for these two families, and *a fortiori* for all others, many grey areas remain.

The development of automatic means to explore possible formal and semantic correspondences between words from different languages and to model their diachronic evolution

[1] http://starling.rinet.ru/babel.php?lan=en

[2] http://pielexicon.hum.helsinki.fi

[3] The Indo-European database in *The Tower of Babel* is based on the Pokorny dictionary, which is now outdated. Moreover, the authors of this database defend non-consensual views on genetic relationships between traditional language families. These views are generally rejected by the scientific community yet still influence some of their etymological proposals.

[4] This applies to the *PIElexicon*, although the justification of such a statement lies beyond the scope of this paper.

would therefore constitute an important step forward for the linguistic sub-fields involved, while raising difficult algorithmic challenges. It would also contribute to the development of resources and tools for the automatic processing of documents written in older forms of the languages, for which they already exist for their modern variant (for instance, documents in Old or Middle English, which cannot be properly processed by tools dedicated to contemporary English). This direction of research should take advantage of the outcome of previous etymological investigations, which should therefore be encoded in the form of formalised electronic lexical resources.

To achieve this goal, we need to find a large-scale source of etymological information, to automatically extract this information from it, and to represent it in a structured or even normalised form. In this paper, we describe a first attempt at carrying out such an enterprise. We rely on the (English) *Wiktionary*,[5] an online collaborative dictionary, whose syntax is semi-structured and which includes relatively detailed and fairly reliable etymological information.

The remainder of this paper is structured as follows. After a brief overview of previous work related to ours (Section 2) and a brief sketch of the various types of etymological relations between lexemes (Section 3), we describe how etymological information is represented in Wiktionary articles and how we extracted and partially structured this information (Section 4). In Section 5, we explain how we transformed this information into a database of lexemes and a database of etymological relations between these lexemes. We provide in Section 6 quantitative information about these two databases, their export formats—including an etymology-oriented extension of the LMF standard[6] currently under discussion—and a manual evaluation of their quality. We conclude this work in Section 7 by discussing possible follow-ups to this work, including possible direct use cases for our etymological databases.

Both databases are freely available under an LGPL-LR license.

2. Related work

Previous work related to ours is threefold: efforts towards the standardisation of etymological information, development of existing databases, and the above-mentioned EtymWordNet.

Since etymological information is only exceptionally taken into account in electronic lexical resources, their structured representation is not yet the subject of recommendations concerning their standardisation. In this regard, the working paper published by Bowers & Romary (2016) reflects the state of ongoing research. It builds on several previous initiatives, including the work by Salmon-Alt (2006). It proposes a set of general principles for the representation of etymological information in electronic dictionaries encoded in TEI. It is based on a relatively broad typology of the underlying phenomena, which covers standard inheritance (what etymologists refer to as *recto itinere*, "in direct line"), borrowing, metaphor, metonymy, composition and grammaticalisation. Some of these mechanisms are not etymological in nature, but are rather lexical creation mechanisms. We shall come back in Section 6 on several limitations of this proposal in its current state.

[5] https://en.Wiktionary.org/
[6] Lexical Markup Framework. See below for details.

Few freely available electronic dictionaries make use of structured representations of etymological information. We have already mentioned *The Tower of Babel* and the *PIElexicon*. Another example is the *Germanic Lexicon Project*[7] by S. Crist, whose representation format can also be considered as a predecessor of the propositions made by Bowers & Romary (2016). However, the various free dictionaries distributed in this framework are only weakly structured: the systematic extraction of etymological relations would be a non-trivial task. This is not the case for the *World Loanword Database*, which, for 1,460 carefully selected meanings, provides one or more lexemes in 41 languages, each associated with a probability level that it results from a borrowing, as well as its possible source lexeme. But the inventory of the 41 languages covered reflects the typological and non-etymological positioning of the project. In any case, it is far from a widely covered resource, and, of course, only borrowing mechanisms are covered, to the exclusion of any other etymological mechanism.

Closer to our work, de Melo (2014) has made available EtymWordNet, which, like in our work, was automatically extracted from the Wiktionary (although in a three-year-old version). However, and despite extensive coverage, the EtymWordNet can not be used as it is for the computerisation of comparative and historical linguistics because of two fundamental limitations. Firstly, the mechanisms at play are not distinguished (for example, no distinction between inheritance, borrowing and morphological derivation). Secondly, and even more importantly, its basic units are lemmas, not lexemes: senses are ignored.

We are not aware of previous works that resulted in a large-scale formalised etymological database at the lexeme level, as is necessary in etymological lexicology (see Section 3) and makes a distinction between etymological mechanisms. That is the purpose of our work.

3. Etymological and lexical creation mechanisms

The extraction and formalisation of etymological information requires a model of this type of information. The first question that arises is that of the basic unit. As recalled by Buchi (2016: 346), only the lexeme can play this role—in our case a lexeme is defined by a citation form, a language identifier and an English gloss.[8] A relation between a lexeme and another lexeme can correspond to changes in the language (diachronic change in the case of inheritance, synchronic change in the case of a borrowing), the citation form (phonetic but also morphological changes) and the meaning (semantic shifts).

The second important question is the nature of the etymological relations between lexemes. Following again Buchi (2016: 346–347), an elementary etymological relation must concern directly related lexemes: there should not be any intermediate lexemes between those involved in the relation. In the case of a *recto itinere* inheritance, and given an inventory of language identifiers, an elementary relation must therefore involve a lexeme in a given language and another lexeme, or several lexemes, in the immediately preceding language or language state.[9] In the case of a borrowing, a direct relation simply involves

[7] http://lexicon.ff.cuni.cz/texts/pgmc_torp_about.html

[8] This also covers the case of place names, person names, people/tribe names, and other proper names.

[9] Fr. *manger* < Mid. Fr. *manger* is therefore a direct relation, contrarily to Fr. *manger* < Old Fr. *mengier* and Fr. *manger* < Late Lat. *manducāre*, which are indirect relations.

the target lexeme and its source lexeme.[10] Using direct relations whenever possible is necessary to be able to correctly specify the nature of the etymological relation involved.[11]

The third question to address for formalising etymological relations is that of the different types of etymological mechanisms. Although we do not cover all of them in this work, we make use of the following typology:

- Inheritance (with phonetic change in most cases, with or without semantic or morphological change); as is customary, we shall note this type of relation as follows: *target lexeme* < *source lexeme*;
- Borrowing; we shall note this type of relation as follows: *target lexeme* ← *source lexeme*;[12]
- Lexical creation
  - Morphological derivation
    - Suffixal derivation; it will be noted as follows: *target lexeme* $<_s$ *base* + *suffix*;
    - Prefixal derivation; it will be noted as follows: *target lexeme* $<_p$ *prefix* + *base*;
    - Other cases (including analogy-based derivation); they will be noted as follows: *target lexeme* $<_d$ *element* + ... + *element*;
  - Morphological composition, noted as follows: *target lexeme* $<_c$ *component* + ... + *component*;
  - Portmanteau word creation, not covered in this work;
  - Truncation and other phenomena, not covered in this work.

To this inventory we shall add a special cognation relation, which will allow us to relate two lexemes (within the same language or in two different languages) that have a common or partly common etymology (in general, at least a same "root"). It will be noted $lexeme_1$ // $lexeme_2$.

4. Extraction and structuration of Wiktionary's etymological information

4.1 Etymological information in Wiktionary

Wiktionary is a collaborative multilingual dictionary. It is organised into articles, which each contain one or more homonymous lexical entries[13] concerning lexemes from one or more languages.

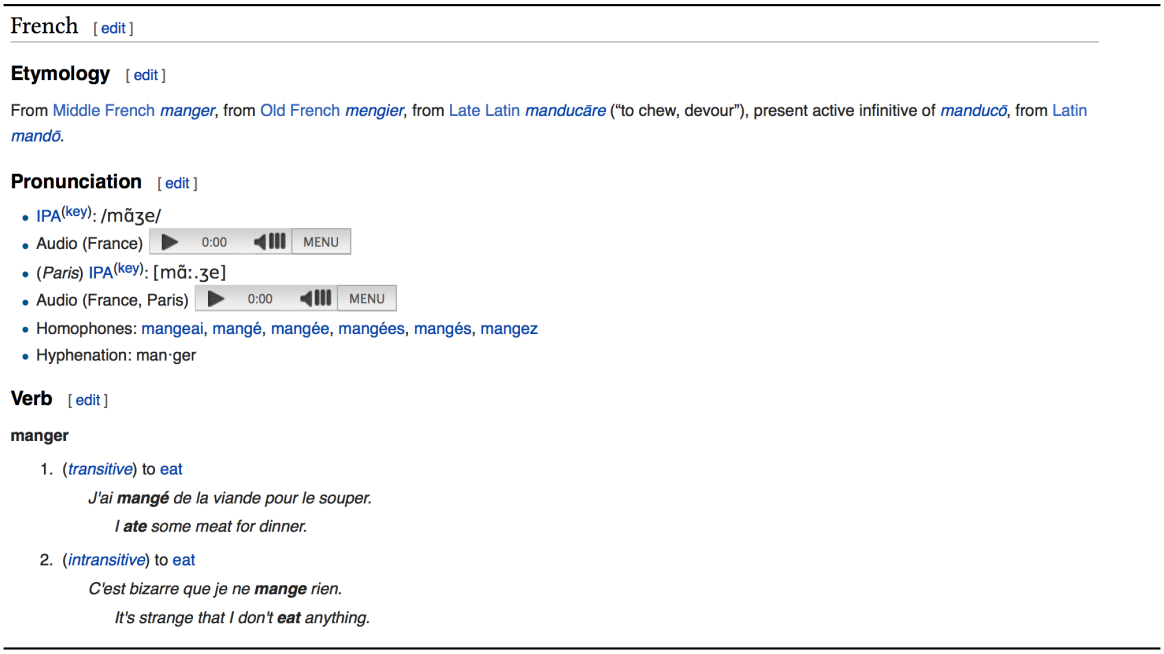
We used the 01/01/2017 dump. It contains nearly 5.5 million articles, more than 40,000 of which are redirect pages. These entries contain a total of 894,453 etymological records.

[10] For instance, relations such as Fr. *abricot* 'apricot' < Esp. *albaricoque* 'id.' and Fr. *abricot* < Port. *albricoque* 'id.' are possible direct relations (both are plausible). The Spanish and Portuguese words are borrowings from Ar. *al-barqūq* 'id.', itself a borrowing from Med. Gr. βερικόκκια 'apricot tree', derived from Ancient Gr. πραικόκιον 'apricot', itself a borrowing from Lat. *praecoquum* 'early (fruit)'. Therefore, a relation such as Fr. *abricot* < Lat. *praecoquum* would be correct but not direct.

[11] Going back to the example introduced in the previous footnote, it would be difficult to assign a simple type to the etymological relation between Lat. *praecoquum* et Fr. *abricot*, as it covers several steps of different natures.

[12] We include in this category all cases of learned loans such as Fr. *oculaire* 'ocular' ← Lat. *ocularis* 'id.'.

[13] Two lexical entries are homonymous if they share the same citation form, independently of the language or part-of-speech of the two underlying lexemes.



French [edit]

Etymology [edit]

From Middle French *manger*, from Old French *mengier*, from Late Latin *manducāre* ("to chew, devour"), present active infinitive of *manducō*, from Latin *mandō*.

Pronunciation [edit]

- IPA(key): /mɑ̃ʒe/
- Audio (France) 0:00 MENU
- (*Paris*) IPA(key): [mɑ̃ː.ʒe]
- Audio (France, Paris) 0:00 MENU
- Homophones: mangeai, mangé, mangée, mangées, mangés, mangez
- Hyphenation: man·ger

Verb [edit]

manger

1. (*transitive*) to eat
   *J'ai **mangé** de la viande pour le souper.*
   *I **ate** some meat for dinner.*
2. (*intransitive*) to eat
   *C'est bizarre que je ne **mange** rien.*
   *It's strange that I don't **eat** anything.*

Figure 1: Part of a Wiktionary entry

This dump is in a semi-structured format: the structuration into articles is encoded in XML and includes metadata for each article; the content of each article is coded using the so-called "wiki syntax", in which the plain text is supplemented by typographical markers (different levels of titles, lists, etc.) and templates allowing the coding of certain information in a systematic way. For example, the template `link` (or `l`) can be used to encode a form that is a link to the article it is the title of. Thus, `{{link|fr|chaise||chair|g=f}}` will be rendered on the Wiktionary site as chaise *f* ("chair"), where the feminine gender is indicated (`g=f`) and where the word chair is a hyperlink to the section of the Wiktionary article "chaise" concerning the French lexems (`fr`).[14]

Figure 1 shows part of the Wiktionary article "*manger*". The corresponding source code is shown in Figure 2.

Finally, "Descendants" sections are sometimes included. They list the descendants of the lexeme at hand, without any further precision on the nature of the etymological relation (inheritance, borrowing).

[14] The language inventory used by Wiktionary is based on the ISO-639-1 to ISO-639-3 standards, with extensions when needed. For more details, cf. https://en.Wiktionary.org/wiki/Wiktionary:Languages. Based on the correspondence between language codes and language names, we also set up a system for the automatic abbreviation of language names as well as a system for the identification of language (codes) based on their usual names or abbreviations as used in the Wiktionary articles. Thus, "OFr.", "Old Fr." or "Old French" are correctly interpreted as reflecting the `fro` language code, which can then be transformed into its standard English abbreviation, " OFr. "

```
==French==

===Etymology===
From {{inh|fr|frm|manger}}, from {{inh|fr|fro|mengier}}, from {{inh|fr|LL.|manducāre||to chew,
devour}}, present active infinitive of {{m|la|manducō}}, from {{inh|fr|la|mandō}}.

(...)

===Verb===
{{fr-verb}}

# {{lb|fr|transitive}} to [[eat]]
#: ''J'ai '''mangé''' de la viande pour le souper.''
#:: ''I '''ate''' some meat for dinner.''
# {{lb|fr|intransitive}} to [[eat]]
#: ''C'est bizarre que je ne '''mange''' rien.''
#:: ''It's strange that I don't '''eat''' anything.''
#: '''''Manger''' au restaurant.''
#:: ''To '''eat''' in a restaurant.''
```

Figure 2: Source code corresponding to the article part shown in Figure 1 (the pronunciation-related part is not shown)

4.2 Extraction and structuration

We first converted the Wiktionary dump into an XML file using a series of regular expressions.[15] This XML file is a set of lexical entries that correspond approximately to lexemes. It contains only entries for which Wiktionary provides etymological information in a dedicated section. It contains is 831,988 entries. Each of them includes the content of this etymological section in an `<etymology/>` tag, in which all forms, especially but not only those mentioned using *templates*, are represented by an XML element `<form/>`. Whenever several `<form/>` are used together (affixed derivation, composition), their combination is harmonised using the symbol "+" (see above). Whenever several alternate forms are listed (variants, principal parts...), they are separated using the symbol "∼". These apparently simple standardisation steps are made complex by the variety of situations, the richness of the available templates and the multiplicity of ways used by Wiktionary contributors to represent etymological information.

If a section listing descendants is available, they are all converted into `<form/>` elements and are included in a dedicated `<descendants/>` element within the `<etymology/>`.

All forms mentioned in the article but outside the etymological section or the descendant section are also extracted in a special section `<forms/>`. This is because these forms, especially those associated with a gloss, might prove useful in the next steps of the extraction process.

Whenever possible, the lexeme at hand is associated with a gloss. If it is an English lexeme, its citation form is considered as its own gloss. In all other cases, we try to extract one or several glosses based on the definitions provided in the article.

[15] This XML format is a working format. It is not intended at this stage to be suitable for TEI compatibility. We shall return in Section 6.2 on how we exported etymological information to an extended TEI format.

From the source code corresponding to the French verb *manger*, shown in Figure 2, our structuration process outputs the entry given in Figure 3.

```
<entry id="manger#French">
  <header><form lang="Fr." l="fr" sense="to eat; food; foodstuff">manger</form></header>
  <etymology>
From <form lang="MFr." l="frm" trgl="fr" trglang="Fr." type="inherited">manger</form>, from <form
lang="OFr." l="fro" trgl="fr" trglang="Fr." type="inherited">mengier</form>, from <form lang="LL."
l="la-lat" sense="to chew, devour" trgl="fr" trglang="Fr." type="inherited">manducāre</form>, present
active infinitive of <form lang="Lat." l="la">manducō</form>, from <form lang="Lat." l="la" trgl="fr"
trglang="Fr." type="inherited">mandō</form>.
  </etymology>
  <forms>
    <form lang="Fr." l="fr">gramen</form>
    <form lang="Fr." l="fr">magner</form>
  </forms>
</entry>
```

Figure 3: Output of our structuration process for the input source code shown in Figure 2

5. Construction of the etymological database

The output of the structuration process described in the previous section is much easier to exploit than the original Wiktionary dump. However, several challenges remain. The main one is of course that the etymological information is given in plain English, apart from the `<form/>` elements. Another one is that, from one article to the other, a same lexeme can associated with different glosses, if any.

To address these challenges, we proceed in several steps. First, we defined a number of regular patterns for infering the gloss of a non-glossed form based on its context.[16] In such a case, the corresponding `<form/>` element is updated accordingly.

We then process all entries and the etymological information they contain, in order to create triples of the form (target lexeme, source lexeme or source lexeme sequence, type of the relation). We now have to merge synonymous lexemes as much as possible. For instance, if the triples we built involve lexemes such as Fr. *bêtement* 'stupidly, idiotically', Fr. *bêtement* '(no gloss)' and Fr. *bêtement* 'stupidly, foolishly', these three lexemes need to be merged into a lexeme Fr. *bêtement* 'stupidly, idiotically, foolishy'. To achieve this goal, we iterate the following steps until stability:

- If a gloss-less lexeme has the same language and the same citation form as exactly one (glossed) lexeme, then these lexemes are merged.
- If two glossed lexemes have the same language, the same citation form, and at least one gloss in common (cf. 'stupidly, foolishly' vs. 'stupidly, idiotically'), then they are merged (in this example, it creates a lexeme with the gloss 'stupidly, foolishly, idiotically', as mentioned above);
- All triples encoding etymological relations are then updated accordingly.

[16] Coming back to our French running example *manger*, the phrase "From Middle French *manger*, from Old French *mengier*...", although it contains no glosses, makes it possible to associate the gloss of the head lexeme *manger*, namely 'to eat', to MFr. *manger* et OFr. *mangier*.

In order to restrict as much as possible our set of etymological relations to direct relations, we remove any relation between two lexemes $lexeme_1$ et $lexeme_3$ such that there exists a relation between $lexeme_1$ and an intermediate lexeme $lexeme_2$ and a relation between this $lexeme_2$ and $lexeme_3$.[17]

Finaly, the type of certain relations is corrected, in order to indicate as precisely as possible cases of borrowing or morphological derivation rather than inheritance, this latter case still remaining the default one.

The outcome of this extraction process is twofold: a set of lexemes, only some of them being glossed, and a set of etymological relations involving a target lexeme, one or more source lexemes (two or more in case of composition or affixal derivations) and a relation type. Here are a few real examples concerning French lexemes:

- Fr. *gobelet* 'goblet' < OFr. *gobel* 'goblet; cup; beaker; tumbler'
- Fr. *maudire* 'to curse' < OFr. *maudire* ∼ *maldire* 'to curse'
- Fr. *éponger* 'to sponge; to absorb' $<_s$ Fr. *éponge* 'sponge' + Fr. *-er*
- Fr. *idéologie* 'ideology' $<_d$ Fr. *idéo-* + Fr. *-logie*
- Fr. *acajou* 'cashew' ← Port. *acajú* 'cashew'
- Fr. *car* 'car; coach' ← E *car*

6. Results and evaluation

6.1 Quantitative information

The initial extraction process described at the beginning of Section 5 has produced almost 1.2 million lexemes, 62,056 lexeme sequences and 548,935 etymological relations between two lexemes or between a lexeme and a sequence of lexemes.[18] A few dozen iterations of the lexeme merging algorithm merged 199,185 lexemes and 289 lexeme sequences, resulting in 975,473 distinct lexemes, 61,809 distinct lexeme sequences and 519,348 distinct relations. After discarding 5,149 non-diret relations, the final number of relations is 514,199.

The lexemes obtained belong to 2311 distinct languages, the best represented of which are, in decreasing order, English (257,978 lexemes), Latin (65,981), French (32,044), Italian (28,028), and Ancient Greek (21,077). Among these lexemes, 659,567 (68%) have a gloss.

Among the 514,199 relations, 452,041 relate two lexemes, whereas other relations relate a lexeme and a lexeme sequence. There are 90,511, cognation relations, all other 423,673 relations being direct relations. Finally, 318,883 relations involved glossed lexemes only.

Note that we could have easily created many more cognation relations by adding relations sharing a (direct or indirect) ancestor in our database.

6.2 Etymological chain inference and TEI export

We have developed an export module for our etymological relation database that encodes data in the TEI format proposed by Bowers & Romary (2016). In this format, direct

[17] The same mechanism applies when $lexeme_3$ is not a unique lexeme but a sequence of lexemes.
[18] In these figures and in all figures below, lexemes invovled in zero relations are not counted.

relations wan be exported in the form of simple <etym/> elements, associated with the type of the relation at hand.

For a given lexeme, it can also be interesting to have not only its direct etymon but also its etymological history in as exhaustive a way as possible. In fact, such an etymological chain is often provided in the etymological information included in Wiktionary articles, as exemplified in Figures 1 and 2. In order to re-build these etymological chains (or derivations) from our relation database, one can simply recursively retrieve relations involving each etymon involved. For instance, from Fr. *manger* 'to eat' < MFr. *manger* 'to eat' and MFr. *manger* 'to eat' < OFr. *mengier* 'to eat', one can re-build the chain Fr. *manger* 'to eat' < MFr. *manger* 'to eat' < OFr. *mengier*. This is how we created etymological chains, before encoding them in TEI.

We had to extend the proposal by Bowers & Romary (2016) in four directions, which could serve as a source of inspiration for its further improvement:

- This proposition does not cover the cogation relation. We therefore introduce an additional relation type (*type="cognate"*) to the element <etym/>.
- It does not allow to refer to another lexical entry providing relevant etymological relation. We therefore introduce a new type (*type="reference"*) to the <etym/> element, within which a direct reference to the relevant lexical entry can be included with an <xr/> (TEI element used for cross-references).
- Bowers & Romary (2016) do not provide any way to encode etymological chains. We simply used a special <etym/> element, which, using a dedicated attribute, indicates that it contains a sequence of etymological relations, each of them being represented by a specific <etym/> element within the global <etym/> element.
- In their document, Bowers & Romary (2016) do not allow for alternative etymological hypotheses, something which is frequent in our database. In this case, we also make use of a special <etym/> element, which indicates using a dedicated attribute that it contains alternative hypotheses, each of them represented by a distinct <etym/>.

In the two last cases, the recursivity of <etym/> elements allows for any possible combinations, such as an etymological chain starting with two "certain" steps followed by an alternative between two different etymological sub-chains.

6.3 Manual evaluation

The evaluation of our etymological relation database could be carried out with four different questions in mind:

1. What is the quality of the etymological information provided by the Wiktionary?
2. What are the errors caused by our extraction and structuration process?
3. What are the errors introduced by our gloss inference and lexeme merging algorithms? Conversely, what is the coverage of these algorithms?
4. Which errors result from the fact that non-typed relations are interpreted by default as inheritance relations?

A detailed answer to the first question is not straightforward, and falls beyond the scope of this paper. An informal study of the etymological information found in a random set

```
<entry xml:id="sla-pro:gostinŭ:guest" xml:lang="sla-pro">
  <form type="lemma">
    <orth>gostinŭ</orth>
  </form>
  <sense>
    <cit type="translation" xml:lang="en">
      <oRef>guest<oRef>
    </cit>
  </sense>
  <etym type="suffixalDerivation">
    <cit type="etymon">
      <oRef xml:lang="sla-pro">gostĭ</oRef>
      <gloss>guest</gloss>
      <etym type="inheritance">
        <cit type="etymon">
          <oRef xml:lang="ine-pro">gʰóstis</oRef>
          <gloss>stranger, guest, host, someone with whom one has reciprocal duties of
                 hospitality</gloss>
        </cit>
      </etym>
    </cit>
    <cit type="etymon">
      <oRef xml:lang="sla-pro">-inŭ</oRef>
    </cit>
  </etym>
</entry>
```

Figure 4: Example of a TEI-formatted entry (cognation relations are not shown)

of articles showed that this information is usually reliable. Only Proto-Indo-European etyma sometimes reflect of a somewhat obsolete knowledge of the field. Nevertheless, it can be considered that etymological information in Wiktionary can generally be trusted, and often reflect the most recent and consensual scientific literature, which are often cited in the references.

The precision and recall of our gloss inference and lexeme merging algorithms are easier to evaluate. We first focused on the recall of the merging algorithm. We randomly selected 50 (language, citation form) pairs among the 124,775 ones (out of 941,757) that correspond to more than one entries. We then extracted all entries for these 50 pairs, and have manually annotated the relevance of their co-existence (as opposed to merging them). In almost all cases, additional merges would have been relevant, but our algorithm was not able to perform these merges. It is therefore an obvious direction for future improvements. Conversely, in order to evaluate the precision of our merging algorithm and that of our gloss inference algorithm, we randomly extracted 100 glossed forms and checked the quality and coherence of their glosses. Out of these 100, we identified two extraction errors (both caused by an unusual use of the "wiki" syntax by contributors), a partial error (some of the glosses are correct, one of them is an easily dismissable "wiki" code fragment), a transcription misinsterpreted as a gloss, and a (correct) definition misinterpreted as a gloss. All other glossed forms were fully correct. Therefore, there are only a few errors, which are almost never caused by our merging and glossing algorithms—yet the extraction and structuration algorithm could be slightly improved.

Finally, we evaluated the etymological relations themselves based on a random set of size 100. Among them, 78 are correct, 18 have type "inheritance" whereas they encode

borrowings, three of them have other relation typing errors, and only one is errineous because of an error while extracting the article. The 18 errors of type "inheritance" instead of "borrowing" are the result of the fact that inheritance is the default relation type, used when the latter is not explicitely provided in the Wiktionary. A finer description of the relations between languages would make it possible to automatically correct these examples. This is something we will do in the near future.

6.4 Comparison with EtymWordnet

EtymWordNet de Melo (2014), freely available without an explicit license,[19] is an etymological database extracted from Wiktionary, although from a dump dating back to 2013. In this resource, contrarily to the one we built, relations are not typed with a sufficient granularity (it only distinguishes between a cognacy relation and a generic etymological origin relation).[20] Moreover, it relates non-glossed citation forms (rather than lexemes). Nevertheless, it is the only resource that is comparable with ours. We therefore evaluated our etymological relation database with respect to this resource.

EtymWordNet contains 473,433 direct yet non-typed etymological relations as well as 538,558 cognacy relations. As mentioned above, many cognation relations can be added based on other relations. The most interesting etymological information is provided by these other relations, which are unfortunately not distinguished within EtymWordNet. Another issue with EtymWordNet is that derivation and composition relations are not modelled in a satisfying way. For instance, (American) English *monophthongize* is the source of two independent etymological relations, one with *-ize* and the other one with *monophthong*.

To make the comparison possible, we had to transform our relations (excluding cognation relations) so that they follow the same model as EtymWordNet. Unsurprisingly, this slightly lowers the number of relations to 559,614. Among them, 464,542 (83%) are not found in EtymWordNet. Conversely, 378,361 relations are only found in EtymWordNet. But among these 378,361 relations, 333,369 (88%) relate forms from the same language: they are derivation or composition relations, extracted from other parts of the Wiktionary articles than the etymological part we exploited (especially the "derived terms" sections). Such relations are less interesting from an etymological point of view. Among the other missing relations, a large (yet hard to quantify) number of them are almost identical to relations that are included in our database, differing only by diacricts added in Wiktionary since 2013. Overall, this comparison shows that our database is significantly richer than EtymWordNet—and recall that our database relates (mostly glossed) lexemes with typed relations correctly representing inheritance, borrowings, derivation and composition, whereas EtymWordNet relates (non-glossed) citation forms with non typed relations (apart from the notion of cognacy relation).

7. Future work: improvement and use of our etymological database

The work presented here shall be improved in two ways. Firstly, patterns used for information extraction from etymological sections in Wiktionary articles can be extended,

[19] http://www1.icsi.berkeley.edu/~demelo/etymwn/
[20] We ignore relations of the type "orthographic variant".

improved, refined. Secondly, the lexeme merging algorithm can be enriched so as to merge lexemes which are not merged yet, mostly because variation of the following types:

- formal variations: differences in transcription or notation, form with or without stress information,[21] complete citation form vs. truncated citation form vs. principal parts;[22]
- variation in glosses[23] (for instance using WordNet or distributional similarity information).

The way we gloss lexemes that have no gloss in Wiktionary can also be improved, for instance by better taking advantage of their context of occurrences and by using external bilingual or multilingual resources.

A model of the phylogenetic relations between languages would also help replacing indirect relations with direct ones, either using simple heuristics or based on a (partial) model of phonetic (and maybe morphological) change. For instance, the relation Fr. *chapitre* 'chapter' < OFr. *chapitre* 'chapter' could be replaced by a relation MFr. *chapitre* 'chapter' < OFr. *chapitre* and a relation MFr. *chapitre* 'chapter' < OFr. *chapitre*, simply by knowing that, given our language inventory, the immediate ancestor of French is Middle French, whose immediate ancestor is Old French. It could help extending the lexicons for a number of intermediate languages with attested words, which could be validated using external lexical resources, or even unattested words.

Finally, it would be useful to extract the etymological information available in other Wiktionary editions, especially its French edition, the Wiktionnaire. Our databases are only affected by the language of the original information source at the level of glosses. We could automatically replace French glosses extracted from the Wiktionnaire by English glosses, for example by exploiting the English translations provided in the Wiktionnaire articles themselves.

In addition to lexicon extension for intermediate languages, as mentioned above, the resource presented in this article could be used as a starting point for research in computational historical linguistics, as suggested in the introduction. It may also be subjected to automated internal consistency checks, for example by automatically extracting phonetic laws and verifying their systematic applicability, modulo the analogical levelling phenomena. In the long term, this could also allows the construction or the automatic completion of large-coverage etymological dictionaries.

8. References

Bowers, J. & Romary, L. (2016). Deep encoding of etymological information in TEI. URL https://hal.inria.fr/hal-01296498. Working draft.

Buchi, E. (2016). Etymological dictionaries. In P. Durkin (ed.) *The Oxford Handbook of Lexicography*. Oxford University Press, pp. 338–349.

de Melo, G. (2014). Etymological Wordnet: Tracing the History of Words. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, LREC 2014. Reykjavik, Iceland, pp. 1048–1054.

[21] Gr. πλάξ 'flat stone' vs. Gr. πλαξ 'flat stone' are not merged yet.

[22] PIE *deh₂mo-* '(pas de glose)' and PIE *deh₂mos* 'people' are not merged yet.

[23] Fr. *aise* 'ease' and Fr. *aise* 'satisfaction, joy' are not merged yet, which prevents a further merging with Fr. *aise* '(pas de glose)' because of the resulting spurious ambiguity.

Salmon-Alt, S. (2006). Data structures for etymology: towards an etymological lexical network. *BULAG*, 31, pp. 1–12.

ORGANIZERS

/instituut voor de Nederlandse taal/

Lexical Computing

trojína INSTITUTE FOR APPLIED SLOVENE STUDIES

PROGRAMME SPONSOR

idm

SPONSORS

Dictionary.com

Oxford Dictionaries

Collins

SUPPORTING INSTITUTIONS

KDICTIONARIES

globaLex

Amsterdam University Press