Electronic lexicography in the 21st century: **post-editing lexicography**

Proceedings of the eLex 2021 conference

edited by

Iztok Kosem
Michal Cukr
Miloš Jakubíček
Jelena Kallas
Simon Krek
Carole Tiberius

virtual, 5–7 July 2021

elex.link/elex2021

Electronic lexicography in the 21st century (eLex 2021):
Post-editing lexicography
Proceedings of the eLex 2021 conference.

edited by

Iztok Kosem
Michal Cukr
Miloš Jakubíček
Jelena Kallas
Simon Krek
Carole Tiberius

published by

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

proofreading by

Paul Steed
Dean DeVos
Jan Nagel

Brno, July 2021

ISSN 2533-5626

eLex 2021

ORGANIZERS

LEXICAL COMPUTING

cjvt

/instituut voor de Nederlandse taal/

Univerza *v Ljubljani*

Institute of the Estonian Language

elexis european lexicographic infrastructure

elex.link/elex2021

CONFERENCE COMMITTEES

Organising Committee

Iztok Kosem
Michal Cukr
Miloš Jakubíček
Jelena Kallas
Simon Krek
Carole Tiberius

Scientific Committee

Andrea Abel
Špela Arhar Holdt
Kristian Blensenius
Gerhard Budin
Nicoletta Calzolari
Lut Colman
Paul Cook
Margarita Correia
Gilles-Maurice de Schryver
María José Dominguez Vazquez
Patrick Drouin
Edward Finegan
Thierry Fontenelle
Polona Gantar
Yongwei Gao
Radovan Garabik
Zoe Gavriilidou
Alexander Geyken
Kris Heylen
Aleš Horák
Miloš Jakubíček
Maarten Janssen
Jelena Kallas
Ilan Kernerman
Maria Khokhlova
Annette Klosa-Kückelhaus
Svetla Koeva
Kristina Koppel
Iztok Kosem
Vojtěch Kovář
Simon Krek
Michal Kren
Tanara Zingano Kuhn
Margit Langemets
Lothar Lemnitzer
Robert Lew
Pilar León Araúz
Henrik Lorentzen
Stella Markantonatou
John P. McCrae
Amalia Mendes
Michal Boleslav Měchura
Julie Miller
Monica Monachini
Orion Montoya
Sara Može
Christine Möhrs
Chris Mulhall
Carolin Müller-Spitzer
Lionel Nicolas
Sussi Olsen
Vincent Ooi
Jordi Porta
Adam Rambousek
Laurent Romary
Hindrik Sijens
Emma Sköldberg
Nicolai Hartvig Sørensen
Egon Stemle
Vít Suchomel
Kristina Štrkalj Despot
Arvi Tavast
Carole Tiberius
Yukio Tono
Lars Trap Jensen
Agnes Tutin
Tamas Varadi

elex.link/elex2021

TABLE OF CONTENTS

Corpus-based Methodology for an Online Multilingual Collocations Dictionary: First Steps

Adriane Orenha-Ottaiano[1], Marcos Garcia [2], Maria Eugênia Olímpio de Oliveira Silva[3], Marie-Claude L'Homme[4], Margarita Alonso Ramos[5], Carlos Roberto Valêncio[6], William Tenório[7]

[1] São Paulo State University (UNESP), Brazil
[2] Universidade de Santiago de Compostela, Galiza, Spain
[3] University of Alcalá, Spain
[4] OLST, Université de Montréal, Québec, Canada
[5] Universidade da Coruña, Spain
[6] São Paulo State University (UNESP), Brazil
[7] São Paulo State University (UNESP), Brazil

E-mail: adriane.ottaiano@unesp.br, marcos.garcia.gonzalez@usc.gal, eugenia.olimpio@uah.es, mc.lhomme@umontreal.ca.ca, margarita.alonso@udc.es, carlos.valencio@unesp.br, williamtenoriotenorio@gmail.com

Abstract

This paper describes the first steps of a corpus-based methodology for the development of an online Platform for Multilingual Collocations Dictionaries (PLATCOL). The platform is aimed to be customized for different target audiences according to their needs. It covers various syntactic structures of collocations that fit into the following taxonomy: verbal, adjectival, nominal, and adverbial. Part of its design, layout and methodological procedures are based on the *Bilingual Online Collocations Dictionary Platform* (Orenha-Ottaiano, 2017). The methodology also relies on the combination of automatic methods to extract candidate collocations (Garcia et al., 2019a) with careful post-editing performed by lexicographers. The automatic approaches take advantage of NLP tools to annotate large corpora with lemmas, PoS-tags and dependency relations in five languages (English, French, Portuguese, Spanish and Chinese). Using these data, we apply statistical measures (Evert et al., 2017; Garcia et al., 2019b) and distributional semantics strategies to select the candidates (Garcia et al., 2019c) and retrieve corpus-based examples (Kilgarriff et al., 2008). We also rely on automatic definition extraction (Bond & Foster, 2013) so that collocations can be more effectively organized according to their specific senses.

Keywords: collocations; collocations dictionary; online platform; automatic extraction; lexicography

1. Introduction

In the past two decades, collocations have been high on the agenda of foreign language teaching and learning (Nesselhauf, 2005; Alonso-Ramos, 2008, 2019; Laufer, 2011; Orenha-Ottaiano, 2021; Torner & Bernal, 2017, among others). Despite this fact, when it comes to the translation of collocations, the number of studies that can contribute to better comprehension of the difficulties regarding the complexity of translation of such combinations is not as significant (Kenny, 2001; Bernardini, 2007; Gregorio-Godeo & Molina, 2011; Orenha-Ottaiano, 2009, 2012, forthcoming).

Additionally, even though several authors emphasise the importance of compiling dictionaries with a special focus on collocations or for the building of specific collocations dictionaries (Alonso-Ramos, 2001; Atkins & Rundell, 2008; Moon, 2008; Orenha-Ottaiano, 2013, 2015, 2017; Kilgarriff, 2015, etc.), the number of online or electronic collocations dictionaries available is still scarce, especially when it comes to bilingual or multilingual collocations dictionaries for general language.

The work described in this paper aims to fill this gap. We describe a methodology for the design and compilation of an online platform for multilingual collocations dictionaries (English, Portuguese, French, Spanish and Chinese). The collection of relevant collocations is corpus-based and semi-automated (automatic extraction with human validation). Furthermore, the design of the platform takes into consideration users' needs as suggested by the principles of the function theory of lexicography (Bothma & Tarp, 2012; Fuertes-Olivera & Tarp, 2014; Tarp, 2015).

Besides the introduction, the paper is structured as follows. Section 2 addresses the motivational aspects for the development of a corpus-based methodology of multilingual collocations dictionaries and an online platform. Section 3 outlines the methodological steps used in this research. Section 4 explores the Multilingual Collocations Dictionary's structure and design. Finally, Section 5 presents the concluding remarks and highlights some ideas for further work.

2. Motivation

One of the main motivations for carrying out this research is that collocations require specific pedagogical attention. Concerning lexicographical work, excellent monolingual collocations dictionaries for learners of English as a second or foreign language are available, such as the *Longman Collocations Dictionary and Thesaurus* (2013), *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), with the last two are only available in paper format.

In Portuguese, to the best of our knowledge, the only online and corpus-based dictionary of collocations is the one developed by Orenha-Ottaiano (2017). As it is bi-directional, and users can consult it both as a monolingual (either Portuguese or English) or as a bilingual (English-Portuguese and Portuguese-English).

In Spanish, the *Diccionario combinatorio práctico del español contemporáneo* (Bosque, 2006) is a corpus-based dictionary for native or foreign language speakers of Spanish, which focuses not only on collocations but also on other phraseologisms, such as idioms (*locuciones fijas*). The *Diccionario de colocaciones del español* (*DiCE*; Alonso-Ramos, 2004) is available online and encodes collocations according to the principles of the Meaning-Text Theory.

In French, Beauchesne's the *Dictionnaire des Cooccurrences* (2001) is an example of a printed and online monolingual collocations dictionary, but it is not corpus-based. The *DiCouèbe* (Jousse & Polguère, 2005) is an online French combinatorial dictionary in which collocations are all encoded with Lexical Functions.

In Chinese, we can mention the *Modern Chinese Collocation Dictionary* (Mei, 1999) and *Dictionary of Chinese Common Word Collocations* (Yang, 1990).

As far as bilingual dictionaries are concerned, as previously mentioned, Orenha-Ottaiano (2016, 2017) built an online platform of bilingual Collocations Dictionary (English-Portuguese and Portuguese-English), which has recently been changed into a platform of multilingual collocations dictionaries, as discussed in this paper. Alegro et al. (2010) published a printed dictionary containing 3,000 adjectival collocations (Portuguese-English), but it is neither corpus-based nor in an electronic or online format.

The *DiCoEnviro* (L'Homme et al., 2018) and the *DiCoInfo* (L'Homme, 2008) are online terminological dictionaries in English, French and Spanish (a few Portuguese, Italian and Chinese terms are also listed) that focus on specialized terms, encodes specialized collocations and explain the meaning of collocates using the system of lexical functions (Mel'čuk, 1996).

Finally, another bilingual dictionary worth mentioning is *The Oxford Collocations Dictionary* (English-Chinese), both printed and app versions.

A lot of research has taken place on corpus-based and online bilingual or multilingual collocations dictionaries in other languages, such as the Dictionary of Collocations of European Portuguese (Pereira & Mendes, 2002), a dictionary of Italian collocations (Spina, 2010), an investigation on the automatic construction of a multilingual dictionary of collocations (Garcia et al., 2019a), and a bilingual English-Italian dictionary of collocations (Berti & Pinnavaia, 2014), among others. Nevertheless, there is still a gap in the availability or publication of online dictionaries themselves as they are research proposals and have not been published yet.

Another motivational aspect of this project concerns the possibility of developing a platform offering a higher degree of customisation of the structure of the dictionaries. It aims at the development of an innovative lexicographical methodology and model for a multilingual collocations dictionary, as well as the design of a collocations software and platform, the PLATCOL[1]. Moreover, it targets the setting up of a useful and large

[1] The Platform for Multilingual Collocations Dictionaries (PLATCOL) is the practical result of the project *A phraseographical methodology and model for an online corpus-based Multilingual Collocations Dictionary Platform*, sponsored by The São Paulo Research Foundation (FAPESP). It is a two-year project with a partnership between São Paulo State University (Brazil), responsible for English and Portuguese languages, the University of Montréal (French), University of Granada (Chinese), University of Coruña and University of Alcalá (Spanish), and University of Santiago de Compostela, for the automatic retrieval of corpus data.

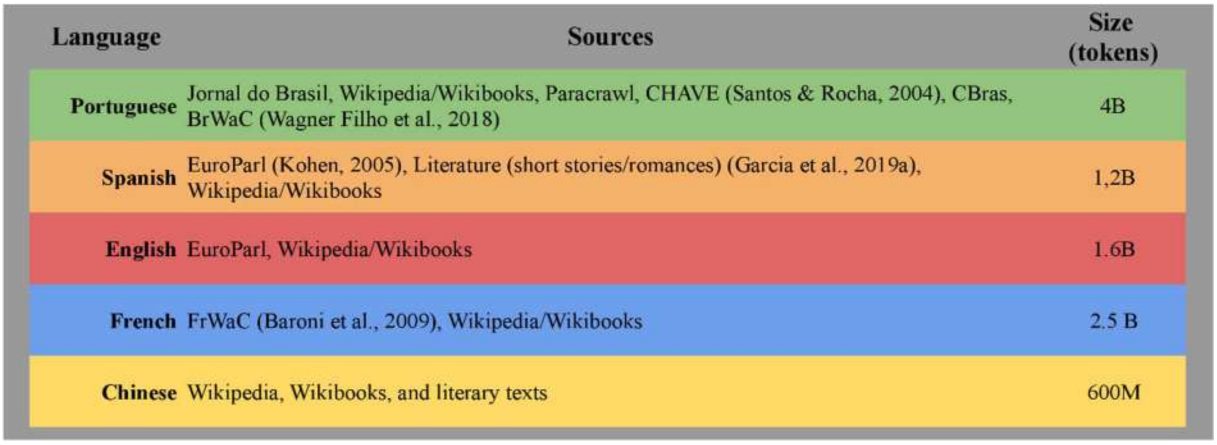
resource for semi-automatic collocations retrieval, as well as automatic extraction of good examples, definitions and translation.

3. Methodology

The methodology to build the dictionary is based on the automatic approach described in Garcia *et al.* (2019a), enriched with sense information of the bases and a manual review and validation of the extracted data made by lexicographers.

3.1 Corpora

We compiled a large corpus for each of the five languages of the project using different source data, as Table 1 below shows:



Language	Sources	Size (tokens)
Portuguese	Jornal do Brasil, Wikipedia/Wikibooks, Paracrawl, CHAVE (Santos & Rocha, 2004), CBras, BrWaC (Wagner Filho et al., 2018)	4B
Spanish	EuroParl (Kohen, 2005), Literature (short stories/romances) (Garcia et al., 2019a), Wikipedia/Wikibooks	1,2B
English	EuroParl, Wikipedia/Wikibooks	1.6B
French	FrWaC (Baroni et al., 2009), Wikipedia/Wikibooks	2.5 B
Chinese	Wikipedia, Wikibooks, and literary texts	600M

Table 1: Corpora Size and Sources

The corpora were parsed with UDPipe (Straka & Straková, 2017) using the latest models (v2.7) trained on the UD corpora (de Marneffe *et al.*, 2021). Previous to this syntactic analysis, we tokenized and PoS-tagged the data using the same UDPipe models for English and French, LinguaKit (Gamallo *et al.*, 2018) for Portuguese and Spanish, and the Stanford CoreNLP suite (Manning *et al.*, 2014) for the Chinese texts.

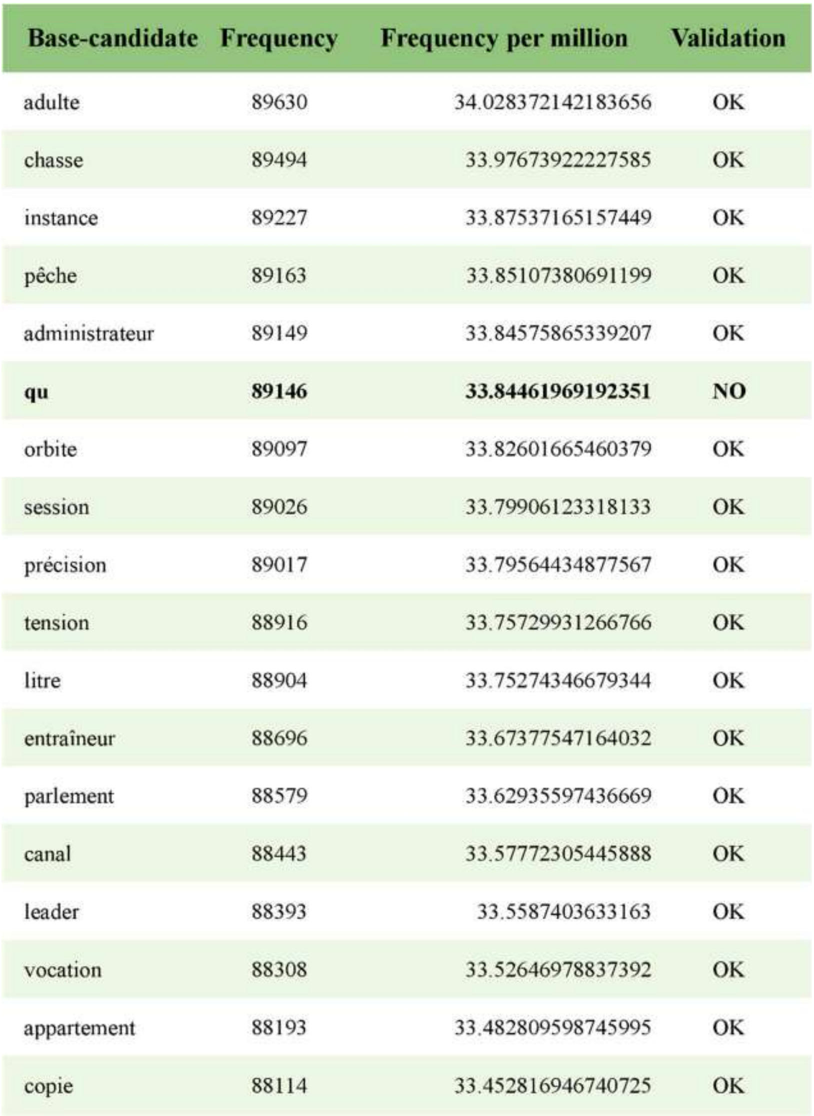
3.2 Definition and extraction of keywords

We focus on collocation types with three morphosyntactic classes of bases: nouns, verbs, and adjectives. Due to the large size of the corpora, we attempt to extract basic vocabulary lists for each class and language. Therefore, we automatically extracted the lemmas of the nouns with a minimum frequency of one occurrence per million tokens in each corpus, annotating them as *known* or *unknown* if they appear in large lexica[2]. We used the dictionaries provided by FreeLing (Padró & Stanilovsky, 2012) for each language (English, Portuguese, French and Spanish), except for Chinese. We didn't use any lexicon for Chinese because we are not aware of any free dictionary for this language.

[2] Due to the lower frequency of verbs and adjectives, we used frequency=>0.5 in these cases.

After the automatic extraction, which took place for each language separately, the lists of keywords were submitted to the lexicographers to filter out noise (e.g., lemmas with typos, entries wrongly processed, etc.) and to select the most frequent lemmas, then used to extract candidate collocations. Besides, each keyword has been enriched with the potential senses present in WordNet, using the Open Multilingual WordNet (Bond & Foster, 2013) by means of the interface provided by the NLTK package (Bird & Klein, 2009).

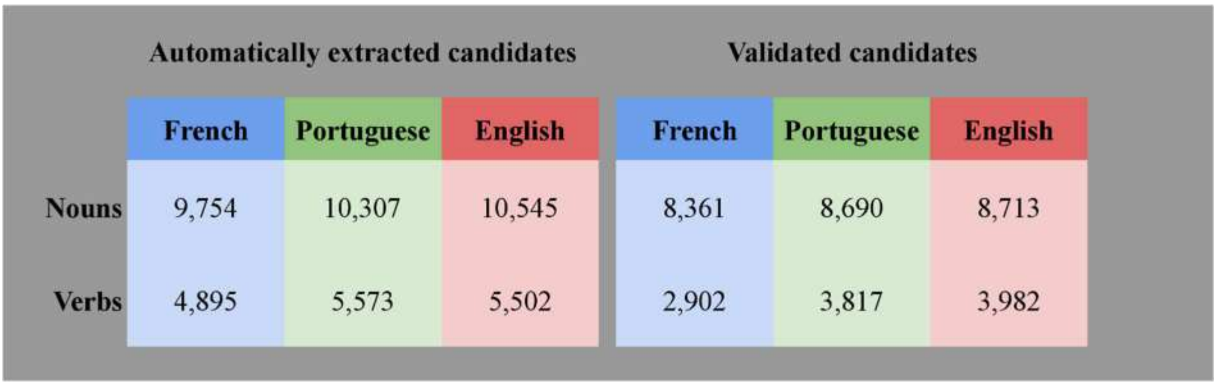
Table 2 shows a sample of keywords in French as an example, sorted by descending order of frequency. Candidates marked NO by lexicographers were removed from the list.



Base-candidate	Frequency	Frequency per million	Validation
adulte	89630	34.028372142183656	OK
chasse	89494	33.97673922227585	OK
instance	89227	33.87537165157449	OK
pêche	89163	33.85107380691199	OK
administrateur	89149	33.84575865339207	OK
qu	**89146**	**33.84461969192351**	**NO**
orbite	89097	33.82601665460379	OK
session	89026	33.79906123318133	OK
précision	89017	33.79564434877567	OK
tension	88916	33.75729931266766	OK
litre	88904	33.75274346679344	OK
entraîneur	88696	33.67377547164032	OK
parlement	88579	33.62935597436669	OK
canal	88443	33.57772305445888	OK
leader	88393	33.5587403633163	OK
vocation	88308	33.52646978837392	OK
appartement	88193	33.482809598745995	OK
copie	88114	33.452816946740725	OK

Table 2: Results of validation in French

After having manually validated the base candidates in each language separately, we reached the following results for English, French and Portuguese, shown in Table 3.



	Automatically extracted candidates			Validated candidates		
	French	Portuguese	English	French	Portuguese	English
Nouns	9,754	10,307	10,545	8,361	8,690	8,713
Verbs	4,895	5,573	5,502	2,902	3,817	3,982

Table 3: Number of automatically extracted and validated candidates

As can be noted in Table 3, about 15% of nouns were discarded in French, 16% in Portuguese, and 18% in English. As for the verbs, 40% of them were discarded in French, 32% in Portuguese, and 28% in English. These results highlight the importance of post-editing in all lexicographical phases.

3.3 Identification of collocations and example sentences

Following Garcia *et al.* (2017) we extract pairs of the target dependency relations using the manually validated keywords and restricting the potential collocates for their morphosyntactic category. Thus, for noun bases we extract the following syntactic relations:[3] *obj* (verb-noun collocations), *nsubj* (instances of noun-verb), *obl* (verb-preposition-noun), *amod* (adjective-noun), and *nmod* and *compound* (both including noun-noun or noun-prep-noun instances). For verb bases we extract *xcomp* (verb-adjective collocations) and *advmod* (verb-adverb). Finally, for adjective bases, we extract *advmod* examples (adjective-adverb candidates).

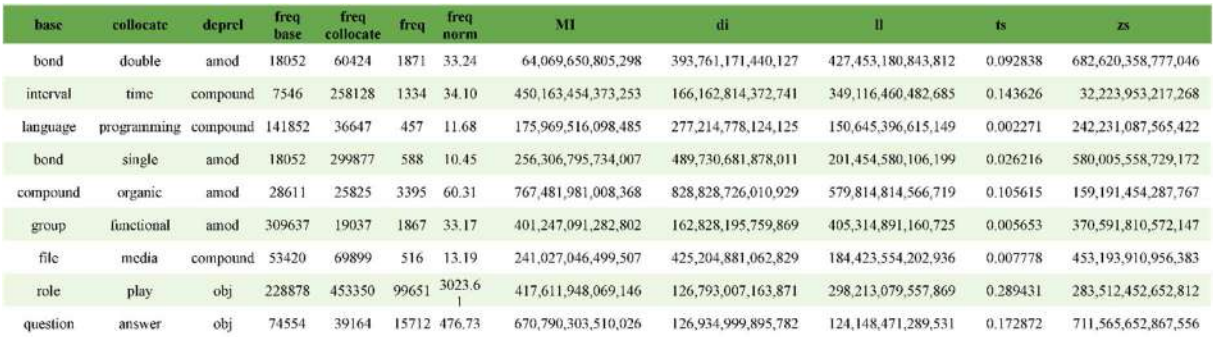
For each triple (base;collocate;relation) we follow the syntactic co-occurrence method described in Evert (2008) to compute, apart from frequency data, the following statistical values: PMI, Dice, log-likelihood, t-score, z-score, ^{2}, and simple-ll (together with ΔP (Gries, 2013). In order to reduce the large size of the candidates sets we remove those combinations with a normalized frequency lower than one per million, and sort the remaining ones by t-score (Garcia *et al.*, 2019b).

Then, we collect up to eight sentences for each candidate collocation, selected by a set of GDEX-inspired heuristics (Kilgarriff et al., 2008). We have implemented a basic strategy using some of the proposals of Kosem et al. (2019a) for English and for Portuguese (the latter were also used for the other romance languages): sentences with less than six tokens are discarded, and those with more than 30 tokens are incrementally penalized. Furthermore, sentences with punctuation, proper nouns, words with more

[3] https://universaldependencies.org/u/dep/all.html

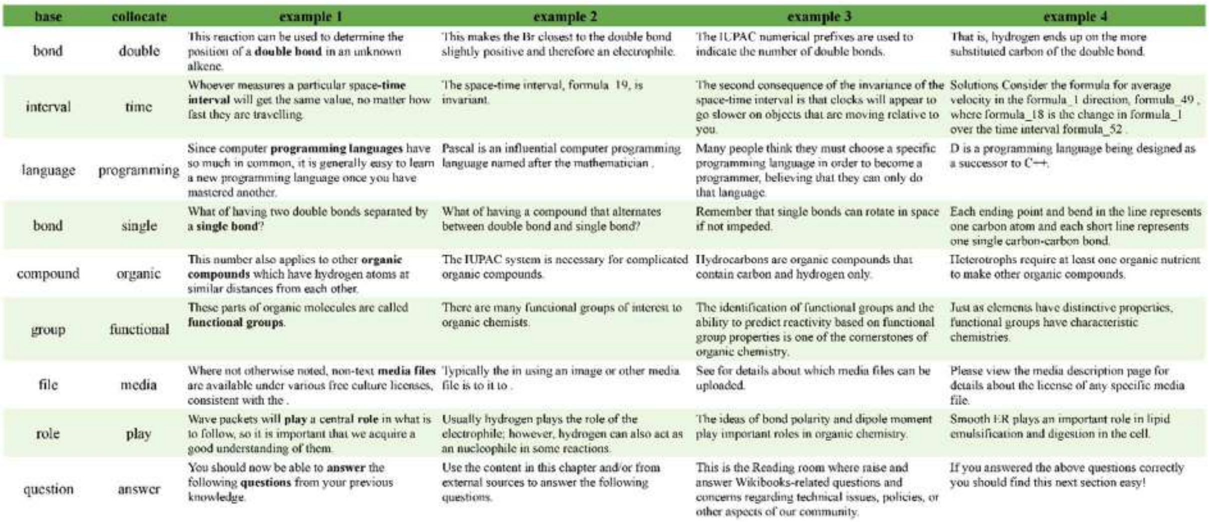
than 12 characters, and strange characters (e.g., in other alphabets and encodings) are also penalized. Other heuristics in the literature were not implemented as they require language-specific resources or are computationally very expensive.

This automatically extracted information is then used by language experts to select the collocations for the final resource. For each candidate, the lexicographers decide which combinations are going to be incorporated into the dictionary, and select the appropriate sense for the base and a set of five examples to be shown on the platform. The tables below show examples of automatically retrieved data in English (Tables 4 and 5) and in Portuguese (Tables 6 and 7) from *noun* bases, showing collocates, frequencies, some of the statistical score results and examples (four out of eight) – the first example has collocations highlighted manually.



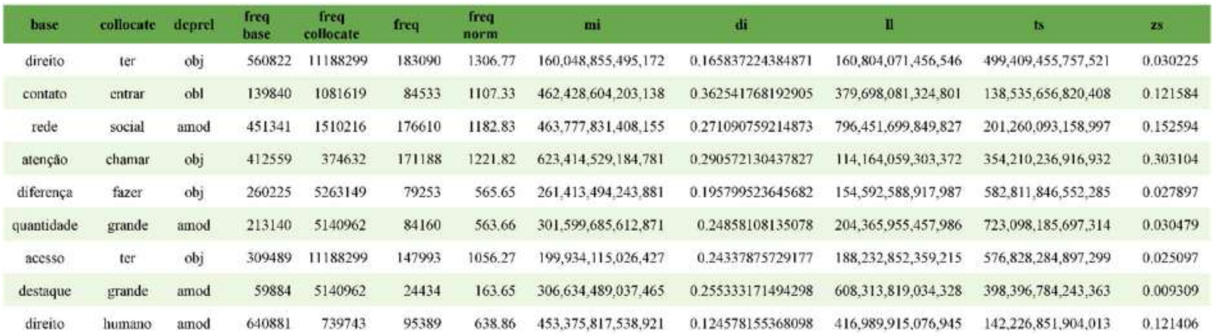
base	collocate	deprel	freq base	freq collocate	freq	freq norm	MI	di	ll	ts	zs
bond	double	amod	18052	60424	1871	33.24	64,069,650,805,298	393,761,171,440,127	427,453,180,843,812	0.092838	682,620,358,777,046
interval	time	compound	7546	258128	1334	34.10	450,163,454,373,253	166,162,814,372,741	349,116,460,482,685	0.143626	32,223,953,217,268
language	programming	compound	141852	36647	457	11.68	175,969,516,098,485	277,214,778,124,125	150,645,396,615,149	0.002271	242,231,087,565,422
bond	single	amod	18052	299877	588	10.45	256,306,795,734,007	489,730,681,878,011	201,454,580,106,199	0.026216	580,005,558,729,172
compound	organic	amod	28611	25825	3395	60.31	767,481,981,008,368	828,828,726,010,929	579,814,814,566,719	0.105615	159,191,454,287,767
group	functional	amod	309637	19037	1867	33.17	401,247,091,282,802	162,828,195,759,869	405,314,891,160,725	0.005653	370,591,810,572,147
file	media	compound	53420	69899	516	13.19	241,027,046,499,507	425,204,881,062,829	184,423,554,202,936	0.007778	453,193,910,956,383
role	play	obj	228878	453350	99651	3023.61	417,611,948,069,146	126,793,007,163,871	298,213,079,557,869	0.289431	283,512,452,652,812
question	answer	obj	74554	39164	15712	476.73	670,790,303,510,026	126,934,999,895,782	124,148,471,289,531	0.172872	711,565,652,867,556

Table 4: Automatically retrieved data from the English corpus – base = noun



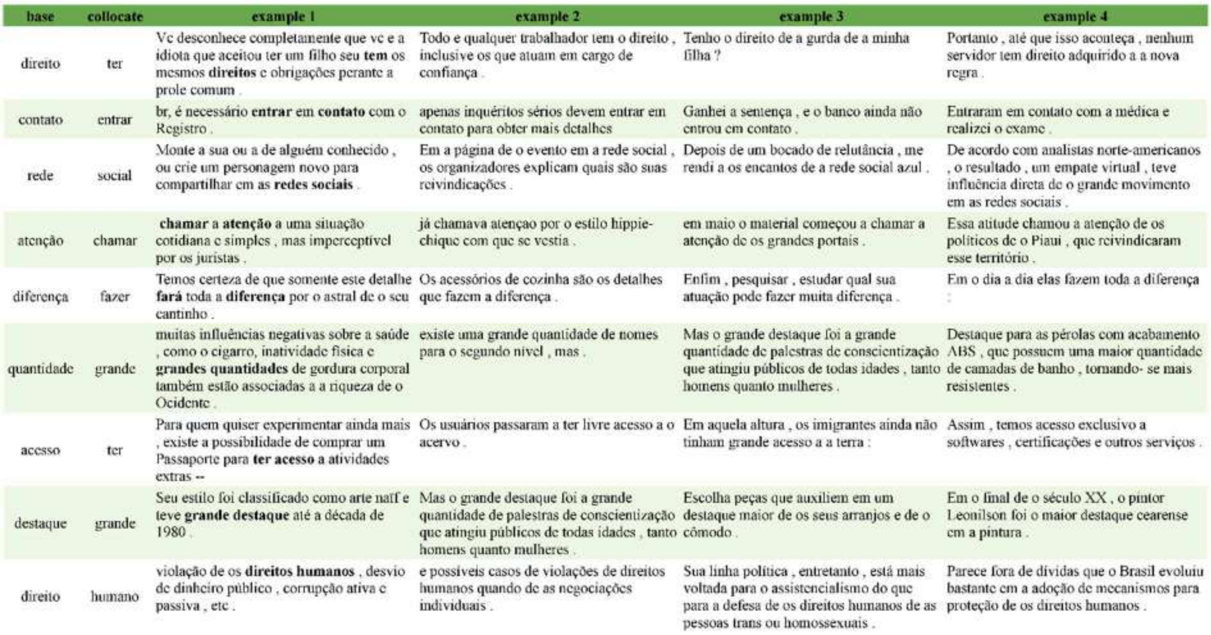
base	collocate	example 1	example 2	example 3	example 4
bond	double	This reaction can be used to determine the position of a **double bond** in an unknown alkene.	This makes the Br closest to the double bond slightly positive and therefore an electrophile.	The IUPAC numerical prefixes are used to indicate the number of double bonds.	That is, hydrogen ends up on the more substituted carbon of the double bond.
interval	time	Whoever measures a particular space-**time interval** will get the same value, no matter how fast they are travelling.	The space-time interval, formula 19, is invariant.	The second consequence of the invariance of the space-time interval is that clocks will appear to go slower on objects that are moving relative to you.	Solutions Consider the formula for average velocity in the formula_1 direction, formula_49 , where formula_18 is the change in formula_1 over the time interval formula_52 .
language	programming	Since computer **programming languages** have so much in common, it is generally easy to learn a new programming language once you have mastered another.	Pascal is an influential computer programming language named after the mathematician .	Many people think they must choose a specific programming language in order to become a programmer, believing that they can only do that language.	D is a programming language being designed as a successor to C++.
bond	single	What of having two double bonds separated by a **single bond**?	What of having a compound that alternates between double bond and single bond?	Remember that single bonds can rotate in space if not impeded.	Each ending point and bend in the line represents one carbon atom and each short line represents one single carbon-carbon bond.
compound	organic	This number also applies to other **organic compounds** which have hydrogen atoms at similar distances from each other.	The IUPAC system is necessary for complicated organic compounds.	Hydrocarbons are organic compounds that contain carbon and hydrogen only.	Heterotrophs require at least one organic nutrient to make other organic compounds.
group	functional	These parts of organic molecules are called **functional groups**.	There are many functional groups of interest to organic chemists.	The identification of functional groups and the ability to predict reactivity based on functional group properties is one of the cornerstones of organic chemistry.	Just as elements have distinctive properties, functional groups have characteristic chemistries.
file	media	Where not otherwise noted, non-text **media files** are available under various free culture licenses, consistent with the .	Typically the in using an image or other media file is to it to .	See for details about which media files can be uploaded.	Please view the media description page for details about the license of any specific media file.
role	play	Wave packets will **play** a central **role** in what is to follow, so it is important that we acquire a good understanding of them.	Usually hydrogen plays the role of the electrophile; however, hydrogen can also act as an nucleophile in some reactions.	The ideas of bond polarity and dipole moment play important roles in organic chemistry.	Smooth ER plays an important role in lipid emulsification and digestion in the cell.
question	answer	You should now be able to **answer** the following **questions** from your previous knowledge.	Use the content in this chapter and/or from external sources to answer the following questions.	This is the Reading room where raise and answer Wikibooks-related questions and concerns regarding technical issues, policies, or other aspects of our community.	If you answered the above questions correctly you should find this next section easy!

Table 5: Automatically retrieved data from the English corpus - examples



base	collocate	deprel	freq base	freq collocate	freq	freq norm	mi	di	ll	ts	zs
direito	ter	obj	560822	11188299	183090	1306.77	160,048,855,495,172	0.165837224384871	160,804,071,456,546	499,409,455,757,521	0.030225
contato	entrar	obl	139840	1081619	84533	1107.33	462,428,604,203,138	0.362541768192905	379,698,081,324,801	138,535,656,820,408	0.121584
rede	social	amod	451341	1510216	176610	1182.83	463,777,831,408,155	0.271090759214873	796,451,699,849,827	201,260,093,158,997	0.152594
atenção	chamar	obj	412559	374632	171188	1221.82	623,414,529,184,781	0.290572130437827	114,164,059,303,372	354,210,236,916,932	0.303104
diferença	fazer	obj	260225	5263149	79253	565.65	261,413,494,243,881	0.195799523645682	154,592,588,917,987	582,811,846,552,285	0.027897
quantidade	grande	amod	213140	5140962	84160	563.66	301,599,685,612,871	0.24858108135078	204,365,955,457,986	723,098,185,697,314	0.030479
acesso	ter	obj	309489	11188299	147993	1056.27	199,934,115,026,427	0.24337875729177	188,232,852,359,215	576,828,284,897,299	0.025097
destaque	grande	amod	59884	5140962	24434	163.65	306,634,489,037,465	0.255333171494298	608,313,819,034,328	398,396,784,243,363	0.009309
direito	humano	amod	640881	739743	95389	638.86	453,375,817,538,921	0.124578155368098	416,989,915,076,945	142,226,851,904,013	0.121406

Table 6: Automatically retrieved data from the Portuguese corpus – base = noun



base	collocate	example 1	example 2	example 3	example 4
direito	ter	Vc desconhece completamente que vc e a idiota que aceitou ter um filho seu **tem** os mesmos **direitos** e obrigações perante a prole comum .	Todo e qualquer trabalhador tem o direito , inclusive os que atuam em cargo de confiança .	Tenho o direito de a gurda de a minha filha ?	Portanto , até que isso aconteça , nenhum servidor tem direito adquirido a a nova regra .
contato	entrar	br, é necessário **entrar** em **contato** com o Registro .	apenas inquéritos sérios devem entrar em contato para obter mais detalhes	Ganhei a sentença , e o banco ainda não entrou em contato .	Entraram em contato com a médica e realizei o exame .
rede	social	Monte a sua ou a de alguém conhecido , ou crie um personagem novo para compartilhar em as **redes sociais** .	Em a página de o evento em a rede social , os organizadores explicam quais são suas reivindicações .	Depois de um bocado de relutância , me rendi a os encantos de a rede social azul .	De acordo com analistas norte-americanos , o resultado , um empate virtual , teve influência direta de o grande movimento em as redes sociais .
atenção	chamar	**chamar** a **atenção** a uma situação cotidiana e simples , mas imperceptível por os juristas .	já chamava atençao por o estilo hippie-chique com que se vestia .	em maio o material começou a chamar a atenção de os grandes portais .	Essa atitude chamou a atenção de os políticos de o Piaui , que reivindicaram esse território .
diferença	fazer	Temos certeza de que somente este detalhe **fará** toda a **diferença** por o astral de o seu cantinho .	Os acessórios de cozinha são os detalhes que fazem a diferença .	Enfim , pesquisar , estudar qual sua atuação pode fazer muita diferença .	Em o dia a dia elas fazem toda a diferença :
quantidade	grande	muitas influências negativas sobre a saúde , como o cigarro, inatividade física e **grandes quantidades** de gordura corporal também estão associadas a a riqueza de o Ocidente .	existe uma grande quantidade de nomes para o segundo nível , mas .	Mas o grande destaque foi a grande quantidade de palestras de conscientização que atingiu públicos de todas idades , tanto homens quanto mulheres .	Destaque para as pérolas com acabamento ABS , que possuem uma maior quantidade de camadas de banho , tornando- se mais resistentes .
acesso	ter	Para quem quiser experimentar ainda mais , existe a possibilidade de comprar um Passaporte para **ter acesso a** atividades extras --	Os usuários passaram a ter livre acesso a o acervo .	Em aquela altura , os imigrantes ainda não tinham grande acesso a a terra :	Assim , temos acesso exclusivo a softwares , certificações e outros serviços .
destaque	grande	Seu estilo foi classificado como arte naïf e teve **grande destaque** até a década de 1980 .	Mas o grande destaque foi a grande quantidade de palestras de conscientização que atingiu públicos de todas idades , tanto homens quanto mulheres .	Escolha peças que auxiliem em um destaque maior de os seus arranjos e de o cômodo .	Em o final de o século XX , o pintor Leonilson foi o maior destaque cearense em a pintura .
direito	humano	violação de os **direitos humanos** , desvio de dinheiro público , corrupção ativa e passiva , etc .	e possíveis casos de violações de direitos humanos quando de as negociações individuais .	Sua linha política , entretanto , está mais voltada para o assistencialismo do que para a defesa de os direitos humanos de as pessoas trans ou homossexuais .	Parece fora de dívidas que o Brasil evoluiu bastante em a adoção de mecanismos para proteção de os direitos humanos .

Table 7: Automatically retrieved data from the Portuguese corpus - examples

The volume of the automatically retrieved data is very large. We set a filter of 20 occurrences per million, in the same syntactic dependence, following Evert (2008). This filter has given, on average, 20,000 candidates with base = name, and 8,000 with base = verb, for example. The post-editing phase is still in progress and may last a few months as data have been manually validated, evaluated and also revised by at least two lexicographers. As collocations are being revised, they are directed to the following phase of automatic translation into other languages, as described in the next section, according to the pairs we have previously set (please see subsection *4.3*)

3.4 Translation of collocations

Once the monolingual collocations are inserted in the platform, we will use an unsupervised approach to retrieve candidate translations among the languages of the project. The strategy, inspired by Garcia et al. (2019c), can be summarized as follows:

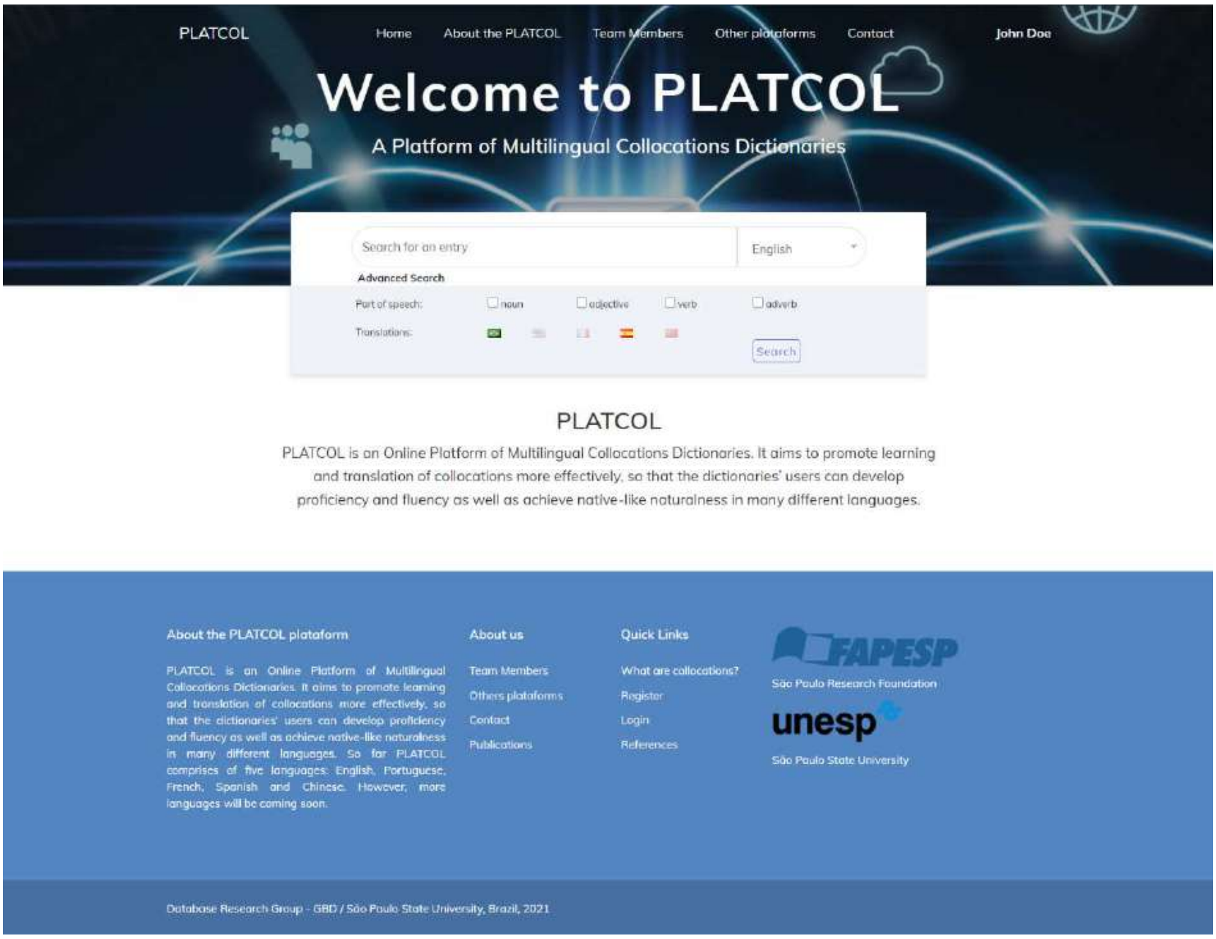
We first train monolingual *word2vec* models (Mikolov et al., 2013) using processed corpora and representing each word as a pair of lemma and PoS-tag (e.g., “house_NOUN”). Then, these models are mapped in a shared vector space with *vecmap* (Artetxe et al., 2018). Finally, we create a compositional vector for a given collocation in language A, and search for similar candidates (in terms of cosine similarity) in language B (Garcia et al., 2019c). The candidate translations are ranked by the confidence of the models, and they will be manually validated by lexicographers in further work.

4. The Multilingual Collocations Dictionary Structure and Design

The Multilingual Collocations Dictionaries[4] (PLATCOL) proposed here aim at fulfilling users’ needs regarding language encoding, and, as such, are considered to be a production dictionary. Besides helping users produce more authentic texts, PLATCOL also has the purpose of developing users’ collocational competence, which is intrinsically connected with fluency. The wider the repertoire of collocations, the greater fluency a learner can achieve. Moreover, the platform is intended to have an easy-to-use layout that offers the possibility of being customized.

Since foreign language learners or dictionary users in general encounter challenges in using collocations in their native language, and PLATCOL is also designed to display monolingual dictionaries. Thus, it will serve as a monolingual, bilingual or multilingual dictionary (English, Portuguese, French, Spanish and Chinese), also taking into account that collocations are automatically activated for each language covered by the platform, as the presentation screen of PLATCOL’s prototype illustrates (Figure 1).

[4] We use the term *dictionaries* as we mean that users can opt to activate monolingual, bilingual or even multilingual dictionaries, according to their needs and languages they want to search for.



Welcome to PLATCOL

A Platform of Multilingual Collocations Dictionaries

PLATCOL

PLATCOL is an Online Platform of Multilingual Collocations Dictionaries. It aims to promote learning and translation of collocations more effectively, so that the dictionaries' users can develop proficiency and fluency as well as achieve native-like naturalness in many different languages.

About the PLATCOL plataform

PLATCOL is an Online Platform of Multilingual Collocations Dictionaries. It aims to promote learning and translation of collocations more effectively, so that the dictionaries' users can develop proficiency and fluency as well as achieve native-like naturalness in many different languages. So far PLATCOL comprises of five languages: English, Portuguese, French, Spanish and Chinese. However, more languages will be coming soon.

Database Research Group - GBD / São Paulo State University, Brazil, 2021

Figure 1: Screenshot of PLATCOL's Presentation Screen Prototype

The new site is under construction, as it will be adjusted to the new languages (French, Spanish and Chinese)[5], with a more ambitious and interactive design as well as more detailed and enhanced lexicographical features and methodology.

4.1 User Profile and Needs

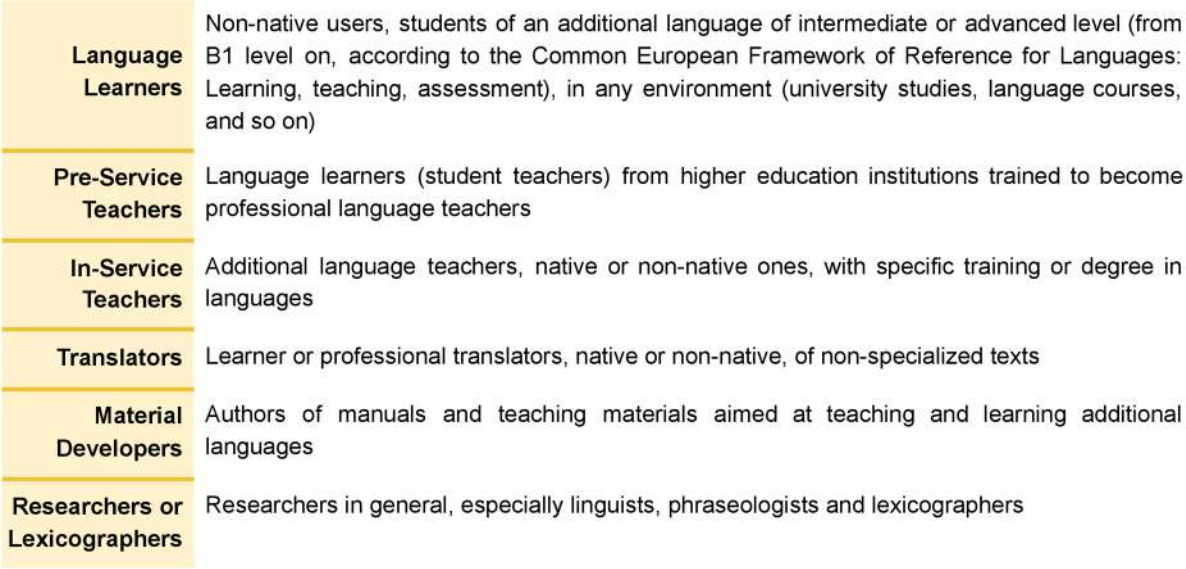
In any lexicographic work, reference is made to the following topics: typology of users, their needs and skills. Thus, in many studies, users' "problem" and needs are the main focus. However, as Fuertes Olivera and Tarp (2014) clearly state, this concern does not bear fruit, since it does not materialise in concrete theoretical and practical decisions, but instead researchers tend to approach the problem in a more general way and do not go into further discussion. Consequently, it is proposed that a better approach is to differentiate between two types of lexicography: a contemplative and a transformative one.

[5] A site used to host the *Bilingual Collocations Dictionary* (Orenha-Ottaiano 2017) and was modified for PLATCOL (http://www.institucional.grupogbd.com/dicionario/index?locale=pt), where users can find information about the platform. However, a new software is being developed under the new methodology and an updated microstructure will be inserted in the near future.

In contemplative lexicography, dictionaries are analysed and users questioned about their use of existing dictionaries to date. In transformative lexicography, theoretical analyses of the potential user situations, the respective user conditions and needs are used to develop new approaches for compiling new dictionaries, typically monofunctional dictionaries (Bergenholtz, Bothma & Gouws, 2011: 34-35).

Generally speaking, the first type can be related to the so-called general theory of lexicography; the second type, in turn, is linked to functional theory. Our proposal is in line with this last perspective and thus the following constitute essential points that guide the development of the platform:

a) The prior definition of the users' profiles to which the proposal is addressed, a crucial step before its elaboration. These are the profiles that have already been defined:



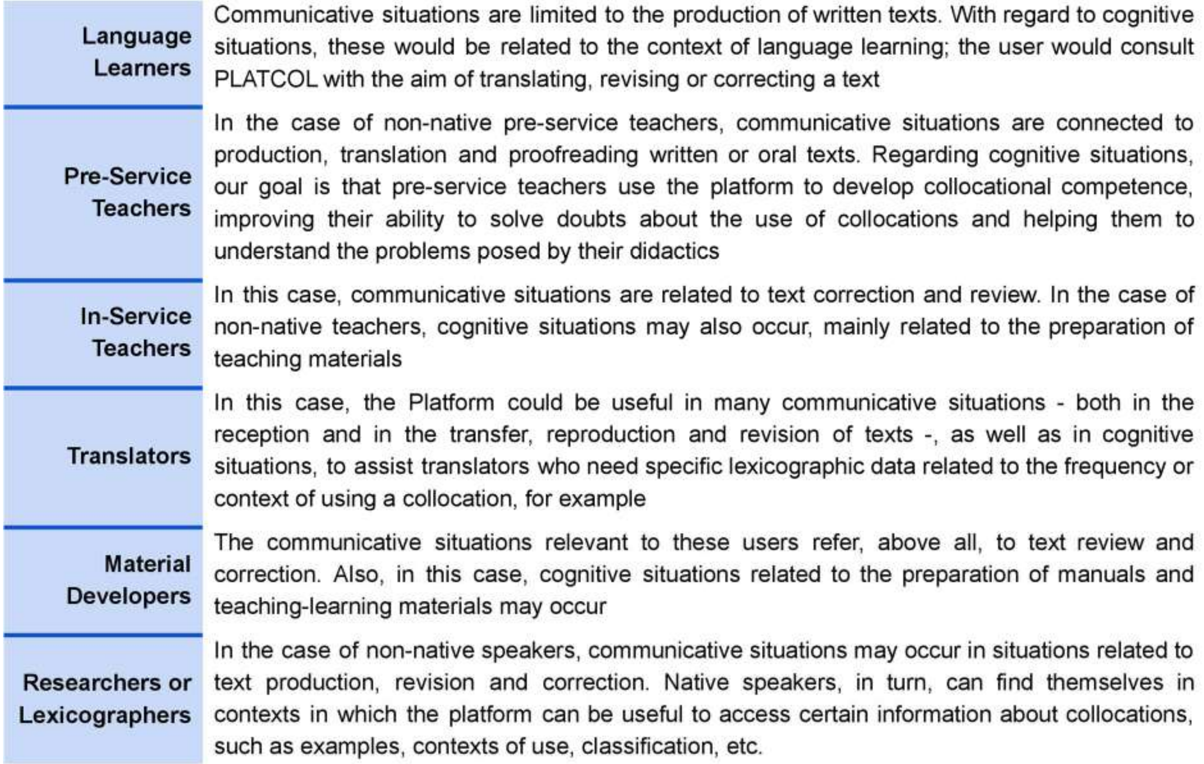
Language Learners	Non-native users, students of an additional language of intermediate or advanced level (from B1 level on, according to the Common European Framework of Reference for Languages: Learning, teaching, assessment), in any environment (university studies, language courses, and so on)
Pre-Service Teachers	Language learners (student teachers) from higher education institutions trained to become professional language teachers
In-Service Teachers	Additional language teachers, native or non-native ones, with specific training or degree in languages
Translators	Learner or professional translators, native or non-native, of non-specialized texts
Material Developers	Authors of manuals and teaching materials aimed at teaching and learning additional languages
Researchers or Lexicographers	Researchers in general, especially linguists, phraseologists and lexicographers

Table 8: User profiles.

b) The consideration of specific extra-lexicographic or social situations that would motivate the use of the platform: "to determine which type of needs a specific type of user may have in each type of situation" (Bergenholtz & Tarp, 2003:173):

We start from the idea that the different target audiences of a lexicographic work have a series of information and consultation needs (Fuertes Olivera & Tarp, 2014). These needs can only be met if users have quick and easy access to a set of lexicographic data prepared according to their profile. This way, users should be able to extract the information they need, so that they can employ it later, according to their purposes. These purposes, in turn, are always related to the extra-lexicographic contexts and situations that gave rise to these needs (Tarp, 2015).

Considering the profile of potential users of the platform, we acknowledge that the lexicographically relevant social situations, among the four defined within functional theory, are as follows: 1. Communicative, in which users try to solve problems related to production, reception, translation, proofreading and correction of written or oral texts; and 2. Cognitive, when users need or want to expand their knowledge of something. This typology could be applied to the profile of all indicated users; however, recognizing the limitations of the proposal, it is necessary to establish some restrictions, as Table 9 shows.



Language Learners	Communicative situations are limited to the production of written texts. With regard to cognitive situations, these would be related to the context of language learning; the user would consult PLATCOL with the aim of translating, revising or correcting a text
Pre-Service Teachers	In the case of non-native pre-service teachers, communicative situations are connected to production, translation and proofreading written or oral texts. Regarding cognitive situations, our goal is that pre-service teachers use the platform to develop collocational competence, improving their ability to solve doubts about the use of collocations and helping them to understand the problems posed by their didactics
In-Service Teachers	In this case, communicative situations are related to text correction and review. In the case of non-native teachers, cognitive situations may also occur, mainly related to the preparation of teaching materials
Translators	In this case, the Platform could be useful in many communicative situations - both in the reception and in the transfer, reproduction and revision of texts -, as well as in cognitive situations, to assist translators who need specific lexicographic data related to the frequency or context of using a collocation, for example
Material Developers	The communicative situations relevant to these users refer, above all, to text review and correction. Also, in this case, cognitive situations related to the preparation of manuals and teaching-learning materials may occur
Researchers or Lexicographers	In the case of non-native speakers, communicative situations may occur in situations related to text production, revision and correction. Native speakers, in turn, can find themselves in contexts in which the platform can be useful to access certain information about collocations, such as examples, contexts of use, classification, etc.

Table 9: User profiles related to lexicographically relevant social situations and some restrictions.

c) the determination of the platform's lexicographic functions:

A lexicographic function must be understood as “the assistance provided by the dictionary to meet a certain type of user’s specific needs in a certain type of extra-lexicographical situation”[6] (Fuertes Olivera & Tarp, 2008: 80, the translation is ours). Our proposal must be considered to be multifunctional, since, according to the extra-lexicographic situations discussed, it must fulfill two functions: a communicative and cognitive one. Given the recommendations of functional theory and considering that users' abilities in dictionary use cannot be determined in advance, we must ensure that access to information is quick and easy.

For this reason, the dictionaries' macrostructure includes a systematic introduction and

[6] “...la asistencia que presta el diccionario para satisfacer el tipo específico de necesidades que tiene un determinado tipo de usuarios en un determinado tipo de situación extra-lexicográfica” (Fuertes Olivera & Tarp, 2008: 80)

usage guide. Likewise, the design of the dictionaries' microstructure has been made taking into account users' profile and needs. The features here described about users' needs and profiles are based on our considerable experience of translation, translation training, foreign language teaching and teacher training. In the near future, we intend to carry out research on users' needs among the target groups.

4.2 Dictionaries' microstructure

The compilation of a collocations dictionary, an already complex task, becomes even more challenging when multiple languages are taken into consideration. The organization of the microstructure, as explained below, is especially daunting.

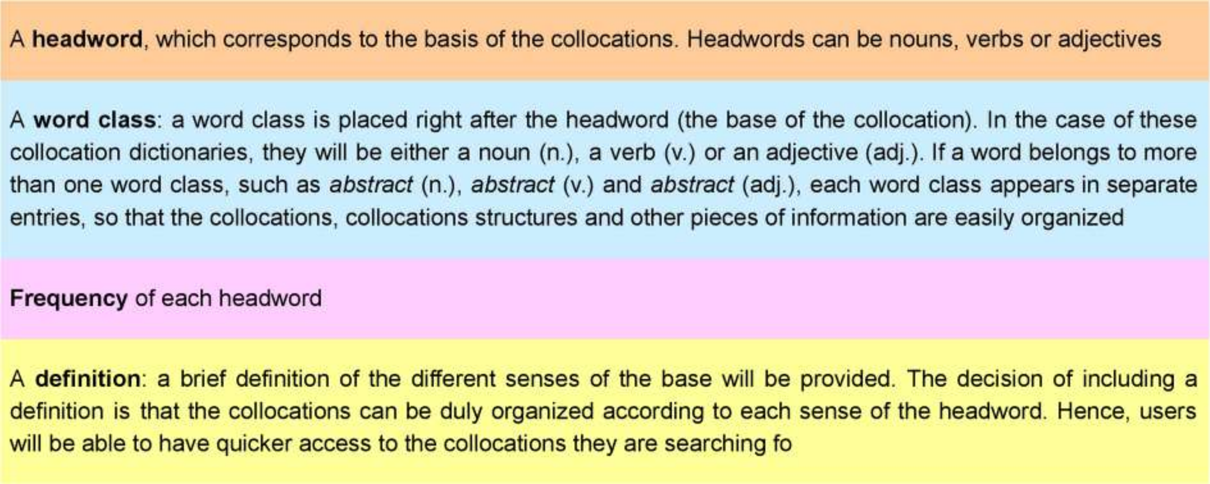
PLATCOL's entries include nouns, verbs, and adjectives which correspond to the bases of the collocations (see more about the collocations structures in this section).

In a collocations dictionary, the headwords can be organized according to at least two different principles. One of the views in the treatment of collocations is statistically based. Collocations are defined under a statistical approach iwith regard to their frequent co-occurrence. This way, the headword can be either the base or the collocate, depending on the frequency of co-occurrence in the corpus.

The other view follows Hausmann's approach (1985, 1989), using the concept of the base, the element usually known by users, and of the collocate, the element they are searching for, that is to say, what learners and translators, for example, need to find.

In this project, we opted for the latter view (Hausmann 1985, 1989), claiming that it is more user-friendly and effective with regard to most user profiles, besides being the starting point for most users. Moreover, users will be able to perform either base or collocate searches in the platform search bar.

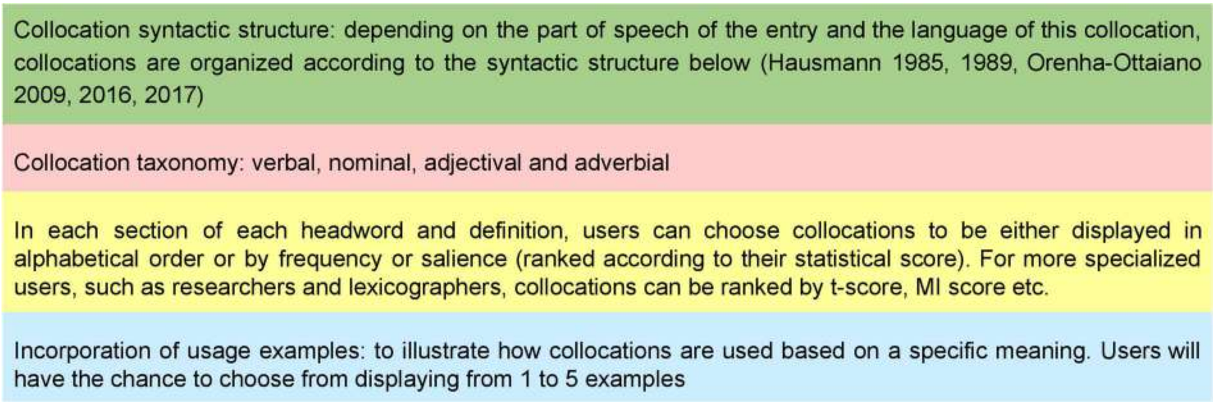
The entries of the multilingual collocation dictionaries consist of the following elements:



A **headword**, which corresponds to the basis of the collocations. Headwords can be nouns, verbs or adjectives
A **word class**: a word class is placed right after the headword (the base of the collocation). In the case of these collocation dictionaries, they will be either a noun (n.), a verb (v.) or an adjective (adj.). If a word belongs to more than one word class, such as *abstract* (n.), *abstract* (v.) and *abstract* (adj.), each word class appears in separate entries, so that the collocations, collocations structures and other pieces of information are easily organized
Frequency of each headword
A **definition**: a brief definition of the different senses of the base will be provided. The decision of including a definition is that the collocations can be duly organized according to each sense of the headword. Hence, users will be able to have quicker access to the collocations they are searching fo

Table 10: Entry elements of the Multilingual Collocation Dictionaries

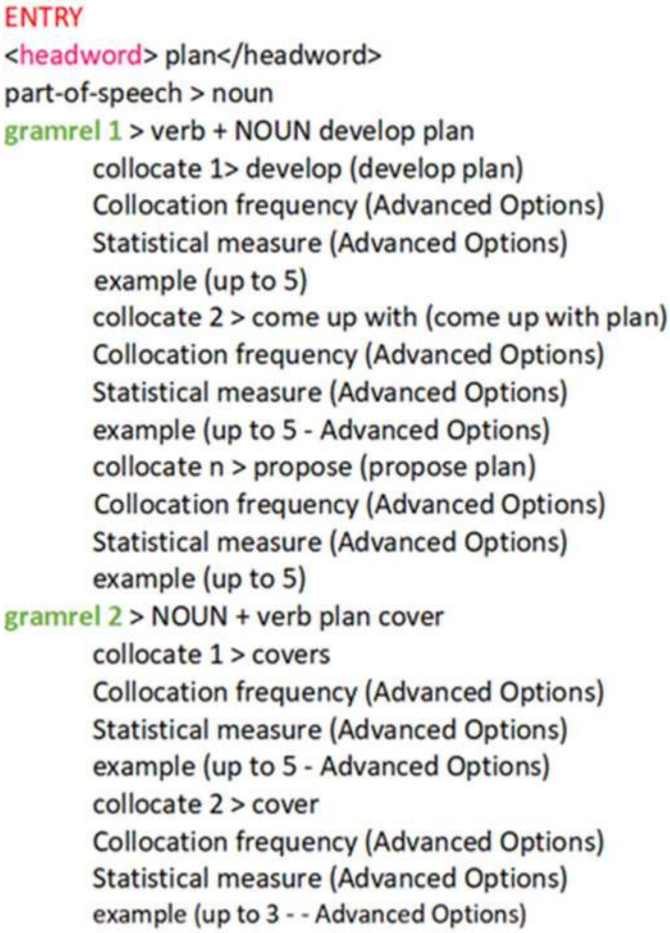
The collocations are structured as follows:



Collocation syntactic structure: depending on the part of speech of the entry and the language of this collocation, collocations are organized according to the syntactic structure below (Hausmann 1985, 1989, Orenha-Ottaiano 2009, 2016, 2017)
Collocation taxonomy: verbal, nominal, adjectival and adverbial
In each section of each headword and definition, users can choose collocations to be either displayed in alphabetical order or by frequency or salience (ranked according to their statistical score). For more specialized users, such as researchers and lexicographers, collocations can be ranked by t-score, MI score etc.
Incorporation of usage examples: to illustrate how collocations are used based on a specific meaning. Users will have the chance to choose from displaying from 1 to 5 examples

Table 11: Collocations' organization

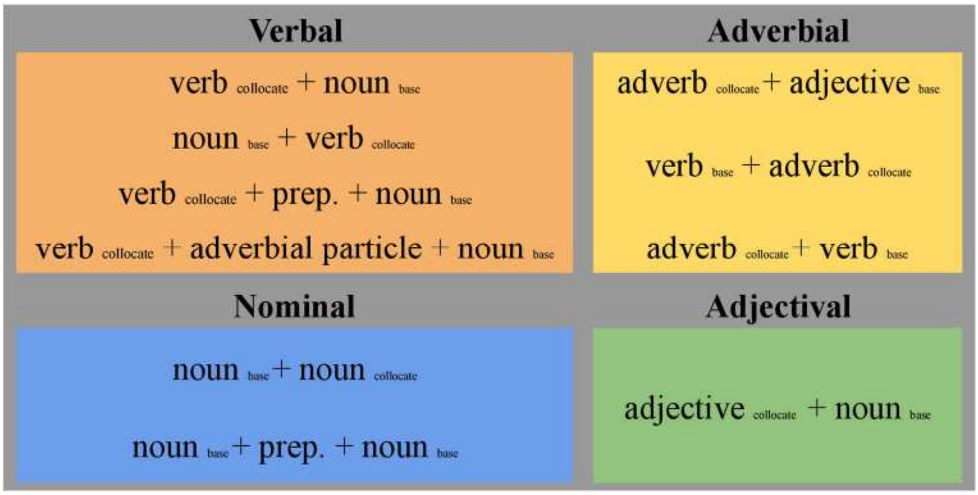
Below, Table 12 shows a summarized entry structure:



```
ENTRY
<headword> plan</headword>
part-of-speech > noun
gramrel 1 > verb + NOUN develop plan
        collocate 1> develop (develop plan)
        Collocation frequency (Advanced Options)
        Statistical measure (Advanced Options)
        example (up to 5)
        collocate 2 > come up with (come up with plan)
        Collocation frequency (Advanced Options)
        Statistical measure (Advanced Options)
        example (up to 5 - Advanced Options)
        collocate n > propose (propose plan)
        Collocation frequency (Advanced Options)
        Statistical measure (Advanced Options)
        example (up to 5)
gramrel 2 > NOUN + verb plan cover
        collocate 1 > covers
        Collocation frequency (Advanced Options)
        Statistical measure (Advanced Options)
        example (up to 5 - Advanced Options)
        collocate 2 > cover
        Collocation frequency (Advanced Options)
        Statistical measure (Advanced Options)
        example (up to 3 - - Advanced Options)
```

Table 12: Microstructure adapted and expanded from Orenha-Ottaiano et al. (2020).

According to the type of collocation and language, the collocations will have the following syntactic structures applied to English:



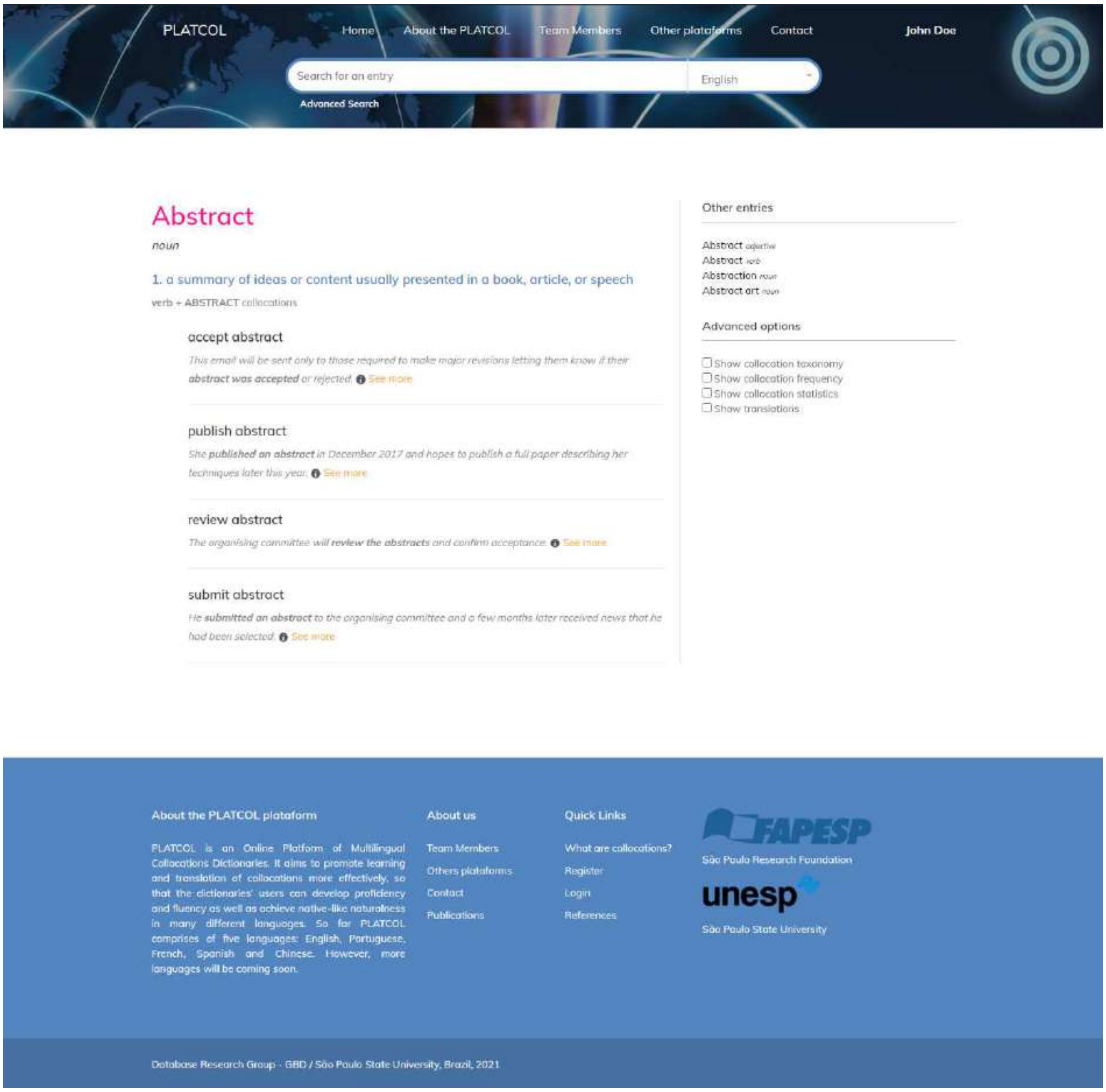
Verbal	Adverbial
$\text{verb}_{\text{collocate}} + \text{noun}_{\text{base}}$ $\text{noun}_{\text{base}} + \text{verb}_{\text{collocate}}$ $\text{verb}_{\text{collocate}} + \text{prep.} + \text{noun}_{\text{base}}$ $\text{verb}_{\text{collocate}} + \text{adverbial particle} + \text{noun}_{\text{base}}$	$\text{adverb}_{\text{collocate}} + \text{adjective}_{\text{base}}$ $\text{verb}_{\text{base}} + \text{adverb}_{\text{collocate}}$ $\text{adverb}_{\text{collocate}} + \text{verb}_{\text{base}}$
Nominal	**Adjectival**
$\text{noun}_{\text{base}} + \text{noun}_{\text{collocate}}$ $\text{noun}_{\text{base}} + \text{prep.} + \text{noun}_{\text{base}}$	$\text{adjective}_{\text{collocate}} + \text{noun}_{\text{base}}$

Table 13: Collocations' Taxonomy and Syntactic Structures.

The syntactic structures or order of the elements of collocations may vary from one language to the other. For example, adjectival collocations in Portuguese, Spanish and French can have two different syntactic structure orders, depending on the meaning the speaker wishes to convey:

$\text{Noun}_{\text{base}} + \text{Adjective}_{\text{collocate}}$
$\text{Adjective}_{\text{collocate}} + \text{Noun}_{\text{base}}$

Users will then have free access to PLATCOL's basic microstructure dictionaries, without having to sign in (as shown in Figure 2).



Abstract

noun

1. a summary of ideas or content usually presented in a book, article, or speech

verb + ABSTRACT collocations

accept abstract

*This email will be sent only to those required to make major revisions letting them know if their **abstract was accepted** or rejected.* See more

publish abstract

*She **published an abstract** in December 2017 and hopes to publish a full paper describing her techniques later this year.* See more

review abstract

*The organising committee **will review the abstracts** and confirm acceptance.* See more

submit abstract

*He **submitted an abstract** to the organising committee and a few months later received news that he had been selected.* See more

Other entries

Abstract *adjective*
Abstract *verb*
Abstraction *noun*
Abstract art *noun*

Advanced options

☐ Show collocation taxonomy
☐ Show collocation frequency
☐ Show collocation statistics
☐ Show translations

About the PLATCOL plataform

PLATCOL is an Online Platform of Multilingual Collocations Dictionaries. It aims to promote learning and translation of collocations more effectively, so that the dictionaries' users can develop proficiency and fluency as well as achieve native-like naturalness in many different languages. So far PLATCOL comprises of five languages: English, Portuguese, French, Spanish and Chinese. However, more languages will be coming soon.

Database Research Group - GBD / São Paulo State University, Brazil, 2021

Figure 2: Screenshot of basic structure of an entry.

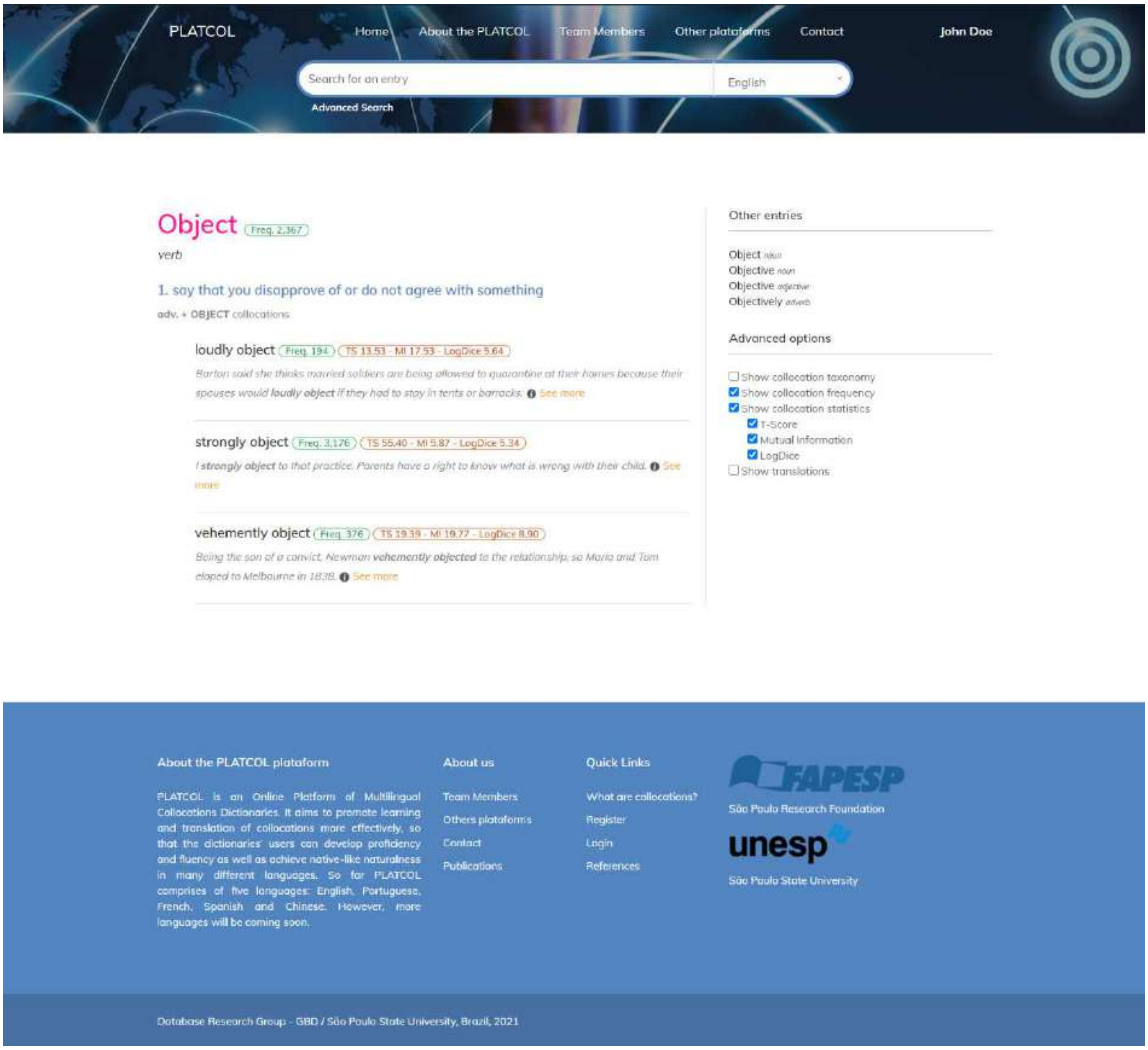
Besides the basic microstructure, *Advanced options* will be available if a user opts to sign in, according to their profile.

A new dictionary structure will be available so users can choose from items in a *Menu* containing the following elements:

□	Collocation frequency
□	Collocation's statistical information: it provides users with statistical measures so that they can check or analyze the results of each collocation's frequency of co-occurrence
□	Taxonomy of Collocation
□	Translation of collocations

Table 14: Dictionary's menu options.

Figures 3 and 4 show the dictionary structure generated by the items chosen from a menu in *Advanced options*.



Object (Freq. 2,367)

verb

1. say that you disapprove of or do not agree with something

adv. + OBJECT collocations

loudly object (Freq. 194) (TS 13.53 - MI 17.53 - LogDice 5.64)

*Barton said she thinks married soldiers are being allowed to quarantine at their homes because their spouses would **loudly object** if they had to stay in tents or barracks.* See more

strongly object (Freq. 3,176) (TS 55.40 - MI 5.87 - LogDice 5.34)

*I **strongly object** to that practice. Parents have a right to know what is wrong with their child.* See more

vehemently object (Freq. 376) (TS 19.39 - MI 19.77 - LogDice 8.90)

*Being the son of a convict, Newman **vehemently objected** to the relationship, so Maria and Tom eloped to Melbourne in 1838.* See more

Other entries

Object *noun*
Objective *noun*
Objective *adjective*
Objectively *adverb*

Advanced options

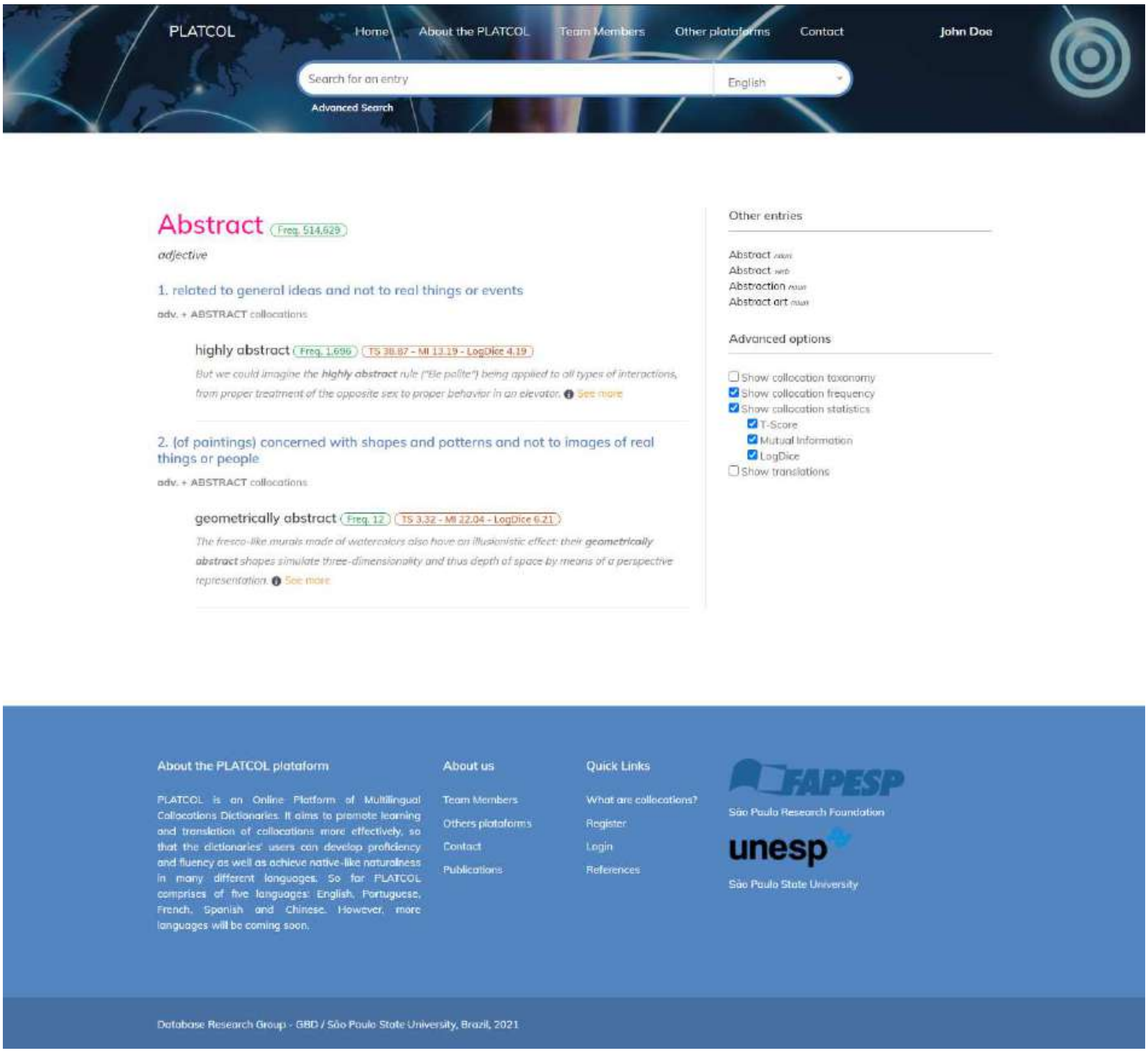
☐ Show collocation taxonomy
☑ Show collocation frequency
☑ Show collocation statistics
  ☑ T-Score
  ☑ Mutual Information
  ☑ LogDice
☐ Show translations

About the PLATCOL plataform

PLATCOL is an Online Platform of Multilingual Collocations Dictionaries. It aims to promote learning and translation of collocations more effectively, so that the dictionaries' users can develop proficiency and fluency as well as achieve native-like naturalness in many different languages. So far PLATCOL comprises of five languages: English, Portuguese, French, Spanish and Chinese. However, more languages will be coming soon.

Database Research Group - GBD / São Paulo State University, Brazil, 2021

Figure 3: Screenshot of the advanced option microstructure (the entry is a verb).



Abstract Freq. 514,629

adjective

1. related to general ideas and not to real things or events

adv. + ABSTRACT collocations

highly abstract Freq. 1,696 TS 38.87 - MI 13.19 - LogDice 4.19

But we could imagine the ***highly abstract*** *rule ("Be polite") being applied to all types of interactions, from proper treatment of the opposite sex to proper behavior in an elevator.* See more

2. (of paintings) concerned with shapes and patterns and not to images of real things or people

adv. + ABSTRACT collocations

geometrically abstract Freq. 12 TS 3.32 - MI 22.04 - LogDice 6.21

The fresco-like murals made of watercolors also have an illusionistic effect: their ***geometrically abstract*** *shapes simulate three-dimensionality and thus depth of space by means of a perspective representation.* See more

Other entries

Abstract *noun*
Abstract *verb*
Abstraction *noun*
Abstract art *noun*

Advanced options

□ Show collocation taxonomy
☑ Show collocation frequency
☑ Show collocation statistics
☑ T-Score
☑ Mutual Information
☑ LogDice
□ Show translations

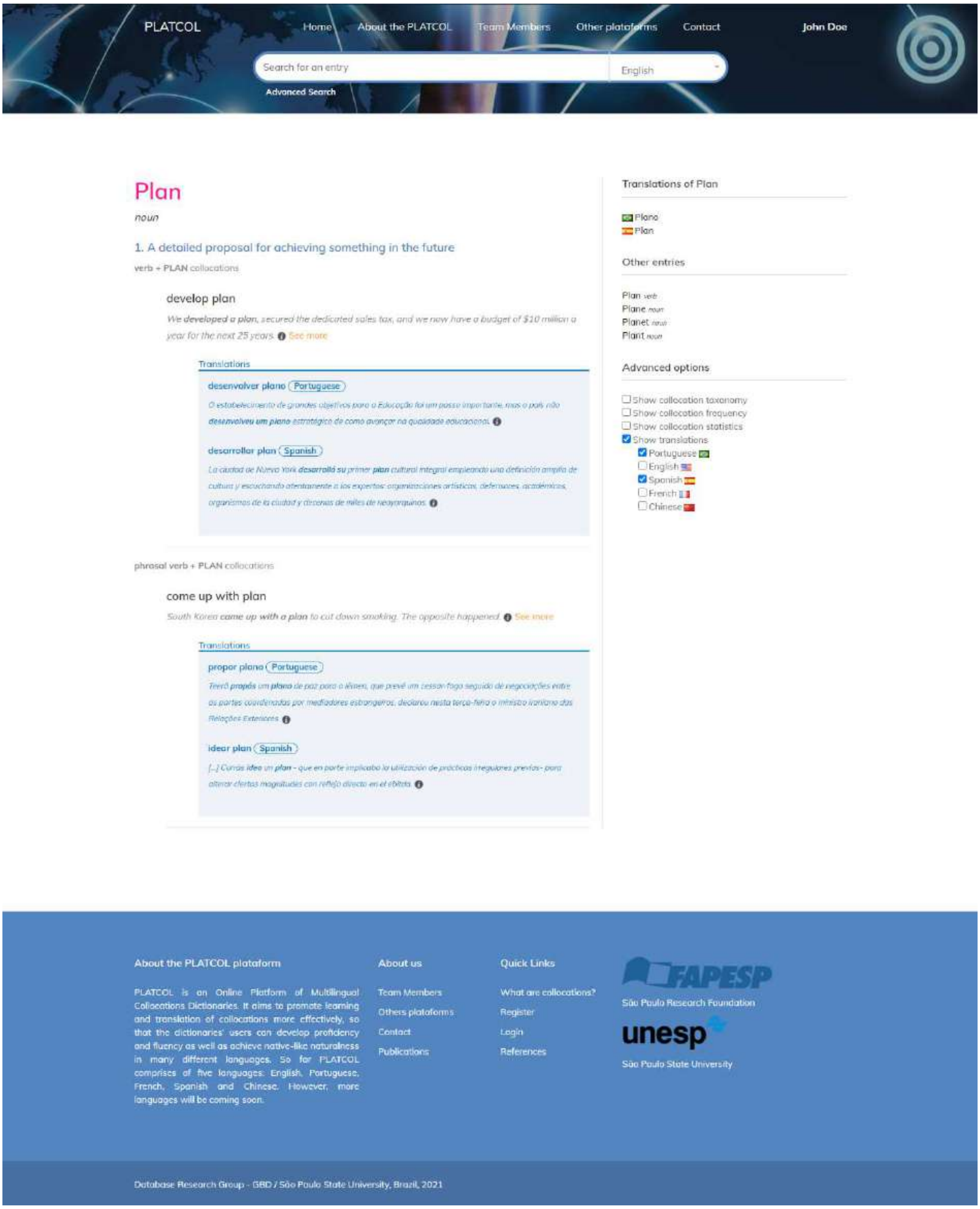
About the PLATCOL plataform

PLATCOL is an Online Platform of Multilingual Collocations Dictionaries. It aims to promote learning and translation of collocations more effectively, so that the dictionaries' users can develop proficiency and fluency as well as achieve native-like naturalness in many different languages. So far PLATCOL comprises of five languages: English, Portuguese, French, Spanish and Chinese. However, more languages will be coming soon.

Database Research Group - GBD / São Paulo State University, Brazil, 2021

Figure 4: Screenshot of the advanced option microstructure (the entry is an adjective).

Additionally, a user may opt to click on *Advanced options* and choose to see the translation equivalents of the sought entry (*plan*) and its collocations in, for example, two more languages of the platform, Portuguese and Spanish (Figure 5).



Plan

noun

1. A detailed proposal for achieving something in the future

verb + PLAN collocations

develop plan

We ***developed a plan****, secured the dedicated sales tax, and we now have a budget of $10 million a year for the next 25 years.* See more

Translations

desenvolver plano (Portuguese)

O estabelecimento de grandes objetivos para a Educação foi um passo importante, mas o país não ***desenvolveu um plano*** *estratégico de como avançar na qualidade educacional.*

desarrollar plan (Spanish)

La ciudad de Nueva York ***desarrolló su*** *primer* ***plan*** *cultural integral empleando una definición amplia de cultura y escuchando atentamente a los expertos: organizaciones artísticas, defensores, académicos, organismos de la ciudad y decenas de miles de neoyorquinos.*

phrasal verb + PLAN collocations

come up with plan

South Korea ***came up with a plan*** *to cut down smoking. The opposite happened.* See more

Translations

propor plano (Portuguese)

Teerã ***propôs*** *um* ***plano*** *de paz para o Iêmen, que prevê um cessar-fogo seguido de negociações entre as partes coordenadas por mediadores estrangeiros, declarou nesta terça-feira o ministro iraniano das Relações Exteriores.*

idear plan (Spanish)

[...] Curras ***ideó*** *un* ***plan*** *- que en parte implicaba la utilización de prácticas irregulares previas - para alterar ciertas magnitudes con reflejo directo en el ebitda.*

Translations of Plan

Plano
Plan

Other entries

Plan *verb*
Plane *noun*
Planet *noun*
Plant *noun*

Advanced options

☐ Show collocation taxonomy
☐ Show collocation frequency
☐ Show collocation statistics
☑ Show translations
☑ Portuguese
☐ English
☑ Spanish
☐ French
☐ Chinese

About the PLATCOL plataform

PLATCOL is an Online Platform of Multilingual Collocations Dictionaries. It aims to promote learning and translation of collocations more effectively, so that the dictionaries' users can develop proficiency and fluency as well as achieve native-like naturalness in many different languages. So for PLATCOL comprises of five languages: English, Portuguese, French, Spanish and Chinese. However, more languages will be coming soon.

About us

Team Members
Others plataforms
Contact
Publications

Quick Links

What are collocations?
Register
Login
References

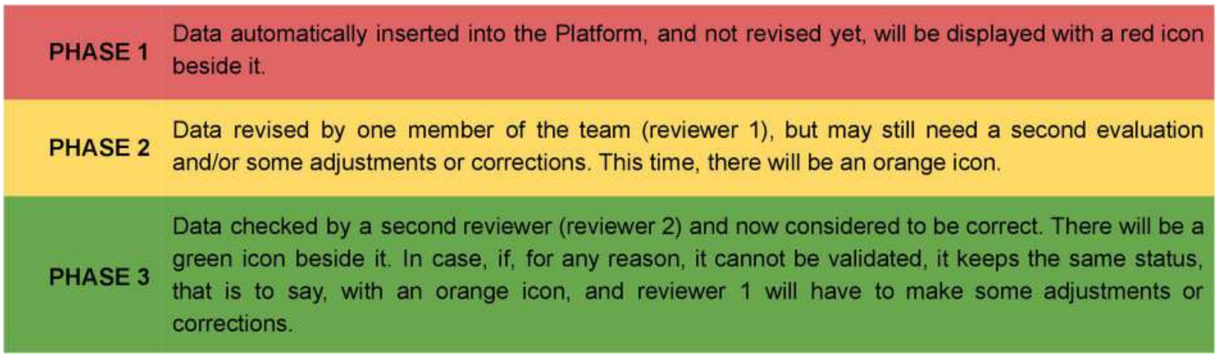
FAPESP
São Paulo Research Foundation
unesp
São Paulo State University

Database Research Group - GBD / São Paulo State University, Brazil, 2021

Figure 5: Screenshot of a user's choice for a translation equivalent of the entry *plan*.

Of course, future developments of the platform will take into account user feedback.

With respect to post-editing and validation of entry structures, the research will undertake the following three phases (traffic lights phases), indicating to users their status:



PHASE 1	Data automatically inserted into the Platform, and not revised yet, will be displayed with a red icon beside it.
PHASE 2	Data revised by one member of the team (reviewer 1), but may still need a second evaluation and/or some adjustments or corrections. This time, there will be an orange icon.
PHASE 3	Data checked by a second reviewer (reviewer 2) and now considered to be correct. There will be a green icon beside it. In case, if, for any reason, it cannot be validated, it keeps the same status, that is to say, with an orange icon, and reviewer 1 will have to make some adjustments or corrections.

Table 15: Phases for post-editing and validation of entry structures

This strategy allows users to have access to all entries, collocations and automatically extracted data without having to wait until the whole validation process is over.

As this is an ongoing project, some methodological aspects as well as macro and microstructure decisions may still be changed or reshaped, with a view to best adjust the platform to the new languages investigated as well as to users' different lexicographical needs. Matters regarding the number of collocations or the amount of data to be displayed on the collocation dictionaries' screen as well as types of filter (Kosem et al., 2019b), aiming to help users find relevant information according to their profile and needs, are still being investigated and will be further discussed in future work.

4.3 Dictionary typology and directionality

Regarding the coverage of languages, the platform can display monolingual, bilingual or multilingual dictionaries. With regard to directionality, collocations are retrieved from all corpora languages and will be automatically translated and post-edited in the following directions:

- from English into Portuguese;
- from Portuguese into English;
- from Spanish into Portuguese;
- from Spanish into English;
- from Chinese into Spanish.

These directions serve only for research purposes. It is worth mentioning that another pair or group of languages can be chosen since the corresponding settings are manually entered into the system, regardless of the automatic retrieval process. Once a collocation in a given language is registered, translations into other languages can also be manually defined in the system.

Once translation pairs between collocations are identified and registered in the system, making up a multilingual database, it becomes possible to identify and automatically suggest new translations among other languages. This process occurs through an inference-based algorithm, built from an inference hypothesis related to the composition of multiple translation dictionaries: if word A translates into word B which in turn translates into word C, what is the probability that C is a translation of A? Studies developed under this hypothesis (e.g. Mausam et al., 2010), presented significant results in relation to the analysis via inference of translation pairs between different languages. In this process, the algorithm performs the analysis of previously registered translations, identifies other translation pairs via inference, and shows lexicographers the possibilities of translations, who must analyze the reliability and quality of the translation found.

For example, the collocations "develop a plan", in English, and "desenvolver um plano", in Portuguese, are equivalents. Similarly, the collocations "desenvolver um plano", in Portuguese, and "desarrollar un plan", in Spanish, also have a translation relationship. This way, even if it has not been previously identified in the automatic extraction process, the relationship between the collocations "develop a plan", in English, and "desarrollar un plan", in Spanish, will be automatically inferred.

4.4 The Dictionaries and CEFR levels

Second language teachers have classified collocations into different CEFR levels, but this classification is not common in collocation dictionaries. Even in learners' English dictionaries which include the level of CEFR, such as *Cambridge*, the level is assigned to the headword, but there is no information about the collocations under the headword. For example, the noun *crime*, assigned as B1. There is no information about collocations such as *commit crime*, *charged of crimes* or *alleged crimes* which appear as examples and do not seem to belong to the same level. We are interested in the relevance of collocations for all levels and, therefore, this dictionary should include collocations for all CEFR learners.

This claim leads to the challenge of establishing criteria to assign collocations to a specific level. There are different approaches. The *English Vocabulary Profile* (Capel, 2010) adds data from learner corpora to frequency information obtained from English corpora or vocabulary lists to determine the lexicon non-native speakers should know at a given level. DICI-A (*Dizionario delle Collocazioni Italiane per Apprendenti*), on the other hand, takes a corpus of native speakers as a reference point (Spina, 2016) and uses a set of parameters to determine the level of collocations it includes: the frequency and dispersion of a collocation in the corpus, its function (expressions with descriptive meaning versus marks of textual organization and pragmatic elements) and the topic with which the collocation in question is associated. As for Spanish collocations, García-Salido and Alonso (2018) choose frequency in the corpus to level the collocations of the *DiCE*, but taking as a point of departure the collocations included in the *Plan Curricular del Instituto Cervantes* (Instituto Cervantes, 1997-2016). By means of analysis of a sample of collocations included in both the dictionary and the *Plan Curricular del Instituto Cervantes*, a negative correlation was found between the levelling proposed for those collocations in the *Plan Curricular* and the corpus frequency; that is, higher levels correspond to lower frequencies, and vice versa.

A challenge for assigning CEFR levels in a multilingual collocation dictionary is to find the equivalence between different languages. For instance, according to frequency criterion, a given collocation in a language could be assigned to B1 level, however, its equivalent in another language could be classified into a lower or higher one, according to the same criterion. For example, even though the collocations *black coffee*, *café solo*, *café noir*, and *café preto* could be considered translation equivalents, they are not found equally in different language corpora and may not be assigned to the same CEFR level.

5. Conclusion and further work

This paper outlined a corpus-based methodology for the development of the Online Platform for a Multilingual Collocations Dictionary, PLATCOL. It described the lexicographical features developed to compile PLATCOL's collocations dictionaries and presented their macro and microstructure.

We also discussed the automatic approaches to annotate corpora with lemmas, PoS-tags and dependency relations in the five languages of PLATCOL. Automatic methods to extract candidate collocations were also explained as well as statistical measures and distributional semantics strategies to select the candidates described, highlighting the relevance of post-edition in the lexicographical process.

The collocations dictionaries' prototypes were presented to illustrate PLATCOL's customized design, layout and lexicographical features, stressing the importance of developing an innovative customization methodology tailored to users' needs and specifically designed for a collocations dictionary. Hence, we hope to contribute to future lexicographical and phraseological/phraseographical research.

For future work, we will take advantage of the strategy presented by Garcia *et al.* (2019c) to gather candidate translations for each selected collocation. This approach generates lists of bilingual collocation equivalents, which will be then reviewed by those lexicographers with a good proficiency in each language pair, approving those proper equivalents which have been automatically extracted by the system, and providing new translations when necessary.

6. Acknowledgements

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

7. References

Alonso-Ramos, M. (2001). Construction d'une base de données des collocations bilingue français-espagnol. *Langages,* 35(143), pp. 5-27.

Alonso-Ramos, M. (2004). *DiCE: Diccionario de Colocaciones del Español.* Universidade da Coruña. Accessed at: http://dicesp.com. (12 April 2021).

Alonso-Ramos, M. (ed.). (2008). Papel de los diccionarios de colocaciones en la enseñanza de español como L2. In E. Bernal & J. DeCesaris (eds.) *Proceedings of the XIII EURALEX International Congress*, Barcelona: IULA/Documenta Universitaria, pp. 1215-1230.

Alonso-Ramos, M. & García-Salido, M. (2019). Testing the Use of a Collocation Retrieval Tool Without Prior Training by Learners of Spanish. *International Journal of Lexicography*, 32(4), pp. 480-497.

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

Artetxe, M., Labaka, G. & Agirre, E. (2018). A robust self-learning method for fully unsu-pervised cross-lingual mappings of word embeddings. In I. Gurevych & Y. Miyao (eds.) *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* Melbourne, Australia, pp. 789–798. Available at: https://www.aclweb.org/anthology/P18-1073.pdf.

Baroni, M., Bernardini, S., Ferraresi, A. & Zanchetta, E. (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3), pp. 209-226.

Beauchesne, J. (2001). *Dictionnaire des cooccurrences.* Montréal: Guérin.

Benson, M., Benson, E. & Ilson, R. (1997). *The BBI Combinatory Dictionary of English.* Amsterdam/Philadelphia: John Benjamins.

Bergenholtz, H. & Tarp, S. (2003). Two opposing theories: On H.E. Wiegand's recent discovery of lexicographic functions. *Hermes. Journal of Linguistics,* 31, pp. 171-196.

Bergenholtz, H., Bothma, T. & Gouws, R.H. (2011). A model for integrated dictionaries of fixed expressions. In I. Kosem & K. Kosem (eds.) *Electronic lexicography in the 21st century: New Applications for New Users. Proceedings of eLex 2011.* Bled, Slovenia, pp. 34-42. Available at: https://elex2011.trojina.si/elex2011_proceedings.pdf.

Bernardini, S. (2007). Collocations in Translated Language: Combining Parallel, Comparable and Reference Corpora. In M. Davies, P. Rayson, S. Hunston & P. Danielsson (eds.) *Proceedings of the Corpus Linguistics Conference* (CL2007). Birmingham, UK, pp. 1-16. Available at: http://ucrel.lancs.ac.uk/publications/CL2007/paper/15_Paper.pdf.

Berti, B. & Pinnavaia, L. (2014). Creating a Bilingual Italian-English Dictionary of Collocations. In A. Abel, C. Vettori & N. Ralli (eds.) *Proceedings of the XVI Euralex International Congress*: The User in Focus. Bolzano/Bozen, pp. 515-524. Available at: http://euralex2014.eurac.edu/en/callforpapers/Documents/EURALEX_Part_3.

pdf.

Bird, S., Loper, E. & Klein, E. (2009). *Natural Language Processing with Python.* Sebastopol, CA*:* O'Reilly Media Inc. Available at: http://www.datascienceassn.org/sites/default/files/Natural%20Language%20Processing%20with%20Python.pdf.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5, pp. 135-146.

Bond, F. & Foster, R. (2013). Linking and extending an open multilingual wordnet. In H. Schuetze, P. Fung & M. Poesio (eds.) *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* Sofia, Bulgaria, pp. 1352-1362. Avalaible at: https://www.aclweb.org/anthology/P13-1133.pdf.

Bosque, I. (2006). *Diccionario combinatorio práctico del español contemporáneo.* Madrid: SM.

Bothma, T.J.D. & Tarp, S. (2012). Lexicography and the Relevance Criterion. *Lexikos*, 22, pp. 86-108.

Capel, A. (2010). A1-B1 Vocabulary: Insights and issues arising from the English Profile Wordlists Project. *English Profile Journal*, 1, pp. 1-11.

Corpas Pastor, G. (2017) Collocations in e-Bilingual Dictionaries: from Underlying Theoretical Assumptions to Practical Lexicography and Translation Issues. In S. Torner & E. Bernal (eds.) *Collocations and other Lexical Combinations in Spanish. Theoretical and Applied Approaches.* London: Routledge, pp. 139-160,

de Marneffe, M.C., Manning, C.D., Nivre, J. & Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, 47(2), pp. 1-52. DOI: https://doi.org/10.1162/coli_a_00402.

Evert, S. 2008. Corpora and collocations. In A. Lüdeling & M. Kytö (eds.) *Corpus Linguistics. An International Handbook*, v. 2. Mouton de Gruyter: Berlin, pp. 1212–1248.

Evert, S., Uhrig, P., Bartsch, S. & Proisl, T. (2017). E-VIEW-affilation–A large-scale evaluation study of association measures for collocation identification. In I. Kosem, C. Tiberius, M. Jakubíček, J. Kallas, S. Krek & V. Baisa (eds.) *Proceedings of eLex 2017–Electronic lexicography in the 21st century*: *Lexicography from Scratch.* Leiden, the Netherlands, pp. 531-549. Available at: https://elex.link/elex2017/proceedings/eLex_2017_Proceedings.pdf.

Fuertes-Olivera, P.A. & Tarp, S. (2014). *Theory and Practice of Specialised Dictionaries. Lexicography versus Terminography.* Berlín/Boston: Walter de Gruyter.

Fuertes-Olivera, P.A. & Tarp, S. (2008). La Teoría Funcional de la Lexicografía y sus consecuencias para los diccionarios de economía del español. *Revista de Lexicografía*, XIV, pp. 75-95.

Gamallo, P., Garcia, M., Piñeiro, C., Martinez-Castaño, R. & Pichel, J. C. (2018). LinguaKit: a Big Data-based multilingual tool for linguistic analysis and information extraction. In Institute of Electrical and Electronics Engineers (ed.). *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS).* Valencia, Spain, pp. 239-244.

Garcia, M., García-Salido, M. & Alonso-Ramos, M. (2017). Using bilingual word-embeddings for multilingual collocation extraction. In S. Markantonatou, C. Ramisch, A. Savary & V. Vincze (eds.) *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017).* Valencia, Spain, pp. 21–30. Available at: https://www.aclweb.org/anthology/W17-1703.pdf.

Garcia, M., García-Salido, M. & Alonso-Ramos, M. (2019a). Towards the automatic construction of a multilingual dictionary of collocations using distributional semantics. In I. Kosem, T. Z. Kuhn, M. Correia, J. P. Ferreira, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek & C. Tiberius (eds.) *Proceedings of eLex 2019: Smart Lexicography.* Sintra, Portugal, pp. 747-762. Available at: https://elex.link/elex2019/wp-content/uploads/2019/09/eLex_2019_42.pdf.

Garcia, M., García-Salido, M. & Alonso-Ramos, M. (2019b). A comparison of statistical association measures for identifying dependency-based collocations in various languages. In A. Savary, C. Parra Escartín, F. Bond, J. Mitrović & V. B. Mititelu (eds.) *Proceedings of the Joint Workshop on Multiword Expressions and WordNet* (MWE-WN 2019). Florence, Italy, pp. 49-59. Available at: https://www.aclweb.org/anthology/W19-5107.pdf.

Garcia, M., García-Salido, M. & Alonso-Ramos, M. (2019c). Weighted compositional vectors for translating collocations using monolingual corpora. In G. Corpas Pastor & R. Mitkov (eds.) *Computational and Corpus-Based Phraseology.* Cham, Switzerland: Springer, pp. 113-128.

Gries, S.Th. (2013). 50-something years of work on collocations. *International Journal of Corpus Linguistics*, 18(1), pp.137–165.

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, pp. 118-129.

Hausmann F.J. (1989). Le dictionnaire de collocations. In F.J. Hausmann, O. Reichmann, H.E. Wiegand & L. Zgusta (eds) *Wörterbücher: ein internationales Handbuch zur Lexicographie. Dictionaries. Dictionnaires.* Berlin/New-York: De Gruyter, pp. 1010-1019.

Jousse, A.L. & Polguère, A. (2005). *Le DiCo et sa version DiCouébe. Document descriptif et manuel d'utilisation.* Université de Montréal: Observatoire de linguistique Sens-Texte (OLST).

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M. & Rychly, 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, pp. 425–432.

Koehn, P. (2005). Europarl: a parallel corpus for Statistical Machine Translation. In *Proceedings of the 10th Machine Translation Summit.* Phuket, Thailand, pp. 79–

86. Available at: https://homepages.inf.ed.ac.uk/pkoehn/publications/europarl-mtsummit05.pdf.

Kosem, I., Koppel, K., Kuhn, T. Z., Michelfeit, J. & Tiberius, C. (2019a). Identification and automatic extraction of good dictionary examples: the case(s) of GDEX. *International Journal of Lexicography*, 32(2), pp. 119–137.

Kosem, I., Krek, S., Gantar, P., Arhar Holdt, Š., Čibej, J. & Laskowski, C. (2019b). Collocations Dictionary of Modern Slovene. Proceedings of the 18th EURALEX International Congress: lexicography in global contexts. Ljubljana: Ljubljana University Press, Faculty of Arts, pp. 989-997.

Laufer, B. (2011). The Contribution of Dictionary Use to the Production and Retention of Collocations in a Second Language. *International Journal of Lexicography*, 24(1), pp. 29–49.

L'Homme, M.C. (2008). Le DiCoInfo. Méthodologie pour une nouvelle génération de dictionnaires spécialisés. *Traduire,* 217, pp. 78-103.

L'Homme, M.C., Robichaud, B. & Prévil, N. (2018). Browsing the Terminological Structure of a Specialized Domain: A Method Based on Lexical Functions and their Classification. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis & T. Tokunaga (eds.) *11th Language Resources and Evaluation, LREC 2018.* Miyazaki, Japon, pp. 3079-3086. Available at: https://www.aclweb.org/anthology/L18-1487.pdf.

Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J. & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In K. Bontcheva & J. Zhu (eds.) *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations.* Baltimore, Maryland, pp. 55-60.

Mausam, S., Etzioni, S., Weld, O., Reiter, D.S., Skinner, K. Sammer, M. & Vessier, S. (2010). Panlingual lexical translation via probabilistic inference. *Artificial Intelligence*, 174(9-10), pp. 619-637.

Mayor, M. (ed.) (2013). *Longman Collocations Dictionary and Thesaurus.* Harlow: Pearson Education.

Mei, J. (ed.) (1999). *Xiandai hanyu dapei cidian* (1st ed.). Shanghai: hanyu da cidian chuban she.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). Efficient estimation of word representations in vector space. In Y. Bengio & Y. LeCun (eds.) *Workshop Proceedings of the International Conferenceon Learning Representations (ICLR 2013).* Scottsdale, AZ, USA. Available at: https://arxiv.org/pdf/1301.3781v3.pdf.

Orenha-Ottaiano, A. (2020). The creation of an Online English Collocations Platform to help develop collocational competence. *PHRASIS. Rivista di Studi Fraseologici e Paremiologici,* 4, pp. 59-81.

Orenha-Ottaiano, A. (forthcoming). Escolhas colocacionais a partir de um Corpus de Aprendizes de Tradução e a importância do desenvolvimento da competência colocacional. *Cadernos de Fraseoloxía Galega.*

Orenha-Ottaiano, A. (2017). The compilation of an Online Corpus-Based Bilingual Collocations Dictionary: motivations, obstacles and achievements. In I. Kosem, C. Tiberius, M. Jakubíček, J. Kallas, S. Krek & V. Baisa (eds.) *Proceedings of eLex 2017–Electronic lexicography in the 21st century*: *Lexicography from Scratch.* Leiden, the Netherlands, pp. 458-473. Available at: https://elex.link/elex2017/wp-content/uploads/2017/09/paper27.pdf.

Orenha-Ottaiano, A. (2013). The proposal of an electronic bilingual dictionary based on corpora. In O. Karpova (ed.) *Life Beyond Dictionaries. X International School on Lexicography.* Florence, Italy, pp. 405-408.

Orenha-Ottaiano, A., Kuhn, T.Z. & Valêncio, C.R. (2020). The building of an Online Platform for Monolingual Dictionaries of Academic Collocations in Portuguese and English. Paper presented at the *56th Linguistics Colloquium*, online.

Oxford Collocations Dictionary (English-Chinese) (2nd ed.). (2006). Oxford: Oxford University Press.

Oxford Collocations Dictionary (English-Chinese) App version. Accessed at: https://play.google.com/store/apps/details?id=hk.com.oup.dicts&amp;hl=en_US&amp;gl=US.

Padró, Ll. & Stanilovsky, E. (2012). FreeLing 3.0: Towards Wider Multilinguality. In N. Calzolari, K. Choukri, T. Declerck, M.U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis (eds.) *Proceedings of the Language Resources and Evaluation Conference (LREC 2012).* Istanbul, Turkey, pp. 2473-2479. Available at: http://www.lrec-conf.org/proceedings/lrec2012/pdf/430_Paper.pdf.

Pereira, L.A.S. & Mendes, A. (2002). An Electronic Dictionary of Collocations for European Portuguese: Methodology, Results and Applications. In A. Braasch & C. Povlsen (eds.) *Proceedings of the 10th EURALEX International Congress*, v. II. Copenhagen, Denmark, pp. 841-849.

Rundell, M. (2010). *Macmillan Collocations Dictionary for Learners of English.* Oxford: Macmillan Publishers Ltd.

Santos, D. & Rocha, P. (2004). The key to the first CLEF in Portuguese: Topics, questions and answers in CHAVE. In C. Peters, P. Clough, J. Gonzalo, G.J.F. Jones, M. Kluck & B. Magnini (eds.) *Multilingual Information Access for Text, Speech and Images, 5th Workshop of the Cross-Language Evaluation Forum* (CLEF 2004), Bath, UK, pp. 821-832.

Spina, S. (2010). The Dictionary of Italian Collocations: Design and Integration in an Online Learning Environment. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner & D. Tapias (eds.) *Conference Proceedings of the International Conference on Language Resources and Evaluation* (LREC 2010). Valletta, Malta, pp. 3202-3208. Available at: http://www.lrec-conf.org/proceedings/lrec2010/pdf/681_Paper.pdf.

Spina, S. (2016). Learner corpus research and phraseology in Italian as a second language: The case of the DICI-A, a learner dictionary of Italian collocations. In B. Sanromán Vilas (ed.) *Collocations Cross-Linguistically. Corpora, Dictionaries and Language Teaching* (Mémoires de la Société Néophilologique de Helsinki).

Helsinki: Société Néophilologique, pp. 219–244.

Straka, M. & Straková, J. (2017). Tokenizing, POS-tagging, lemmatizing and parsing UD 2.0 with UDPipe. In J. Hajič & D. Zeman (eds.) *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies.* Vancouver, Canada, pp. 88-99. Available at: https://www.aclweb.org/anthology/K17-3009.pdf.

Tarp, S. (2015). La teoría funcional en pocas palabras. *Estudios de Lexicografía*, 4, pp. 31-42.

Torner, S. & Bernal, E. (eds.) (2017). *Collocations and Other Lexical Combinations in Spanish.* London: Routledge.

Wagner Filho, J. A., Wilkens, R., Idiart, M. & Villavicencio, A. (2018). The brWaC Corpus: A New Open Resource for Brazilian Portuguese. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis & T. Tokunaga (eds.) *11th Language Resources and Evaluation, LREC 2018.* Miyazaki, Japon, pp. 4339-4344. Available at: https://www.aclweb.org/anthology/L18-1686.pdf.

Wiegand, H.E. (1984). On the Structure and Contents of a General Theory of Lexicography. In R.R.K. Hartmann (ed.) *LEXeter'83 Proceedings. Papers from the International Conference on Lexicography. Exeter, 9-12 September 1983.* Tübingen: Niemeyer, pp. 13-30.

Yang, T. (1990). *Hanyu changyongci dapei cidian.* (1st ed.). Beijing: Waiyu jiaoxue yu yanjiu chuban she.

Visualising Lexical Data for a Corpus-Driven Encyclopaedia

Santiago Chambó[1], Pilar León-Araúz[2]

[1, 2] Department of Translation and Interpreting,
University of Granada, Buensuceso, 11, 18002 Granada (Spain)
E-mail: santiagochambo@ugr.es, pleon@ugr.es

Abstract

The Humanitarian Encyclopedia (HE) is an ongoing corpus-driven project that aims at defining and documenting the dynamics of 129 concepts that are particularly controversial, fuzzy or ill-defined within the humanitarian action domain, thus enhancing communication in a sensitive area. In the HE, each entry is created according to an approach that combines corpus-driven knowledge with expert knowledge. Concept entries are authored by field experts who are provided with a Linguistic Analysis Report (LAR) created by a team of linguists. In LARs, HE linguists support their claims by i) presenting, quantifying and categorising textual data and by ii) making comparisons among subcorpora, which are created based on the corpus metadata (i.e. document type, region, organisation type, publication year). This article presents the visualisations created by HE linguists to represent both semantic information (i.e., conceptual combinations and non-hierarchically related concepts) and quantifiable concordance and collocational data. This includes approaches to disaggregating measures according to different kinds of subcorpus types and strategies to represent collocational intersections among subcorpora (i.e., collocates occurring in multiple subcorpora) as well as collocates unique to each subcorpus. Other concept-specific visualisations were also designed and are examined in this article.

Keywords: lexical data; visualisation; concept

1. The Humanitarian Encyclopedia: a Corpus-Driven Project with Lexical Data Visualisations

The Humanitarian Encyclopedia (HE; https://humanitarianencyclopedia.org/home) is an ongoing corpus-driven project that aims at defining and documenting the dynamics of 129 concepts that are particularly controversial, fuzzy or ill-defined within the humanitarian action domain. In the humanitarian domain there are many stakeholders (i.e. academics, practitioners, decision-makers) who do not always share a consensual understanding of humanitarian concepts, such as VULNERABILITY, RESILIENCE or AID DEPENDENCE. Although, at least theoretically, they all share common principles and values, even the very notion of HUMANITARIANISM raises controversial issues. The humanitarian sector is thus a highly dynamic domain due to different factors, such as history, academic and professional disciplines, culture, religion, organisational cultures and contexts, which are the reasons behind both its richness and controversies. Conceptual controversies raise operational, political, societal and educational challenges that can hinder the effectiveness of humanitarian action in a global world.

In this context, the initiative of the HE aims to cover an existing gap in the humanitarian sector contributing to the public good. As acknowledged on its website there is a current need for "creating a common understanding and formulation of the key humanitarian concepts to build bridges and promote an open dialogue to improve collective humanitarian action".

In the HE, each entry is created according to an approach that combines corpus-driven knowledge with expert knowledge. Concept entries are authored by field experts who are provided with a Linguistic Analysis Report (LAR) stored in the Linguistic Analysis Portal for the Humanitarian Encyclopedia (https://sites.google.com/view/humanitarianencyclopedia). A team of linguists is in charge of producing LARs for each concept based on data extracted from a corpus of humanitarian texts. Every LAR provides an overview of how a concept is understood explicitly and implicitly in humanitarian discourse and proposes a definitional template for it. Each LAR is generally composed of the following elements:

- Frequencies, which allow experts to see the regions, document types, years and organisation types where the concepts appear more relevant.

- Definitions, whether standardised and authoritative (if found in the corpus), or *ad hoc* (based on implicit categorisation), together with a summary of definitional elements and a comparison based on corpus metadata.

- Related concepts: indicating how concepts change their relational behaviour based on organisation type, geographical regions or time (e.g. causes and consequences, affected population, subtypes classified on different conceptual dimensions, ways of managing humanitarian concepts, etc.).

- Frequent collocations, mostly nouns, adjectives and verbs, showing other surrounding concepts in the corpus, which allow experts to understand the different facets of the concept over time and across organisations.

- Synonyms and antonyms, where applicable, together with the sources from which they were extracted.

- Usage over time, where applicable, according to both the HE corpus and Google Ngram Viewer.

- Trends, debates and controversies surrounding each concept, which is one of the richest elements and requires extensive manual curation.

HE linguists decided to include visualisations to aid their own analyses and make lexical data more accessible and thought-provoking for HE authors, which, due to space limitations, is the focus of this paper.

Projects driven by lexical data require visualisation strategies that facilitate data interpretation and enable knowledge transfer (Allen, 2017). Firstly, making sense of any kind of data without the support of graphical representations constitutes a cognitively challenging task, and linguistic data is no different (Siirtola et al., 2010). Secondly, in a multidisciplinary project where linguists and field experts interact, the visualisation of lexical data serves as an intermediary between both stakeholders.

The remainder of this paper is structured as follows. Section 2 describes the materials and methods used by the HE linguists. Section 3 presents the visualisations created to support lexical data interpretation. In Section 4 conclusions and future lines of research are presented.

2. Materials and Methods

This section describes the materials and methods used to create datasets of lexical data and to build visualisations based on such datasets.

2.1 Materials

2.1.1 Sketch Engine

Sketch Engine (www.sketchengine.eu) is a browser-based software that enables users to build, analyse and query corpora (Kilgarriff et al., 2004). It contains many tools and functionalities that can be combined. Table 1 provides a summary of the main tools and functionalities used for the purposes of this work.

Tool	**Description**	**Functionality**	**Description**
Concordance	Queries a corpus and return results in context, which can be sorted, filtered and processed with many additional functionalities. Complex searches are conducted with CQL[1].	Hide Sub-Hits filter	Removes sub-hits from matches obtained with queries containing ranges (e.g., {1,3}), only keeping the longer results.
		Frequency	Computes frequencies from results, generating frequency reports.
		Collocations	Compute collocations from results.
Word Sketch	Provides a summary of a search term's collocates and other surrounding words. Results categorised by grammar relations defined by a file containing a set of rules known as sketch grammar.	-	-

Table 1: Main tools and functionalities used in Sketch Engine.

2.1.2 The HE Corpus

The HE Corpus is a collection of 4,824 humanitarian documents published between 2004 and 2019, which amount to a total of 84,926,707 tokens and 71,201,157 words. Documents are tagged with metadata according to the type and subtype of issuing organisation, region, year of publication and document type. These are referred to in Sketch Engine as text types. Table 2 contains all metadata fields and values associated with each document save for organisation subtype because it is not used in the visualisations described in this paper.

The corpus was uploaded onto Sketch Engine and processed with a custom sketch grammar that combines Sketch Engine's default sketch grammar for English with

[1] Corpus Query Language (CQL), as referred to in Sketch Engine documentation, is a concordance notation that allows users to search corpora for complex grammatical and lexical patterns. It is based to a large extent on the Corpus Query Processor language (or QQP-syntax) implemented in Corpus Workbench and developed by Christ et. al (1999).

the EcoLexicon Semantic Sketch Grammar (León-Araúz & San Martín, 2018) and an unpublished set of rules for multi-word term extraction (see Section 4.5).

Text Types	**Classes**
Organisation Type	**NGO** (Non-Government Organisations), **NGO_Fed** (NGO Federations), **IGO** (Intergovernmental Organisations), **RC** (Red Cross/Crescent), **Net** (Networks), **Found** (Foundations/Funds), **State** (Government/State Entities), **RE** (Religious Entities), **C/B**, Project and WHS
Region	**Africa**, **Asia**, **CCSA** (Caribbean, Central and South America), **MENA** (Middle East and North Africa), **North_America**, **Oceania**
Year	Between **2004** and **2019**
Document Type	**General_Document**, **Activity_Report**, **Strategy**

Table 2: Pertinent text types in the HE Corpus

2.1.3 Tableau

Tableau (www.tableau.com) is a commercial data visualisation software, which has been used in previous projects to visualise linguistic data (Allen, 2017; Desagulier, 2019). It interprets datasets in multiple formats and provides the user with a graphic interface that enables him or her to create visualisations by combining a wide range of options. The visualisations described in this paper were created with Tableau Desktop. To embed our visualisations on the website where the LARs are published, each visualisation has to be uploaded onto a Tableau Public profile (https://public.tableau.com/).

2.1.4 Google Data Studio

Google Data Studio (datastudio.google.com) is a browser-based visualisation solution similar to Tableau. It is solely used to create filtrable and searchable tables (see Section 4.7) because Tableau does not offer such visualisation option.

2.1.5 Spreadsheet software

To create datasets in supported formats, we used Microsoft Excel for the visualisations built with Tableau, and Google Sheets for Google Data Studio.

2.2 Methods

This subsection provides a brief overview of the methods used to extract data from the HE Corpus with Sketch Engine and to create the datasets in a way that can be interpreted correctly by Tableau. For clarity, specific steps and procedures for each visualisation are described in Section 4.

Data is extracted from the HE Corpus through two methods with the Sketch Engine querying functionalities. The first method entails querying the corpus with CQL expressions by using the Concordance tool and its processing options (see Table 1 in Section 3.1.1). With this method, we aim at creating datasets that contain string value fields for lexical units and associated measures. This method also enables us to conduct restricted searches in specific portions of the corpus (i.e., subcorpora) by specifying document metadata in our CQL queries. A second method uses the Word Sketch functionality to query the corpus. Data is therefore collected from specific grammatical relation reports.

Data Fields	**Data Type in Tableau**	**Description**
Lexical units	Dimension (string)	Any word or words extracted by querying the corpus (e.g., concordance matches, multi-word expressions, collocates, contexts, etc.) to be displayed in a visualisation.
Organisation type	Dimension (string)	Metadata values from documents in the corpus
Year	Date	
Document type	Dimension (string)	
Region	Dimension (string)	
Frequency (absolute frequency)	Measure (whole number)	Number of occurrences in the corpus
Relative frequency	Measure (decimal percentage)	Subcorpus frequency divided by the frequency of a query in the entire corpus; expressed as a percentage (Kilgarriff et al., 2015)
logDice	Measure (decimal number)	Score expressing typicality of extracted collocations; independent of corpus size and recommended to compare phenomena among subcorpora (Rychlý, 2008)

Table 4: Data fields used to build the visualisations

In Sketch Engine, each query generates a report that we treat with spreadsheet software to create CSV files with a data structure that can be processed by Tableau. These datasets are built by processing data fields and values from reports for single queries obtained from Sketch Engine, as well as combining results from multiple reports. Table 4 details all the data fields sourced from Sketch Engine reports and used to build datasets.

3. Visualisations

This section presents the visualisations created to support lexical data interpretation in LARs. Each subsection is organised around the datasets used to build each visualisation. Visualisations built with the same dataset are discussed in the same subsection. Unfortunately, due to length constraints, we will not provide detailed instructions of how each visualisation was built on Tableau.

By default, all LARs contain at least six visualisations, namely:

- a frequency histogram, disaggregating frequency by year of publication, organisation type, region and document type;
- a map, representing absolute frequency and relative frequency by region;
- a collocation histogram, showing the collocates by year with the highest logDice score;
- a dual axis bar and line chart, representing relative frequency and absolute frequency by year, region, organisation type and document type;
- a unique collocate packed bubble chart, representing collocates unique to each organisation type and their logDice scores; and
- a bar chart, representing collocates shared by more than two organisation types.

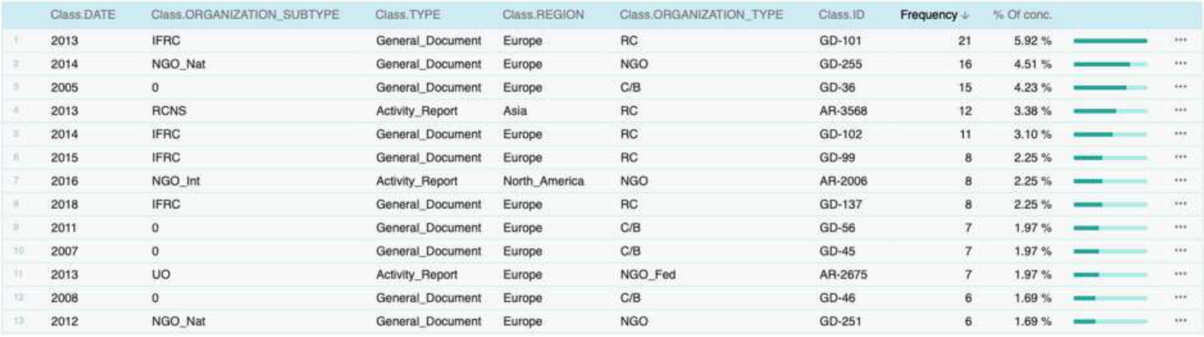
Additional visualisations are created depending on the nature of each concept entry. This article also covers the following *ad hoc* visualisations:

- square treemaps, detailing conceptual combinations and coordinated concepts;
- a histogram, representing manually curated contexts to represent conceptual development across time; and
- sortable and searchable tables containing manually curated contexts.

3.1 Frequency Histogram

A frequency histogram represents the frequency of a search term disaggregated by year of publication. This only requires a simple dataset that can be easily obtained from Sketch Engine. With it, our histogram can also disaggregate yearly frequencies by organisation type, region and document type.

To begin, we query the corpus with the Concordance tool by using the CQL expression [lemma_lc="x"] where x is the term or list of terms designating a concept. We then use the Frequency functionality to compute the frequencies of the search words in the concordance lines. Lastly, we select the Line Details pre-set, which generates a report detailing every document in the corpus that contains the search term. Each record represents a document and details all its text type metadata, frequency and a percentage of the total concordances (see Figure 1). This report is exported as a CSV file.



	Class.DATE	Class.ORGANIZATION_SUBTYPE	Class.TYPE	Class.REGION	Class.ORGANIZATION_TYPE	Class.ID	Frequency ↓	% Of conc.	
1	2013	IFRC	General_Document	Europe	RC	GD-101	21	5.92 %	···
2	2014	NGO_Nat	General_Document	Europe	NGO	GD-255	16	4.51 %	···
3	2005	0	General_Document	Europe	C/B	GD-36	15	4.23 %	···
4	2013	RCNS	Activity_Report	Asia	RC	AR-3568	12	3.38 %	···
5	2014	IFRC	General_Document	Europe	RC	GD-102	11	3.10 %	···
6	2015	IFRC	General_Document	Europe	RC	GD-99	8	2.25 %	···
7	2016	NGO_Int	Activity_Report	North_America	NGO	AR-2006	8	2.25 %	···
8	2018	IFRC	General_Document	Europe	RC	GD-137	8	2.25 %	···
9	2011	0	General_Document	Europe	C/B	GD-56	7	1.97 %	···
10	2007	0	General_Document	Europe	C/B	GD-45	7	1.97 %	···
11	2013	UO	Activity_Report	Europe	NGO_Fed	AR-2675	7	1.97 %	···
12	2008	0	General_Document	Europe	C/B	GD-46	6	1.69 %	···
13	2012	NGO_Nat	General_Document	Europe	NGO	GD-251	6	1.69 %	···

Figure 1: Frequency Line Details report for LEAVE NO ONE BEHIND

The resulting raw CSV file requires minimal treatment with spreadsheet software because the target data structure mirrors the one generated by Sketch Engine, as can be seen in Figure 1. This means creating a spreadsheet with each row representing a document, six columns containing text type metadata and a seventh column containing frequency values. The percentage of total concordances is discarded. After treatment, the CSV file is ready to be added on Tableau as a data source.

In Tableau, fields for text type metadata are set as dimensions, whereas frequency is set as a measure. Figure 2 shows the default view of our frequency histogram as published in the LAR for LEAVE NO ONE BEHIND. On the right there are three toggle options that allow users to further disaggregate frequencies by increasing the number of axes.

Field experts can thus observe that LEAVE NO ONE BEHIND appears mostly in documents published in Europe, followed by North America. Overall, the top five contributors in terms of occurrences are IGO, NGO, NGO_Fed, Net and State organisations. IGO documents generate more than half of all occurrences in the HE Corpus. Contributions from other organisation types are significantly smaller.

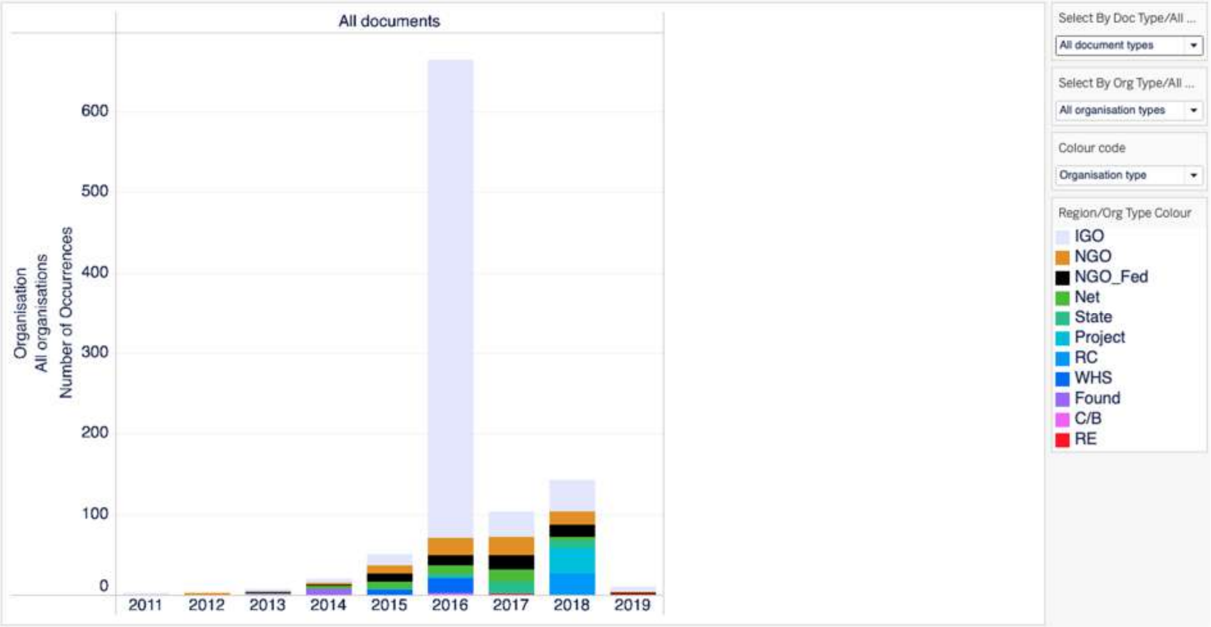

Select By Doc Type/All ...
All document types
Select By Org Type/All ...
All organisation types
Colour code
Organisation type

Figure 2: Default view of the frequency histogram for LEAVE NO ONE BEHIND

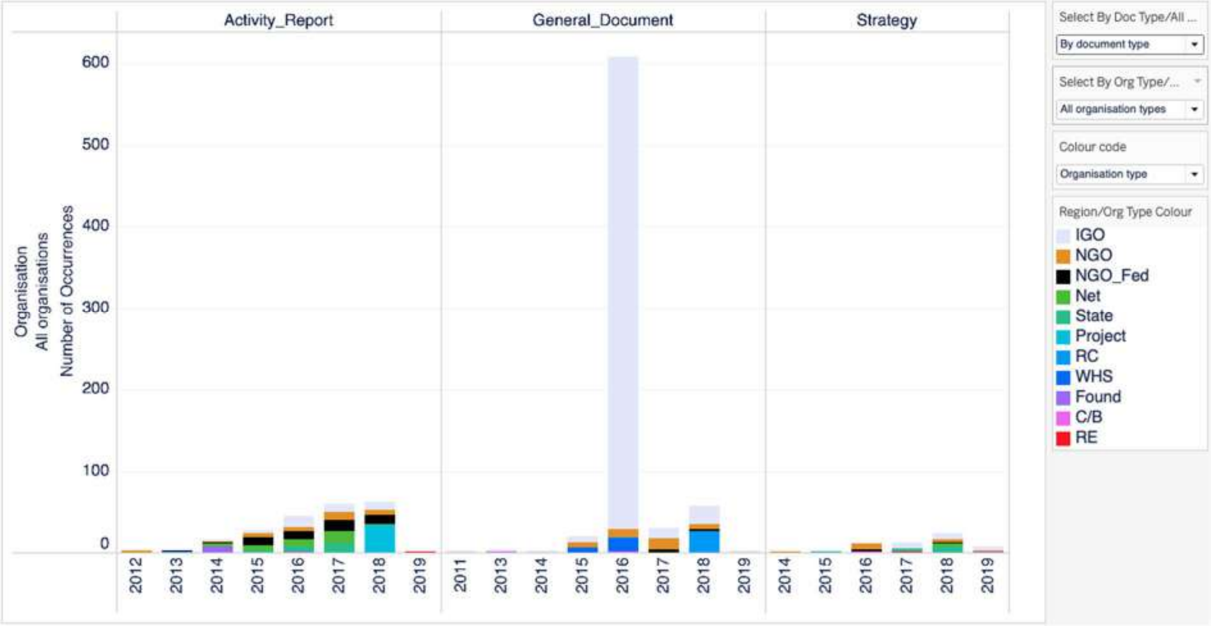

Select By Doc Type/All ...
By document type
Select By Org Type/...
All organisation types
Colour code
Organisation type

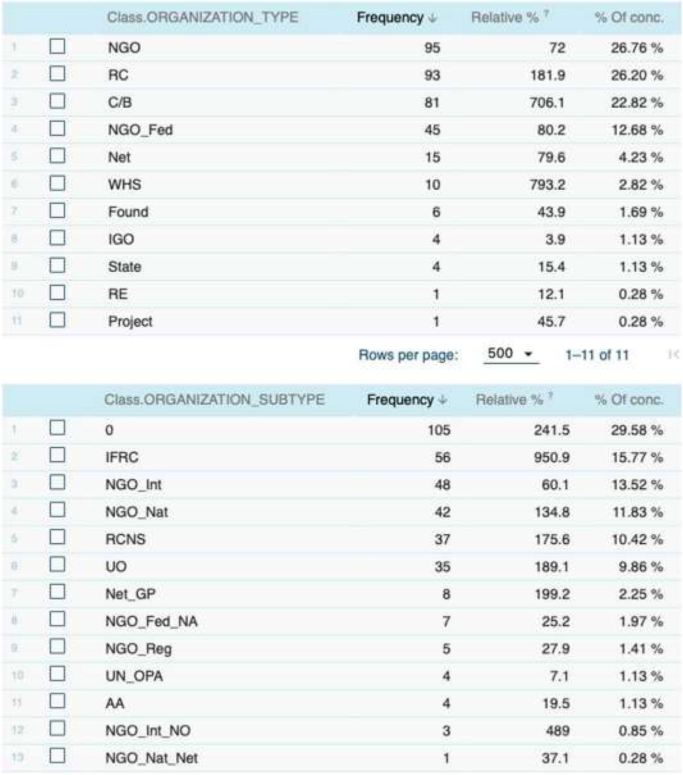
Figure 3: A dynamic axis view disaggregating
yearly frequencies by document type.

3.2 Map and Relative Frequency Bar Charts

Comparing absolute frequency and relative frequencies can be achieved by building a dataset for each concept. It can be used to create two visualisations. The first is a map representing absolute and relative frequencies by region. The second constitutes a set of bar charts that focus on comparing absolute and relative frequencies disaggregated by year, organisation type, region and document type.

As with the dataset described in Section 4.1, we queried the corpus with the Concordance tool by using CQL. In the Frequency functionality, we used instead the Text Types pre-set, which generates a report detailing the absolute frequency, relative frequency and percentage of total concordances for each text type in the corpus. Figure 4 shows part of this report with values for organisation types and subtypes.

Text type reports as CSV files require more treatment with spreadsheet software. Each record in the report corresponds to a kind of text type, and this is not specified in the raw CSV file. This means that all text types are contained in the same column. For this reason, a new column has to be added to disambiguate text types. Figure 5 illustrates the spreadsheet treatment process to obtain a data structure that can be interpreted correctly by Tableau.

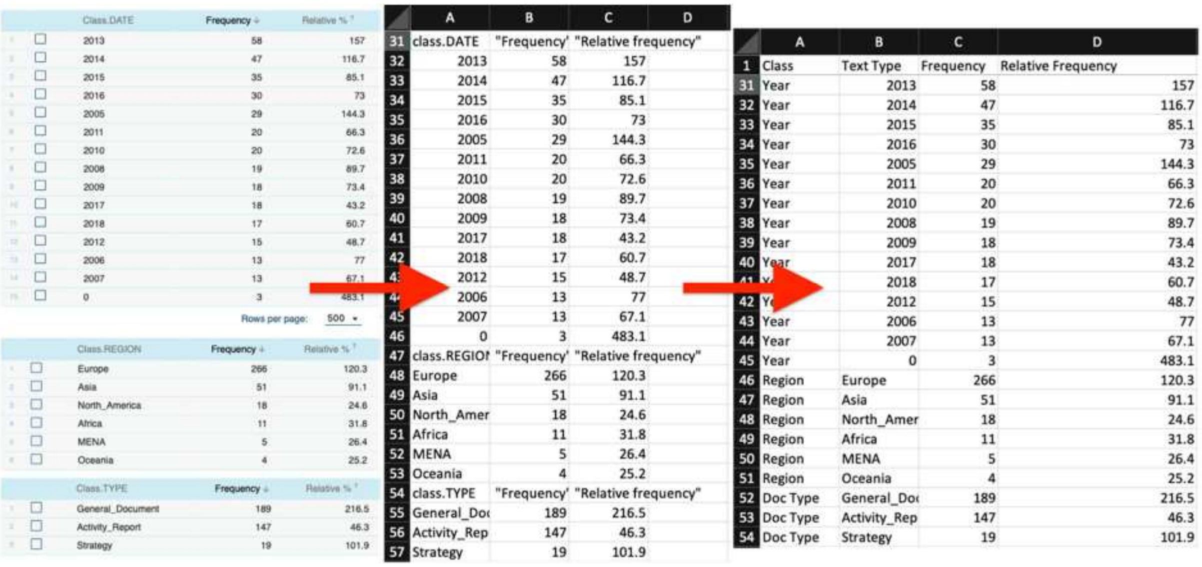


		Class.ORGANIZATION_TYPE	Frequency ↓	Relative % ?	% Of conc.
1	□	NGO	95	72	26.76 %
2	□	RC	93	181.9	26.20 %
3	□	C/B	81	706.1	22.82 %
4	□	NGO_Fed	45	80.2	12.68 %
5	□	Net	15	79.6	4.23 %
6	□	WHS	10	793.2	2.82 %
7	□	Found	6	43.9	1.69 %
8	□	IGO	4	3.9	1.13 %
9	□	State	4	15.4	1.13 %
10	□	RE	1	12.1	0.28 %
11	□	Project	1	45.7	0.28 %

Rows per page: 500 ▾ 1–11 of 11

		Class.ORGANIZATION_SUBTYPE	Frequency ↓	Relative % ?	% Of conc.
1	□	0	105	241.5	29.58 %
2	□	IFRC	56	950.9	15.77 %
3	□	NGO_Int	48	60.1	13.52 %
4	□	NGO_Nat	42	134.8	11.83 %
5	□	RCNS	37	175.6	10.42 %
6	□	UO	35	189.1	9.86 %
7	□	Net_GP	8	199.2	2.25 %
8	□	NGO_Fed_NA	7	25.2	1.97 %
9	□	NGO_Reg	5	27.9	1.41 %
10	□	UN_OPA	4	7.1	1.13 %
11	□	AA	4	19.5	1.13 %
12	□	NGO_Int_NO	3	489	0.85 %
13	□	NGO_Nat_Net	1	37.1	0.28 %

Figure 4: Text type report for HUMANITARIANISM



		Class.DATE	Frequency ↓	Relative %
1	□	2013	58	157
2	□	2014	47	116.7
3	□	2015	35	85.1
4	□	2016	30	73
5	□	2005	29	144.3
6	□	2011	20	66.3
7	□	2010	20	72.6
8	□	2008	19	89.7
9	□	2009	18	73.4
10	□	2017	18	43.2
11	□	2018	17	60.7
12	□	2012	15	48.7
13	□	2006	13	77
14	□	2007	13	67.1
15	□	0	3	483.1

Rows per page: 500 ▾

		Class.REGION	Frequency ↓	Relative %
1	□	Europe	266	120.3
2	□	Asia	51	91.1
3	□	North_America	18	24.6
4	□	Africa	11	31.8
5	□	MENA	5	26.4
6	□	Oceania	4	25.2

		Class.TYPE	Frequency ↓	Relative %
1	□	General_Document	189	216.5
2	□	Activity_Report	147	46.3
3	□	Strategy	19	101.9

	A	B	C	D
31	class.DATE	"Frequency'	"Relative frequency"	
32	2013	58	157	
33	2014	47	116.7	
34	2015	35	85.1	
35	2016	30	73	
36	2005	29	144.3	
37	2011	20	66.3	
38	2010	20	72.6	
39	2008	19	89.7	
40	2009	18	73.4	
41	2017	18	43.2	
42	2018	17	60.7	
43	2012	15	48.7	
44	2006	13	77	
45	2007	13	67.1	
46	0	3	483.1	
47	class.REGIOI	"Frequency'	"Relative frequency"	
48	Europe	266	120.3	
49	Asia	51	91.1	
50	North_Amer	18	24.6	
51	Africa	11	31.8	
52	MENA	5	26.4	
53	Oceania	4	25.2	
54	class.TYPE	"Frequency'	"Relative frequency"	
55	General_Do(	189	216.5	
56	Activity_Rep	147	46.3	
57	Strategy	19	101.9	

	A	B	C	D
1	Class	Text Type	Frequency	Relative Frequency
31	Year	2013	58	157
32	Year	2014	47	116.7
33	Year	2015	35	85.1
34	Year	2016	30	73
35	Year	2005	29	144.3
36	Year	2011	20	66.3
37	Year	2010	20	72.6
38	Year	2008	19	89.7
39	Year	2009	18	73.4
40	Year	2017	18	43.2
41	Y	2018	17	60.7
42	Y	2012	15	48.7
43	Year	2006	13	77
44	Year	2007	13	67.1
45	Year	0	3	483.1
46	Region	Europe	266	120.3
47	Region	Asia	51	91.1
48	Region	North_Amer	18	24.6
49	Region	Africa	11	31.8
50	Region	MENA	5	26.4
51	Region	Oceania	4	25.2
52	Doc Type	General_Do(	189	216.5
53	Doc Type	Activity_Rep	147	46.3
54	Doc Type	Strategy	19	101.9

Figure 5: Spreadsheet treatment for a text type report

In Tableau, fields for Class and Text Type are set as dimensions, while absolute frequency and relative frequency are set as measures. As shown in Figure 6, our map represents, for each HE region, frequency with solid colour bubbles and relative frequency with a ring around each bubble. Tableau comes with a great deal of predefined geographical units for disaggregation such as countries, US states, Canadian provinces, European NUTS, among others. However, these do not match HE regions. By means of calculated fields, we linked our HE regions to a country whose location on the map serves as a good anchor point for each bubble. For example, the Europe bubble is anchored to Denmark, whereas the CCSA (Central Caribbean and South America) bubble is anchored to Bolivia.

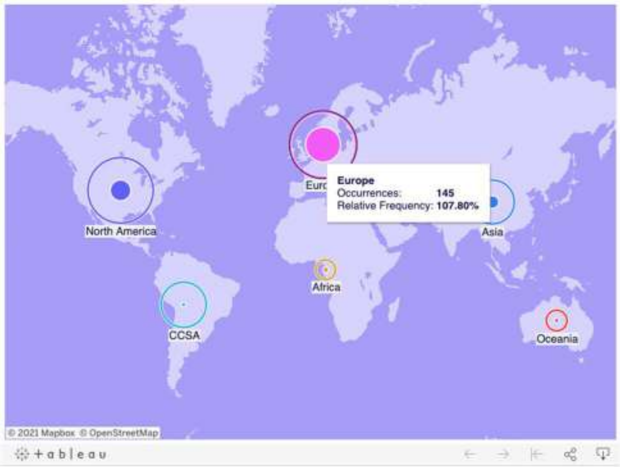


Figure 6: A map for HUMANITARIANISM

The same dataset can be used to build visualisations that compare relative frequency with absolute frequency, disaggregating by text type, namely year, region, organisation and document type. This entails building four different bar charts, which can be presented together with a Tableau story. Figure 7 shows the default view of this story, a histogram representing relative frequency as bars and absolute frequency as a superimposed line. To view the other three disaggregation options, users can use the buttons located at the top.

Exploring the visualisation in Figure 7 in detail sheds light on the temporal evolution of PARTICIPATION. Collectively, its occurrences were highest in 2015, whereas 2013 saw the highest relative frequency with nearly 160 %; European general documents generated the greatest number of occurrences; and the top five organisation types with the highest relative frequency of participation are WHS, C/B, RC, NGO_Fed and Net.

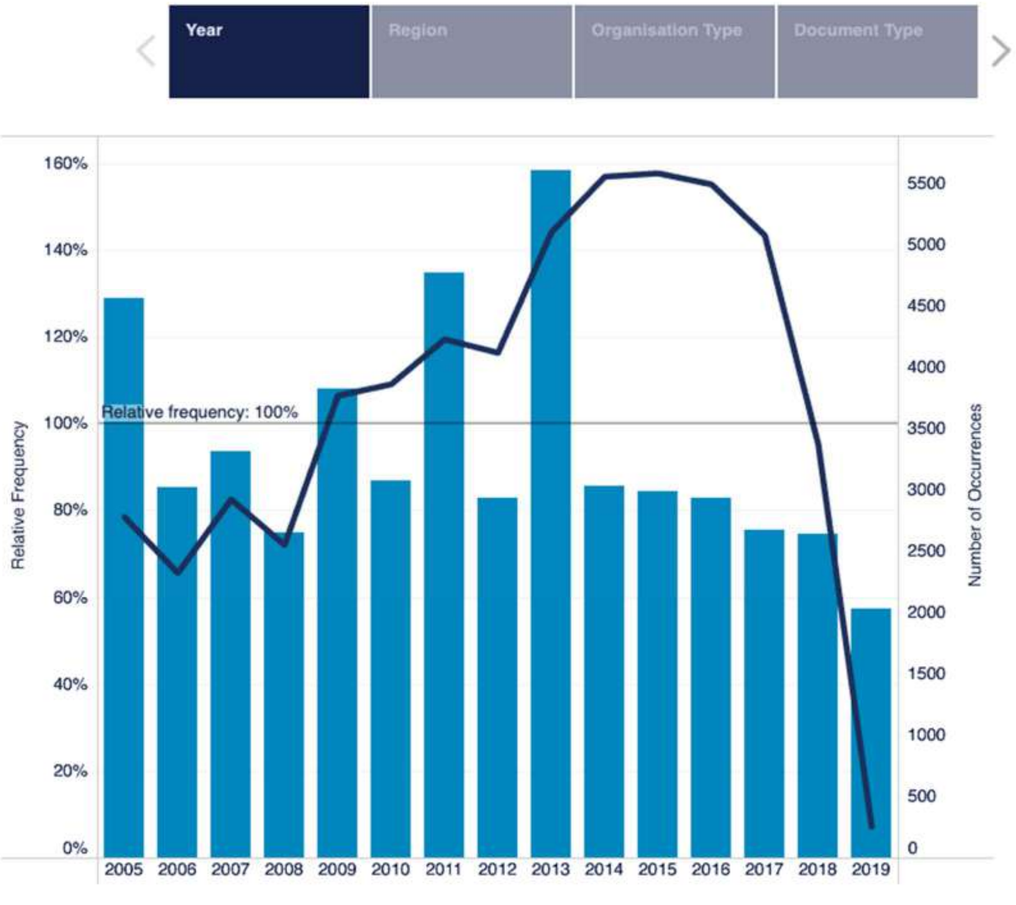


Figure 7: A Tableau story showing a histogram comparing absolute and relative frequency for PARTICIPATION

3.3 Top Yearly Collocate Histogram

To explore the evolution of collocates for a given search expression (which can be highly informative when analysing concept dynamics), a dataset can be built by conducting multiple queries in Sketch Engine. This dataset has to contain the necessary information to disaggregate collocates by year and organisation type.

To begin, we conduct multiple queries so as to obtain collocate reports for each year. This can be achieved by specifying document metadata in each CQL query. Firstly, we query the corpus with the Concordance tool by using the expression [lemma_lc="x"] within <class(DATE="y")>, where x is the term or terms designating a concept, and y is a year of publication. Once a list of concordances is generated, we then select the Collocations functionality, which computes collocations of the search term or terms. Even though Sketch Engine encourages the use of its Word Sketch tool for this purpose, there are unfortunately two issues with this. The first is that it can only be used with lemma tags, which means that capitalised occurrences are automatically discarded. The second is that it does not work well with multi-word expressions, which is the case for many of HE concepts. In the Collocations functionality, we set a range of -3, 3 and select lemma (lowercase) as the computation attribute. Finally, a collocational report is generated, which contains all extracted collocates and a set of measures. This step has to be repeated 15 times, changing the year of publication for each query. For the purposes of our dataset, we are interested in the collocates and their corresponding logDice score.

For yearly organisation type-specific collocation reports, we query the corpus with [lemma_lc="x”] within <class(DATE="y") & (ORGANIZATION TYPE ="z" >, where z is the code of the five organisation types with the highest absolute frequency. As with organisation type-unspecific reports, collocational reports are generated through the Collocations functionality with the same settings. This task has to be performed 75 times, with all possible year-organisation type combinations.

Before all individual collocational reports are combined into a single spreadsheet, we curate collocates manually to remove prepositions, truncated words and other empty expressions from the lists. To ease the process, a stop word list is used, which is continuously fed with removed collocates from previous tasks. With spreadsheet software, records from organisation type-unspecific and specific reports are added in the file with an additional column. Furthermore, a second column is added to indicate the year of publication. This leaves us with a single CVS containing collocates by year for the entire corpus, as well as collocates by year disaggregated by organisation type.

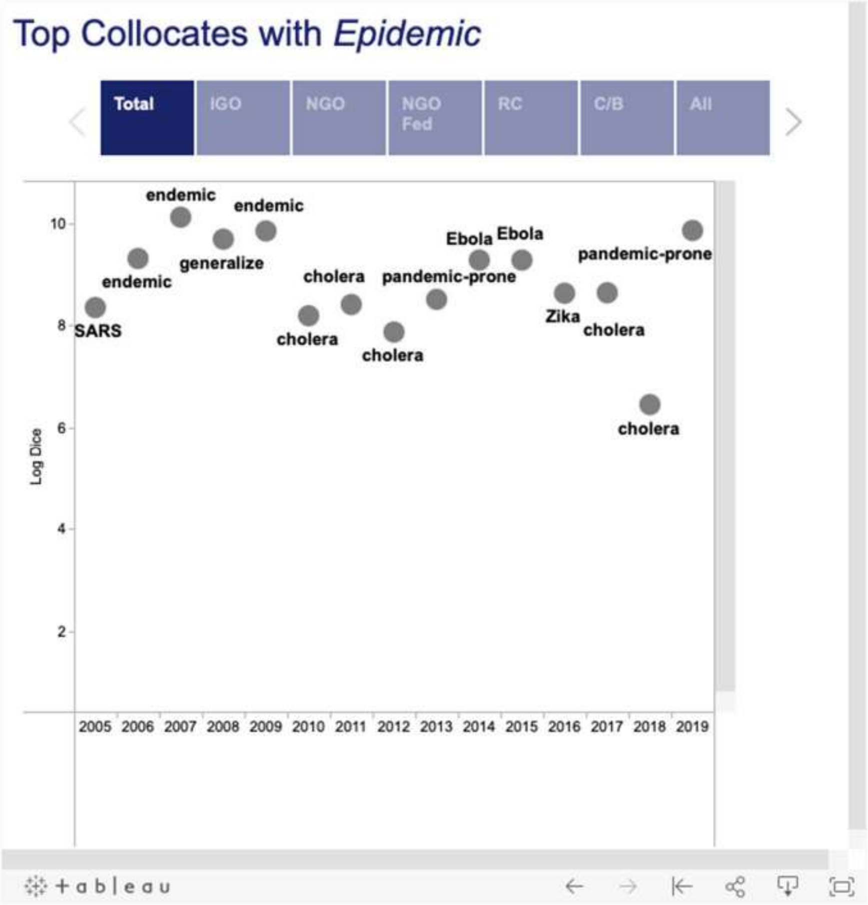


Figure 8: Histogram showing yearly top collocates for EPIDEMIC across the entire corpus

With such a dataset, we can create collocational histograms that allow users to see the top collocate for each year for the entire corpus (i.e., the whole set of concordances), as well as disaggregated by the top five organisation types (i.e., the five organisation types that generate the highest frequencies).

For instance, in the case of EPIDEMIC, epidemic types are the most salient collocates over the years (*SARS* in 2005, *cholera* in 2010-2012, 2017 and 2018, *Ebola* in 2014-15), *zika* in 2016). In 2006-7 and 2009 *endemic* stands out and the single verb in the selection is *generalize* (2008). More recently (2019), *pandemic-prone* is the top collocate, which reflects current concerns about epidemics. *Pandemic-prone* and *pandemic* seem to have been relevant for IGOs and RC for longer (the top collocate in 2013 and 2019 for IGOs' and 2010, 2013 and 2016 for RC). IGO's top collocates related to epidemic types also include *meningitis* (2007, 2009), whereas NGOs show more interest in *malaria* (2007) and *AIDS* (2008, 2009). The only top collocates related to epidemic management are found in texts by NGOs and C/Bs: *combat*, *forecasting* and *prevention*. And the only collocates related to causes are mentioned by NGO_Feds and NGOs: *miningococal* and *waterborne*.

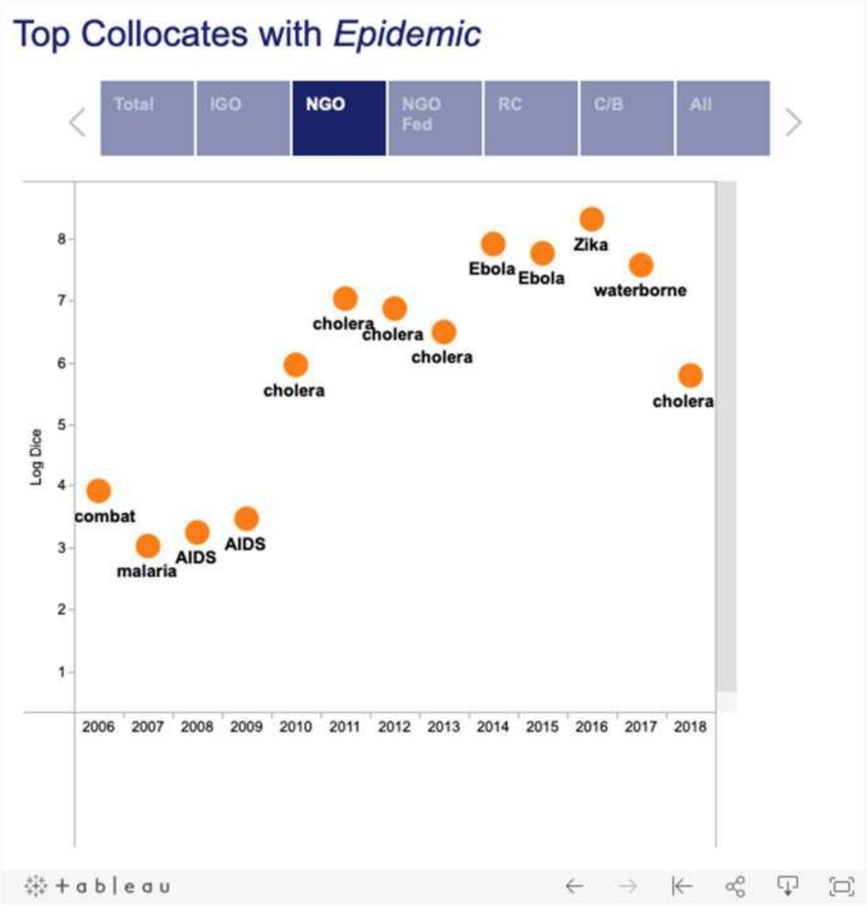


Figure 9: A histogram showing top yearly collocates for EPIDEMIC in NGO documents

The building process in Tableau is similar to that of the relative frequency bar charts (see Section 4.2) in that it requires multiple visualisations be presented together with the Story functionality. Fields for collocates, years and organisation types are set as dimensions, whilst logDice is set as a measure. As can be seen in Figure 8, collocates are presented as colour circles placed in a histogram at varying heights based on their logDice score. Figure 9 shows the top yearly collocates obtained from NGO documents.

3.4 Unique and Shared Collocates

Reporting on collocates that are unique to a single organisation type constitutes a way of ascertaining what a given organisation says about a concept that others do not. Examining which collocates are shared by multiple organisations can help identify what common areas among organisations when discussing a certain concept.

A dataset for this purpose can be built by corpus querying with a similar method seen in Section 4.3. However, in this case, we use the CQL expression [lemma_lc="x"]within<class(ORGANIZATION_TYPE="y")>, which does not specify a year of publication. The rest of the extraction process in Sketch Engine is identical. The corpus has to be queried five times for each of the five organisation types with the highest frequencies.

To combine the five collocational reports into CSV, a column is added in the spreadsheet to specify the organisation type from which each record was obtained. After this process, we have a dataset that enables us to compare collocates among organisation types. In the same workbook in Tableau, the fields for collocates and organisation types are set as dimensions, whereas logDice is set as a measure. Collocates unique to each organisation type can be well represented with a packed bubble chart (Figure 10). By means of a conditional set, collocates found in more than one organisation type can be filtered out.

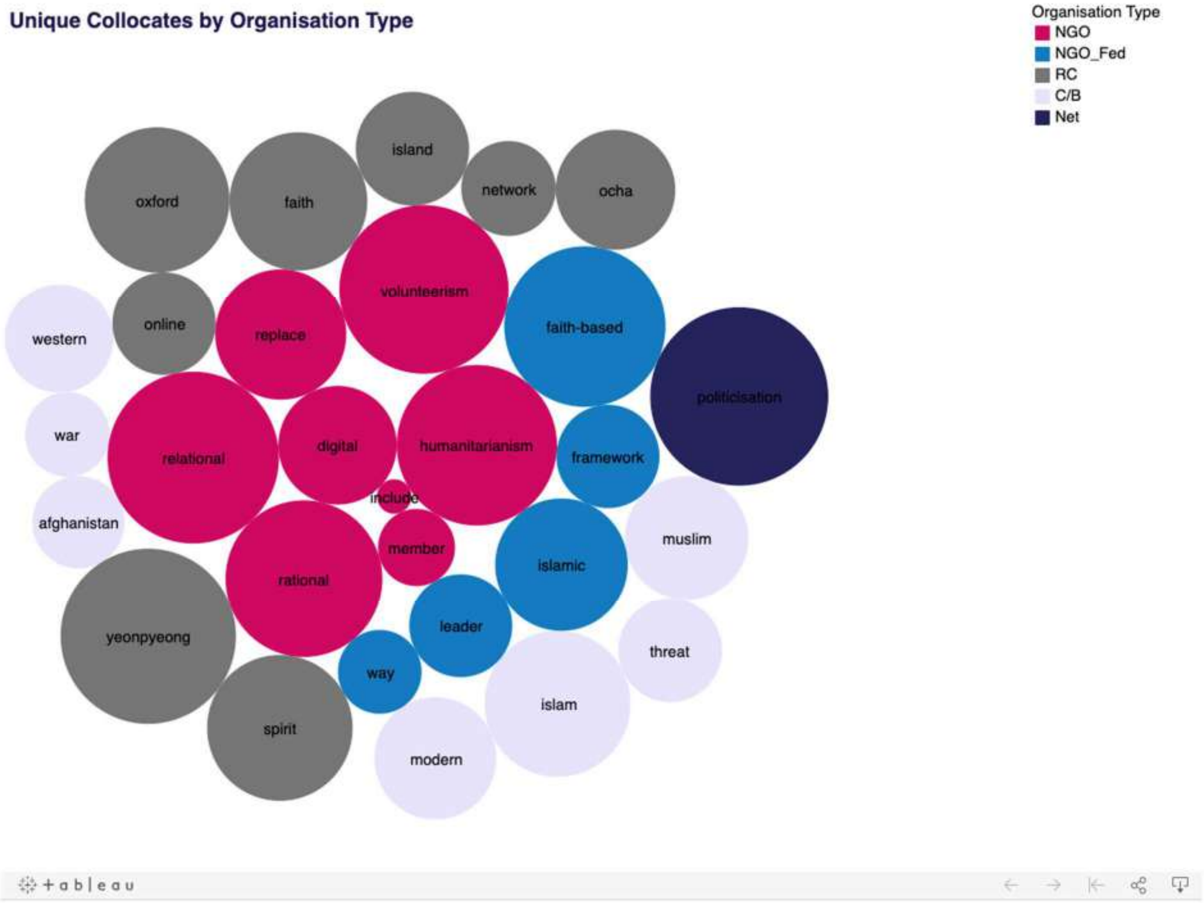


Figure 10: Unique collocates for HUMANITARIANISM

Collocates shared by multiple organisation types would be optimally represented by Venn diagrams. Here, shared collocates can be understood as the collocates that constitute intersections between organisation types, i.e., intersections between subcorpora. However, Tableau does not offer an option to build Venn diagrams. For this reason, we resorted to bar charts, which serve as a good alternative. With a parameter, a dynamic conditional set and a filter, we can create a bar chart that shows which collocates are shared by two or more organisation types. As can be seen in Figure 11, each collocate is represented by a bar that can be divided into multiple colour sections. The colour of each section represents an organisation type, while its size represents the collocate's logDice score within that given type. Thanks to a filtering parameter, users can filter collocates by the number of organisation types in which they appear.

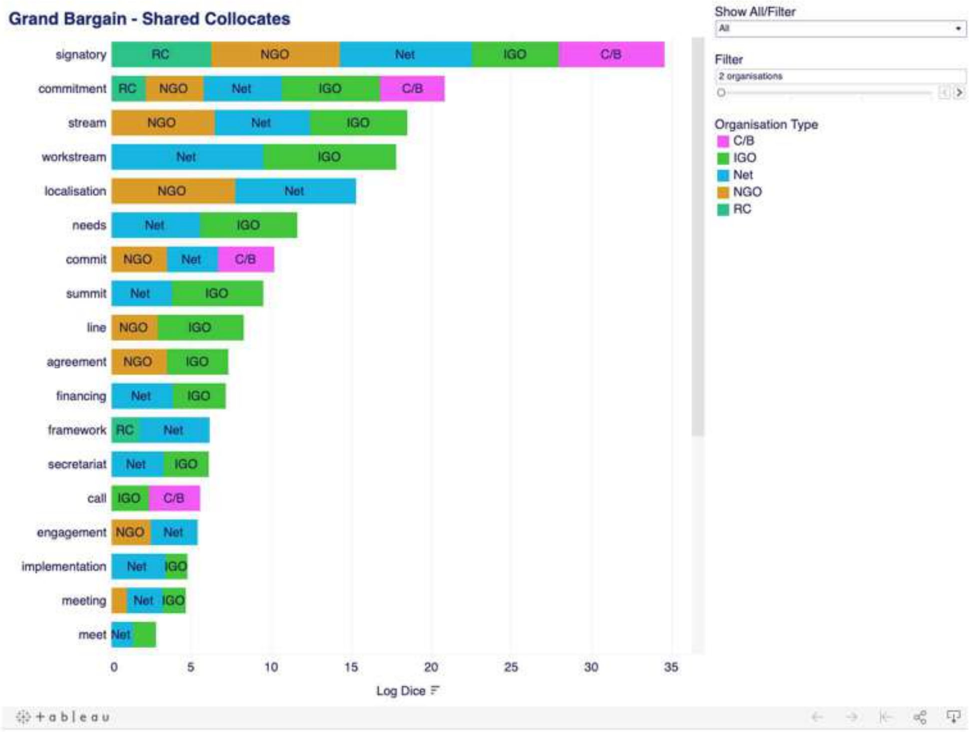


Figure 11: Shared collocates for GRAND BARGAIN

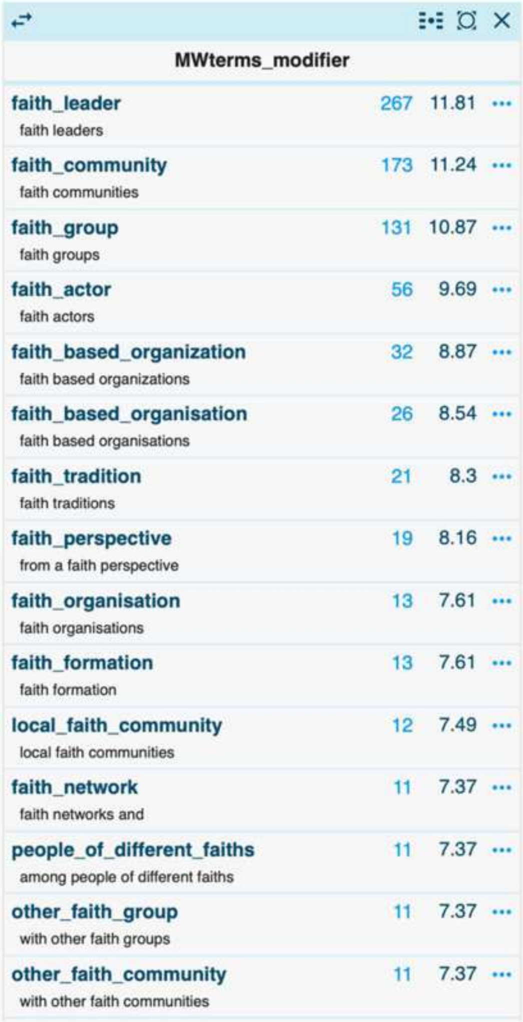
Thanks to the visualisations in Figure 10 and 11, linguists can inform experts about the collocating trends of *humanitarianism*. For instance, NGO documents feature the unique collocates of *relational*, *volunteerism*, *rational*, *replace*, *digital*, *member* and *include*, whereas C/B documents contain *Islam*, *Muslim*, *modern*, *Western*, *threat*, *Afghanistan*, *war* and *project*, pointing to very different concerns. The top collocates shared by two organisations are *value* and *crisis*, whereas the only collocate shared by three organisations is *development*. No collocates were found to be shared by either four or five organisations, which could indicate that large discrepancies are found in the conceptualisation of HUMANITARIANISM.

3.5 Square Treemaps

Square treemaps are an interesting option for the categorisation of multiple elements, as well as measures associated with said elements. This section will examine two case uses of this visualisation option within Tableau.

3.5.1 Representing compound concepts

Complex nominals are phrases consisting of a head noun modified by other elements, such as other nouns, adjectives and prepositional phrases. They are considered as instantiations of conceptual combinations, whereby compound concepts are formed by pre-existing simpler ones (Cabezas-García & Chambó, in press). Analysing the understanding of a concept in a given domain requires looking at the conceptual combination that it forms. Square treemaps are an effective way of representing such information.



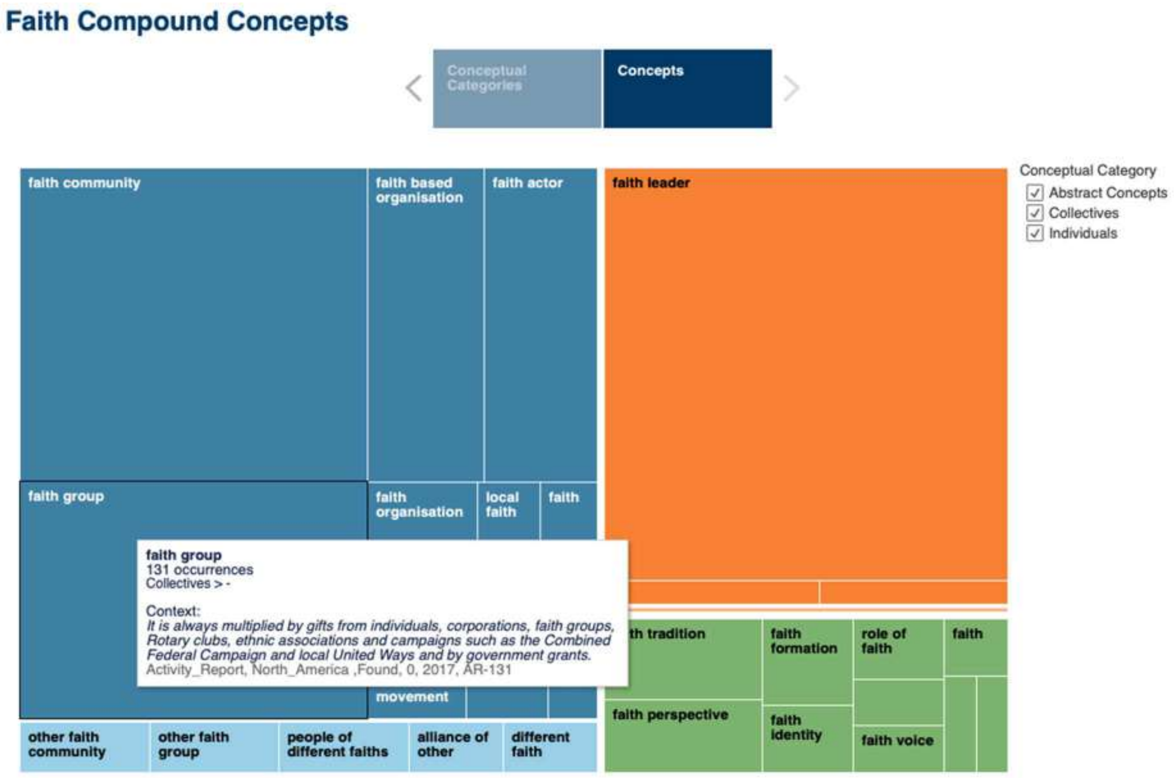
MWterms_modifier

faith_leader faith leaders	267	11.81	···
faith_community faith communities	173	11.24	···
faith_group faith groups	131	10.87	···
faith_actor faith actors	56	9.69	···
faith_based_organization faith based organizations	32	8.87	···
faith_based_organisation faith based organisations	26	8.54	···
faith_tradition faith traditions	21	8.3	···
faith_perspective from a faith perspective	19	8.16	···
faith_organisation faith organisations	13	7.61	···
faith_formation faith formation	13	7.61	···
local_faith_community local faith communities	12	7.49	···
faith_network faith networks and	11	7.37	···
people_of_different_faiths among people of different faiths	11	7.37	···
other_faith_group with other faith groups	11	7.37	···
other_faith_community with other faith communities	11	7.37	···

Figure 11: Word Sketch for MWTs for FAITH

Given that FAITH is designated by a monolexical term, we used a modified version of Sketch Engine's default sketch grammar, which is the backbone of the Word Sketch tool. This custom sketch grammar is able to extract the multi-word terms (MWTs) in which the search term appears as both as a head or a modifier. On the one hand, MWTs with *faith* as a head constitute hyponyms of FAITH (e.g., CHRISTIAN FAITH, ISLAMIC FAITH, LOCAL FAITH, etc.), which can also be classified according to different facets. On the other hand, MWTs with *faith* as a modifier constitute conceptual combinations in which faith intervenes (e.g., FAITH LEADER, FAITH COMMUNITY, FAITH IDENTITY, etc.), which would point to non-hierarchical relations and event participants. To represent the conceptual compounds with faith contained in the HE corpus, we extracted the MWTs with *faith* as a modifier (Figure 12).

All extracted MWTs with their frequencies were transferred into a spreadsheet and classified into conceptual categories by creating additional columns. Separately, another spreadsheet was manually populated with sample contexts from the HE corpus for each MWT, together with each context's metadata.



faith group
131 occurrences
Collectives > -

Context:
It is always multiplied by gifts from individuals, corporations, faith groups, Rotary clubs, ethnic associations and campaigns such as the Combined Federal Campaign and local United Ways and by government grants.
Activity_Report, North_America ,Found, 0, 2017, AR-131

Figure 12: Square treemap providing
a summary of compound concepts with FAITH

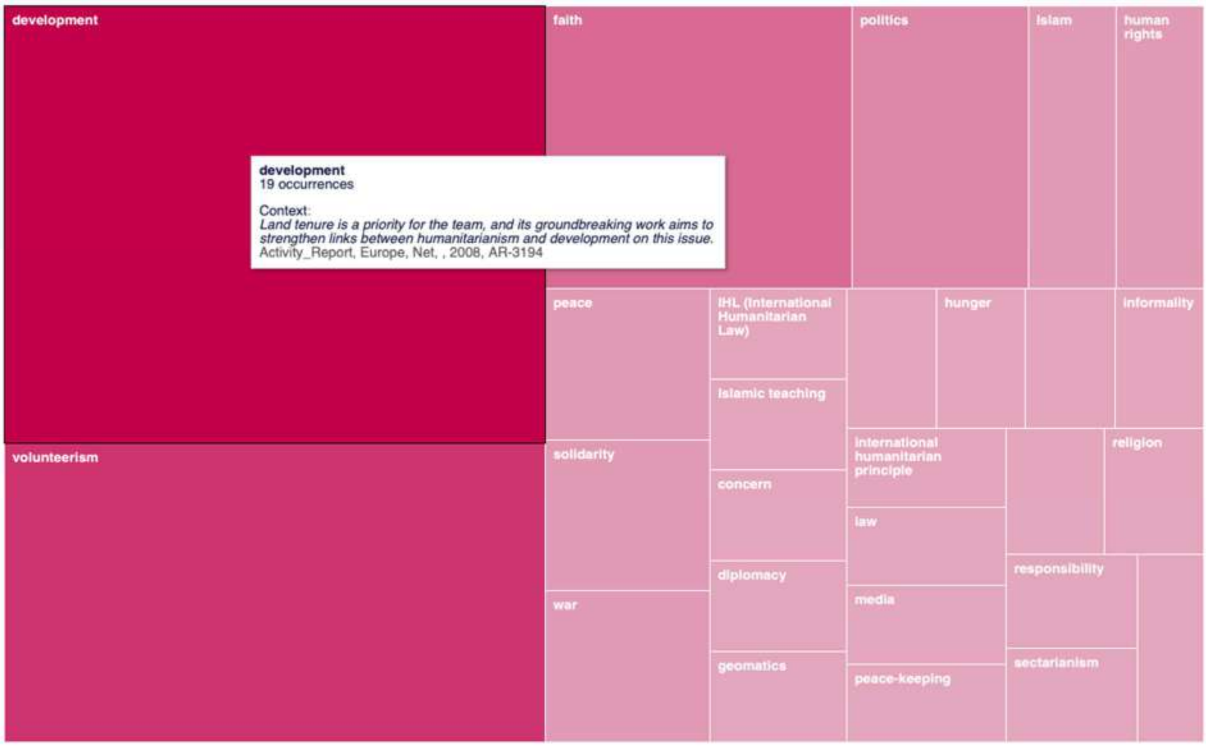
In Tableau, both spreadsheets are joined with a union. The frequency for each MWT is set as the measure, while compound concepts and context metadata are set as dimensions. In a square treemap, each compound concept is symbolised by a rectangle whose size represents its frequency in the corpus. As can be seen in Figure 12, when the user hovers a rectangle, a tooltip provides a sample context as well as the details of the document from which it was sourced.

3.5.2 Representing coordinated concepts

In text, associated concepts may also appear in coordination. The Word Sketch functionality can extract expressions linked to a search term through coordination with the conjunctions *and* and *or*. However, it is not powerful enough to extract coordinated MWTs. For this reason, in order to create a dataset containing all coordinated concepts with HUMANITARIANISM, we queried the corpus with the following two CQL expressions:

- [tag="N.*|J.*"]{1,3} within ([lemma_lc="humanitarianism"]
  [word="and|/or"]
  ([tag="N.*|J.*"]{1,3}within[tag!="N.*|J.*"][tag="N.*|J.*"]{1,3}[tag!="N.*"]))
- [tag="N.*|J.*"]{1,3} within
  (([tag="N.*|J.*"]{1,3}within[tag!="N.*|J.*"][tag="N.*|J.*"]{1,3}[tag!="N.*"])
  [word="and|/or"] [lemma_lc="humanitarianism"])

In brief, the above expressions extract both single-word and multiword expressions coordinated with humanitarianism. Concordances were filtered with the Hide Sub-Hits quick filtering functionality, which removes concordances including partial hits. This is bound to occur when using ranges (e.g., {1,3}) to capture complex nominals. Both sets of concordances were computed using the Frequency functionality, which generated two report containing full coordinated expressions on both sides of our search term.



development
19 occurrences

Context:
Land tenure is a priority for the team, and its groundbreaking work aims to strengthen links between humanitarianism and development on this issue.
Activity_Report, Europe, Net, , 2008, AR-3194

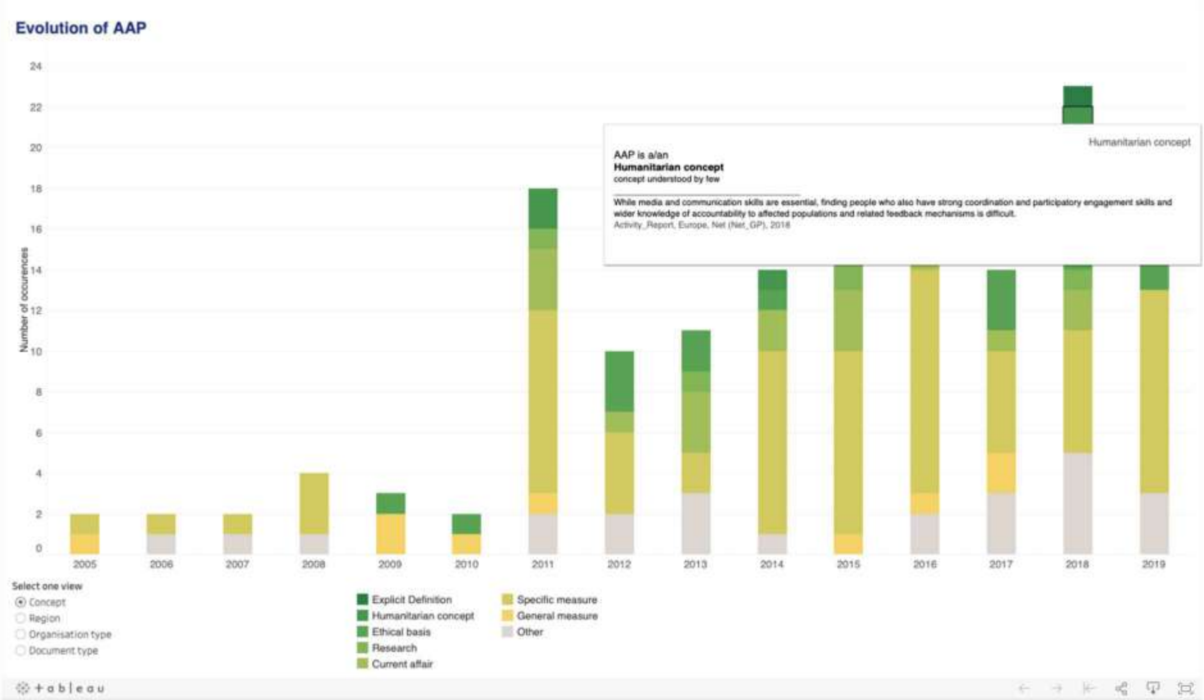
Figure 12: Coordinated concepts with HUMANITARIANISM

Both reports were combined into a single spreadsheet containing frequencies for each expression. As with the case use described in Section 4.5.1, a separate spreadsheet with context samples was also built. Similarly, both data sources were joined with a union in Tableau and visualised using the treemap functionality. The resulting visualisation provides a summary of the concepts coordinated with HUMANITARIANISM (Figure 12). The analysis of conceptual compounds reveals that humanitarian discourse is concerned with notional discussions about the concept of HUMANITARIANISM. Other important aspects include the *constituent elements* of humanitarianism (e.g. core values, language, activities, practice, etc.) and those *processes that affect humanitarianism* (e.g. demilitarisation, politicisation, sanctification, etc.).

3.6 Conceptual Development Histogram

Some concepts can be so specific that it pays to represent their development over time. This is usually the case for compound concepts that generate a handful of knowledge-rich contexts, which can be curated manually and classified into descriptive categories. The compound concept of ACCOUNTABILITY TO AFFECTED POPULATIONS is a highly specialised humanitarian concept that is formed by AFFECTED POPULATION, a concept with constitutes a fully-fledged entry in the HE. Numerous occurrences of its acronym – AAP – Indicate that the concept has solidified.

A low number of occurrences allows a linguist to download and classify statements manually into multiple categories, thus creating a heavily textual dataset. Using a similar method as described in Section 4.1., a histogram can be built to represent categorised contexts by year as shown in Figure 13.



AAP is a/an
Humanitarian concept
concept understood by few

While media and communication skills are essential, finding people who also have strong coordination and participatory engagement skills and wider knowledge of accountability to affected populations and related feedback mechanisms is difficult.
Activity_Report, Europe, Net (Net_GP), 2018

Figure 13: A histogram representing the evolution of ACCOUNTABILITY TO AFFECTED POPULATIONS (AAP)

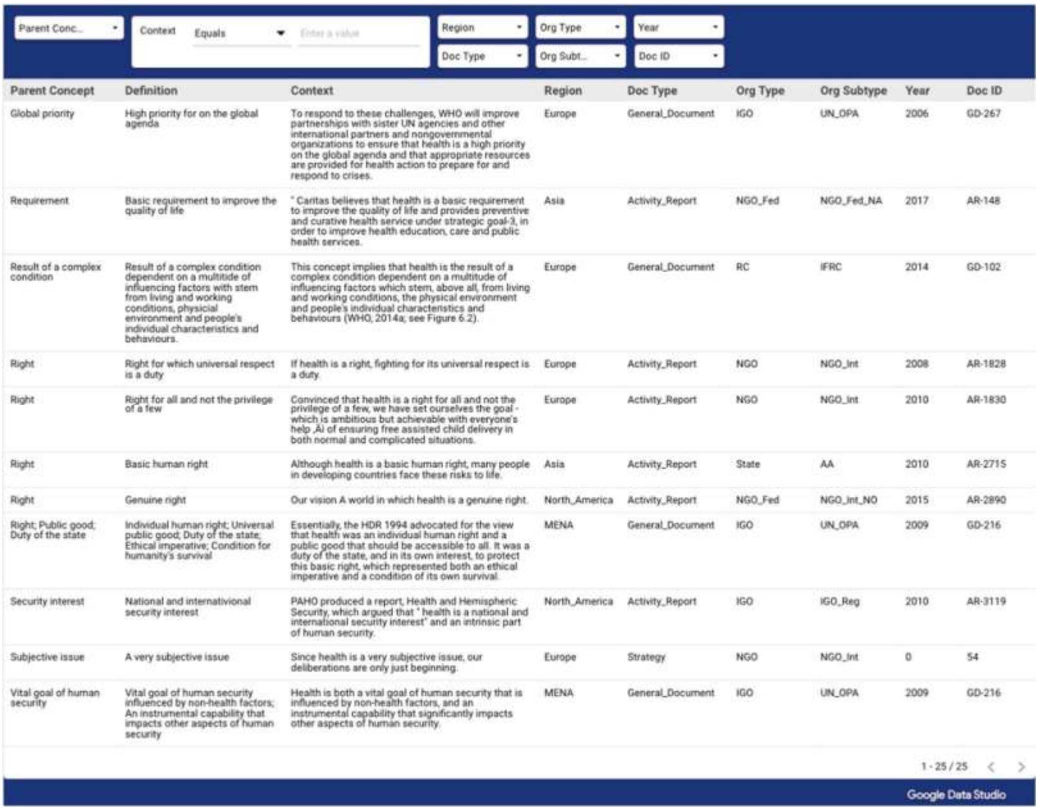
All contexts were classified into eight statement categories based on what organisations say about it, namely: general measure, specific measure, current affair, research, ethical basis, humanitarian concept, explicit definition and other. Contexts in the general measure and specific measure categories describe measures taken or that could be taken by organisations to increase ACCOUNTABILITY TO AFFECTED POPULATIONS. Contexts categorised as a current affair describe the concept as an ongoing concern in the humanitarian domain. The research category includes contexts stating that research is either needed or being conducted to increase and/or better understand ACCOUNTABILITY TO AFFECTED POPULATIONS. The ethical basis category consists of contexts in which the concept is described as an organisational value or principle. Contexts in the humanitarian concept category state that it constitutes a humanitarian concept, whilst the explicit definition category contain an authoritative definition for the concept as found in the corpus. Lastly, contexts classified as other mostly include statements merely claiming that a given organisation works towards AAP, as well as other marginal cases. This statement classification system makes it possible to represemt how AAP develops from vague mentions to more specific and defined mentions. In the visualisation, hovering over each bar section to reveals more details about each mention, including the context.

In 2011, APP began to attract attention as evidenced by a dramatic increase in occurrences. Documents published in this year contain the two first mentions pointing out that AAP is an ill-defined concept. It was not until 2015 that a progressive increase surpasses the value for 2011. Mentions of conceptual vagueness reappeared in 2015 and 2016. The greatest number of occurrences were obtained from documents published in 2018, when a change in proportion of statement categories can be observed. 2018 saw the only explicit definition of AAP as well as a considerable increase of occurrences classified as ethical basis. 2019 also experiences a change in proportion, with most statements being from the specific measure and ethical basis categories. It also ceased to be referred to as a cross-cutting issue or key current challenge.

Analysis suggests that ACCOUNTABILITY TO AFFECTED POPULATIONS crystallised as a concept in 2018. This is when its first definition appeared, and the greatest number of organisations claimed their adherence to it as a principle.

3.7 Filterable and Searchable Tables

Sometimes linguistic reporting for certain concepts may require presenting entire sets of manually curated contexts. Reporting on explicit definitions is perhaps the first step when describing the conceptualisation of a notion. For example, the HE corpus contains many definitions for the concept of HEALTH in varying degrees of explicitness. As shown in Figure 14, these were presented in a sortable and searchable table built with Google Data Studio. These are designed to allow users to search and filter contexts.



Parent Concept	Definition	Context	Region	Doc Type	Org Type	Org Subtype	Year	Doc ID
Global priority	High priority for on the global agenda	To respond to these challenges, WHO will improve partnerships with sister UN agencies and other international partners and nongovernmental organizations to ensure that health is a high priority on the global agenda and that appropriate resources are provided for health action to prepare for and respond to crises.	Europe	General_Document	IGO	UN_OPA	2006	GD-267
Requirement	Basic requirement to improve the quality of life	" Caritas believes that health is a basic requirement to improve the quality of life and provides preventive and curative health service under strategic goal-3, in order to improve health education, care and public health services.	Asia	Activity_Report	NGO_Fed	NGO_Fed_NA	2017	AR-148
Result of a complex condition	Result of a complex condition dependent on a multitide of influencing factors with stem from living and working conditions, physicial environment and people's individual characteristics and behaviours.	This concept implies that health is the result of a complex condition dependent on a multitude of influencing factors which stem, above all, from living and working conditions, the physical environment and people's individual characteristics and behaviours (WHO, 2014a; see Figure 6.2).	Europe	General_Document	RC	IFRC	2014	GD-102
Right	Right for which universal respect is a duty	If health is a right, fighting for its universal respect is a duty.	Europe	Activity_Report	NGO	NGO_Int	2008	AR-1828
Right	Right for all and not the privilege of a few	Convinced that health is a right for all and not the privilege of a few, we have set ourselves the goal - which is ambitious but achievable with everyone's help ‚Äì of ensuring free assisted child delivery in both normal and complicated situations.	Europe	Activity_Report	NGO	NGO_Int	2010	AR-1830
Right	Basic human right	Although health is a basic human right, many people in developing countries face these risks to life.	Asia	Activity_Report	State	AA	2010	AR-2715
Right	Genuine right	Our vision A world in which health is a genuine right.	North_America	Activity_Report	NGO_Fed	NGO_Int_NO	2015	AR-2890
Right; Public good; Duty of the state	Individual human right; Universal public good; Duty of the state; Ethical imperative; Condition for humanity's survival	Essentially, the HDR 1994 advocated for the view that health was an individual human right and a public good that should be accessible to all. It was a duty of the state, and in its own interest, to protect this basic right, which represented both an ethical imperative and a condition of its own survival.	MENA	General_Document	IGO	UN_OPA	2009	GD-216
Security interest	National and internativional security interest	PAHO produced a report, Health and Hemispheric Security, which argued that " health is a national and international security interest" and an intrinsic part of human security.	North_America	Activity_Report	IGO	IGO_Reg	2010	AR-3119
Subjective issue	A very subjective issue	Since health is a very subjective issue, our deliberations are only just beginning.	Europe	Strategy	NGO	NGO_Int	0	54
Vital goal of human security	Vital goal of human security influenced by non-health factors; An instrumental capability that impacts other aspects of human security	Health is both a vital goal of human security that is influenced by non-health factors, and an instrumental capability that significantly impacts other aspects of human security.	MENA	General_Document	IGO	UN_OPA	2009	GD-216

Figure 12: A filterable and searchable table showing a selection of explicit definitions for HEALTH

These tables are presented in a separate subpage and are mainly intended as supportive evidence for a linguist's claims in the many pages and body of his or her LAR. By analysing explicit definitions, the linguist concludes that definitions are built on three distinct conceptualisations of HEALTH: health as a state or condition, health as human right; and health as a fundamental component.

These tables are also used in order to store and provide an interactive access to debates and controversies, widely found in humanitarian discourse and manually curated and categorised by HE linguists.

3.8 Recapitulation

Table 5 provides a summary of all the visualisation types discussed in this paper. It also contains a link to each visualisation on Tableau public and Google Data Studio from which it can be freely downloaded.

Visualisatio n	**Purpose**	**Dimensions**	**Measures**
Frequency Histogram	To display the evolution of frequency over time, allowing users to disaggregate yearly frequencies by organisation type, document type and region.	Year, Organisation type, Region, Document type	Frequency
Map	To display the geographical distribution of absolute frequency and relative frequency among regions.	Region	Frequency, Relative frequency
Relative Frequency Bar Chart	To compare yearly absolute frequencies and relative frequencies, allowing users to explore other distributions by organisation type, document type and region.	Year, Organisation type, Region, Document type	Frequency, Relative frequency
Top Yearly Collocate Histogram	To compare most significant collocates over time and among organisation types.	Lexical unit (collocate), Year, Organisation type	logDice
Unique Collocates	To show collocates unique to organisation types.	Lexical unit (collocate), Organisation type	logDice
Shared Collocates	To show collocates shared by two or more organisation types.	Lexical unit (collocate), Organisation type	logDice
Compound Concept Treemap	To provide a summary of the lexical compounds in which a search expression intervenes, arranged by semantic categorisation and frequency.	Lexical unit (compound), Category, Context, Year, Organisation type, Region, Document type	Frequency
Coordinated Concept Treemap	To provide a summary of the lexical units appearing in coordination with a search expression, arranged by frequency.	Lexical unit, Context, Year, Organisation type, Region, Document type	Frequency
Conceptual Evolution	To display the evolution of the conceptualisation of a notion by	Context, Hypernym, Context, Category, Year,	None

Histogram	arranging and categorising contexts with varying degrees of definitional precision.	Organisation Type, Region, Document Type	
Filterable and Searchable Table	To display a set of manually curated contexts.	Hypernym, Definitional element, Context, Region, Organisation type, Organisation subtype, Document type, Document ID	None

Table 5: Summary of visualisations

4. Conclusions and future work

In this paper we have shown how data visualisation can have a two-fold role in a corpus-driven project. It can assist linguists for the interpretation of corpus information in a field where they are not experts, but it can also be especially useful when serving as intermediary with field experts. Field experts, who are not familiar with corpus linguistics or raw lexical data, can benefit from interactive visualisations because they can freely interact with the data in a more intuitive fashion and build their own claims, complementing those offered by the team of linguists.

Most entries in the HE are expected to be written by external experts. Nonetheless, linguist-expert interaction is still limited to an in-house humanitarian at the HE. At the time of writing, only one LAR has been used to build a sample entry, which served to validate the LAR-building process. This will also provide external experts with a reference for guidance when writing their own entries. In addition, linguists are also interacting with another in-house expert who is in charge of compiling a list of concept-specific research questions. Sometimes, these questions may be answered by querying the corpus. This form of linguist-expert interaction provides linguists with concept-specific tasks and therefore contributes to shaping each LAR by adding particularised sections. As content production is expected to scale up, we will soon have more data on linguist-expert interaction, which will prompt a new line of research and provide us with a new way to improve our data visualisation skills.

In parallel, our efforts are currently centred on designing visualisations that represent collocational intersections between subcorpora more satisfactorily. For example, Venn diagrams with RStudio have the potential to replace our current packed bubble charts in future LARs. Additionally, we are working on a system to query the HE Corpus through Sketch Engine's API. At present, collocational data is only being extracted from the top five organisation types with the greatest number of occurrences of a given term. To create histograms, collocates are also disaggregated by year of publication. More meaningful comparisons between subcorpora could be drawn if collocational data were further disaggregated by every type of corpus metadata, i.e. increasing granularity. With our current manual approach, our top yearly collocate histogram for one concept requires a total of 90 queries through Sketch Engine's graphic user interface. Using Sketch Engine's API will not only remove manual querying tasks from our workflow, but it will also provide us with richer and more comprehensive datasets.

5. Acknowledgements

This research was carried out as part of project FFI2017-89127-P, Translation-Oriented Terminology Tools for Environmental Texts (TOTEM), funded by the Spanish Ministry of Economy and Competitiveness.

6. References

Allen, W. (2017). Making corpus data visible: Visualising text with research intermediaries. Corpora, 12(3), pp. 459–482. Available at: https://doi.org/10.3366/cor.2017.0128

Cabezas-García, M. & Chambó, S. (in press). Multi-Word Term Variation: Prepositional and Adjectival Complex Nominals in Spanish. SJAL (Spanish Journal of Applied Linguistics).

Christ, O., Schulze, B.M., Hofmann, A., & König, E. (1999). The IMS Corpus Workbench: Corpus Query Processor (CQP) - User's Manual. Institute for Natural Language Processing. Stuttgart: University of Stuttgart.

Guillaume D. (2019), Mapping lexical variation with Tableau software. Around the word. Accessed at: https://corling.hypotheses.org/2853 (25 March 2021)

Humanitarian Encyclopedia (2020). Available at: https://humanitarianencyclopedia.org/wp-content/uploads/2020/05/HE-brochure_finalMay2020.pdf

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. EURALEX. Lorient, France, pp 105–115.

León-Araúz, P. & San Martín, A. (2018) The EcoLexicon Semantic Sketch Grammar: from Knowledge Patterns to Word Sketches. In I. Kernerman & S. Krek (eds.) Proceedings of the LREC 2018 Workshop "Globalex 2018 – Lexicography & WordNets. Miyazaki, Japan, pp. 94–99. Available at: http://lexicon.ugr.es/pdf/Leon-Arauz2018.pdf

Linguistic Analysis Portal for the Humanitarian Encyclopedia Available at: https://sites.google.com/view/humanitarianencyclopedia

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

Siirtola, H., Räihä, K. J., Säily, T., & Nevalainen, T. (2010). Information visualization for corpus linguistics: Towards interactive tools. International Conference on Intelligent User Interfaces, Proceedings IUI, pp. 33–36. Available at: https://doi.org/10.1145/2002353.2002365

Towards the ELEXIS data model: defining a common vocabulary for lexicographic resources

Carole Tiberius[1], Simon Krek[2], Katrien Depuydt[1], Polona Gantar[3], Jelena Kallas[4], Iztok Kosem[2], Michael Rundell[5]

[1] Instituut voor de Nederlandse Taal, Leiden, The Netherlands
[2] Jožef Stefan Institute, Ljubljana, Slovenia
[3]Faculty of Arts, University of Ljubljana, Slovenia
[4]Institute of the Estonian Language, Tallinn, Estonia
[5]Lexical Computing, Brno, Czech Republic
E-mail: {carole.tiberius,katrien.depuydt}@ivdnt.org, {simon.krek,iztok.kosem}@ijs.si,
apolonija.gantar@ff.uni-lj.si, jelena.kallas@eki.ee, michael.rundell@gmail.com

Abstract

In this paper we describe ongoing work on the identification and definition of core lexicographic elements to be used in the ELEXIS data model. ELEXIS is a European infrastructure project fostering cooperation and information exchange among lexicographical research communities. One of the main goals of ELEXIS is to make existing lexicographic resources available on a significantly higher level than is currently the case. Therefore, a common data model is being developed which aims to: a) streamline the integration of lexicographic data into the infrastructure (using the ELEXIFIER tool), b) enable reliable linking of the data in the ELEXIS Dictionary Matrix, and c) provide a basic template for the creation of new lexicographic resources, such that they can automatically benefit from the tools and services provided by the ELEXIS infrastructure. Here we focus on the development of a common vocabulary and report on the results of an initial survey that was conducted to collect feedback from experts in lexicography.

Keywords: data model; common vocabulary; lexicographic resource; interoperability

1. Introduction

Reliable and accurate information on word meaning and usage is of crucial importance in today's information society. The most consolidated and refined knowledge on word meanings can traditionally be found in dictionaries – monolingual, bilingual or multilingual. In each and every European country, elaborate efforts are put into the development of lexicographic resources describing the language(s) of the community. Although confronted with similar problems relating to technologies for producing and making these resources available, cooperation on a larger European scale has long been limited. In addition, standardisation efforts have not been particularly successful within the field of lexicography before the digital age, an observation which was confirmed by the results from the ELEXIS[1] survey on lexicographic practices in Europe (Kallas et al., 2019). More specifically, the results from the survey show that:

- most lexicographic projects use structured data, but some projects are still working with a non-structured data and text format;
- proprietary XML and (customised) TEI are the most commonly used XML formats;
- use of existing standard vocabularies for encoding lexicographic data is not yet common practice at the ELEXIS lexicographic partner institutions. IsoCat, GOLD, and TEI were mentioned.

As a consequence, the lexicographic landscape in Europe is still rather heterogeneous. It is characterised by stand-alone lexicographic resources and there is a significant variation in the level of expertise and resources available to lexicographers across Europe. This situation forms a major obstacle to more ambitious, innovative, transnational, data driven approaches to dictionaries, both as tools and objects of research.

The ELEXIS project aims to overcome these obstacles by developing a sustainable infrastructure for lexicography. To allow all different kinds of dictionary data to be included in the infrastructure and ensure that it will be open to a wide range of lexicographers, common protocols have been developed and a common vocabulary is being defined, which is the topic of this paper. Before we turn to the ongoing work on the ELEXIS data model and more specifically the common vocabulary in section 3, we will first introduce the ELEXIS project in more detail in section 2. In section 4 we discuss the results of a pilot survey that was conducted to get feedback from lexicographic experts on the common vocabulary.

2. ELEXIS

ELEXIS (Krek et al., 2018, 2019; Pedersen et al., 2018; Woldrich et al., 2020) is a Horizon 2020 project dedicated to creating a sustainable infrastructure for lexicography. The main objectives of the infrastructure are to:

1. enable efficient access to high quality lexical data/semantic information in the digital age;
2. bridge the gap between more advanced and lesser-resourced scholarly communities working on lexicographic resources;
3. enable the use of new technology and data in industry in the digital single market.

[1] https://elex.is/

Within ELEXIS, strategies, tools and standards are under development for extracting, structuring and linking lexicographic resources to unlock their full potential for Linked Open Data, NLP and the Semantic Web, as well as in the context of digital humanities. In a virtuous cycle of cross-disciplinary exchange of knowledge and data, a higher level of language description and text processing will be achieved. By harmonising and integrating lexicographic data into the Linked Open Data cloud, ELEXIS will make this data available to AI and NLP for semantic processing of unstructured data, considerably enhancing applications such as machine translation, machine reading and intelligent digital assistance thanks to the ability to scale to wide coverage in multiple languages. This, in turn, will enable the development of improved tools for the production of structured proto-lexicographic data in an automated process, using machine learning, data mining and information extraction techniques, where the extracted data can be used as a starting point for further processing either in the traditional lexicographic process or through crowdsourcing platforms.

Lexicographic data is crucial for realising the ELEXIS infrastructure. Within ELEXIS, data comes from a number of different data providers, i.e.:

- Consortium partners
- Observer institutions
- Other open access resources containing lexicographic data available through, amongst others, CLARIN and DARIAH.

To date, 118 different datasets, e.g. general dictionaries, bilingual dictionaries, thesauri, specialised dictionaries (terminology, dialects), and lemma lists have been collected from 32 ELEXIS partner and observer institutions. A sample list of the datasets can be found in the ELEXIS Deliverable 6.3 Intermediate interoperability report.

Most of these datasets have been compiled within national and regional projects, and as noted they are typically encoded in their own custom data format, i.e. proprietary XML, (customised) TEI, HTML, JSON-LD or are stored in a relational database. A growing number also have API access. To be able to integrate these diverse datasets in the ELEXIS infrastructure a set of common protocols have been developed (McCrae et al., 2019) and different access routes are distinguished into the infrastructure. Data can be contributed either as TEI Lex-0 or Ontolex-Lemon, which are the two data formats supported by ELEXIS. It is also possible to deliver data as proprietary XML or in another format. Proprietary XML data can take advantage of the ELEXIFIER tool which converts custom XML or PDF into TEI Lex-0 (see Section 2.2). Those contributing data in another format can create an implementation of the REST interface according to the specifications provided by ELEXIS (ELEXIS Deliverable 2.2 Interoperable interface for Lemon and TEI resources; McCrae et al., 2019).

Having a set of common protocols ensures what Ide and Pustejovsky (2010) call syntactic interoperability, which “relies on specified data formats, communication protocols, and the like to ensure communication and data exchange. It means that the systems involved can process the exchanged information, but there is no guarantee that the interpretation is the same”. This means that an element labelled ‘example’ in dataset X is not necessarily the same as an element labelled ‘example’ in Y. If we want to be able to link, edit, enrich and publish data from various sources reliably (as envisaged in ELEXIS, see Figure 1), we also need semantic interoperability.

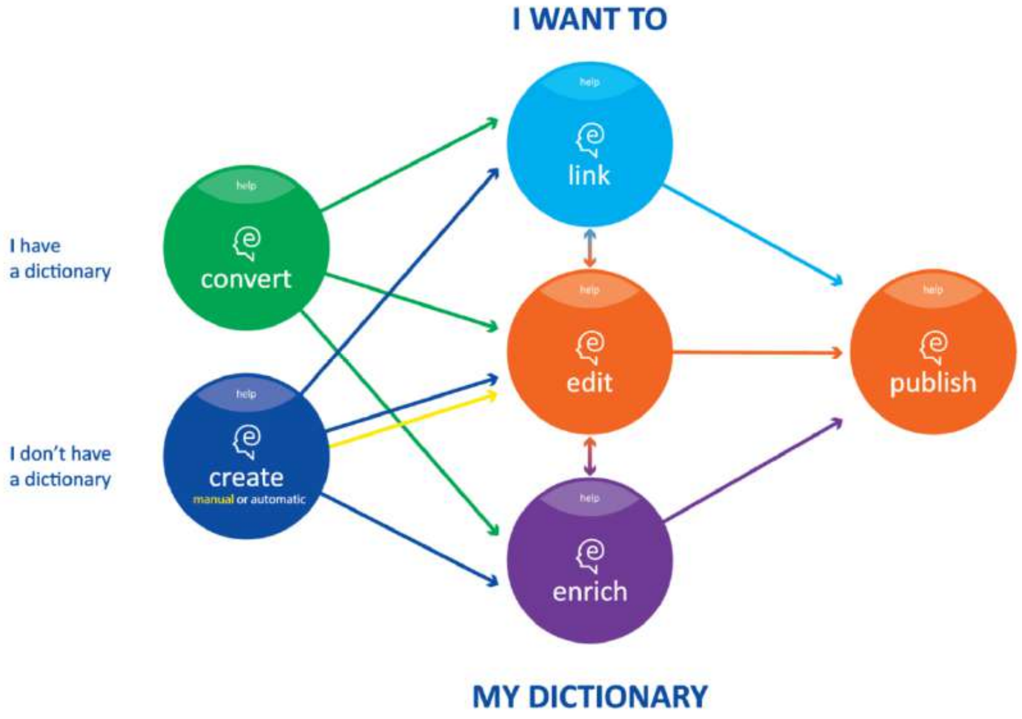


Figure 1: Graphic guide to the ELEXIS Dictionary Tools

According to Ide and Pustejovsky (2010) “semantic interoperability exists when two systems have the ability to automatically interpret exchanged information meaningfully and accurately in order to produce useful results via deference to a common information exchange reference model”. The first step towards such a model is the definition of a common vocabulary (see section 3), which is needed among others in the ELEXIFIER tool and the ELEXIS Dictionary Matrix.

2.1 ELEXIFIER

ELEXIFIER[2] (Repar et al., 2020) is a cloud-based dictionary conversion service for converting legacy dictionaries into a shared data format so that it can be integrated in the ELEXIS infrastructure. It can take lexicographic data in two distinct formats as input: (1) custom XML and (2) PDF. In the custom XML scenario, XPath formalisms

[2] https://elexifier.elex.is/

are used for identifying the core elements in the original dictionary data and transforming these to a TEI Lex-0 compliant format. All information contained in the original dictionary is kept, and only the core elements are transformed to the shared format. The supported elements are the same as those defined in the common vocabulary.

In the PDF scenario a more complex process is needed. The PDF is first transformed in a flat structure using a pdf2xml conversion script (based on https://github.com/kermitt2/pdf2xml). Then, a chunk of the resulting XML file is sent to Lexonomy[3] (Měchura 2017), an online dictionary editing tool for manual annotation. Approximately four pages need to be annotated. The annotated text is then used as the training material for machine learning algorithms that produce the entire dictionary converted to TEI Lex-0 compliant format. Dictionaries that have been transformed using ELEXIFIER, can be edited further in Lexonomy.

2.2 ELEXIS Dictionary Matrix

One of the main results of ELEXIS will be the ELEXIS Dictionary Matrix: a universal repository of linked senses, meaning descriptions, collocations, phraseology, translation equivalents, examples of usage and other types of lexical information found in existing lexicographic resources, monolingual, multilingual, modern, historical etc., available through a RESTful web service developed as part of LEX1 infrastructure. LEX1 is the part of the ELEXIS infrastructure which consists of a set of services and tools dedicated to the automatic segmentation, structuring, alignment and conversion of lexicographic resources to a uniform data format. The existence of common data models and standards that are produced bottom-up from within the lexicographic community fostered by ELEXIS is a necessary condition for successful development of this segment of the infrastructure.

The ELEXIS Dictionary Matrix will be also available as part of the Linguistic Linked Open Data cloud (LLOD), and it will serve as the source for providing links to (particular headwords, senses, etc. in dictionaries available online, through the European Dictionary Portal[4], and included in the matrix.

3. ELEXIS Data Model

To support the development of the Dictionary Matrix, a common data model is being developed which aims to a) streamline the integration of lexicographic data into the infrastructure (using the ELEXIFIER tool, see section 2.1) b) enable reliable linking of the data in the Dictionary Matrix (see section 2.2.), and c) provide a basic template

[3] https://www.lexonomy.eu/

[4] http://www.dictionaryportal.eu/

for the creation of new lexicographic resources, allowing for a smooth integration of new content into the matrix.

The aim of ELEXIS is not to develop a fully-fledged data model. Neither does the project aim to replace existing models. The main goal is to ensure semantic interoperability between lexicographic resources predominantly using their own custom format, focusing on a set of core elements which are necessary for the development of the Dictionary Matrix.

As a first step towards the development of the ELEXIS data model, efforts have been taken to establish a common vocabulary where the main concepts are unambiguously defined.

3.1 ELEXIS Common Vocabulary

As a starting point, a detailed analysis of sample data (provided by ELEXIS lexicographic partners and observer institutions) was carried out resulting in the following core elements: entry, headword, secondary headword, variant headword, part of speech, sense, sense structure, definition, sense indicator, label, example, translation, cross reference, note and inflected form. Table 1 gives an overview of the elements identified and their definitions. The overall strategy was to keep definitions as simple and as unambiguous as possible.

Element	**Definition**
entry	Part of a lexicographic resource which contains information related to at least one headword.
headword	Organising element of an entry in a lexicographic resource. *Note: In printed dictionaries typically at the top of an entry.*
secondary headword	Headword-like lexical item occurring within an entry in a lexicographic resource, for example derived forms, feminine forms, multiword expressions. Often an organising element of a part of an entry.
variant headword	Lexical item representing one of the alternative forms of the headword, for example a spelling or regional variation.
part of speech	Any of the word classes to which a lexical item may be assigned, e.g. noun, verb, adjective, etc.
sense	Part of an entry which groups together information relating to a meaning of a headword (or secondary headword), for example definitions, examples, and translations.

sense structure	Division and ordering of the senses in an entry.
definition	Statement that describes a meaning and permits its differentiation from other meanings within a sense structure of an entry.
sense indicator	Short statement that gives an indication of a meaning and permits its differentiation from other meanings within a sense structure of an entry.
label	Item from a controlled vocabulary indicating some kind of restriction on the use of the lexical item, for example, time, region, domain, register.
example	Instance of a lexical item's usage in a specific sense.
translation	Equivalent in another language of any element in an entry.
cross reference	Element providing any kind of link or reference to another element within or outside the lexicographic resource.
note	Free text remark that can accompany any element in a lexicographic resource.
inflected form	Form of the inflectional paradigm of the headword.

Table 1. ELEXIS core elements

In addition to the core elements, the following terms have been defined as they are used in the definitions of the core elements or they are potentially relevant in the context of ELEXIS:

Term	**Definition**
lexicographic resource	Needs to be defined; see section 4.1.
lexical item	Any word, abbreviation, partial word, or phrase which is described or mentioned in an entry in a lexicographic resource.
word class	A category of words grouped together based on form, meaning or syntactic characteristics.
meaning	The unique semantic, grammatical and/or pragmatic contribution that a headword in a particular sense makes to the overall understanding of an utterance.
controlled vocabulary	Fixed list of items which are used to reduce ambiguity and ensure consistency.
multiword expression	Sequence of lexical items that has properties that may not be predictable

	from the properties of the individual lexical items or their normal mode of combination. For example, collocations, phrasemes, compounds, idiomatic expressions, lexical combinations, and so forth. A multiword expression can have the status of headword or secondary headword in the lexicographic resource.
source language	The language of a lexical item (that is to be translated in another language). [cf. ISO1951:2007]
target language	The language into which a lexical item is to be translated. [cf. ISO1951:2007]

Table 2: Terms used in the definitions of the ELEXIS core elements

The next steps are to refine and finalise the definitions for these core elements and to express the ELEXIS data model in a formalism like UML. This way the serialisations to the two ELEXIS interoperability formats, i.e. Ontolex-Lemon and TEI Lex-0 can be realised.

Work on the ELEXIS data model is done in collaboration with the Lexicographic Infrastructure Data Model and API (LEXIDMA) Technical Committee within OASIS[5].

3.2 Related work

The ELEXIS data model does not stand on its own. In the past decade, several institutions and organisations have started harmonising the internal workflow trying to arrive at a uniform data model to be used for all lexicographic projects within the institution (e.g. Kernerman 2011, Depuydt et al. 2019; Parvizi et al., 2016; Tavast et al., 2018). Other larger initiatives which are particularly relevant to ELEXIS are TEI Lex-0 with a special focus on retrodigitised dictionaries, Ontolex-Lemon, the *de facto* standard for representing lexical information as RDF, and LMF (Lexical Markup Framework) which is being developed by the ISO Technical Committee (TC) 37 titled ‘Language and terminology’.

The ISO 24613 LMF multipart standard is based upon the definition of an implementation-independent metamodel combining a core model with extensions. As such it provides mechanisms that allow the development and integration of a variety of electronic lexical resource types and its scope is therefore much broader than that of the ELEXIS model.

The TEI Lex-0 (Tasovac et al., 2018) initiative aims at establishing a baseline encoding and a target format to facilitate the interoperability of heterogeneously encoded lexical

[5] https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=lexidma

resources. As specified in the TEI Lex-0 rationale[6], TEI Lex-0 should be primarily seen as a format which implements a set of constraints on top of those provided by the TEI Guidelines so that existing TEI dictionaries, once univocally transformed, can be queried, visualised, or mined in a uniform way. Furthermore, TEI Lex-0 aims to stay as aligned as possible with the TEI subset developed in conjunction with the revision of the ISO LMF standard (cf. Romary, 2015), ensuring future interoperability and sustainability.

Ontolex-Lemon (Cimiano et al., 2016) was originally developed to act as a model for the representation of lexical information in ontologies and is now the *de facto* standard for representing lexical information as RDF. It is also widely used to present data from lexicographic resources as Linked Data on the web. However, a mapping of traditional dictionary content to Ontolex-Lemon was not feasible without the development of an additional model, to be able to represent aspects of dictionaries like order and hierarchy of senses, or the fact that there is not always a 1:1 match between a dictionary entry and an ontolex:LexicalEntry (which requires it to have only one part of speech). The Lexicog module[7] is aimed to deal with these issues.

Both TEI Lex-0 and Ontolex-Lemon are supported within ELEXIS and serialisations will be provided from and to both TEI Lex-0 and Ontolex-Lemon. In addition, a tei2ontolex[8] conversion stylesheet has been developed.

4. Survey on the ELEXIS core elements and their definitions

A pilot survey was set up in order to collect feedback from experts in lexicography on the ongoing work on the common vocabulary. The survey was conducted in the autumn of 2020. It was sent to the lexicographic experts on the ELEXIS international advisory board and to the lexicographic partners in the project.

As it was a pilot survey, the goal was primarily qualitative rather than quantitative. Therefore, none of the questions in the survey was made obligatory and additional comments could be given for almost all questions. The survey was implemented in the 1ka survey system[9] which has been used for several other surveys within ELEXIS.

Only the following core elements were included in the pilot – entry, headword, secondary headword, sense, sense structure, definition, translation and example. For each of these a separate section was created in the survey where the relevant definitions were given together with a few extracts from existing dictionaries (see Figures 2-12).

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

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

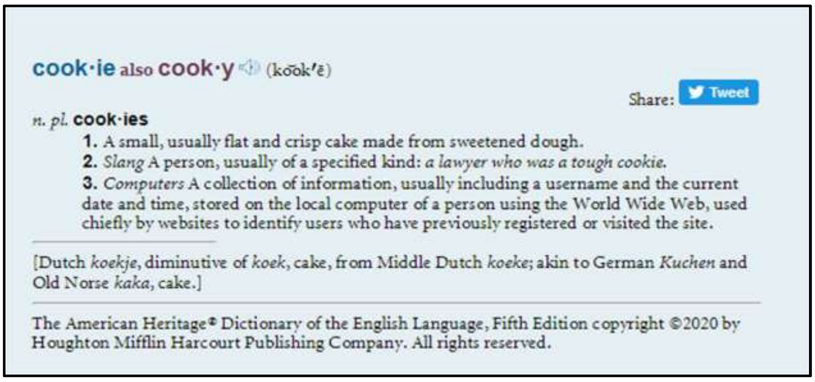
[8] https://github.com/elexis-eu/tei2ontolex

[9] https://www.1ka.si/

The lexicographic experts were then asked to answer questions about the element in relation to these extracts and the definitions provided. In order to get a wide range of examples, extracts were taken from various monolingual, bilingual, general-purpose, and also specialised dictionaries. Average completion time of the survey was 15 minutes, and we received 10 valid responses. Although this is undoubtedly a small number of responses, the results clearly show what the bottlenecks are when trying to define and identify core elements in lexicography. In the remainder of this section we discuss the results from this initial survey.

4.1 Entry

For ‘entry’, three extracts from three completely different dictionaries (traditional, born-digital, and specialised) were given: one from the American Heritage Dictionary[10] (see Figure 2), one from dictionary.com[11] (see Figure 3), and one from The Right Rhymes[12] (see Figure 4), a dictionary of hip-hop language.



cook·ie also **cook·y** (ko͝ok′ē)

n. pl. **cook·ies**

1. A small, usually flat and crisp cake made from sweetened dough.

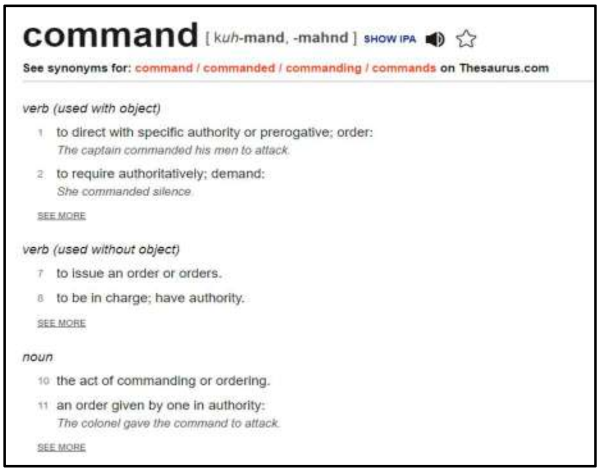
2. *Slang* A person, usually of a specified kind: *a lawyer who was a tough cookie.*

3. *Computers* A collection of information, usually including a username and the current date and time, stored on the local computer of a person using the World Wide Web, used chiefly by websites to identify users who have previously registered or visited the site.

[Dutch *koekje*, diminutive of *koek*, cake, from Middle Dutch *koeke*; akin to German *Kuchen* and Old Norse *kaka*, cake.]

The American Heritage® Dictionary of the English Language, Fifth Edition copyright ©2020 by Houghton Mifflin Harcourt Publishing Company. All rights reserved.

Figure 2: Extract from the American Heritage Dictionary



command [*kuh*-mand, -mahnd]

See synonyms for: command / commanded / commanding / commands on Thesaurus.com

verb (used with object)

1 to direct with specific authority or prerogative; order:
The captain commanded his men to attack.

2 to require authoritatively; demand:
She commanded silence.

verb (used without object)

7 to issue an order or orders.

8 to be in charge; have authority.

noun

10 the act of commanding or ordering.

11 an order given by one in authority:
The colonel gave the command to attack.

Figure 3: Extract from dictionary.com

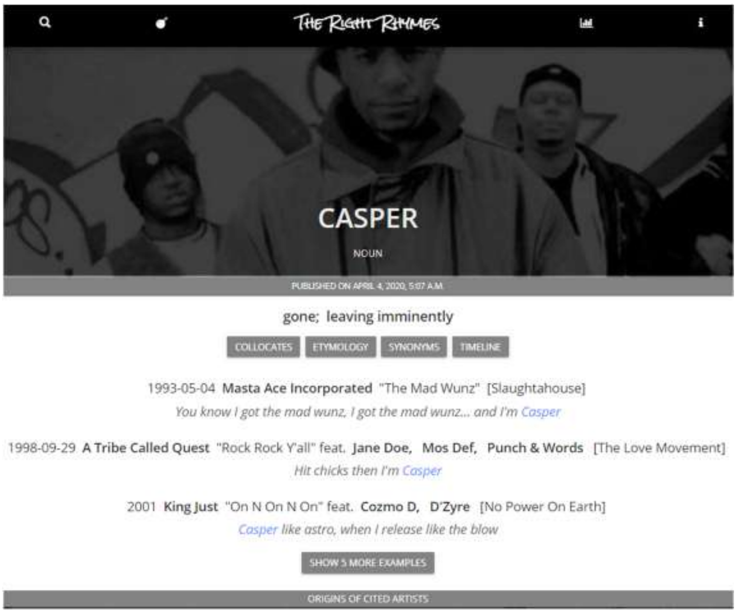
[10] https://www.ahdictionary.com/word/search.html?q=cookie

[11] https://www.dictionary.com/browse/command

[12] https://therightrhymes.com/casper

All experts considered the extract from the American Heritage dictionary in Figure 2 as an ‘entry’ according to the definition provided.

In relation to the extract from dictionary.com, one respondent noted that this should be considered as two entries, one for the verb and one for the noun. Indeed, one of the macrostructural decisions lexicographers need to make relates to what is considered as a homograph and how to treat them.[13]



CASPER

NOUN

PUBLISHED ON APRIL 4, 2020, 5:07 A.M.

gone; leaving imminently

1993-05-04 **Masta Ace Incorporated** "The Mad Wunz" [Slaughtahouse]
You know I got the mad wunz, I got the mad wunz... and I'm Casper

1998-09-29 **A Tribe Called Quest** "Rock Rock Y'all" feat. **Jane Doe, Mos Def, Punch & Words** [The Love Movement]
Hit chicks then I'm Casper

2001 **King Just** "On N On N On" feat. **Cozmo D, D'Zyre** [No Power On Earth]
Casper like astro, when I release like the blow

Figure 4: Extract from The Right Rhymes

There was more disagreement on the extract from The Right Rhymes. One expert felt that it did not fulfil the definition of ‘entry’ because it does not seem to be part of a lexicographic resource and only contains headword and part of speech information. Another respondent also found it difficult to consider this an entry. This shows that there are different views on what counts as a lexicographic resource,[14] and this term also needs to be defined. Some lexicographers/linguists may not consider a dictionary such as The Right Rhymes a lexicographic resource.

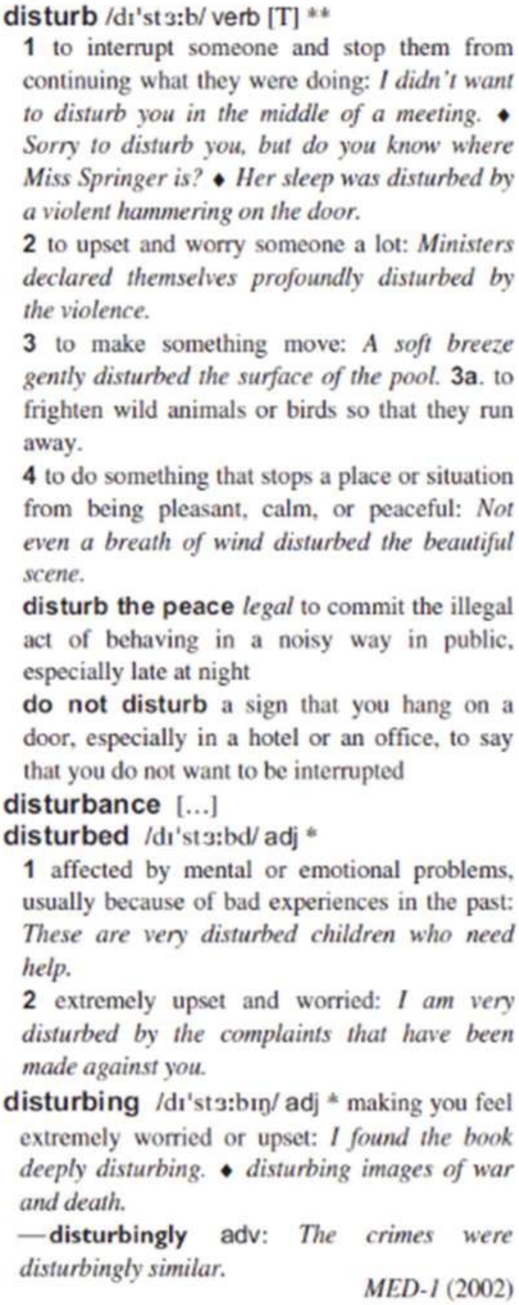
4.2 Headword and secondary headword

The questions on ‘headword’ and ‘secondary headword’ were combined. Again, three extracts from different dictionaries were given: the verb entry for *disturb* from the Macmillan English Dictionary (2002) (see Figure 5), the noun entry for *Katze* ‘cat’ from

[13] See e.g. Atkins and Rundell (2008: 192-193) for criteria that are used in lexicography in relation to homographs to decide whether there should be one entry or more and the discussion in Svensén (2009: 94-102) on the establishment of lemmas.

[14] A lexicographic resource was not yet defined at the time of the survey and thus not included.

the DWDS dictionary[15] (see Figure 6), and the entry for *ohulaada* ‘make smth firm’ from the Webonary Lynyole dictionary[16] (see Figure 7). The experts were asked to indicate whether they considered various lexical items from these extracts as ‘headword’, ‘secondary headword’ or something else.



disturb /dɪˈstɜːb/ verb [T] **
1 to interrupt someone and stop them from continuing what they were doing: *I didn't want to disturb you in the middle of a meeting.* ♦ *Sorry to disturb you, but do you know where Miss Springer is?* ♦ *Her sleep was disturbed by a violent hammering on the door.*
2 to upset and worry someone a lot: *Ministers declared themselves profoundly disturbed by the violence.*
3 to make something move: *A soft breeze gently disturbed the surface of the pool.* **3a**. to frighten wild animals or birds so that they run away.
4 to do something that stops a place or situation from being pleasant, calm, or peaceful: *Not even a breath of wind disturbed the beautiful scene.*
disturb the peace *legal* to commit the illegal act of behaving in a noisy way in public, especially late at night
do not disturb a sign that you hang on a door, especially in a hotel or an office, to say that you do not want to be interrupted
disturbance [...]
disturbed /dɪˈstɜːbd/ adj *
1 affected by mental or emotional problems, usually because of bad experiences in the past: *These are very disturbed children who need help.*
2 extremely upset and worried: *I am very disturbed by the complaints that have been made against you.*
disturbing /dɪˈstɜːbɪŋ/ adj * making you feel extremely worried or upset: *I found the book deeply disturbing.* ♦ *disturbing images of war and death.*
—**disturbingly** adv: *The crimes were disturbingly similar.*

MED-1 (2002)

Figure 4: Extract from the Macmillan English Dictionary taken from Atkins and Rundell (2008: 36). The experts were asked whether *disturb, disturb the peace, do not disturb, disturbance, disturbed, disturbing* and *disturbingly* are a ‘headword’, ‘secondary headword’ or something else.

[15] https://www.dwds.de/

[16] https://www.webonary.org/lunyole/

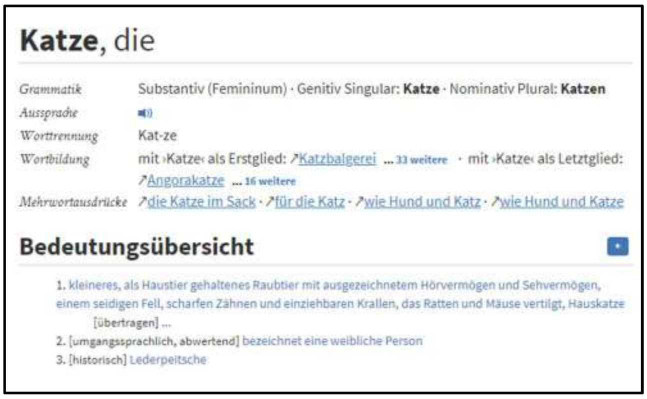
For the extract from the Macmillan English Dictionary (see Table 3) there was complete agreement on *disturb* being a 'headword', but the opinions on the status of *disturb the peace*, *do not disturb* varied significantly. Approximately half of the experts considered these as a 'secondary headword' whereas the other half considered them as something else.

	headword	secondary headword	something else[17]
disturb	10		
disturb the peace		5	4
do not disturb		4	6
disturbed	9	1	
disturbing	9	1	
disturbingly		9	1

Table 3. Experts' decisions on 'headword'/ 'secondary headword'/ something else

When the option 'something else' was chosen, terms such as phrase, collocation, idiom and derivative forms were given to describe the item. It was also mentioned that structurally these items can be considered as '(secondary) headwords' as in the tagging structure they represent discrete blocks, but that conceptually they should be tagged for what they are, e.g. an idiom block, a phrasal verb block or a run-on. It was also pointed out that this type of structural choice (that has been done for search-engine-friendly reasons) divorces the phrase or idiom from its context, from the environment of its source "word".

In the entry for *Katze* (see Figure 5) the results for the hyperlinked items *Katzbalgerei* und *wie Hund und Katze* were mixed. The reason that was given several times for calling these something else was that they look like cross-references to other entries and that the user thus has to go to another page to view them.



Katze, die

Grammatik	Substantiv (Femininum) · Genitiv Singular: **Katze** · Nominativ Plural: **Katzen**
Aussprache	🔊
Worttrennung	Kat-ze
Wortbildung	mit ›Katze‹ als Erstglied: ↗Katzbalgerei ... 33 weitere · mit ›Katze‹ als Letztglied: ↗Angorakatze ... 16 weitere
Mehrwortausdrücke	↗die Katze im Sack · ↗für die Katz · ↗wie Hund und Katz · ↗wie Hund und Katze

Bedeutungsübersicht

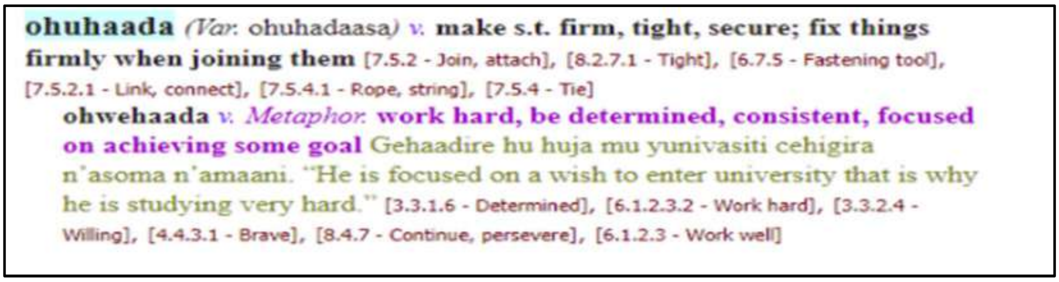
1. kleineres, als Haustier gehaltenes Raubtier mit ausgezeichnetem Hörvermögen und Sehvermögen, einem seidigen Fell, scharfen Zähnen und einziehbaren Krallen, das Ratten und Mäuse vertilgt, Hauskatze
[übertragen] ...
2. [umgangssprachlich, abwertend] bezeichnet eine weibliche Person
3. [historisch] Lederpeitsche

Figure 5: Entry for *Katze* 'cat' in the DWDS dictionary[18]. The experts were asked whether *Katzen* 'cats', *Katzbalgerei* 'scuffle', *wie Hund und Katze* 'like dog and cat' are 'headword', 'secondary headword' or something else.

[17] As it was not made obligatory to check a box for each item, the numbers do not add up.

[18] https://www.dwds.de/wb/Katze

Experts did agree on the third extract containing the entry for *ohuhaada* from the Webonary Lynyola dictionary (see Figure 6), considering *ohuhadaasa* as a 'headword' and *ohwehadaa* as a 'secondary headword'. To the question as to whether there were any other items that could be considered as a 'secondary headword' in this entry, one respondent mentioned *ohuhadaasa* (the form in between brackets given after the 'headword').



ohuhaada *(Var:* ohuhadaasa*)* *v.* **make s.t. firm, tight, secure; fix things firmly when joining them** [7.5.2 - Join, attach], [8.2.7.1 - Tight], [6.7.5 - Fastening tool], [7.5.2.1 - Link, connect], [7.5.4.1 - Rope, string], [7.5.4 - Tie]

ohwehaada *v. Metaphor:* **work hard, be determined, consistent, focused on achieving some goal** Gehaadire hu huja mu yunivasiti cehigira n'asoma n'amaani. "He is focused on a wish to enter university that is why he is studying very hard." [3.3.1.6 - Determined], [6.1.2.3.2 - Work hard], [3.3.2.4 - Willing], [4.4.3.1 - Brave], [8.4.7 - Continue, persevere], [6.1.2.3 - Work well]

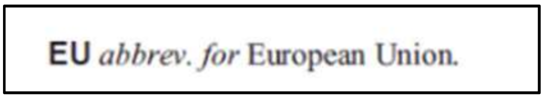
Figure 6: Entry for *ohuhaada* 'make smth firm' in the Lynyole dictionary[19]. The experts were asked whether *ohuhaada* and *ohwehaad*a are 'headword', 'secondary headword' or something else.

These results show that the definition of 'secondary headword' may need to be refined or at least further explained if we want to get a consistent transformation for this element in the ELEXIFIER tool across different datasets.

4.3 Part of Speech

As noted by Svensén (2009: 136), "there is considerable variation between languages, lexicographic traditions and user categories as concerns the occurrence, format and function of part-of-speech indications". This can also be observed in the survey results where experts noted that it is a tricky question as to whether something like *transitive verb* should be considered as a 'part of speech' or as two separate labels. Most respondents noted that strictly speaking *verb* is the 'part of speech' and *transitive* additional information. However, it was also noted that if it is the style of the dictionary to conflate two concepts in a single element, then it is a 'part of speech'. Similar observations were made in relation to *proper noun.*

With part of speech there are clear cases, but there are also some problematic cases, as is illustrated by the extract in Figure 7.



EU *abbrev. for* European Union.

Figure 7: Entry for *EU* in the Collins English Dictionary (2000) (Atkins and Rundell, 2008: 196)

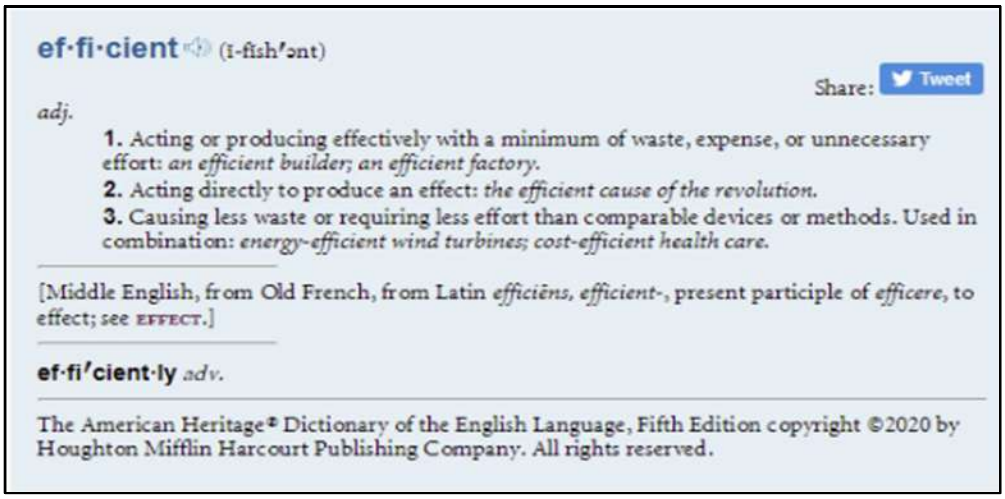
[19] https://www.webonary.org/lunyole?s=ohuhaada

Only three experts considered *abbrev.* as a ‘part of speech’, whereas seven marked it as something else. The reason for this is clearly summarised by one expert:

> “If you want to split hairs, abbreviations, acronyms, etc. aren’t really a separate word class; the underlying part of speech is whatever the thing they’re an abbreviation for is. But in terms of listing this information in the header information of the dictionary, you’ll find that most dictionaries put this kind of indicator inside POS tags.”

4.4 Sense and sense structure

To learn more about the perception of ‘sense’ and ‘sense structure’, we took an extract from the American Heritage Dictionary illustrating the entry for *efficient*[20] (Figure 8).



ef·fi·cient (ĭ-fĭsh′ənt)

adj.

1. Acting or producing effectively with a minimum of waste, expense, or unnecessary effort: *an efficient builder; an efficient factory.*
2. Acting directly to produce an effect: *the efficient cause of the revolution.*
3. Causing less waste or requiring less effort than comparable devices or methods. Used in combination: *energy-efficient wind turbines; cost-efficient health care.*

[Middle English, from Old French, from Latin *efficiēns, efficient-*, present participle of *efficere*, to effect; see EFFECT.]

ef·fi′cient·ly *adv.*

The American Heritage® Dictionary of the English Language, Fifth Edition copyright ©2020 by Houghton Mifflin Harcourt Publishing Company. All rights reserved.

Figure 8: Entry for *efficient* in the American Heritage Dictionary

There was full agreement that the numbers 1., 2. and 3. represent the ‘sense structure’. There was, however, quite some disagreement on what actually constitutes a ‘sense’, as shown in Table 4.

	Sense	**Something else**
2. Acting directly to produce an effect: *the efficient cause of the revolution.*	5	4
2. Acting directly to produce an effect	6	3
Acting directly to produce an effect	6	4
Acting directly to produce an effect: *the efficient cause of the revolution.*	3	5

Table 4: Experts’ decisions on whether the options provided are a ‘sense’ or something else

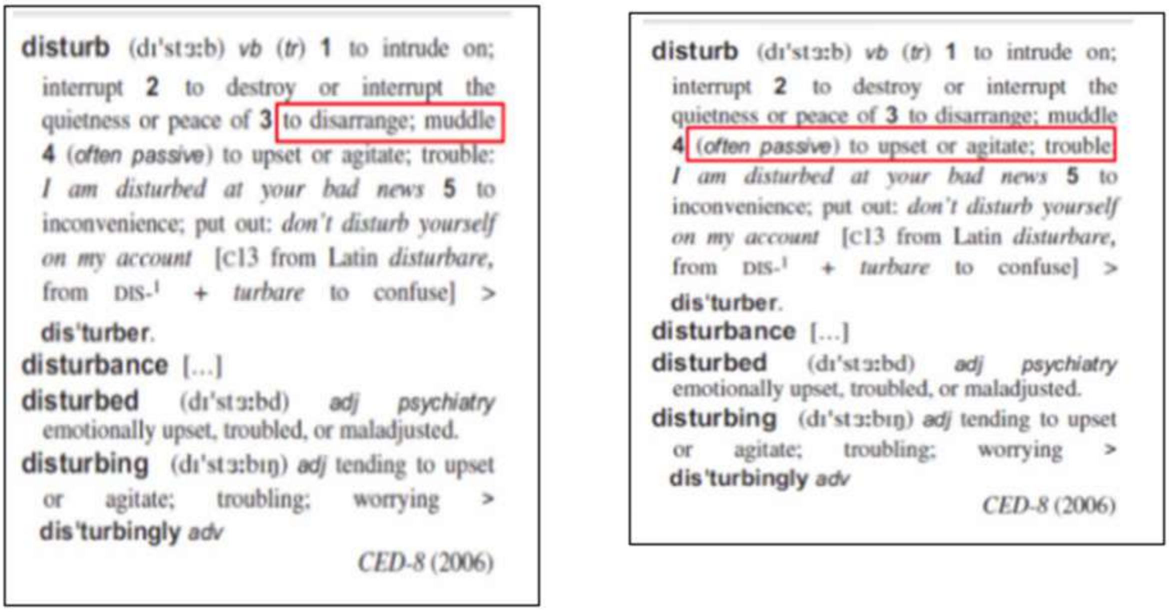
[20] https://www.ahdictionary.com/word/search.html?q=efficient

Four possible variants were provided and there was actually none that all the experts agreed on. Some considered the inclusion of the example necessary for it to be a ‘sense’ (which is in line with the definition provided), others mentioned the presence of a sense number (unless numbering is automatic) and for some, ‘sense’ itself is the definition. The latter was motivated by stating that structurally, explanatory examples are part of the sense and tend to be included in the sense block in a tagging structure. They can illustrate the sense, but they are not truly the sense.

These answers suggest that there is an interplay between how elements are commonly marked in dictionary structures and how lexicographers think about them conceptually.

4.5 Definition

In relation to the ‘definition’ element, we were particularly interested to find out whether information which is sometimes included in brackets is considered as part of the definition or not. Two extracts, both from Atkins and Rundell (2008) were taken, one from the Collins English Dictionary (see Figure 9) one from the Oxford Advanced Learner’s Dictionary (see Figure 10).



disturb (dɪ'stɜːb) *vb* (*tr*) **1** to intrude on; interrupt **2** to destroy or interrupt the quietness or peace of **3** to disarrange; muddle **4** (*often passive*) to upset or agitate; trouble: *I am disturbed at your bad news* **5** to inconvenience; put out: *don't disturb yourself on my account* [c13 from Latin *disturbare*, from DIS-^{1} + *turbare* to confuse] > **dis'turber**.
disturbance [...]
disturbed (dɪ'stɜːbd) *adj psychiatry* emotionally upset, troubled, or maladjusted.
disturbing (dɪ'stɜːbɪŋ) *adj* tending to upset or agitate; troubling; worrying > **dis'turbingly** *adv*

CED-8 (2006)

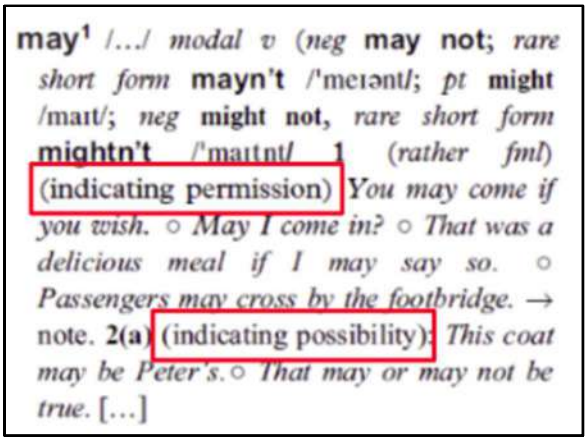
disturb (dɪ'stɜːb) *vb* (*tr*) **1** to intrude on; interrupt **2** to destroy or interrupt the quietness or peace of **3** to disarrange; muddle **4** (*often passive*) to upset or agitate; trouble: *I am disturbed at your bad news* **5** to inconvenience; put out: *don't disturb yourself on my account* [c13 from Latin *disturbare*, from DIS-^{1} + *turbare* to confuse] > **dis'turber**.
disturbance [...]
disturbed (dɪ'stɜːbd) *adj psychiatry* emotionally upset, troubled, or maladjusted.
disturbing (dɪ'stɜːbɪŋ) *adj* tending to upset or agitate; troubling; worrying > **dis'turbingly** *adv*

CED-8 (2006)

Figure 9: Entry for *disturb* in the Collins English Dictionary (2006) (Atkins and Rundell 2008: 36)

The text in the marked red box on the left hand side was considered a ‘definition’ by all lexicographic experts, the text in the marked red box on the right hand side by three only, while the others indicated that the information in brackets is grammatical or usage information.

We also included an extract containing a function word or what Atkins and Rundell (2008:196-198) call a grammatical word entry, as these entries often describe the function rather than the meaning.



may[1] /.../ *modal v* (*neg* **may not;** *rare short form* **mayn't** /ˈmeɪənt/; *pt* **might** /maɪt/; *neg* **might not,** *rare short form* **mightn't** /ˈmaɪtnt/ **1** (*rather fml*) (indicating permission) *You may come if you wish.* ○ *May I come in?* ○ *That was a delicious meal if I may say so.* ○ *Passengers may cross by the footbridge.* → note. **2(a)** (indicating possibility): *This coat may be Peter's.* ○ *That may or may not be true.* [...]

Figure 10: Part of the entry for *may* in The Oxford Advanced Learner's Dictionary (1995) from Atkins and Rundell (2008: 197).

Seven experts would call the parts marked by the red box a 'definition', but three would not, as they considered these as semantic comments or comments on semantic implicatures.

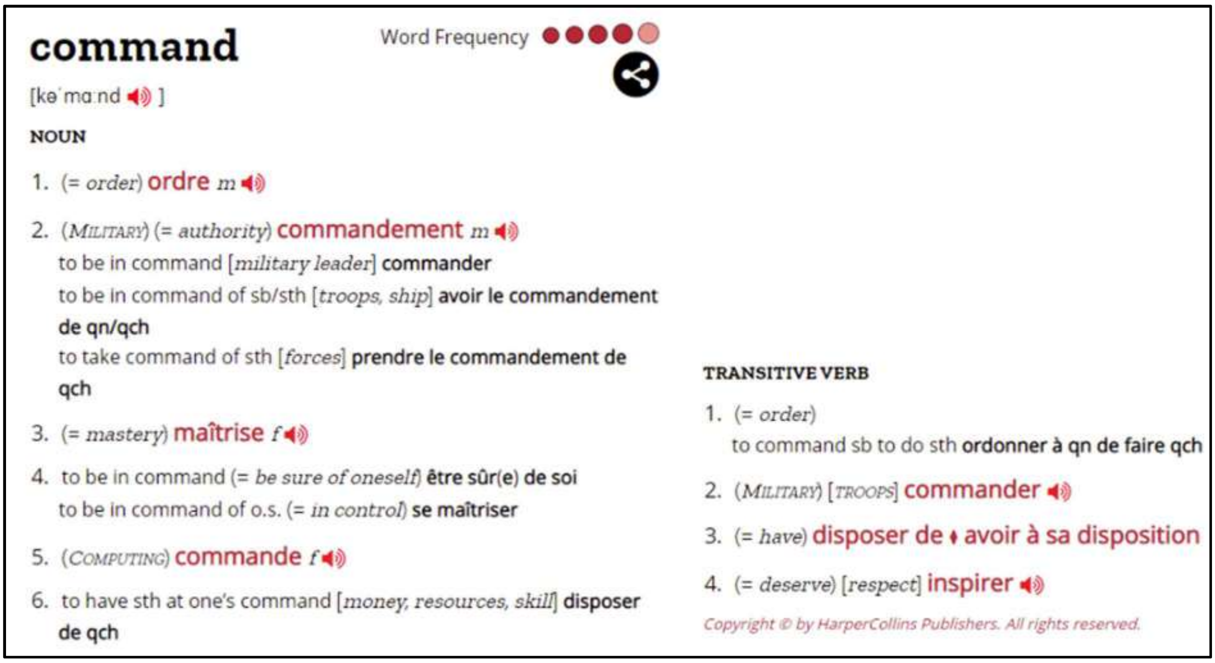
4.6 Translation and Example

For 'translation', an extract from the bilingual English-French Collins Dictionary[21] was selected (see Figure 11).

There was complete agreement among the experts. All considered the three items that were offered *ordre, être sûr(e) de soi*, and *disposer de, avoir à sa disposition* as 'translation'. One noted that the last one actually contains two translations.

For the 'example' element, one extract from a modern dictionary (the Collins Dictionary English-French) and one extract from a historical dictionary (Petit Larousse Illustré) were selected.

[21] https://www.collinsdictionary.com/dictionary/english-french/command



command

Word Frequency ●●●●●

[kəˈmɑːnd]

NOUN

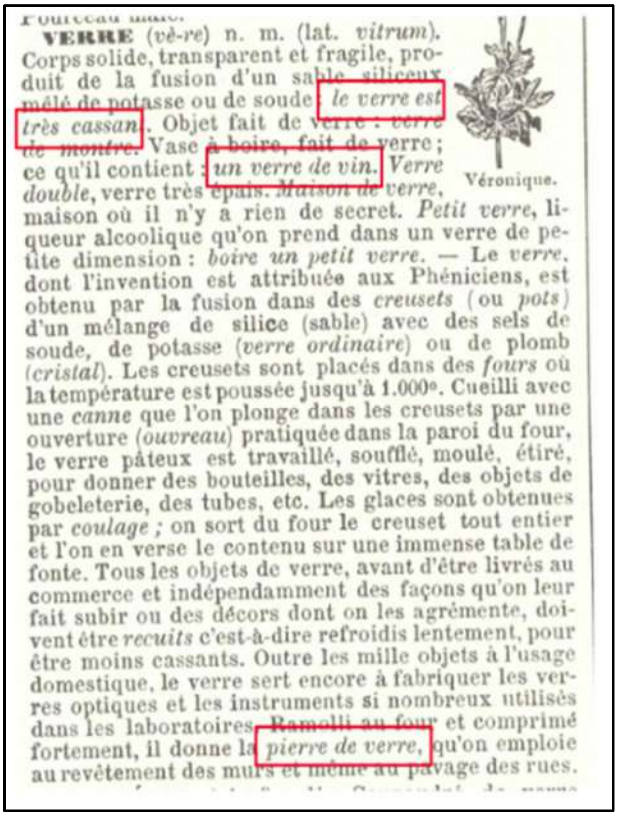
1. (= *order*) ordre *m*
2. (*Military*) (= *authority*) commandement *m*
   to be in command [*military leader*] **commander**
   to be in command of sb/sth [*troops, ship*] **avoir le commandement de qn/qch**
   to take command of sth [*forces*] **prendre le commandement de qch**
3. (= *mastery*) maîtrise *f*
4. to be in command (= *be sure of oneself*) **être sûr(e) de soi**
   to be in command of o.s. (= *in control*) **se maîtriser**
5. (*Computing*) commande *f*
6. to have sth at one's command [*money, resources, skill*] **disposer de qch**

TRANSITIVE VERB

1. (= *order*)
   to command sb to do sth **ordonner à qn de faire qch**
2. (*Military*) [*troops*] commander
3. (= *have*) disposer de ♦ avoir à sa disposition
4. (= *deserve*) [*respect*] inspirer

Copyright © by HarperCollins Publishers. All rights reserved.

Figure 11: Entry for *command* in the English-French Collins Dictionary



Fourreau mâle.

VERRE (*vè-re*) n. m. (lat. *vitrum*). Corps solide, transparent et fragile, produit de la fusion d'un sable siliceux mêlé de potasse ou de soude : *le verre est très cassant*. Objet fait de verre : *verre de montre*. Vase à boire, fait de verre ; ce qu'il contient : *un verre de vin*. *Verre double*, verre très épais. *Maison de verre*, maison où il n'y a rien de secret. *Petit verre*, liqueur alcoolique qu'on prend dans un verre de petite dimension : *boire un petit verre*. — Le *verre*, dont l'invention est attribuée aux Phéniciens, est obtenu par la fusion dans des *creusets* (ou *pots*) d'un mélange de silice (sable) avec des sels de soude, de potasse (*verre ordinaire*) ou de plomb (*cristal*). Les creusets sont placés dans des *fours* où la température est poussée jusqu'à 1.000°. Cueilli avec une *canne* que l'on plonge dans les creusets par une ouverture (*ouvreau*) pratiquée dans la paroi du four, le verre pâteux est travaillé, soufflé, moulé, étiré, pour donner des bouteilles, des vitres, des objets de gobeleterie, des tubes, etc. Les glaces sont obtenues par *coulage* ; on sort du four le creuset tout entier et l'on en verse le contenu sur une immense table de fonte. Tous les objets de verre, avant d'être livrés au commerce et indépendamment des façons qu'on leur fait subir ou des décors dont on les agrémente, doivent être *recuits* c'est-à-dire refroidis lentement, pour être moins cassants. Outre les mille objets à l'usage domestique, le verre sert encore à fabriquer les verres optiques et les instruments si nombreux utilisés dans les laboratoires. Ramolli au four et comprimé fortement, il donne la *pierre de verre*, qu'on emploie au revêtement des murs et même au pavage des rues.

Figure 12: The extract from Petit Larousse Illustré 1905

The answers to the question with the extract from the modern dictionary did not reveal anything unexpected. For the historical dictionary there was a little uncertainty on whether the last item marked by a red box in Figure 12 was an 'example' or something else.

Only seven experts gave an answer for *pierre de verre* and only three of those considered this an 'example'. The "reluctance" to answer may also suggest that some simply did not know what to answer.

The pilot survey clearly showed certain bottlenecks and as such provided useful feedback on the common vocabulary. The elements 'secondary headword', 'part of speech', and 'sense' in particular need further work. The survey also emphasised the importance of supporting the common vocabulary with concrete examples. In the near future, we will extend the survey to all elements from the ELEXIS common vocabulary and to a larger audience.

5. Summary and further work

In this paper we described ongoing work on the ELEXIS data model. We focussed on the description of the common vocabulary and discussed the results of a pilot survey that was conducted among lexicographic experts. In the near future, the pilot survey will be extended to all elements from the common vocabulary and a larger audience so that we get a more complete insight into the understanding of the core elements in the lexicographic community. This will undoubtedly lead to revisions and refinements in the work on the data model.

In the next phase, it will also be necessary to express the ELEXIS data model in a formalism like UML, in order to realise the serialisation to the two ELEXIS interoperability formats, i.e. Ontolex-Lemon and TEI Lex-0. When the model is finished, a full mapping will also be provided with the related models (TEI Lex-0, Ontolex-Lemon and LMF).

The work on the ELEXIS data model and the common vocabulary is ongoing, and a lot remains to be done, but we hope that it will inspire a constructive debate on standardisation in the lexicographic community and related fields.

6. Acknowledgements

The research received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

7. References

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

Bosque-Gil, J & Gracia, J. (eds.) (2019). *The OntoLex Lemon Lexicography Module Final Community Group Report 17 September 2019.* Accessed at: https://www.w3.org/2019/09/lexicog/. (9 April 2021)

Cimiano, P, McCrae, J.P. & Buitelaar, P. (2016) *Lexicon Model for Ontologies: Community Report, 10 May 2016 Specification.* Accessed at: https://www.w3.org/2016/05/ontolex/. (9 April 2021).

Depuydt, K., Schoonheim, T. & de Does, J. (2019) Towards a More Efficient Workflow for the Lexical Description of the Dutch Language. Accessed at: http://videolectures.net/elexisconference2019_depuydt_dutch_language/. (9 April 2021)

Ide, N. & Pustejovsky, J. (2010). What does interoperability mean, anyway? Toward an operational definition of interoperability. *Proceedings of the Second International Conference on Global Interoperability for Language Resources (ICGL 2010).* Hong Kong, Available at: http://www.cs.vassar.edu/~ide/papers/ICGL10.pdf.

ISO 1951:2007 *Presentation/representation of entries in dictionaries – Requirements, recommendations and information.*

ISO/CD 24613-1:2018(E) *Language resource management — Lexical markup framework (LMF) — Part 1: Core model.*

ISO/CD 24613-2:2019(E) *Language resource management — Lexical markup framework (LMF) — Part 2: Machine Readable Dictionary (MRD) model.*

ISO/WD 24613-3:2020(E) *Language resource management — Lexical Markup Framework (LMF) — Part 3: Etymological Extension.*

ISO/WD 24613-4:2020 *Language resource management — Lexical Markup Framework (LMF) — Part 4: TEI serialisation.*

ISO NP 24613-5:2018 *Language resource management — Lexical markup framework (LMF) — Part 5: Lexical base exchange (LBX) serialization.*

Kallas, J., Koeva, S., Langemets, M., Tiberius, C. & Kosem, I. (2019). Lexicographic practices in Europe: Results of the ELEX survey on user needs. In I. Kosem et al. (eds.) *Electronic Lexicography in the 21st Century: Smart Lexicography. Proceedings of the eLex 2019 Conference, 1–3 October 2019, Sintra, Portugal.* Available at: https://elex.link/elex2019/wp-content/uploads/2019/09/eLex_2019_30.pdf.

Kernerman, I. (2011). From Dictionary to Database: Creating a Global Multi-Language Series. In I. Kosem & K. Kosem (eds.) *Electronic Lexicography in the 21st Century. Proceedings of the eLex 2011 Conference, 11-12 November 2011, Bled Slovenia.* Available at: https://elex2011.trojina.si/Vsebine/proceedings/eLex2011-14.pdf.

Kosem, I, Navigli, R., McCrae, J. P. & Jakubíček, M. (2021). Intermediate interoperability report. ELEXIS Deliverable 6.3. Available at: https://elex.is/wp-content/uploads/2021/02/ELEXIS_D6_3_Intermediate_interoperability_report.pdf.

Krek, S., McCrae, J. P., Kosem, I., Wissik, T., Tiberius, C., Navigli, R., & Pedersen, B. (2018). European Lexicographic Infrastructure (ELEXIS). In J. Čibej et al.

(eds.) *Proceedings of the XVIII EURALEX International Congress on Lexicography in Global Contexts, Ljubljana, Slovenia, 17-21 July 2018.* Available at http://doi.org/10.5281/zenodo.2599902.

Krek, S., Declerck, T., McCrae, J.P. & Wissik, T. (2019). *Towards a Global Lexicographic Infrastructure.* Presented at the Language Technology 4 All Conference. Available at http://doi.org/10.5281/zenodo.3607274

McCrae, J.P. (2020). Interoperable interface for Lemon and TEI resources. ELEXIS Deliverable 2.2. Available at: https://elex.is/wp-content/uploads/2020/02/ELEXIS_D2_2_Interoperable_Interface_for_Lemon_and_TEI_resources.pdf.

McCrae, J.P., Tiberius, C., Khan, A.F., Kernerman, I., Declerck, T., Krek, S., Monachini, M. & Ahmadi, S. (2019). The ELEXIS interface for interoperable resources. In I. Kosem et al. (eds.) *Electronic Lexicography in the 21st Century: Smart Lexicography. Proceedings of the eLex 2019 Conference, 1–3 October 2019, Sintra, Portugal. Brno: Lexical Computing CZ, s.r.o.* Available at: https://elex.link/elex2019/wp-content/uploads/2019/09/eLex_2019_37.pdf.

Měchura, M. B. (2017). Introducing Lexonomy: an open-source dictionary writing and publishing system. In I. Kosem et al. (eds.) *Electronic Lexicography in the 21st Century: Lexicography from Scratch. Proceedings of the eLex 2017 conference, 19-21 September 2017, Leiden, The Netherlands.* Available at: https://www.lexonomy.eu/docs/elex2017.pdf.

Parvizi, A., Kohl, M.,Gonzàlez, M. & Saurí, R. (2016). Towards a Linguistic Ontology with an Emphasis on Reasoning and Knowledge Reuse. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Portorož, Slovenia. Available at: https://www.aclweb.org/anthology/L16-1071/.

Pedersen, B. S., McCrae, J. P., Tiberius, C. & Krek, S. (2018). ELEXIS - a European infrastructure fostering cooperation and information exchange among lexicographical research communities. In F. Bond, T. Kuribayashi, C. Fellbaum & P. Vossen (eds.) *Proceedings of the 9th Global WordNet Conference (GWC 2018), Global Wordnet Association, Singapore.* Available at: http://doi.org/10.5281/zenodo.2599954.

Repar, A. & Krek, S. (2020). Tools for the automatic segmentation and identification of lexicographic content. ELEXIS Deliverable 1.3. Available at: https://elex.is/wp-content/uploads/2020/02/ELEXIS_D1_3_Tools_for_the_automatic_segmentation_and_identification_of_lexicographic_content.pdf.

Romary, L. (2015). TEI and LMF crosswalks. *JLCL - Journal for Language Technology and Computational Linguistics*, 30 (1).

Svensén, Bo (2009). *A handbook of lexicography. The theory and practice of dictionary-making.* Cambridge: Cambridge University Press.

Tasovac, T, Romary, L., Banski, P., Bowers, J., de Does, J., Depuydt, K., Erjavec, T., Geyken, A., Herold, A., Hildenbrandt, V., Khemakhem, M., Petrović, S., Salgado,

A. & Witt, A.. (2018). *TEI Lex-0: A baseline encoding for lexicographic data.* Version 0.8.6. DARIAH Working Group on Lexical Resources. Available at https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html.

Tavast, A., Langemets, M., Kallas, J. & Koppel, K. (2018). Unified Data Modelling for Presenting Lexical Data: The Case of EKILEX. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (eds.) *Proceedings of the XVIII EURALEX International Congress: EURALEX: Lexicography in Global Contexts, Ljubljana, 17-21 July 2018.* Ljubljana University Press, Faculty of Arts, pp. 749—761. Available at: https://euralex.org/publications/unified-data-modelling-for-presenting-lexical-data-the-case-of-ekilex/.

Dictionary titles used in the survey:

The American Heritage Dictionary of the English Language. Fifth Edition. Accessed at: https://www.ahdictionary.com. (9 April 2021)

Collins English Dictionary (2000) Fifth Edition, HarperCollins Publishers, Glasgow, UK

Collins English Dictionary (2006) Eight Edition, HarperCollins Publishers, Glasgow, UK

Collins Dictionary English-French. HarperCollins Publishers. Accessed at: https://www.collinsdictionary.com (9 April 2021)

Dictionary.com. Accessed at: https://www.dictionary.com/. (9 April 2021)

DWDS *Digitales Wörterbuch der Deutschen Sprache.* Accessed at: https://www.dwds.de. (9 April 2021)

Lynyole dictionary. Accessed at: https://www.webonary.org. (9 April 2021)

MacMillan English Dictionary for Advanced Learners (2002) First Edition, MacMillan

Oxford Advanced Learner's Dictionary (1995) Fifth Edition, Oxford University Press, Oxford, UK

Oxford-Hachette French Dictionary (1994) First Edition, Oxford University Press, Oxford, UK

Petit Larousse Illustré 1905

The Right Rhymes Dictionary. Accessed at: https://therightrhymes.com. (9 April 2021)

A Word Embedding Approach to Onomasiological Search in Multilingual Loanword Lexicography

Peter Meyer[1], Ngoc Duyen Tanja Tu[1]

[1] Leibniz-Institut für Deutsche Sprache, R5, 6-13 68161 Mannheim Germany
E-mail: meyer@ids-mannheim.de, tu@ids-mannheim.de

Abstract

In this paper we present an experimental semantic search function, based on word embeddings, for an integrated online information system on German lexical borrowings into other languages, the *Lehnwortportal Deutsch* (LWPD). The LWPD synthesizes an increasing number of lexicographical resources and provides basic cross-resource search options. Onomasiological access to the lexical units of the portal is a highly desirable feature for many research questions, such as the likelihood of borrowing lexical units with a given meaning (Haspelmath & Tadmor, 2009; Zeller, 2015). The search technology is based on multilingual pre-trained word embeddings, and individual word senses in the portal are associated with word vectors. Users may select one or more among a very large number of search terms, and the database returns lexical items with word sense vectors similar to these terms. We give a preliminary assessment of the feasibility, usability and efficacy of our approach, in particular in comparison to search options based on semantic domains or fields.

Keywords: onomasiological search; word embeddings; multilingual lexicography; lexical borrowings

1. Introduction

The Lehnwortportal Deutsch (LWPD) is an online platform developed at the Leibniz-Institut für Deutsche Sprache and comprising lexicographical resources on German loanwords in other languages. The LWPD in its entirety realises the concept of a ‘reverse loan dictionary’ that does not focus on the target languages of the borrowing processes, but on the source language. Besides offering a traditional, lemma-based access to the individual dictionaries, the system provides sophisticated portal-wide cross-resource options to search for lexical units (German etyma, corresponding loanwords, variants and derivatives thereof, etc.).

At present, however, onomasiological access is restricted to simple substring-based searches on the word sense definitions for words as provided in the individual dictionaries. Consequently, a genuine semantic search in the LWPD would be more suitable for research questions like “Which languages have a conspicuously high proportion of German loanwords in certain thematic areas, such as *food* and *drinks*?”

In a project funded by the Fritz Thyssen Stiftung the LWPD is currently being substantially revised on both the backend and the user interface levels (Meyer &

Eppinger, 2019). The new edition will go online in early 2022, featuring a number of newly added resources on German borrowings in English, Dutch, French, Portuguese, Hungarian, Czech and Slovak. The new system will offer a much more powerful and simplified way to search the underlying graph database (Meyer, 2014), which represents the portal data as a network of partially cross-resource relationships between lexical units, through an innovative 'query builder' interface (Meyer, 2019). The semantic search function discussed in this paper will be an integral part of the query builder.

Conceptually, the approach presented below differs from hand-crafted semantic domain taxonomies that are used as search features in similar projects (e.g. van der Sijs, 2015; Osservatorio degli Italianismi nel Mondo) and come with many well-known problems:

(a) Semantic domain definitions are inherently vague and cannot be exhaustive, i.e. there is not a (perfectly) suitable domain for every word sense. This usually leads to senses without domain assignment or, equivalently, to the introduction of a semantically unspecified default 'miscellaneous' domain. Assignment of a word sense to multiple domains is frequently possible due to overlap, but is usually not wanted and must be avoided by arbitrary assignment decisions. If domain schemas are explicitly designed for multiple assignments, then this considerably complicates both the manual annotation process and the burden on the part of the user who has to experiment with combinations of (typically rather broad) domains.

(b) An introspection-based manual annotation procedure will inevitably lead to a complex lexicographical practice of domain assignments, especially if maximal inter-annotator agreement is demanded. This actually requires a considerable amount of *reverse engineering* of that (typically opaque) practice on the part of the user, and will prove difficult for word senses that do not fit easily into one of the domains, implying the annotator assigns them according to subjective intuition or some internal conventions.

(c) It is challenging to find a reasonable middle ground between ease of use and sufficient granularity. If the taxonomy is too coarse, the user might get too many search results, which makes the search inefficient. If, on the other hand, the taxonomy is too fine-grained, the number of categories to choose from becomes impractical and confusing, in particular for casual use.

(d) The domain taxonomy is essentially static. If certain domains turn out to yield unsatisfactory (e.g. counterintuitive) results, there is nothing the user can do apart from trying to get further relevant search results by randomly trying other domains. For lexicographers, any revision of the 'boundary' of a domain may turn out to be a time-consuming process as it involves a possibly large number of reassignments.

Our experimental approach, presented in section 2, is an attempt at addressing the problems mentioned above. Section 3 discusses the problem of evaluating this approach with regard to its usability and performance as well as the quality of the search results. In section 4, we briefly summarise the pros and cons of our approach in comparison to domain-based searches.

2. Approach

2.1 Basic idea

In the revised LWPD, lexical items (etyma, loans, derivatives, and so on … figuring in the included dictionaries) can be searched for using any number of search criteria in arbitrary Boolean combinations. Basically, the new semantic search function will allow the user to describe the desired 'range' of meanings by entering words that are, in an intuitive sense, similar in meaning or topic. The user actually selects words from a very large given list of frequently used German words (henceforth: 'search keys') and takes advantage of autosuggest functionality during input. This speeds up typing and gives instantaneous feedback on the availability of search keys. Multiple search keys can be combined with each other to describe different aspects of a semantic 'field'. The query returns words with at least one word sense sufficiently close in semantics to the meanings of all search keys provided.

The list of search keys is meant to be of roughly the same order of magnitude as the active vocabulary of a native German speaker. So far, we have experimented with the 10,000 most frequent verbs, nouns and adjectives from DeReWo. DeReWo is a word frequency list based on DeReKo, the world's largest collection of German-language corpora. Note that the list of search keys available to the user can be altered, even radically, at any time, as will become clear in what follows.

2.2 Technical implementation

The technical implementation of our approach is based on word embeddings (Mikolov et al., 2013), a technique to represent the distributional properties of words in large corpora mathematically through vectors, i.e. lists of numbers. A simple measure, the cosine similarity of two vectors, is supposed to represent the semantic similarity of the respective words (Speer et al., 2018). Thus the semantic similarity between the search key and an LWPD word sense can be calculated by computing the cosine similarity between the vector representations of the two objects. The greater the cosine similarity, the more semantically similar the two words are. The maximum cosine similarity is 1.0, the minimum is -1.0. The semantic search function picks out word senses that have a sufficiently high cosine similarity (i.e., close to 1.0) to the search keys input by the user.

In our project, we use the ConceptNet (CN) NumberBatch pre-computed word embeddings (Speer et al., 2018; we use version 19.08) to map each LWPD lexical unit word sense and each search key to a vector. Note that we could not train custom word embeddings ourselves since we do not have access to the corpus data underlying many of the portal's lexicographical resources. The CN embeddings are trained on multilingual data as well as otherwise known semantic relationships between words. Vectors for all included words of the more than 70 languages present in CN are aligned in one vector space, i.e. similarities can be measured across languages – which is evidently a basic precondition for their use in an LWPD search. As we will see soon, the dataset of embeddings can easily be replaced at a later time, if other pre-computed embeddings turn out to yield better search results.

The basic parts of the database architecture for the semantic search are shown in Figure 1.

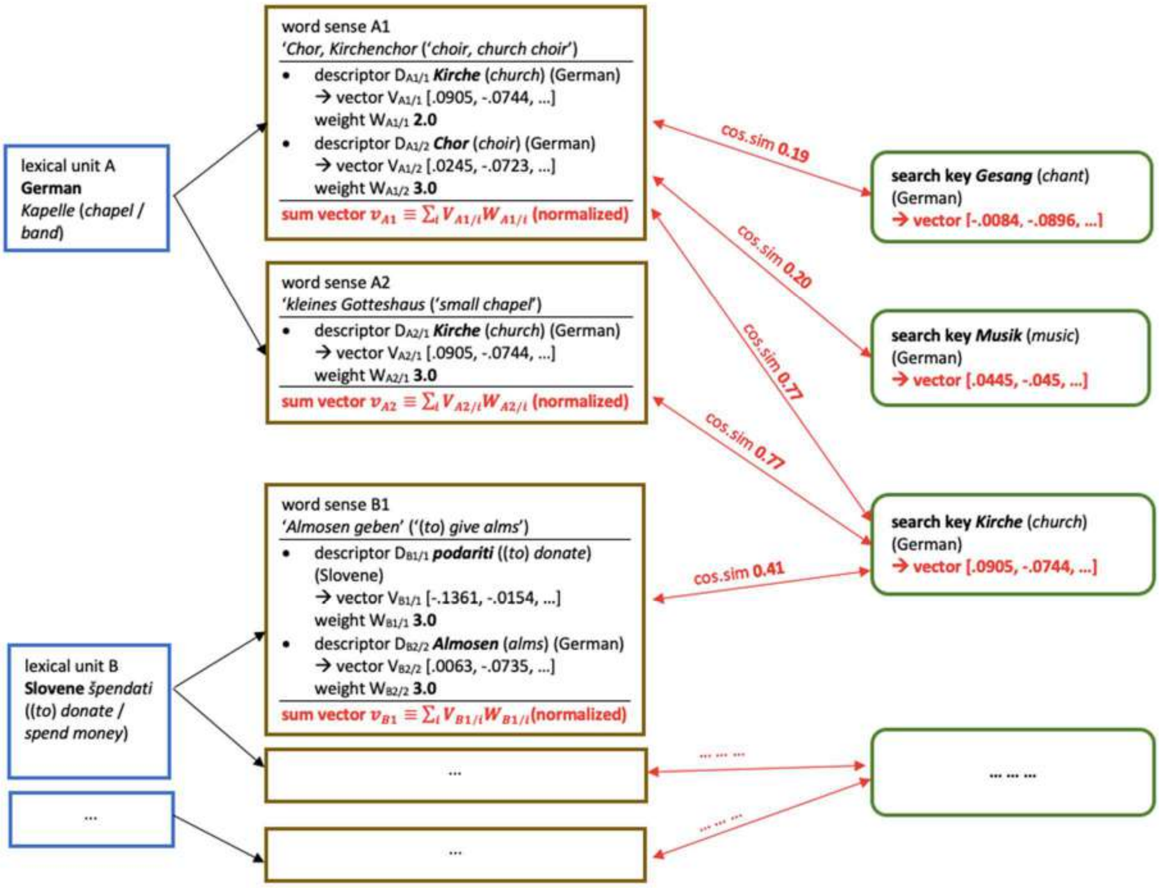

word sense A1
'Chor, Kirchenchor ('choir, church choir')
descriptor $D_{A1/1}$ **Kirche** (church) (German)
→ vector $V_{A1/1}$ [.0905, -.0744, ...]
weight $W_{A1/1}$ **2.0**
descriptor $D_{A1/2}$ **Chor** (choir) (German)
→ vector $V_{A1/2}$ [.0245, -.0723, ...]
weight $W_{A1/2}$ **3.0**
sum vector $v_{A1} \equiv \sum_i V_{A1/i} W_{A1/i}$ (normalized)
lexical unit A
German
Kapelle (chapel / band)
word sense A2
'kleines Gotteshaus ('small chapel')
descriptor $D_{A2/1}$ **Kirche** (church) (German)
→ vector $V_{A2/1}$ [.0905, -.0744, ...]
weight $W_{A2/1}$ **3.0**
sum vector $v_{A2} \equiv \sum_i V_{A2/i} W_{A2/i}$ (normalized)
cos.sim 0.19
cos.sim 0.20
cos.sim 0.77
cos.sim 0.77
cos.sim 0.41
search key Gesang (chant) (German)
→ vector [-.0084, -.0896, ...]
search key Musik (music) (German)
→ vector [.0445, -.045, ...]
search key Kirche (church) (German)
→ vector [.0905, -.0744, ...]
word sense B1
'Almosen geben' ('(to) give alms')
descriptor $D_{B1/1}$ **podariti** ((to) donate) (Slovene)
→ vector $V_{B1/1}$ [-.1361, -.0154, ...]
weight $W_{B1/1}$ **3.0**
descriptor $D_{B2/2}$ **Almosen** (alms) (German)
→ vector $V_{B2/2}$ [.0063, -.0735, ...]
weight $W_{B2/2}$ **3.0**
sum vector $v_{B1} \equiv \sum_i V_{B1/i} W_{B1/i}$(normalized)
lexical unit B
Slovene špendati ((to) donate / spend money)
...
... ...
... ...
... ...

Figure 1: Basic database architecture of assigning embeddings to word senses and search keys.

This architecture is now explained in more detail.

(1) In LWPD, all lexical units are represented as nodes (vertices) in a property graph database. A lexical unit may appear in multiple dictionaries/entries (not shown in Figure 1); this occurs frequently with German etyma.

(2) All word senses of a lexical unit as found in the resources are represented as

separate sense nodes in the graph. There can be considerable overlap between sense definitions if the lexical unit appears in multiple sources. No attempt at unifying these sense definitions is made in the LWPD.

(3) Using an in-house web application, student annotators assign to each word sense in the LWPD at least one word from CN, henceforth called a *descriptor* of the sense. The descriptors are supposed to have meanings that are closely related to the word sense in question. For these assignments, the full range of words covered by CN is available, with a vocabulary size of almost 600,000 items available for German alone. In most cases, a default descriptor is provided in advance; in the most elementary case this is simply the word the word sense is related to. For manual editing, the annotators have a number of tools and rules at their disposal, on which see below. Assigning multiple descriptors helps to overcome the notorious difficulties of word embeddings, in particular the fact that embeddings are not context-sensitive and do not differentiate in cases of polysemy and homonymy. For example, the etymon *Reif* appears in the present LWPD exclusively in the sense of 'hoop, bracelet'; just assigning the CN word *reif* to this sense would obscure the fact that there is a homonymous *Reif* meaning 'hoarfrost' and an adjective *reif* 'ripe, mature' – the latter since CN words are case-insensitive. So a second descriptor like German *ring* 'ring' can help to disambiguate. If multiple descriptors are used, they have to be *labelled* by the annotators according to their function. Labels are selected from a predefined list and include 'disambiguating word with similar meaning', 'hypernym', 'cohyponym' and others. For example, the CN words *bräme* ('trimming'), *verbrämung* ('trimming') and *pelzbesatz* ('fur trimming') might be assigned to the Polish word *bramik* ('fur trimming'). The latter CN word would get the label 'synonym', the first two CN words the label 'hypernym'.

(4) Each descriptor label is mapped onto a number representing the *weight* of the descriptor for the word sense it is assigned to. For example, hypernyms might get mapped to the integer 2 and synonyms to the number 2.5 (if a word sense has only one descriptor, weights play no role; formally, the weight of a solitary descriptor is always 1). This allows us to test (and change between) different mapping schemes in order to find the one that gives optimal results.

(5) The weighted and normalised sum of the vectors belonging to the CN descriptors yield the vector representation of the word sense. Thus, each word sense node in the LWPD graph has one such vector as a property.

(6) The search keys available to the users are selected as explained above, e.g. from a frequency list of lemmatised German words with relevant part of speech. They must be words in CN; but in practice this is not a serious restriction due to size of the CN data. Though it would seem natural not to restrict the available choices at all and use the entire German CN vocabulary, this would

result in a disturbing amount of noise presented to the user. Each search key is represented as a node in the graph which has its CN vector as a property.

(7) The cosine similarity between all word sense vectors and all search key vectors is computed; if it is above a certain threshold, an edge (i.e. a relation) between the word sense and the search key is stored in the graph and assigned the cosine similarity as a property. Consequently, no edge is stored between the word sense and the search key if their cosine similarity is only slightly above 0, which is why there is no edge between the search key *Gesang* ('chant') and the word sense B1 *Almosen geben* ('(to) give alms') in Figure 1**.** The threshold can be defined arbitrarily but should exclude very low similarities in order to reduce noise in the search results; ultimately it is a matter of practical experience.

The annotators follow a complex, tool-guided procedure for assigning descriptors and labels in a meaningful and consistent way. Note that the notion of inter-annotator agreement is ill-defined in this context since the number of plausible alternative assignments is, in general, simply too high. The following remarks give a brief sketch of a still evolving practice.

(a) Default assignments 1: If an LWPD word is contained in CN, the word itself is automatically assigned to all of its word senses as its descriptor. For example, the Slovene word *bager* ('excavator') is contained in CN, so the assigned CN word is *bager.* If the LWPD word has more than one word sense, all its senses are marked for later manual revision, which means they are prioritised for a manual check because it is very likely that further differentiation among the senses is necessary. To give an example, the Hebrew word *Zup* has the two senses 'Suppe' ('soup') and 'Abschmecken einer Flüssigkeit' ('seasoning a liquid'). The first sense could be covered by the German CN word *Suppe* ('soup') corresponding to the etymon of *Zup*, the second one by the CN word *abschmecken* ('(to) season').

(b) Default assignments 2: If an LWPD word *w* is not included in CN, but there is an LWPD word *w** with an etymological or variational relationship to it that is included, then this CN word is taken as the default descriptor for the word senses of *w* (see (a) above for an example). These assignments are marked for manual review later. Information on the relationship between words is available in the LWPD graph database. For example, the Slovene loanword *ravbati* ('(to) rob') is not included in CN, but its German etymon *rauben* ('(to) rob') is, so *rauben* becomes the default descriptor for the senses of *ravbati.*

(c) Flagging of highly polysemous CN descriptors: The in-house tool warns annotators of polysemous descriptors, suggesting the use of additional descriptors for disambiguation purposes. It is not a trivial task to automatise the detection of polysemy. Typical lexicographical resources such as Wiktionary or WordNet-type databases exhibit a level of sense differentiation that is too granular for our purposes. Among the strategies that we are trying out to detect problematic cases of polysemy in German CN words are the following: (i) GermaNet (Hamp & Feldweg, 1997; Henrich & Hinrichs, 2010) partitions its synsets into different 'semantic fields'. If the synsets containing a certain CN word are distributed among multiple semantic fields, then we assume significant polysemy. (ii) Consulting the lemmatisation of a reliable reference dictionary of German such as the DWDS, if the CN word corresponds to multiple headwords, we assume significant polysemy. The identification of significantly polysemous words from other languages is an open issue.

(d) Manual editing: Where default assignments are either not possible or introspectively misleading, appropriate descriptors have to be selected in a 'manual' fashion by searching for CN words that have a close semantic relationship to the LWPD word (e.g. hypernyms, synonyms, etc.), using resources such as OpenThesaurus, DWDS, and Wortschatz Universität Leipzig.

2.3 Performing queries

As explained above, semantic queries for words in the upcoming LWPD are specified by one or more search keys. An autocomplete function makes it easier to find and enter the search keys.

A typical user query may look like this: If you are interested in finding out whether German terms for certain types of dishes have been borrowed in the languages available in the LWPD's dictionary, you can use specific search keys to do so. In a domain-based semantic search, you would first have to make sure that a suitable domain exists. In our semantic search system, you could just use the search keys *Speise* ('dish') and *flüssig* ('liquid') if you want to get terms for liquid dishes present in the LWPD. As a search result you will obtain, among other things, *Suppe* ('soup') and *Mus* ('pulp'). If you are interested in sweet dishes, then you just have to enter *Speise* ('dish') and *süß* ('sweet') as search keys and you obtain among others *Nachtisch* ('dessert'), *Süßigkeit* ('candy') and *Zimtstern* ('star-shaped cinnamon cookie'). Thus, a user can search for very specific word fields without consulting any *a priori* taxonomy.

Technically, the semantic search is part of a traversal of the graph database. The database will search for word sense nodes whose cosine similarity to all of the search key nodes provided by the user is greater than a certain threshold. The search result list contains the LWPD words connected to these word sense nodes. The user may alter the threshold in the query to influence the size of the result set and obtain results that are more or less 'strict'.

A very similar approach has already been successfully used for search engine optimisation (Castro Fernandez et al., 2018; Kuzi et al, 2016; Fernandez et al., 2008) but not for semantic searches of lexicographic resources.

3. Evaluation

3.1 Usability and performance

The quality of a semantic search can be measured in terms of two properties: 1) Usability and 2) performance. (Elbedweihy et al., 2012)

(1) Usability: In our onomasiological search, search queries are entered using natural language search keys, so no query language needs to be learned. It also allows anyone to easily execute semantic search queries without having to read a manual beforehand. In addition to this, due to the autosuggesting input facility, the user does not have to invest much time in finding out which search keys are available at all and in formulating his search queries. In contrast, with a domain-based search, one must first become familiar with the taxonomy before starting a search. Furthermore, the searches are highly flexible. Thus, users can add or alter a search key if they want to filter the results of the previous search or found that the previous search was incomplete.

(2) Performance: The cosine similarities between the LWPD word senses and the search keys are all precomputed and stored in the graph database, if the cosine similarity is above a certain threshold. Since both the cosine similarities and the search keys stored in the graph database are indexed, a traversal from a search key to 'matching' LWPD words is possible in (approximately) constant time, and therefore very fast.

3.2 Quality of the search results

The quality of the search results of many semantic searches is evaluated by comparing the results of different search engines for the same query (e.g. Tümer et al., 2009; Uma Devi & Meera Gandhi, 2015). In our case, however, this is not possible because the data of lexicographical resources with a semantic search function differ from each other, which means that they are trivially providing different search results for the same query.

Moreover, the notion of recall of the search results is ill-defined in the case of the system presented here. The recall is calculated as the quotient of the relevant search results and that of *all* relevant items from the LWPD, i.e. those lexical units from the LWPD that *should* appear in the search results. However, the relevant search results would have to be determined by a human annotator, which has several disadvantages: (a) there are no fixed criteria for deciding whether a lexical item is 'really' a relevant search result, so subjective decisions are necessary; (b) an exhaustive search for relevant search results would be too time-consuming even for a small fraction of search keys.

The precision of the search results seems to be somewhat less problematic and could be tackled in a similar way as in Chauhan et al. (2013) and Mohamed and Shokry (2020). The precision is calculated as the quotient of the relevant search results and the number of all search results. Thus, it indicates the proportion of relevant search results in relation to all search results – it is not necessary to determine *all* possibly relevant items in the LWPD. In practice, however, it is still almost impossible to decide whether a result offered by the system should be considered relevant, e.g. if you select the search key *Speise* ('dish'), is *Koch* ('cook') relevant? What about *Service* ('(coffee) set')? Operationalising the evaluation of search result quality beyond taking samples from user studies is clearly an avenue for future research.

Unfortunately, a thorough evaluation of LWPD's onomasiological search will have to wait until at least a considerable subset of our data is available. We hope to complete the annotation of word senses for all German etyma by the end of 2021.

To get a first impression of the quality of the search results, we conducted a small study on the German etyma that are represented in the LWPD in its current incarnation, simulating possible search queries by looking for suitable words in the lexicographical sense definitions of these etyma. Of the 3,709 'meta-etyma' that serve as headwords in the Dictionary of German Etyma in the present database of the LWPD, 2,074 appear as CN words and also figure as lexical units in at least one GermaNet synset (we used GermaNet 14.0). For each such etymon E, we collected its word sense definitions as given in the LWPD dictionaries. All words in these definitions were POS-tagged and lemmatised with a standalone version[1] of the GATE DictLemmatizer plugin. For 1,668 etyma, at least one lemmatised word W was found that (i) belongs to the NN, ADJA or VV* POS-classes most relevant for searches and (ii) appears both in CN and in at least one GermaNet synset. For each such word W we determined the pair of one synset containing E and one synset containing W that has maximum semantic similarity $S_{E,W}$ according to the information-content-based measure by Lin (1998), assuming that the semantics of words W in a sense definition for a word E bears significant similarity to a word sense of E. The resulting 4,676 pairs turned out to be, in hindsight, a surprisingly noise-free collection of pairs of clearly semantically related terms such that the words W appearing in the definitions for the respective E did indeed very often appear to be good candidates for search keys relevant to E.

We then calculated, for each E-W pair, the CN-based cosine similarity between E and W and compared it to the $S_{E,W}$ measure introduced above. The results are shown in Figure 2. The more similar a word W in the definition of an etymon E is according to GermaNet, the higher, on average, is the cosine similarity between these two words. For highly GermaNet-related words, the average cosine similarity goes up to a

[1] The software is available at http://staffwww.dcs.shef.ac.uk/people/A.Aker/activityNLPProjects.html .

remarkable 0.65. It must be emphasised that these numbers constitute at best anecdotal evidence of the power of our approach to semantic search, but given the fundamentally different ways in which Lin's measure on GermaNet synsets and cosine similarity of word embeddings treat semantic similarity, they nevertheless indicate a basic and non-trivial consistency of search result quality with our theoretical expectations.

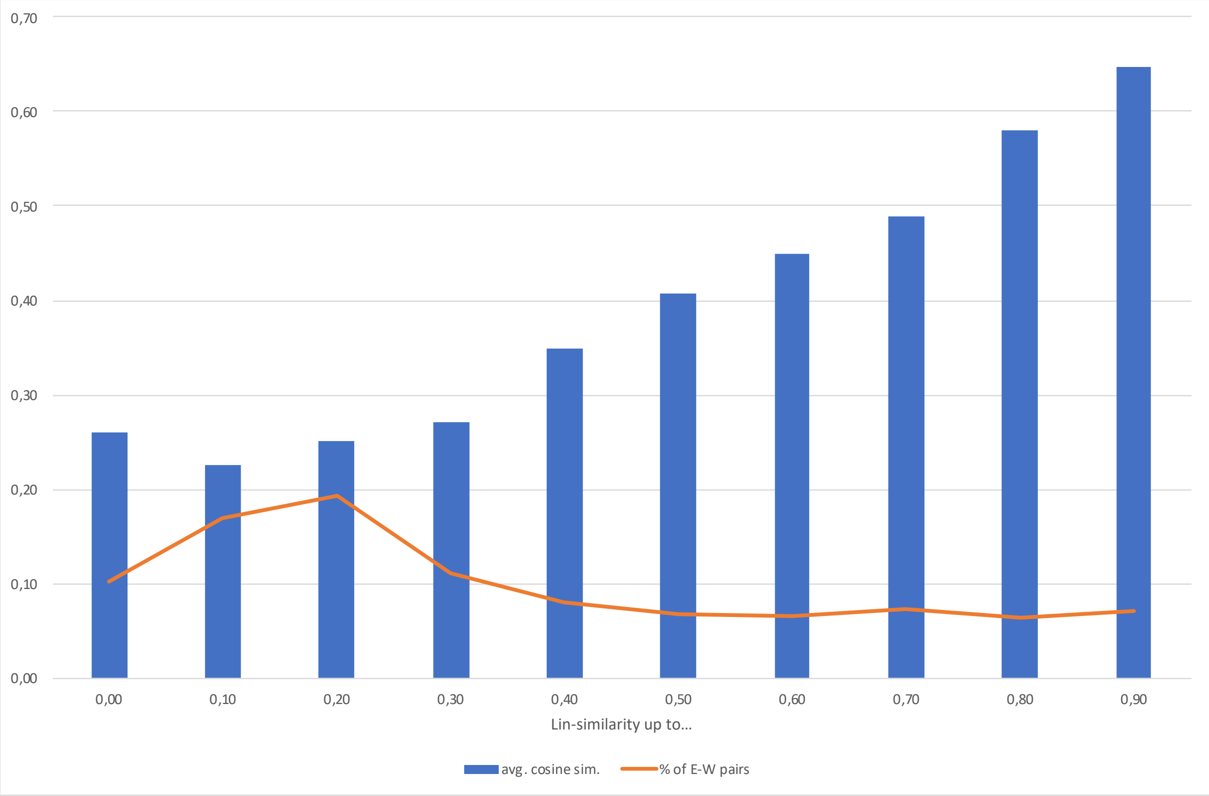


Figure 2: Average cosine similarity (blue bars) between German etyma E in the LWPD and words W in their definitions as a function of the Lin-measure based similarity of the corresponding maximally semantically similar GermaNet synsets (x-axis). The leftmost bar represents a maximum Lin-similarity between 0.0 and 0.1, and so on. The orange line indicates the percentage of E-W pairs falling in the respective class; so for example the eighth column reads "6.5% of all E-W pairs [orange line] have a Lin-similarity between 0.8 and 0.9 [x-axis position] of their respective synsets; the average cosine similarity of E and W in this class is 0.58 [blue bar].".

4. Conclusion

The experimental approach to onomasiological access in a multilingual lexicographical resource outlined in this paper is still in an early stage of implementation. It offers possible solutions to many of the issues of traditional 'domain-based' search strategies, sketched in section 1. Taking up the points listed there, we can wrap up our discussion with the following observations.

(a) Lexicographical annotators gain enormous flexibility in characterising word senses through a huge number of descriptor words. The downside to this is the curious fact that, as noted above, annotator agreement is not a useful validation criterion anymore; in addition, annotators cannot assess the implications of their descriptor assignment choices for future users. It is, however, possible to give the annotators some feedback on the 'effect' their assignments have by showing them which other lexical units in the LWPD the assigned descriptors are semantically similar to and would be retrieved using the assigned descriptors in a query.

(b) Instead of having to reconstruct a lexicographical practice of domain assignments, the user is offered a much more open, even playful access to semantic search. Guided by autocomplete functionality and without prior familiarisation with a system of domains, users can experiment with any combinations of search keys to delimit and change (narrow down or open up) the scope of their queries. Thus, this kind of semantic search fits very well into the concept of the LWPD, since it is a lexical resource aimed at scientists as well as interested laypeople.

(c) The fundamental problem of having to decide on a more or less fixed set or taxonomy/hierarchy of semantic domains in advance of the whole annotation process simply disappears.

(d) As said above, it takes a lot of effort to change the taxonomy in a domain-based search or just redefine the 'boundaries' of a given domain. In contrast, the word embedding approach is highly dynamic. (i) The set of search keys can be altered in any conceivable way any time, including additional languages (as long as the keys are included in CN, which is very likely, because the CN embeddings are trained on a very big database). (ii) The scheme of mapping descriptor labels onto weights can be adjusted as needed. (iii) The pretrained set of multilingual embeddings can be exchanged for another one. In this case, only word senses with descriptors absent from the new embeddings must be annotated anew. It is not to expected that this concerns a sizeable fraction of the word senses. (iv) Of course, assignments for individual word sense can be revised any time. In all cases, all it takes for the changes to take effect is a recomputation of the vectors and cosine similarities in the database.

In the end, the most desirable state of affairs would most certainly that of offering users a combination of different semantic search options. Finding out which option is the best for which usage scenario remains a topic for further research.

5. References

Castro Fernandez, R., Mansour, E., Qahtan, A. & Elmagarmid, A. K. (2018). Seeping Semantics: Linking Datasets Using Word Embeddings for Data Discovery. In IEEE (eds.) *Proceedings of the 34th International Conference on Data*

Engineering, ICDE 2018. Paris, pp. 989–1000. Available at: https://ieeexplore.ieee.org/document/8509314/.

Chauhan, R., Goudar, R., Sharma, R. & Chauhan, A. (2013). Domain ontology based semantic search for efficient information retrieval through automatic query expansion. In R. Kher, N. Gondaliya, M. Bhesaniya, L. Ladid & M. Atiquzzaman (eds.) *Proceedings of the International Conference on Intelligent Systems and Signal Processing, ISSP 2013*. Gujarat, pp. 397–402. Available at: http://ieeexplore.ieee.org/document/6526942/.

ConceptNet NumberBatch. Accessed at: https://github.com/commonsense/conceptnet-numberbatch. (06 April 2021)

DeReKo: *Das deutsche Referenzkorpus* Accessed at: https://www1.ids-mannheim.de/kl/projekte/korpora/. (06 April 2021)

DeReWo: *Die Deutsche Referenzkorpus Wortliste.* Accessed at: https://www1.ids-mannheim.de/kl/projekte/methoden/derewo.html. (06 April 2021)

DWDS: *Digitales Wörterbuch der Deutschen Sprache.* Accessed at: http://www.dwds.de. (06 April 2021)

Elbedweihy, K., Wrigley, S. N., Ciravegna, F., Reinhard, D. & Bernstein, A. (2012). Evaluating semantic search systems to identify future directions of research. In R. García-Castro, L. Nixon & S. Wrigley (eds.) *Proceedings of the Second international Workshop on Evaluation of Semantic Technologies.* Heraklion, Greece, pp. 25-36. Available at: https://www.zora.uzh.ch/id/eprint/63315/.

Fernandez, M., Lopez, V., Sabou, M., Uren, V., Vallet, D., Motta, E. & Castells, P. (2008). Semantic Search Meets the Web. In IEEE (eds.) *Proceedings of the 2008 International Conference on Semantic Computing*. Santa Monica, USA, pp. 253–260. Available at: http://ieeexplore.ieee.org/document/4597199/.

Hamp, B. & Feldweg, H. (1997). Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications. In: *Proceedings of the ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications.* Madrid. Available at: https://www.aclweb.org/anthology/W97-0800.pdf.

Haspelmath, M. & Tadmor, U. (2009). The Loanword Typology project and the World Loanword Database. In: M. Haspelmath & U. Tadmor (eds.): *Loanwords in the World's Languages: A Comparative Handbook*. Berlin: De Gruyter, pp. 1–34.

Henrich, V. & Hinrichs, E. (2010). GernEdiT - The GermaNet Editing Tool. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner & D. Tapias (eds.) *Proceedings of the Seventh Conference on International Language Resources and Evaluation.* Malta, pp. 2228–2235. Available at: https://www.aclweb.org/anthology/L10-1180/.

Kuzi, S., Shtok, A. & Kurland, O. (2016). Query Expansion Using Word Embeddings. In Association for Computing Machinery, New York, NY, United States (eds.) *Proceedings of the 25th ACM International Conference on Information and Knowledge Management*. Indianapolis Indiana USA, pp. 1929–1932. Available at:

https://dl.acm.org/doi/10.1145/2983323.2983876.

Lin, D. (1998). An Information-Theoretic Definition of Similarity. *Proc. of Conf. on Machine Learning*, pp. 296–304.

LWPD: Lehnwortportal Deutsch. Leibniz-Institut für Deutsche Sprache, Mannheim. Accessed at: http://lwp.ids-mannheim.de/. (06 April 2021)

Meyer, P. (2014): Graph-Based Representation of Borrowing Chains in a Web Portal for Loanword Dictionaries. In A. Abel, C. Vettori & N. Ralli (eds.) *Proceedings of the 16th EURALEX International Congress*. Bolzano, pp. 1135–1144. Available at:
http://euralex2014.eurac.edu/en/callforpapers/Documents/EURALEX%202014_gesamt.pdf.

Meyer, P. & Eppinger, M. (2018). fLexiCoGraph: Creating and Managing Curated Graph-Based Lexicographical Data. In: J. Čibej, V. Gorjanc, I. Kosem & S. Krek (eds.): *Proceedings of the XVIII EURALEX International Congress. Lexicography in Global Contexts, 17-21 July, Ljubljana*. Ljubljana: Znanstvena založba, pp. 1017-1022.

Meyer, P. (2019). Leistungsfähige und einfache Suchen in lexikografischen Datennetzen. Ein Query Builder für lexikografische Property-Graphen. In: P. Sahle (ed.): *Digital Humanities: multimedial & multimodal. 6. Tagung des Verbands Digital Humanities im deutschsprachigen Raum e.V. (DHd 2019), Frankfurt am Main, Mainz, 25.3.2019 – 29.3.2019. Konferenzabstracts*. Frankfurt a.M.: Zenodo, pp. 312-314.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). Efficient estimation of word representations in vector space. *CoRR abs*/1301.3781.

Mohamed, E. H. & Shokry, E. M. (2020). QSST: A Quranic Semantic Search Tool based on word embedding. In *Journal of King Saud University - Computer and Information Sciences*.

OpenThesaurus. Accessed at: https://www.openthesaurus.de/. (06 April 2021)

Osservatorio degli Italianismi nel Mondo. Accessed at: http://www.italianismi.org. (06 April 2021)

Speer, R., Chin, J. & Havasi, C. (2018). ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In: arXiv:1612.03975 [cs].

Tümer, D., Shah, M. A. & Bitirim, Y. (2009). An Empirical Evaluation on Semantic Search Performance of Keyword-Based and Semantic Search Engines: Google, Yahoo, Msn and Hakia. In IEEE Computer Society (eds.) *2009 Fourth International Conference on Internet Monitoring and Protection*. Venice/Mestre, Italy, pp. 51–55. Available at: http://ieeexplore.ieee.org/document/5076348/.

Uma Devi, M. & Meera Gandhi, G. (2015). Wordnet and Ontology Based Query Expansion for Semantic Information Retrieval in Sports Domain. In *Journal of Computer Science*, 11(2), pp. 361–371.

van der Sijs, N. (2015). Uitleenwoordenbank, uitleenwoordenbank.ivdnt.org, hosted by the Instituut voor de Nederlandse Taal. Accessed at: http://uitleenwoordenbank.ivdnt.org/. (06 April 2021)

Wortschatz Universität Leipzig. Abteilung *Automatische Sprachverarbeitung* am Institut für Informatik der Universität Leipzig. Accessed at: https://corpora.uni-leipzig.de/ (06 April 2021)

Zeller, J. P. (2015). The semantic fields of German loanwords in Polish. In *Studies in Polish Linguistics*, pp. 153–174.

Using Open-Source Tools to Digitise Lexical Resources for Low-Resource Languages

Ben Bongalon[1], Joel Ilao[2], Ethel Ong[3], Rochelle Irene Lucas[4], Melvin Jabar[5]

[1] Independent Researcher, California, USA
[2,3] College of Computer Studies, De La Salle University, Manila, Philippines
[4] English and Applied Linguistics, De La Salle University
[5] Behavioral Sciences, De La Salle University
E-mail: ben@isawika.org, {joel.ilao, ethel.ong, rochelle.lucas, melvin.jabar}@dlsu.edu.ph

Abstract

Advances in open-source lexicography tools have made it more practical to digitise historical dictionaries and lexical resources. However, most retro-digitisation efforts have catered to dominant languages while ethnic minority and indigenous languages tend to be neglected. In countries with a large number of regional and local languages, such as the Philippines, retro-digitisation is a daunting challenge. Of its 186 languages and 500+ dialects, only a few are known to have e-dictionaries produced. The traditional "top-down" approach simply does not scale, since the community need for language documentation far outstrips the number of motivated linguists, lexicographers and funding entities available. This paper describes a complete tool chain and workflow that we used to digitise a Hanunoo-English dictionary originally published in the 1950s (Conklin, 1953). A trainable OCR engine, Tesseract (Smith, 2007), is used to handle the novel glyphs found in the dictionary. Post-edits were performed to fix OCR errors, extract lexical elements from the transcribed pages, and produce an XML-formatted electronic dictionary containing 5,779 entries. The Lexonomy dictionary editor (Měchura, 2017) was used to edit the entries and host the access-controlled electronic dictionary online.

Keywords: indigenous language; retro-digitisation; electronic lexicography; OCR; LSTM

1. Introduction

Starting with the publication of "Samuel Johnson: A Dictionary of the English Language" on CD–ROM in 1996 (Schneiker, 2009; McDermott 1996), a growing number of projects to digitise historical dictionaries have been launched. The reasons for undertaking these projects vary and include: disseminating resources of "great historical value for European lexicographical heritage" (Salgado, 2019), aiding research to trace "the history of the language" and understand "society's situation at the time of the publication" (Özcan, 2018), providing "valuable information on the first attestations of words, on their variants (ranging e.g. from formal to diachronic or diatopic kinds), on the authors who quote them, and on their etymologies" (Sassolini, 2019).

Having a dictionary in one's mother tongue confers many advantages (SIL, 2020) including:

- Validating the use of the vernacular language and boosting the community's self-esteem
- Promoting literacy and serving as a bridge to mainstream languages
- Helping mother-tongue writers record their oral traditions and author new material
- Helping in creating educational resources in the local language
- Facilitating translation of health bulletins, news and other informational materials

Moreover, when dictionaries are digitised and made available online or as mobile applications, they promote cultural identity and a sense of pride, foster language use in youth (who heavily use mobile apps), and encourage learners around the world to interact and use the language which helps in preserving it.

Despite the numerous benefits of having retro-digitised lexical resources, many speakers of minority and indigenous languages today do not have electronic dictionaries and grammar reference books for their own communities to use. Why is this so? We believe the overall cost of retro-digitisation projects in terms of the time, money and skills required are still too high, making them out of reach for marginalised language communities. Without adequate funding and institutional support, these communities often depend on external partners who happen to express interest in their mother tongue to initiate the projects on their behalf.

Creating dictionaries from scratch takes considerable time and resources. Not only is the initial word collection effort expensive, but even the subsequent phase of producing the dictionary typically requires two people working full time for 12 to 18 months (SIL, 2020). This is where historical dictionaries can play a vital role. Many dictionaries for languages of ethnic minority and indigenous groups have been published in the last 100 years. Often it took years to compile them given the language barriers and extreme difficulty in reaching the target communities, who often lived in remote locations. Thus they contain substantial linguistic and cultural knowledge, and while no doubt many words have shifted in meaning or are no longer used by today's native speakers, core vocabularies are surprisingly resilient to semantic shift and can be used to bootstrap or augment modern dictionary-building initiatives when desired by the community. In other words, retro-digitisation enables ethnic minority and indigenous communities to start building e-dictionaries for their language with less risk, cost and effort.

However, retro-digitisation presents a huge challenge for countries with a large number of minority and indigenous languages. The traditional "top-down" approach where language documentation projects typically require multi-year efforts and sizable budgets simply does not scale (i.e., the number of languages to be documented far outstrips the number of motivated linguists, lexicographers and funding entities available). The Philippines makes for a good example. With 186 languages (Eberhard et al., 2021) and 500+ dialects, it is the 25th most linguistically diverse country in the world (World Atlas, 2009), but almost half of these languages are considered endangered (Eberhard et al., 2021), and thus the need to produce more language resources to revitalise them.

In this paper, we describe our project to retro-digitise a historical dictionary developed for the Hanunoo Mangyan language. Hanunoo (IPA: [hanunuʔɔ]) is spoken by one of the eight Mangyan ethnic groups in Mindoro, an island in the southwestern part of the Philippines. Other languages include Alangan, Iraya, Buhid and Tadyawan (Zorc, 1974). It is classified as an Austronesian language, a sub-classification of Malayo-Polynesian, further sub-classified as a Greater Central Philippine language (South Mangyan) (Eberhard et al., 2021; Blust, 1991). There were approximately 25,100 speakers of Hanunoo Mangyan as of 2010 (Eberhard et al., 2021).

2. Related Work

The Hanunoo Mangyan is a unique ethnolinguistic group in the Philippines as it has its own indigenous system of writing, known as the Surat Mangyan. Their system of writing is said to have descended from the ancient Sanskrit alphabet. There are 18 characters in the syllabary, three of which are vowels; the remaining 15 are written in combination with the vowels (Conklin, 1953). However, the writing system is no longer used in the day-to-day encounters of the Hanunoo Mangyan population.

Prior works in documenting the Hanunoo language are found in literature. Studies on the Hanunoo vocabulary (Scannel, 2015) and Hanunoo and English (Conklin, 1953, 1955, 1962) have been conducted and dictionaries produced. Harold Conklin, an American anthropologist who studied the indigenous Hanunoo culture in the Philippines after serving in the US Army during WWII, authored the “Hanunoo-English Vocabulary” (Conklin, 1953) using field notes from his voluntary fieldwork in Mindoro. It is this dictionary that inspired our retro-digitisation project.

Digitising historical dictionaries has been carried out for various languages including English (Johnson, 1996), German (Christmann, 2003), Portuguese (Simões, 2016; Salgado, 2019b), Turkish (Özcan, 2018). Italian (Sassolini, 2019), French (Salgado, 2019b) and Spanish (Salgado, 2019b). Text capture, the process of converting print pages into text, can be grouped into three approaches. For digital-born dictionaries that were printed from LaTex or tagged PDF documents, the embedded markup in the typesetting files was used directly to create XML-formatted e-dictionaries with minimal processing (Simões, 2016; Salgado, 2019b). Some projects, including the Oxford English Dictionary 2nd Edition and the Deutsches Wörterbuch (Christmann, 2003) relied on brute force, employing typists to manually enter the entire text, in some cases double-keyed to achieve higher accuracy. The third and most common approach is to apply OCR technology to transcribe scanned page images to text (Sassolini, 2019).

The Text Encoding Initiative Guidelines (TEI Consortium, 2016) is a *de facto* standard for digitally encoding all types of written texts, ranging from novels and poetry to mathematical formulae or music notation (Salgado, 2019a). Its “Dictionaries” chapter provides guidelines for encoding human-oriented monolingual and multilingual dictionaries, glossaries and similar documents. TEI-Lex0 (Banski, 2017) is a proposed extension to address representational ambiguities in TEI with a stricter set of encoding rules. It has been used to construct the Nxaʔamxcín (Czaykowska-Higgins, 2014), Portuguese, Spanish, and French Academy Dictionaries. Salgado (2019b) proposed further enhancements to TEI-Lex0, most notably in terms of diatextual labels.

3. Materials and Methods

In this section, we discuss how the Hanunoo dictionary was digitised and published for our target audience. We use the workflow stages defined in the DariahTeach's "Digitizing Dictionaries" course (DariahTeach, 2020) to organise our presentation.

Several post-editing tasks were needed to convert the original book into a user-accessible digital resource. In this retro-digitisation project, we trained an OCR engine to recognise special characters used in the Hanunoo dictionary because out-of-the-box OCR engines did not perform well and thus were put aside. Proofreaders were employed to correct residual errors in the OCR output, and to format the content to conform to an XML schema we defined for semantic markup.

3.1 Planning

Planning was simple given that the project is a loose collaboration between the primary author (independent researcher) and faculty members of the De La Salle University's (Philippines) English and Applied Linguistics, Behavioral Science and Computer Technology departments. We aimed to explore innovative ways to leverage mutual interest in developing electronic lexical resources for the Philippines' indigenous languages.

The immediate goal was to produce a high-quality, digitised version of the Conklin dictionary which could serve as: 1) an accessible historical reference of the Hanunoo language, and 2) an auxiliary source of lexical data to augment recent Hanunoo language documentation projects. To make the e-dictionary accessible to our target users, we published it as a web-based application and shared the data for research and community use by providing the XML source. To ensure a high-quality final output, each page would be proofread. While we did not set a formal project schedule, we discussed a soft target of three to six months.

3.2 Image and Text Capture

Because the Conklin dictionary is out-of-print and rare, we sent our copy to a book-scanning service for non-destructive scanning in order to preserve it. We received an image scan of all the pages as a PDF file, as well as an OCR-ed version in Microsoft Word. However the OCR output had too many transcription errors which the company could not correct, so an alternate OCR solution was needed.

In analysing the transcription errors, we found a systematic pattern. Most were due to two special characters used in the Conklin dictionary that stumped out-of-the-box OCR engines: the ŋ (eng) letter and the ʔ glottal stop symbol. They were often mis-transcribed as ‘g’ and question mark ‘?’ characters, respectively. Another set of common errors were the sporadic omission of diacritical marks on vowels. The ŋ and diacritical mark errors were especially problematic, because being both pervasive and subtle, manually correcting them would have been very labour-intensive and so it is desirable to have them accurately transcribed.

To overcome these errors, we searched for OCR engines that can be trained to recognise new symbols. Of the two that we found, Tesseract (Smith, 2007) and OCRopus (Breuel, undated), we chose the former because it supports many more pre-trained language models[1] and is actively maintained. Moreover, starting with version 4, Tesseract employs Deep Learning technology (LSTM neural networks) for more accurate text recognition.

3.2.1 Training the OCR Engine

Training Tesseract began with finding a pre-trained language model that can recognise the most characters present in the source document’s character set. For Conklin’s dictionary, a reasonable assumption would be to use the Tagalog model (tgl.traineddata), since both Tagalog and Hanunoo are Philippine languages. However, our experiment showed that the Spanish model (spa.traineddata) was a better starting point because it recognised diacritical marks in vowels (á, é, í, ó, ú) more accurately than the Tagalog model.

Next, we strategised on how to handle the ŋ and ʔ special characters. The ŋ (eng) symbol, a ligature of the digraph “ng”, is pervasive in some Philippine languages. Thus we wanted the OCR to recognise ŋ accurately to avoid a massive number of post-corrections. On the other hand, question marks ‘?’ were seldomly used in the vocabulary pages so globally replacing them with a glottal stop symbol yielded very few errors which were easily corrected during proofreading. We will revisit the theme of minimising the production cost in the Discussion section. The key point is that by choosing a good starting language model and allowing for a small number of expected transcription errors, we reduced the OCR training task to recognising just one new character (ŋ).

The high-level steps are described in Table 1. We wrote scripts to execute each step as single-line commands. For reference, the scripts and the detailed steps are available on GitHub[2]. To create the training data, we chose 20 sample pages from the scanned dictionary, preferring pages with Hanunoo words containing ŋ in different positions (first, middle, last letter of the words).

[1] For a list of Tesseract language models, see https://github.com/tesseract-ocr/tessdata

[2] Our project repository can be found at https://github.com/isawika/retro-digitization

Step	Notes

1. Prepare the training data.
Split the PDF document into individual pages.
$ pdftk book.pdf burst
Convert the PDF pages to TIFF format.
$ pdf2tiff *.pdf
	Output: page-01.pdf, page-02.pdf, etc.
	We use TIFF image files because both Tesseract and jTessBoxEditor support it.
	Output: page-01.tiff, page-02.tiff, etc.

2. Create a Tesseract box file for each page. $ for i in *.tif; do ../tessbox.sh $i; done;	A box file contains Tesseract's predicted characters in the page. OCR is performed using a pre-trained Spanish language model.
3. Open each page in jTessBoxEditor, then find and correct the OCR errors.	jTessBoxEditor saves the edits in the box file.
4. Convert each box file into a plain text file. $ for i in *.box; do ../box2lines $i; done;	**Output:** page-01.txt, page-02.txt, etc.

5. Create the training text.
Combine the plain text files from Step 4.
$ cat page*.txt > hanunoo.txt
Prune the file and add to the Spanish training data.
$ cat hanunoo.txt >> spa.training_text
	Multiple experiment runs may be needed to determine the appropriate mix of new and original training data. See 3.2.2 for details.

6. Run the Tesseract fine-tuning procedure. $ tesstrain.sh; combine_tessdata; lstmtraining	For brevity, the full commands are not shown. They mimic the commands in the "Fine Tuning" section of the Tesseract tutorial. [3]

Table 1: Steps for fine-tuning the Tesseract OCR engine

3.2.2 Evaluating the models

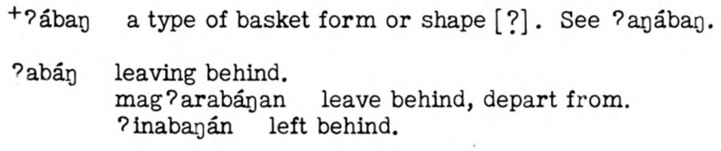
Only a small amount of sample text is needed to fine-tune the OCR engine. For the Conklin dictionary, we found that adding 40 lines of Hanunoo text to the original 68 lines of Spanish training data (*spa.training_text*) yielded the best results. In fact, including more Hanunoo text resulted in more OCR errors. Even more surprising, removing the Spanish text completely and replacing it with Hanunoo text produced a model that performed the worst and generated unknown words ("hallucinations" in Tesseract parlance). In the latter two cases, we believe the resulting neural net models

[3] Training, see https://tesseract-ocr.github.io/tessdoc/tess4/TrainingTesseract-4.00.html

were overfitted to the training data. We ran eight experiments in total, from which we selected the best performing model.

3.2.3 Using the trained model

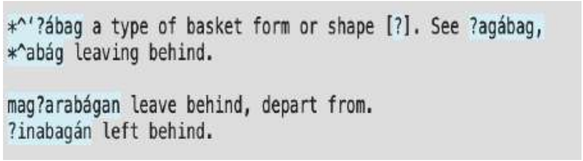
We transcribed the vocabulary pages (N=270) using the best re-trained language model “X3”, then replaced all occurrences of question marks ‘?’ with glottal stop ‘ʔ’ symbols. Figure 2 shows a sample result. All ŋ symbols were recognised. However the glottal stop substitution rule incorrectly replaced the question mark symbol “[ʔ]” in Line 1 (an infrequent error). These need to be fixed in the post-edit step.



```
+ʔábaŋ   a type of basket form or shape [ʔ]. See ʔaŋábaŋ.

ʔabáŋ    leaving behind.
         magʔarabáŋan   leave behind, depart from.
         ʔinabaŋán   left behind.
```

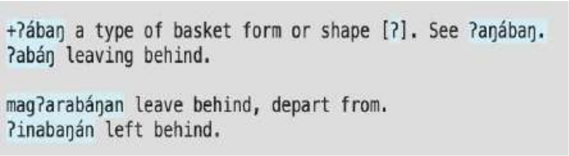
Figure 2a: Source PDF



```
*^'?ábag a type of basket form or shape [?]. See ?agábag,
*^abág leaving behind.

mag?arabágan leave behind, depart from.
?inabagán left behind.
```

Figure 2b: Transcription before training



```
+ʔábaŋ a type of basket form or shape [ʔ]. See ʔaŋábaŋ.
ʔabáŋ leaving behind.

magʔarabáŋan leave behind, depart from.
ʔinabaŋán left behind.
```

Figure 2c: After OCR training & glottal stop replacement

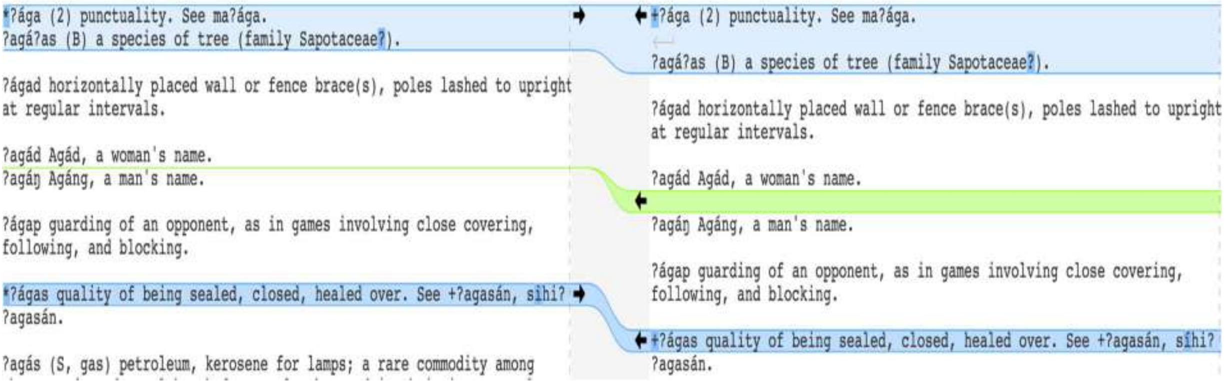
3.2.4 Post-Editing

The transcribed pages needed manual review to correct residual errors. We used UpWork[4] to find freelance proofreaders and had a positive experience. After posting the project for five days, we received 31 bids, screened applicants with a sample task, and hired two freelancers to work in parallel. While we initially planned to hire a third person to provide 2X coverage on 25% of the pages, this proved unnecessary as the quality of the two proofreaders’ work was excellent.

We spot-checked the pages on a MacBook computer using the open-source Meld tool[5], visually comparing the OCR transcript with the proofreaders’ edits and consulting the original PDF page as needed. Figure 3 shows an example output.

[4] see http://upwork.com

[5] see https://meldmerge.org



```
*?ága (2) punctuality. See ma?ága.
?agá?as (B) a species of tree (family Sapotaceae?).

?ágad horizontally placed wall or fence brace(s), poles lashed to upright
at regular intervals.

?agád Agád, a woman's name.
?agáŋ Agáng, a man's name.

?ágap guarding of an opponent, as in games involving close covering,
following, and blocking.

*?ágas quality of being sealed, closed, healed over. See +?agasán, sihi?
?agasán.

?agás (S, gas) petroleum, kerosene for lamps; a rare commodity among
```

```
+?ága (2) punctuality. See ma?ága.

?agá?as (B) a species of tree (family Sapotaceae?).

?ágad horizontally placed wall or fence brace(s), poles lashed to upright
at regular intervals.

?agád Agád, a woman's name.

?agáŋ Agáng, a man's name.

?ágap guarding of an opponent, as in games involving close covering,
following, and blocking.

+?ágas quality of being sealed, closed, healed over. See +?agasán, síhi?
?agasán.
```

Figure 3: Comparison of an OCR output (left) and the proofread page (right). Blue highlights denote modified lines, with the actual changes in dark blue. The green highlight denotes a blank line that was added to separate two dictionary entries.

3.3 Data Modelling and Enrichment

Data modeling and data enrichment were intricately enmeshed in our project and so we discuss them together. First, we analysed the dictionary entries to identify the various semantic elements present to design the encoding schema in Figure 4a. We followed the TEI-Lex0 standard (Banski, 2017) with some deviations for a simpler markup. For example we skipped the use of <form> elements, inserting the <headword> and <pronounce> elements directly under the <entry> node.

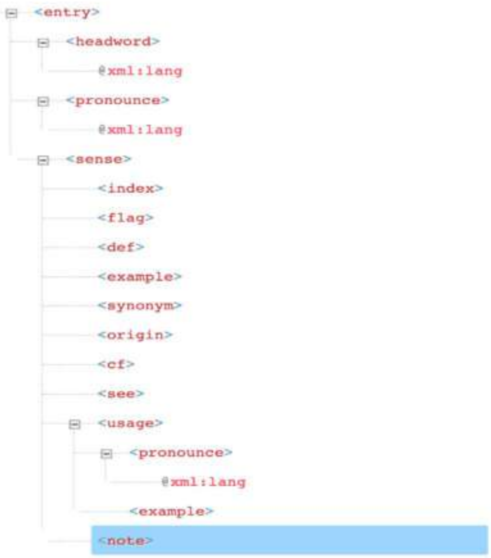
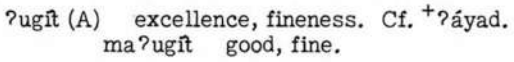
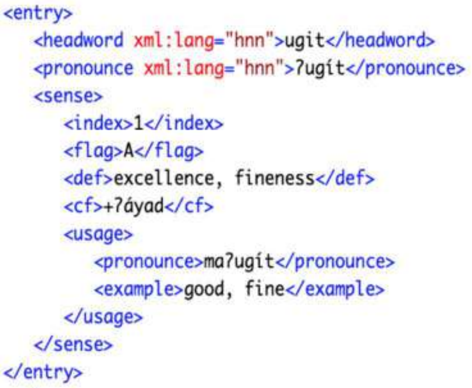


Figure 4a: Schema for Conklin dictionary



?ugît (A) excellence, fineness. Cf. +?áyad.
ma?ugît good, fine.

Figure 4b: An entry in an OCR-ed page



```
<entry>
  <headword xml:lang="hnn">ugit</headword>
  <pronounce xml:lang="hnn">?ugít</pronounce>
  <sense>
    <index>1</index>
    <flag>A</flag>
    <def>excellence, fineness</def>
    <cf>+?áyad</cf>
    <usage>
      <pronounce>ma?ugít</pronounce>
      <example>good, fine</example>
    </usage>
  </sense>
</entry>
```

Figure 4c: Entry formatted in XML

The entries in the Conklin dictionary intermixed references to synonyms, word origins, “c.f.” / “see also” terms or other annotations with the definition body (Figure 4b). We wrote a Python script (*conklin2xml.py*) to unpack them into separate XML elements. The textual flow followed a fairly regular pattern, making it easy to define pattern-extraction rules.

To make the dictionary searchable, the script also created two XML elements for each headword. The <headword> field contained a Romanised form of the word with syllable hyphens and glottal stop symbols removed, and with "ŋ" symbols changed to "ng". The <pronounce> field retained the original orthography. For example:

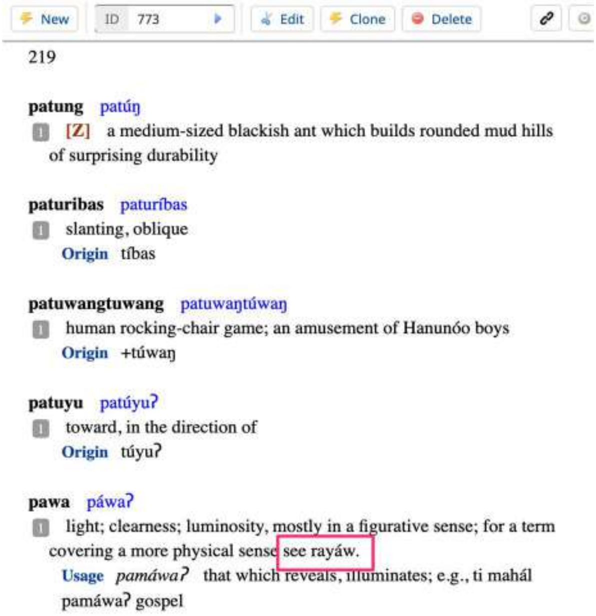
"ʔínaʔ ʔulúŋ" (stepmother) **becomes** <headword>ina ulung</headword>
<pronounce>ʔínaʔ ʔulúŋ</pronounce>

Calling the Python script with the OCR-ed text as input, as shown below, will produce a fully-formatted XML document (Figure 4c):

```
$ conklin2xml.py page021-ocr.txt > page021.xml
```

As in the OCR text capture, the output XML documents contained errors that needed manual correction. In addition, post-edits were needed to undo several "typographical and editorial conventions of the print medium" (Tasovac, 2010), specifically to merge lexical entries that spanned across two pages and to dehyphenate words that wrapped at the end of a text line.

We hired a third freelancer to perform the post-edits, a task that took 14 days to complete. To simplify the editing task, we loaded a specially formatted version of the XML documents into a self-hosted Lexonomy dictionary editing application (Měchura, 2017), where each "entry" embodied a page's worth of lexical entries. This "page view" format significantly aided proofreading because it was easier to visually compare a virtual page against its original PDF source (Figures 5a and 5b) and make corrections to the virtual page (Figure 5c).



219

patung patúŋ
1 [Z] a medium-sized blackish ant which builds rounded mud hills of surprising durability

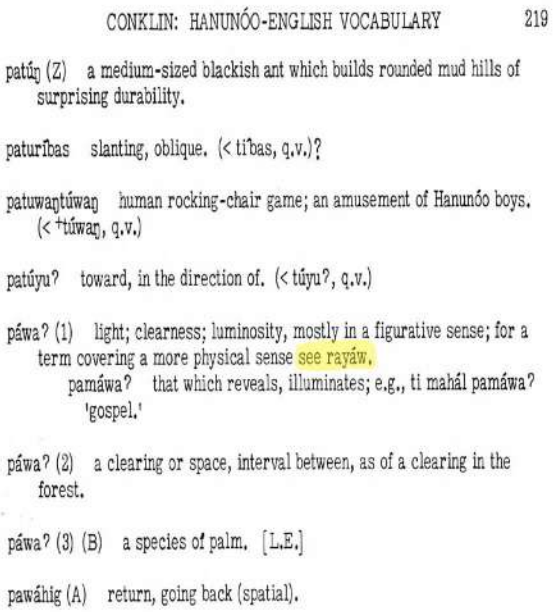
paturibas paturíbas
1 slanting, oblique
Origin tíbas

patuwangtuwang patuwaŋtúwaŋ
1 human rocking-chair game; an amusement of Hanunóo boys
Origin +túwaŋ

patuyu patúyuʔ
1 toward, in the direction of
Origin túyuʔ

pawa páwaʔ
1 light; clearness; luminosity, mostly in a figurative sense; for a term covering a more physical sense see rayáw.
Usage *pamáwaʔ* that which reveals, illuminates; e.g., ti mahál pamáwaʔ gospel

Fig 5a: "Page view" has an entry with dangling text caused by a run-on sentence


CONKLIN: HANUNÓO-ENGLISH VOCABULARY 219

patúŋ (Z) a medium-sized blackish ant which builds rounded mud hills of surprising durability.

paturíbas slanting, oblique. (< tíbas, q.v.)?

patuwaŋtúwaŋ human rocking-chair game; an amusement of Hanunóo boys. (< +túwaŋ, q.v.)

patúyu? toward, in the direction of. (< túyu?, q.v.)

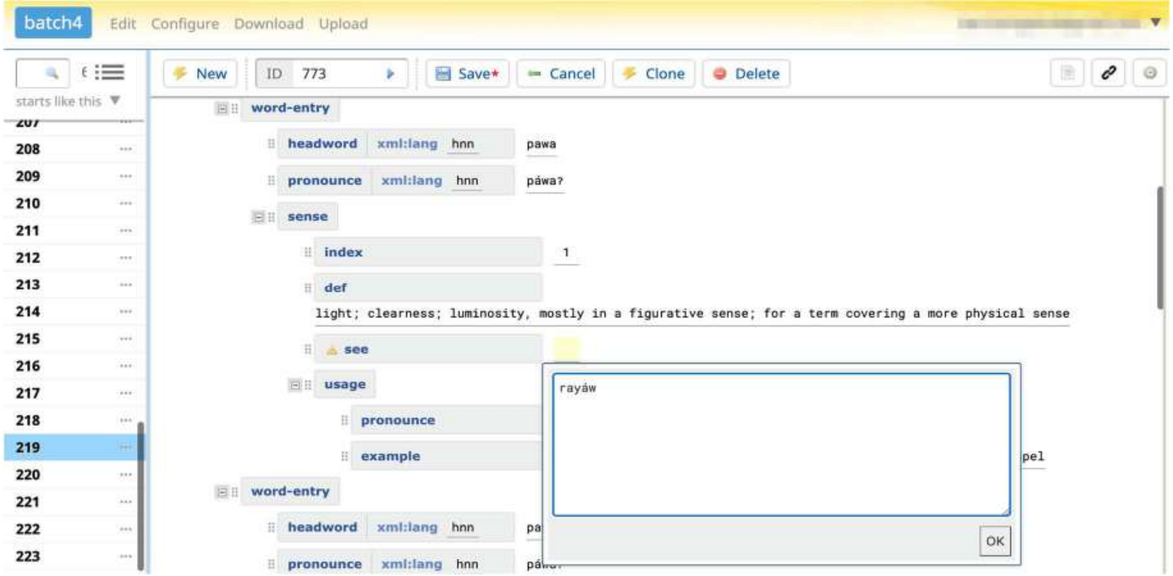
páwa? (1) light; clearness; luminosity, mostly in a figurative sense; for a term covering a more physical sense see rayáw.
pamáwa? that which reveals, illuminates; e.g., ti mahál pamáwa? 'gospel.'

páwa? (2) a clearing or space, interval between, as of a clearing in the forest.

páwa? (3) (B) a species of palm. [L.E.]

pawáhig (A) return, going back (spatial).

Figure 5b: Source PDF (error highlighted)



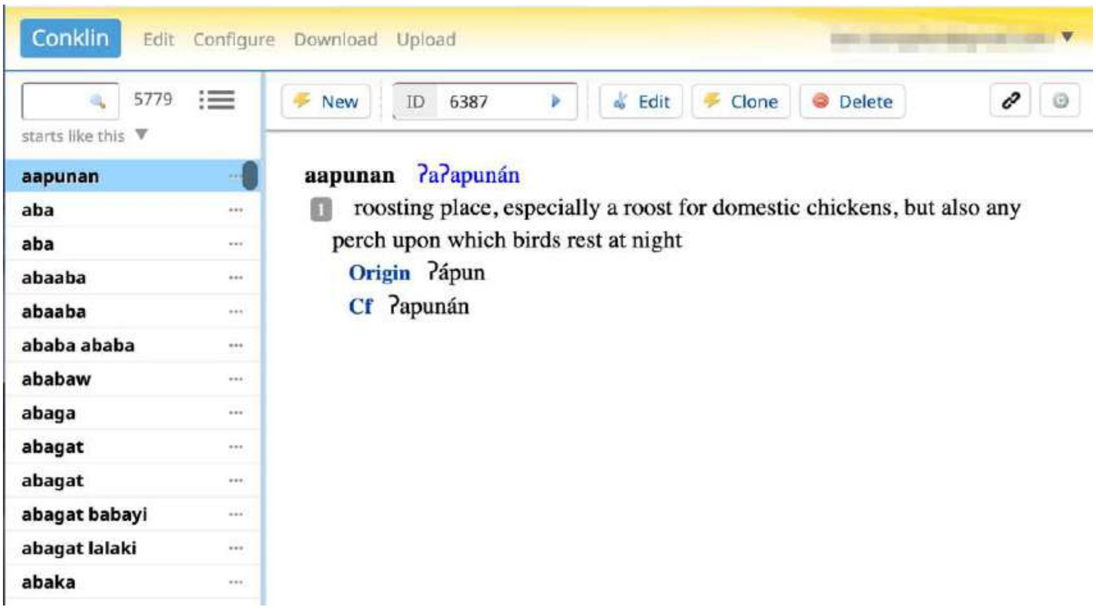
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223

word-entry
headword xml:lang hnn pawa
pronounce xml:lang hnn páwaʔ
sense
index 1
def
light; clearness; luminosity, mostly in a figurative sense; for a term covering a more physical sense
see
usage
pronounce
example
rayáw
pel
word-entry
headword xml:lang hnn pa
pronounce xml:lang hnn páw

Figure 5c: Entry is fixed by splitting the reference into a "see" XML element

3.4 Publishing

After the data enrichment edits were completed, the XML documents were downloaded from the Lexonomy[6] platform. The documents were reformatted to detach the individual dictionary entries from the page frames and were re-uploaded. The resulting e-dictionary contains a total of 5,779 headwords, as shown in Figure 6.



aapunan
aba
aba
abaaba
abaaba
ababa ababa
ababaw
abaga
abagat
abagat
abagat babayi
abagat lalaki
abaka

aapunan ʔaʔapunán
1 roosting place, especially a roost for domestic chickens, but also any perch upon which birds rest at night
Origin ʔápun
Cf ʔapunán

Figure 6: Format of the dictionary after the entries were detached from the page frames

The Hanunoo-English dictionary is online and access-controlled with individual permissions granted in consultation with representatives of the Mangyan community. The same Lexonomy platform used for editing is used to publish the e-dictionary.

[6] see https://www.lexonomy.eu

4. Discussion

Despite its introduction over 20 years ago, retro-digitisation technology is still immature. While numerous projects have documented their workflow and tools to share knowledge, there are no clear guidelines to help lexicographers figure out which solution is best for their needs. Often they must find out by trial and error. This is an inconvenience for larger and better-funded organisations but a barrier for the resource-constrained, many of whom represent or support ethnic minority and indigenous communities. We thus aimed to help address this issue by introducing a complete digitisation workflow that leverages open-source tools to eliminate or significantly reduce software expenses, and by sharing techniques that contribute to best practices for digitising lexical resources.

In implementing our project, we observed some limitations in the tools we used:

- The Tesseract training program (tesstrain.sh) randomly shuffles the input training data which unpredictably varies the performance of the trained model. To compensate, we ran experiments multiple times to obtain the best model for a given training setup.
- Lexonomy does not support limiting user access to a subset of a dictionary. To prevent proofreaders from accidentally overwriting others' work, we created separate dictionaries containing only the entries each one was responsible for.
- Lexonomy has no built-in support for the "page view" editing as described in Section 3.3. We jerry-rigged it by temporarily reformatting the XML document.
- There appears to be a lack of data interoperability among lexicography tools from different providers. For example, an organisation that wants to use SIL's Dictionary App Builder[7] to create a mobile version of their Lexonomy e-dictionary would first need to build a custom translator.

We admit that the workflow we propose still includes steps that may be challenging and intimidating to less technical users. Training the Tesseract OCR remains to be an art and needs to be simplified. Similarly converting the OCR-ed dictionary pages into XML documents requires someone skilled in writing Python scripts. For the latter, tools such as GROBID-dictionaries (Khemakhem, 2017) which allow users to specify the transformation rules by giving examples can enable laypeople to do the task.

There are also aspects of our method that require further exploration. While our solution worked well for digitising the Hanunoo-English dictionary, we do not know how generalisable it is. Questions include: *How likely will other projects be able to find a good OCR language model as a starting point? How does the number of unknown characters in the source's alphabet affect training complexity? What conditions make it possible to achieve high recognition accuracy on mixed-language text with a single*

[7] See https://software.sil.org/dictionaryappbuilder

language model? In our case, we obtained surprisingly excellent transcription quality for both Hanunoo and English text from a language model that we did not train with English text included.

Digitising the Hanunoo-English dictionary presented some ethical concerns. While the dictionary itself became public domain when its US copyright expired, the vocabulary it contains is considered property of the Mangyan people. Therefore publishing it online requires their Free, Prior and Informed Consent (FPIC) as mandated in the Philippine Indigenous Peoples' Rights Act of 1997 (IPRA, 1997) because "the copyright to their indigenous language has no expiration" (private communication). There is also the question of whether our team is guilty of treating "language as data" (Bird, 2020). In this regard, Bird seems to level criticism against researchers who employ "zero resource" techniques that automatically "discover the language" from audio recordings or transcriptions without further input from linguists, speakers or previously developed language resources. Our project takes a completely opposite approach, reusing and repurposing linguistic knowledge that Conklin and several members of the Hanunoo tribe meticulously documented 70 years ago. However, due to these concerns we took the measured approach of making the e-dictionary available only to the Mangyan community and for limited research. While the Mangyan people are reluctant to publicly share their vocabulary online for fear of cultural misappropriation, they supported and participated in building the vocabulary for an earlier e-dictionary project initiated by the De La Salle University research team (Uy, 2020). In that project the community acknowledged the importance of digitising their language for preservation purposes, affirming their openness to change.

5. Conclusion and Further Work

We presented a tool chain and detailed workflow for digitising a historical dictionary which required the use of a trainable OCR engine to recognise special characters. While the technique was successfully demonstrated in one dictionary, we believe it is applicable to other similar projects. In designing the workflow, we aimed to lower the bar to retro-digitisation in order to encourage more paper dictionaries for other languages to be digitised. We also hope to give minority and indigenous communities an easier way to build and shape their own language resources so help them become more active participants in the digital age.

We plan to host the Hanunoo e-dictionary online indefinitely given the modest cost of hosting (US$800 to $2,500 per year). We will seek volunteers and explore support options for maintaining the dictionary content and the website. Our group intends to expand the research to the other Mangyan languages, namely Buhid, Tawbuwid, Alangan, Iraya and Tadyawan, and possibly to other Philippine indigenous languages.

In doing so, we anticipate some challenges ahead. First, data availability is a concern because there may be fewer printed lexical materials and native speakers available to build a dictionary for the other indigenous languages. Related to this is the issue of combining digital resources for the same language. Various sources are likely to differ in levels of organisation, from unstructured (narratives, poems) to structured (dictionaries), and some materials may even incorporate the orthographies of the Mangyan indigenous writing scripts. These informational mismatches must be reconciled, with a suitable XML dictionary schema developed, so that content can be merged. Third, maintaining and growing the e-dictionaries will require more robust data management processes to enable faster, distributed content creation without sacrificing data quality. As an example, we would like to harness crowdsourcing to build dictionaries more rapidly but with appropriate submissions screening and review processes in place. Another issue is that when working with indigenous groups, securing the appropriate ethical approvals for research takes time and this can significantly delay or curtail the data gathering process. Finally, funding grants for language documentation is difficult in the Philippines given the limited government support for such research endeavours. Despite these challenges, we remain determined to pursue these projects and leverage the open-source, retro-digitisation solution we developed.

6. Acknowledgements

The research team wishes to express their gratitude to the following agencies: Department of Science and Technology (DOST) - National Research Council of the Philippines (NRCP), and Summer Institute of Linguistics (SIL) Manila, Philippines.

7. References

Banski, P., Bowers, J., & Erjavec, T. (2017). *TEI-Lex0 Guidelines for the Encoding of Dictionary Information on Written and Spoken Forms*. HAL Archives.

Bird, S. (2020). Decolonising Speech and Language Technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 3504-3519. Barcelona, Spain.

Blust, R. (1991). The Greater Central Philippines Hypothesis. *Oceanic Linguistics*, 30(2), pp. 73-129. University of Hawai'i Press. https://doi.org/10.2307/3623084. Available at: https://www.jstor.org/stable/3623084 (22 March 2021)

Breuel, T. The OCRopus Open Source OCR System. Accessed at: https://github.com/ocropus/ocropus.github.io (22 March 2021)

Christmann, R. & Schares, T. (2003). Towards the User: The Digital Edition of the Deutsche Wörterbuch by Jacob and Wilhelm Grimm. *Literary and Linguistic Computing*, 18(1), pp. 11–22. https://doi.org/10.1093/llc/18.1.11

Conklin, H. (1953). *Hanunóo-English Vocabulary.* Berkeley: University of California Press.

Czaykowska-Higgins (2014). Using TEI for an Endangered Language Lexical Resource: The Nxaʔamxcín Database-Dictionary Project. Available at: https://scholarspace.manoa.hawaii.edu/bitstream/10125/4604/8/czaykowska.pdf

DariahTeach (2017). *Digitizing Dictionaries* course. Accessed at: https://teach.dariah.eu/mod/page/view.php?id=343 (22 March 2021)

Eberhard, D., Simons, G. & Fennig, C. (2021). *Ethnologue: Languages of the World. Twenty-fourth edition.* Dallas, Texas: SIL International. Available at: http://www.ethnologue.com

Harrison, K.D., Lillehaugen, B.D., Fahringer, J., & Lopez, F.H. (2019). Zapotec Language Activism and Talking Dictionaries. In I. Kosem, T. Zingano Kuhn, M. Correia, J. P. Ferreria, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek & S. & C. Tiberius (eds.) *Electronic Lexicography in the 21st Century. Proceedings of the eLex 2019 Conference*, pp. 31-50. Brno: Lexical Computing CZ, s.r.o. Available at: https://works.swarthmore.edu/fac-linguistics/252

IPRA (1997). "The Indigenous Peoples' Rights Act of 1997". Republic Act No. 8371. Philippine Official Gazette. October 29, 1997. Available at: https://www.officialgazette.gov.ph/1997/10/29/republic-act-no-8371/ (5 April 2021)

Jabar, M., Lucas, R., Collado, Z., & Regadio, C. (2019). An Ethnolinguistic Vitality Study of the Hanunoo Mangyan Language. *Terminal Report*, De La Salle University, Philippines.

Johnson, S. & McDermott, A. (1996). *A Dictionary of the English Language on CD–ROM.* Cambridge, England and NY, USA: Cambridge University Press.

Khemakhem, M., Foppiano, L., & Romary, L. (2017). Automatic Extraction of TEI Structures in Digitized Lexical Resources Using Conditional Random Fields. In I. Kosem et al. (eds.) *Electronic Lexicography in the 21st Century: Proceedings of eLex 2017 Conference.* Leiden, Netherlands. Hal-01508868v2

Měchura, M. (2017). Introducing Lexonomy: An Open-source Dictionary Writing and Publishing System. In I. Kosem et al. (eds.) *Electronic Lexicography in the 21st Century: Proceedings of eLex 2017 Conference.*

OED (2019). Digitizing the OED: the making of the Second Edition. OED Blog, 15 January 2019. Available at: https://public.oed.com/blog/digitizing-the-oed-the-making-of-the-second-edition/ (5 April 2021)

Özcan, E. (2018). Retro-digitizing Turkish Dictionaries Using GROBID-dictionaries. *Lexical Data Masterclass Symposium.* Berlin: Germany. (HAL-01969337)

Postma, A. (1986). *Primer to Mangyan Script (1st ed.).* Oriental Mindoro, Philippines: Mangyan Research Center.

Postma, A. (2002). *Primer to Mangyan Script (1st Rev. ed.).* Oriental Mindoro, Philippines: Mangyan Heritage Center.

Postma, A. (2013). *Primer to Mangyan script (2nd Rev. ed).* Oriental Mindoro, Philippines: Mangyan Heritage Center.

Salgado, A., Costa, R., & Tasovac, T. (2019a). Improving the Consistency of Usage Labelling in Dictionaries with TEI Lex-0. *Lexicography ASIALEX* 6, pp. 133–156. https://doi.org/10.1007/s40607-019-00061-x

Salgado, A., Costa, R., Tasovac, T., & Simões, A. (2019b). TEI Lex-0 In Action: Improving the Encoding of the Dictionary of the Academia das Ciências de Lisboa. In Kosem, I., Zingano Kuhn, T., Correia, M., Ferreria, J. P., Jansen, M., Pereira, I., Kallas, J., Jakubíček, M., Krek, S. & Tiberius, C. (eds.) *Electronic Lexicography in the 21st Century. Proceedings of the eLex 2019 Conference.*

Sassolini, E., Khan, A.F., Biffi, M., Monachini, M., & Montemagni, S. (2019). Converting and Structuring a Digital Historical Dictionary of Italian: A Case Study. In In I. Kosem, T. Zingano Kuhn, M. Correia, J. P. Ferreria, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek & S. & C. Tiberius (eds.) *Electronic Lexicography in the 21st Century. Proceedings of the eLex 2019 Conference.*

Schneiker, C., Seipel, D., & Wegstein, W. (2009). Schema and Variation: Digitizing Printed Dictionaries. In *Proceedings of the ACL-IJCNLP 2009 Third Linguistic Annotation Workshop (LAW 2009)*, pp. 82-89.

Schreibman, S., Agiatis, B., Clivaz, C., Ďurčo, M., Huang, M., Papaki, E., Scagliola, S., Tasovac, T. & Wissik, T. (2016). #dariahTeach: online teaching, MOOCs and beyond. *Digital Humanities 2016: Conference Abstracts.*

SIL. (2020-21). *Dictionary-Making and Lexicography Course.* Accessed at: https://sites.google.com/sil.org/dls-course (27 March 2021)

Simões, A., Almeida, J.J., & Salgado, A. (2016). Building a Dictionary using XML Technology. In *Proceedings of the 5th Symposium on Languages, Applications and Technologies* (SLATE'16). https://doi.org/10.4230/OASIcs.SLATE.2016.14

Smith, R. (2007). An Overview of the Tesseract OCR Engine. In *Proceedings of the 9th International Conference on Document Analysis and Recognition (ICDAR 2007)*, pp. 629-633. Curitiba, Brazil. https://doi.org/10.1109/ICDAR.2007.4376991.

Tasovac, T. (2010). Reimagining the dictionary, or why lexicography needs digital humanities. *Digital Humanities*, pp. 254–256. Center for Computing in the Humanities, Kings College London.

TEI Consortium, eds. (2016). TEI P5: Guidelines for Electronic Text Encoding and Interchange. TEI Consortium. Available at: http://www.tei-c.org/Guidelines/P5/ (13 February 2017)

Uy, D. (2020). Hanunoo Mangyan Project: Saving Languages through Technology, *The La Sallian.* Accessed at: https://thelasallian.com/2020/03/10/50467/ (22 March 2021)

World Atlas. (2009). Countries where the most languages are spoken. Accessed at: https://www.worldatlas.com/articles/the-most-linguistically-diverse-countries-in-the-world.html (4 April 2021)

Compiling an Estonian-Slovak Dictionary with English as a Binder

Michaela Denisová[1]

[1] Masaryk University, Žerotínovo nám. 617/9, 601 77 Brno
E-mail: michaeladenisova@gmail.com

Abstract

For such a rare language combination as Estonian-Slovak, it is complicated to find study materials designated for Slovaks learning Estonian, especially a bilingual dictionary, an essential language study resource. However, building a bilingual dictionary from scratch requires a lot of work and effort. The half-automatic computational methods and available open-source language resources offer a possible solution for this complicated task. One approach is to merge two already existing dictionaries that share a common language to derive a new language pair dictionary. However, as words are polysemous, many mistakes could occur while attempting so. Therefore, it is required to edit the aligned translations afterwards.

This article describes the process of compiling the Estonian-Slovak dictionary created from English-Estonian and English-Slovak dictionaries. English was chosen as an intermediate language, as it is a well-resourced language, and all materials are easy to find. Various automatic techniques were applied in the editing step to decrease the number of incorrectly aligned translations. Finally, the techniques used and quality of the dictionary were manually evaluated on a random sample of 1,000 translations.

The final version of the dictionary consists of 138,779 translations, and the Estonian headword list covers about 85% of basic Estonian vocabulary, which contains around 5,000 lemmas. The correct translations form approximately 40% of the dictionary. Additionally, a web application is being developed for this dictionary.[1]

Keywords: bilingual dictionaries; (semi)automatic compilation; intermediate language; Estonian; Slovak

1. Introduction

This project was created as a master’s thesis to provide more learning materials for Slovak students who learn Estonian at the department of Baltic Studies at Masaryk University. Students struggle with a lack of study resources in their mother tongue, especially at the beginning, and it is challenging to find an accurate translation. Therefore, this project assumed that a dictionary is one of the most crucial study materials when acquiring a new language, as students look up a foreign word in the dictionary to understand its meaning and use it correctly. Still, it is difficult to translate directly into the learners' mother tongue for a low-resource language pair, such as Estonian-Slovak. Many students thus use another major language as an intermediary, which provides more materials. However, this could lead them to incorrect translations

[1] https://estonian-slovak-dictionary.herokuapp.com (23 March 2021).

and cause mistakes in language usage. Unfortunately, creating a new bilingual dictionary is a non-trivial task, especially for rare language combinations. Using automatic methods and available open-source language resources could solve this problem. One option is to derive a new bilingual dictionary from existing dictionaries with well-resourced language as their common language. In this project, English-Estonian and English-Slovak open-source dictionaries were merged to create a new Estonian-Slovak dictionary. The direction from Estonian to Slovak was preferred as it may be more critical for the learners to grasp the meaning of the foreign words at first.

As words are polysemic, the incorrect translations are also aligned by this method. For example, *piť* (verb) ~ *drink* (verb, noun) ~ *jook* (noun). For this problem, several solutions were introduced, e.g. inverse consultation (Tanaka & Umemura, 1994) or inverse consultation combined with distributional similarity (Saralegi et al., 2011). The inverse consultation method was based on extracting translations via intermediate language and repeating the same process with the obtained words in the reverse direction. Meanings included in the acquired intersection were considered correct. In the latter method, the distributional similarity was computed from the custom-built parallel corpora to retrieve the distances between the translations.

The above mentioned approaches have been used and improved over the years in various projects, for instance, in a Japanese-French dictionary (Tanaka & Umemura, 1994), Korea-Japanese dictionary (Shirai & Yamamoto, 2001), Basque-Chinese dictionary (Saralegi et al., 2011), or in a project dealing with automatic generation of several bilingual dictionaries (Ordan et al., 2017).

The difference is that those approaches focused more on the process of combining the dictionaries. In this project, the merging is not as crucial as the automatic correction of the aligned translations after merging. Besides, the techniques applied and the quality of the resulting dictionary were manually evaluated to provide precise and accurate results for each technique separately and the whole dictionary.

This article is structured as follows. The second section explains the nature of the chosen dictionaries, and the following one focuses on the compilation process. The fourth section deals with techniques applied for automatic correction of the incorrectly aligned translations after the process of merging. Techniques are divided into separate subsections: alignment based on the part of speech, comparison with WordNet and Google Translation datasets, comparison of Estonian headword list with the EKI Combined Dictionary, and comparison of named entities. After that, the results of the evaluation are given. Finally, the conclusion is drawn, and new ideas are outlined.

2. Dictionaries

For this project, three types of open-source dictionaries were used, one English-Estonian dictionary[2] obtained from the Estonian Language Institute, and two English-Slovak dictionaries, one from online dictionary platform dict.cc[3] where authors can contribute and share their dictionaries and the other one from DictionaryForMIDs[4], which is a free multi-purpose dictionary designed for cell phones, PDAs, or PCs.

The English-Estonian dictionary was a large dictionary with an extensive vocabulary, which contained 83,089 headwords. On the other hand, the English-Slovak dictionary obtained from DictionaryForMIDs had only 26,070 English headwords. Therefore, it was necessary to include the second English-Slovak dictionary to create a greater intersection with English words from the English-Estonian dictionary, so the resulting compiled Estonian-Slovak dictionary would contain more entries. The English-Slovak dictionary from dict.cc had 25,025 English headwords that belonged to the general vocabulary and words from specific fields, such as anatomy or biology. It contained explanatory notes and abbreviations as well, but those were eliminated in the text pre-processing phase.

As a result, these two English-Slovak dictionaries together had 41,516 English headwords. They overlapped in less than 10,000 headwords.

3. Merging Dictionaries

The first step required to merge English-Estonian and English-Slovak dictionaries was to extract their common English word list. According to this list, every Estonian translation was aligned with every Slovak translation, which caused an exponential increase of the translations, and it aligned together words with different meanings (see Figure 1). This merging created two dictionary directions: Estonian to Slovak and Slovak to Estonian. As mentioned above, in the next steps, only the Estonian-Slovak direction was processed.

The first version of the Estonian-Slovak dictionary contained 34,674 Estonian headwords.

It was necessary to perform manual control on a random sample of 1,000 translations to analyse the main mistakes after merging, so the solutions could be adjusted accordingly. It was also essential to find out the proportion of the correct and incorrect translations. According to the control performed, only around 25% were correctly aligned translations. The remaining translations consisted of mistakes.

[2] https://www.eki.ee/litsents/ (21 March 2021).

[3] https://www.dict.cc/ (21 March 2021).

[4] https://sourceforge.net/projects/dictionarymid/ (21 March 2021).

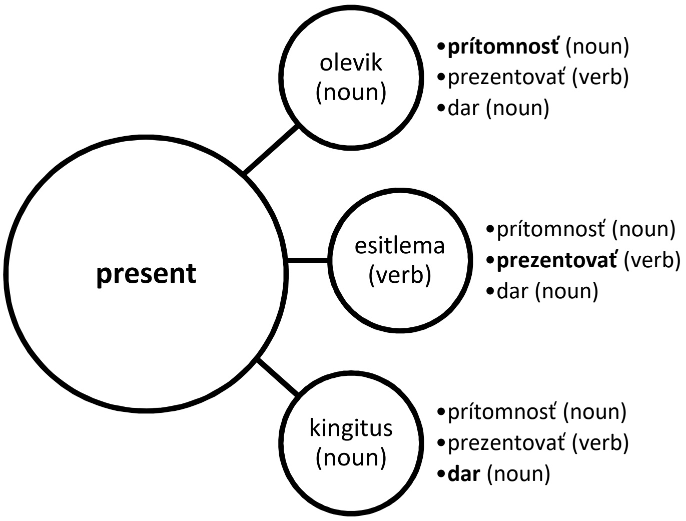


Figure 1: Aligning the Estonian words *olevik*, *esitlema, and kingitus* with the Slovak words *prítomnosť*, *prezentovať* and *dar* according to the English word 'present'

As it was assumed, most of the mistakes were caused by the ambiguity of the words, for instance, the diversity of the parts of speech typical for the English words (present (noun) vs to present (verb)) (see Figure 1). Moreover, it included the words describing nationalities, language groups and countries as in the word 'Italian', which could be in Estonian *itaallane* (nationality) or *itaalia keel* (Italian language). These types of mistakes occurred in 75% of all incorrectly aligned translations in the control group.

Other translations, around 7%, consisted of misspelled words, rarely used words, non-lemmas, proper nouns, or foreign words from other languages.

The analysis of Estonian headword lists revealed that there were incorrect headword candidates. 14% of them were multiword expressions. Multiword expressions were considered as word sequences with some unpredictable properties (Parmentier & Waszczuk, 2019). They were manually detected during the control, and this group included mainly expressions untypical for learners' dictionaries, e.g. in Estonian *graafiliselt esitama* 'to present graphically' or *tuhast puhastama* 'to clean from the ash'.

2.5% of the Estonian headwords contained a hyphen, usually prefixes such as *eba-* 'un' or *silbi-* 'syllabic'. These headwords were easily removable, but the question was whether and which of these headwords are relevant for the learners, and thus worth keeping in the dictionary.

There are several possible explanations for why those incorrect headword candidates appeared in the headword list. The first is that English headwords' descriptive translations became a headword in Estonian and Slovak while merging the dictionaries. Other possibilities are that mistakes were made during the text pre-processing, or potential mistakes were already in the original dictionaries.

In the next section, the techniques applied for solving the errors mentioned above are stated.

4. Applied Techniques

This section describes the techniques which were applied after merging the dictionaries. All those techniques were adjusted to the mistakes revealed during the manual control described in the third section. They were chosen to efficiently eliminate as many incorrectly aligned translations or incorrect Estonian headword candidates as possible while maintaining the correct translations.

4.1 Alignment based on the part of speech

The first technique assumed that translations have the same part of speech in both languages. This means that the Slovak translation of an Estonian noun is a noun as well, and all word pairs with a different part of speech are incorrectly aligned together. This solution addressed the problem with the word classes' diversity of the English words (see Figure 1).

EstNLTK library version 1.4.1.[5] was used for annotating Estonian headwords, and the web application Slovak POS Tagger[6] developed by the Slovak University of Technology in Bratislava was chosen for Slovak translations.

The main problem while using this technique was with the accuracy of the libraries. Tagging libraries give more accurate results when the context of the word is available. However, there were no contextual words in this case, and the morphological analysis proposed only one part of the speech tag that could but did not have to be the correct one. For instance, Estonian verbs in a past passive participle form (e.g., *teatud* 'done') are translated into Slovak as adjectives, but the EstNLTK tagger marked them in different cases, either as verbs or adjectives, e.g. *tagatud* 'guaranteed' as a verb, *maetud* 'buried' as an adjective.

Another problem was that the Slovak tagger did not recognise around 13% of all Slovak translations and marked them as unknown, which reduced the number of aligned translations to compare. Between those words were rarely used words, inflected word forms or multiword expressions containing spelling errors. However, any multiword expressions could not be included in the part of speech comparison because they received a POS tag according to the first word in the expression. Although usually, the last word determines the part of speech of the whole expression.

[5] https://github.com/estnltk/estnltk (21 March 2021).

[6] http://morpholyzer.fiit.stuba.sk:8080/PosTagger/. The accuracy when choosing a single tag is 65%. (21 March 2021).

While comparing tags between aligned translations, only nouns, verbs, adjectives, numerals, and pronouns were considered, since the other word classes groups were more likely to contain mistakes and incorrect differences between the tags given by the libraries. Moreover, only Estonian headwords with more than one Slovak translation were included. This measure was taken because if the dictionary contained Estonian headwords aligned with a single Slovak translation with a different part of speech, the automatic comparison would remove it from the dictionary. This would result in the loss of the headwords while the objective was not only to remove aligned translations, but also to maintain the vocabulary.

As a result of this technique, around 25% of all aligned translations were removed from the dictionary (in contrast to the number of aligned translations occurring in the dataset before the part of speech comparison), which was a satisfactory result.

4.2 Comparison with WordNet and Google Translation datasets

The second technique that was applied was the extraction of new bilingual datasets and comparison of the results across them. One of the available language resources for both languages was WordNet.[7] WordNet is a network that connects words according to their semantic relationships, while every word carries its own index. According to this index, words can be looked up in wordnets in different languages. Although there are limitations of WordNet (Pedersen & Braasch, 2009), in this project it was considered a trustworthy language resource since it was made manually by lexicographers (compared to half-automatically derived resources).

Estonian WordNet[8] and Slovak WordNet[9] were used for these purposes. Words with the same index, which indicated an equivalent synonym, were matched together and thus created a new Estonian-Slovak dictionary with 6,829 translations. In comparison to the original Estonian-Slovak dictionary, only 1,254 translations occurred in both dictionaries. It was a very small number, as in the first extracted dictionary were 156,180 translations.[10]

WordNet can be used in several different ways. One option is to measure the distances between the words computed via the Open Multilingual WordNet module in the NLTK library[11] or EstNLTK. However, the similarities measurement works within one language, not across the languages. Additionally, the intersection between Estonian and

[7] https://wordnet.princeton.edu/ (7 April 2021).

[8] https://www.cl.ut.ee/ressursid/teksaurus/index.php?lang=et (21 March 2021).

[9] https://korpus.sk/WordNet.html (21 March 2021).

[10] After splitting the translation into the format – one Estonian headword with one Slovak translation per row.

[11] https://www.nltk.org/ (28 March 2021).

Slovak WordNets was trifling in terms of receiving reasonable results.

Another option for this technique to work was to extract another dictionary so the results could be more accurate. Using Google Translate API[12] appeared as a good option. All Estonian headwords from the original dictionary and the WordNet dictionary were extracted and translated via the Google Translate API library into the Slovak language. The result was a third Estonian-Slovak dictionary with 45,178 translations.

The Google API derived dataset was manually checked on a control group consisted of around 700 randomly chosen translations. The reason was to assess the quality for the next steps. Different types of mistakes were revealed; for example, headwords translated using the same word (*tõtlikult – tõtlikult*)[13] or headwords translated into languages other than Slovak (*ebaloomulikkus – unnaturalness*). Additionally, errors caused by polysemy appeared. For instance, the Estonian word *sepikoda* 'forge shop' was translated into Slovak as *falšovať* 'to fake', where both words came from the English word 'forge' containing both meanings. The percentage of correct translations in the control group was around 55%, slightly more accurate than the original Estonian-Slovak dictionary.

These three datasets were sufficient to make comparisons. The idea was to give a score to every translation according to its occurrences in the datasets. If the translation occurred only in the Google dataset or only in the original one, it received a score of 0.25. The score for the WordNet dataset was the highest - 0.5. Thus, if the translation was in all three datasets, it got a score of 1. If found in the WordNet and Google datasets it obtained a score of 0.75, etc. The logic behind this was that WordNet is the most trustworthy resource since it was compiled manually, whereas the other datasets were automatically derived.

The success rate of this technique depends on how many resources are available to compare. Naturally, most of the translations received a score of 0.25, and the smallest group consisted of translations with a score of 1 (see Table 1). On the other hand, this technique could serve as an indicator for users as to what extent they can rely on a current translation. Moreover, the score indicates which group of translations should be corrected when manually post-editing. Additionally, each score group was manually checked on a random sample of 500 translations, and the results are described in more detail in Section 5.

[12] https://cloud.google.com/translate/docs/ (21 March 2021).

[13] Those were easily removed.

The number of word pairs	**Score**
178,678	0.25
15,194	0.5
1314	0.75
502	1

Table 1: Comparison with WordNet and Google Translation datasets.

4.3 Comparison of headword list in EKI Combined Dictionary

This technique aimed to eliminate words that are not usually given as a headword in general-purpose dictionaries, e.g. proper names, inflected word forms, misspelled words, abbreviations, or foreign words from other languages. The EKI Combined Dictionary[14] and its user interface Sõnaveeb (Tavast et al., 2019) were used for these purposes. The EKI Combined Dictionary contains rich linguistic information about Estonian words. This technique's objective was to look up every Estonian headword automatically. Suppose the headword could not be found in the EKI Combined Dictionary; in that case, it could be eliminated from the dictionary since this technique assumed that the EKI Combined Dictionary contains all relevant words for users or learners.

As a result, 10,014 Estonian headwords were not found in the EKI Combined Dictionary. Manual control was performed on a random sample of 1,000 translations. The biggest group consisted of multiword expressions, around 72%. The rest of them made up foreign words from other languages, e.g. 'capriccio' or 'curling', proper names or rarely used words.

The problem with the inflected word forms persisted, as automatic searching through the EKI Combined Dictionary allows to look up a lemma of an inflected word form. The words found in the EKI Combined Dictionary also contained words with a comparison score of 0.75 or even 1 (see Section 4.2.). This was exactly 123 words (e.g., *varastaja* 'thief', *aadlinaine* 'noblewoman'), so the decision was to keep such words in the dictionary as headwords.

4.4 Comparison of named entities

This last technique focused on the polysemy problem between words referring to

[14] https://metashare.ut.ee/repository/browse/the-eki-combined-dictionary-2021/af363d08857111eba6e4fa163e9d4547c858d4634fcb44eea7b56db3e452675c/ (21 March 2021).

nationalities, countries, or language groups. For instance, in English, the word 'Italian' refers to the nationality (Italian men or Italian woman) or language. This problem could be solved by using a named entity recognition library which decides about every name if it is either person (PER), location (LOC) or organisation (ORG), e.g. the Estonian word *Itaallane* – PER (Italian men). The Estonian headword tag could then be compared to its Slovak translations' tags; when the tags differ, it is an incorrectly aligned translation.

Estonian headwords were tagged by the libraries EstNLTK and Polyglot.[15] Polyglot was also used for classifying Slovak translations.

This technique was the most unsuccessful because libraries gave different results and marked the same words with different tags. For example, some countries were classified as a location, while others as an organisation, even in some cases when the translation was correct (see Table 2). An interesting choice was for the word *European Union*, which received in Slovak language organisation tag and a location tag in Estonian. Differences occurred between the libraries used for the same language. For example, Polyglot tagged the Estonian word *Mars* as a person while EstNLTK as a location.

Another problem was that the EstNLTK library did not tag nationalities, and Polyglot tagged only a few exceptions, which significantly limited the group of translations compared. Polyglot classified 1,477 Estonian headwords and 928 Slovak translations, and the EstNLTK library marked 935 Estonian headwords. Due to the small number of translations that could be compared and significant differences between the given tags, the results were not considered.

Estonian headword	**EstNLTK library**	**Slovak translation**	**Polyglot library**
• *Filipiinid* (‘Philippines’)	• ORG	• *Filipíny* (‘Philippines’)	• LOC
• *Gruusia* ('Georgia')	• LOC	• *Georgia* ('Georgia')	• ORG
• *Somaali* ('Somalia')	• PER	• *Somálsky* ('Somalian’)	• LOC

Table 2: Results of the comparison of named entities

[15] https://polyglot.readthedocs.io/en/latest/ (21 March 2021).

5. Evaluation

The resulting automatically derived dictionary consisted of 138,779 translations (28,873 Estonian headwords), and it was evaluated from two points of view. Firstly, what kind of vocabulary it contained and, secondly, which of the applied techniques helped improve the dictionary's quality and what types of mistakes persisted.

The Estonian headword lists were compared to the lemma list of the Balanced Corpus of Estonian and then to the lemma lists in each sub-corpus individually from the same corpus.[16] This corpus comprises texts from newspapers, literature, and academic texts. 92% of the words with a frequency over 5,000 were included in the Estonian headword list. When looking at the various genres, including journalism, fiction, and scientific texts, the percentages of Estonian headwords included in the dictionary with occurrences over 1,000 and 5,000 were in the range from 88% up to 94%. The results are stated below in Table 3.

	Headwords with frequency over 5,000	**Headwords with frequency over 1,000**
• Whole corpus	• 92%	• 90%
• Journalism sub-corpus	• 94%	• 91%
• Fiction sub-corpus	• 90%	• 88%
• Scientific sub-corpus	• 93%	• 92%

Table 3: The percentage of Estonian headwords from the dictionary contained in different sub-corpora frequency lists

Since written language varies from the spoken language, the comparison with the wordlist extracted from the Basic Estonian Dictionary[17] provided a more accurate picture. This dictionary was compiled for learners at A2 to B1 CEFR levels and covers the basic Estonian vocabulary. Around 85% of headwords from the Basic Estonian Dictionary occurred in the Estonian-Slovak dictionary. Missing words were, for instance, zodiac signs or the word 'me'.

When assessing the headword list with regard to the part of speech representation, the biggest group was made up nouns and adjectives, around 67% and 12%, respectively. Those were followed by verbs with approximately 10%. The remaining 11% consisted

[16] https://www.cl.ut.ee/ressursid/sagedused1/index.php?lang=en (21 March 2021).

[17] http://www.eki.ee/dict/psv/ (21 March 2021).

of adverbs, pronouns, interjections, numerals, etc.

For the second evaluation, 1,000 translations were randomly chosen and manually controlled. This control revealed that around 40% of translations are correctly aligned, which is 15% more than during the first control before post-processing. The most frequent mistakes were still related to polysemy. Specifically, incorrectly aligned translations with the same or unknown part of speech and the persisting problem with nationalities, countries, and languages. The percentage of incorrect translations caused by polysemy was approximately 92% in this control group. Compared to the initial control, it is 17% more, which means that the percentage of other mistakes decreased.

Other errors that methods could not eliminate were related to multiword expressions, misspelled words and inflected word forms in Slovak translations, Estonian non-lemmas, and translations in other languages than Slovak (e.g., English, Czech etc.).

As a result, the most successful technique appeared to be alignment based on the part of speech, where the number of wrongly connected translations fell by around 25%. This percentage could be increased by using a more accurate tagger.

A comparison with WordNet and Google Translation datasets also gave good results. Each group with a different score (1, 0.75, 0.5, 0.25) was manually evaluated on a random sample of 500 translations.[18] The manual evaluation confirmed that the given score corresponds with the error percentage in the group. The results are stated in Table 4 below. Overall, the given score can be valuable for dictionary users as an appropriateness indicator or for further dictionary development.

Score	**The Percentage of Errors**
• 1	• 0.59%
• 0.75	• 4.6%
• 0.5	• 15%
• 0.25	• 66.4%

Table 4: The percentage of errors in each score group

On the other hand, the technique using named entity recognition failed because of immense differences between the results for the exact words given by different taggers.

[18] All translations from group with score 1 were checked.

An overview of all applied methods and results is provided in Table 5.

Method	**Impact**
• Alignment based on the part of speech	• 25% incorrectly aligned word pairs removed
• Comparison with WordNet and Google Translation datasets	• See Table 4
• Comparison with the EKI Combined Dictionary	• 24% of Estonian headwords were removed
• Comparison of named entities	• The method failed due to the immense differences between tags given by tagging libraries

Table 5: Applied techniques and their results summarisation.

6. Conclusion and Future Works

This article introduced the Estonian-Slovak dictionary, automatically derived from two already existing dictionaries that shared English as their common language. Merging of the dictionaries produced many incorrectly aligned translations, where most of the errors were caused by polysemy.

Several techniques were applied to reduce the number of incorrectly aligned translations: alignment based on the part of speech, comparison with WordNet and Google Translation datasets, comparison of Estonian headword list with the EKI Combined Dictionary, and comparison of named entities. The best approach turned out to be comparing the part of speech tags between the aligned translations. In contrast, tagging word pairs with named entity recognition feature failed due to the different tags.

In the end, the quality of the dictionary was evaluated. The evaluation revealed that the dictionary consists of 138,779 translations and the Estonian headword list covers 85% of the basic Estonian vocabulary. Regarding the quality of the translations, around 40% of the translations are correct, while in the remaining roughly 60% some errors persisted, mostly caused by polysemy.

There are several options to increase accuracy. Firstly, the scoring technique could be extended by another dictionary extracted from Estonian-Slovak parallel corpora. Estonian corpora could be used to control if translations contain all relevant meanings. On top of that, the web application for this Estonian-Slovak dictionary was built, and its further development is planned.

7. Acknowledgements

I express my gratitude to my supervisor, Dr. Kadri Muischnek, MA, Ph.D., and Jelena Kallas, MA, Ph.D. for their guidance and valuable advice.

8. References

Ordan, N., Gracia J. & Kernerman I. (2017). Auto-generating Bilingual Dictionaries. In I. Kosem et al. (eds.) e*Lex 2017 Proceedings*, pp. 474-484.

Pedersen, B. S. & Braasch, A. (2009). What do we need to know about humans? A view into the DanNet database. *Proceedings of the 17th Nordic Conference of Computational Linguistics NODALIDA 2009.* Editors: Kristiina Jokinen and Eckhard Bick, pp. 158–166.

Saralegi, X., Manterola I. & San Vicente I. (2012). Building a Basque-Chinese Dictionary by Using English as Pivot. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12),* L12-1006, pp. 1443–1447.

Shirai, S. & Yamamoto K. (2001). Linking English Words in Two Bilingual Dictionaries to Generate Another Language Pair Dictionary. *Proc. of ICCPOL*, pp. 174-179.

Tanaka, K. & Umemura, K. (1994). Construction of a Bilingual Dictionary Intermediated by a Third Language. *COLING '94: Proceedings of the 15th conference on Computational linguistics*, pp. 297-303.

Tavast, A., Langemets, M., Kallas, J. & Koppel, K. (2018). Unified data modelling for presenting lexical data: The Case of EKILEX. In J. Čibej V. Gorjanc, I. Kosem & S. Krek (eds.) *Proceedings of the XVIII EURALEX International Congress. EURALEX: Lexicography in Global Contexts, Ljubljana, 17–21 July 2018. Ljubljana: Ljubljana University Press, Faculty of Arts*, pp. 749—761.

Parmentier, Y. & Waszczuk, J. 2019. Preface. In Yannick Parmentier & Jakub Waszczuk (eds.), *Representation and parsing of multiword expressions: Current trends*, iii–ix. Berlin: Language Science Press.

Websites:

Basic Estonian Dictionary. Accessed at: http://www.eki.ee/dict/psv/. (21 March 2021)

dict.cc. Accessed at: http://www.dict.cc. (21 March 2021)

DictionaryForMIDs. Accessed at: https://sourceforge.net/projects/dictionarymid/. (21 March 2021)

EKI Combined Dictionary. DOI: 10.15155/3-00-0000-0000-0000-08979L. Accessed at: https://metashare.ut.ee/repository/browse/the-eki-combined-dictionary-2021/af363d08857111eba6e4fa163e9d4547c858d4634fcb44eea7b56db3e452675c/. (21 March 2021)

English-Estonian dictionary. Accessed at: https://www.eki.ee/litsents/. (21 March 2021)

EstNLTK. Accessed at: https://github.com/estnltk/estnltk. (21 March 2021)

Estonian-Slovak dictionary web application. Accessed at: https://estonian-slovak-dictionary.herokuapp.com/index/ee. (23 March 2021)

Estonian WordNet. Accessed at: https://www.cl.ut.ee/ressursid/teksaurus/index.php?lang=et. (21 March 2021)

Frequency lists of The Balanced Corpus of Estonian. Accessed at: https://www.cl.ut.ee/ressursid/sagedused1/index.php?lang=en. (21 March 2021)

Google translation API. Accessed at: https://cloud.google.com/translate/docs/ (21 March 2021)

Natural Language Toolkit. Accessed at: https://www.nltk.org/ (28 March 2021)

Polyglot. Accessed at: https://polyglot.readthedocs.io/en/latest/. (21 March 2021)

Slovak POS Tagger. Accessed at: http://morpholyzer.fiit.stuba.sk:8080/PosTagger/. (21 March 2021)

Slovak Wordnet. Accessed at: https://korpus.sk/WordNet.html. (21 March 2021)

Sõnaveeb. Accessed at: https://sonaveeb.ee/. (21 March 2021)

WordNet. Accessed at: https://wordnet.princeton.edu/. (7 April 2021)

The Distribution Index Calculator for Estonian

Ene Vainik[1], Ahti Lohk[1], Geda Paulsen[1, 2]

[1] Institute of the Estonian Language, Roosikrantsi 6, Tallinn 10119, Estonia
[2] Uppsala University, Thunbergsvägen 3 L, Uppsala 75126, Sweden
E-mail: Ene.Vainik@eki.ee, Ahti.Lohk@eki.ee, Geda.Paulsen@eki.ee

Abstract

Lexicographers working with such morphologically rich languages as Estonian face the task of detecting the lexicographic status of some word forms that look like case forms of nouns but can behave as function words to a certain degree. Hence, a measurable criterion for making a word form an autonomous headword is needed. The present paper describes the idea and development of a tool called the Distribution Index Calculator (DIC) for Estonian. It is a web-based application which finds the frequency data of word forms and lemmas from an annotated corpus and retrieves a statistic called the Distribution Index (DI). The DI indicates the relative prominence of a word form as compared to its expected normative level of salience. The application is described in detail and some illustrations of its performance are provided. The evaluation of its quality is as follows: a higher than critical level of DI can be trusted as an indicator of the relative autonomy of a word form, while a lower than critical level of DI does not preclude such autonomy. The DIC thus gives relative heuristics rather than absolute ratings or true-value decisions.

Keywords: language technology; lexicography; morphology; distribution of case forms; the Estonian language

1. Introduction

There is an endless source of candidates for new dictionary headwords in the era of e-dictionaries and automated compilation processes. This is so not only because of such obvious neologisms as *koroonaviirus* ‘coronavirus’ and *karjaimmuunsus* ‘herd immunity’, but also because of the effort to present fairly established word forms as autonomous headwords in a dictionary. The latter holds when such autonomy is justified, i.e. when the lexical items serve a function or meaning distinguishable from the base word (e.g. Blensenius & Martens, 2019).

Lexicographers working with such morphologically rich languages as Estonian face a specific task: to detect the lexicographic status of word forms that look like case forms of nouns but can behave as function words to a certain degree (e.g. *sõnul* : is it the noun *sõna* ‘word’ in plural adessive or the indecomposable adposition *sõnul* ‘according to (someone’s) claim’ (Karelson, 2005; Paulsen et al., 2019)). The task is to establish the degree of emancipation of such word forms from the noun paradigm, and thus provide a justification for upgrading them to the status of independent headwords in dictionaries. A similar task in languages lacking case form morphology is, for example, establishing the lexicographic status of plural forms or derivatives. Practical decisions about whether to include a word form as a headword or not have to be made by lexicographers daily. Hence, a measurable (synchronic) criterion for word form emancipation is needed.

We can now introduce the first working prototype of the DIC[1]. Below, we refer to the theoretical underpinnings briefly, and describe the idea behind the statistic and its calculation. We also give the details of its realisation as an eight-line pseudocode and present some illustrations of how it works. The evaluation of the results was carried out as an experiment comparing the results of the DIC with the decisions made by lexicographers. The problems and future directions of development are also discussed.

1.1 Some notes about the theoretical background

The ubiquitous process of grammaticalisation offers a theoretical explanation for the phenomenon of developing new function words out of case forms of nouns (Grünthal, 2003; Habicht et al., 2011). A process called lexicalisation could be considered at play as well, as far as we talk about the emergence of new lexical units: the stand-alone headwords in a dictionary (for more references and discussion see Paulsen et al., 2021).

In Estonian, there are both already fossilised lexemes (e.g. *kõrval* 'beside' in (1b)) and (continually new) forms on their way to the status of lexical items (e.g. *äärel* 'on the edge' in (2b)) (see e.g. Karelson, 2005; Paulsen et al., 2019), which require the attention of a lexicographer:

(1) a. *Koera **kõrva-l** istub kärbes.*
dog.GEN ear-ADE sit-3SG fly
'A fly is sitting on the dog's ear.'

b. *Laps istub koera **kõrval***
child sit-3SG dog.GEN aside
The child is sitting next to the dog.'

(2) a. *Mees kõnnib katuse **ääre-l**.*
man walk-3SG roof.GEN edge-ADE
'The man is walking on the edge of the roof.'

b. *Valitsus on kokkukukkumise **äärel**.*
government is-3SG collapse.GEN edge
'The government is on the brink of collapse.'

It has been established that there are two types of processes that take place in the grammaticalisation of a lexical item: 1) semantic change from a referential meaning to a grammatical meaning (Hopper & Traugott, 2003: 1 (also called bleaching, see Heine, 2005: 578-579)), and 2) increase in the usage of a word form (see e.g. Feltgen et al., 2017). The two processes appear simultaneously. We can only think of the frequency of usage being a prerequisite for the semantic change. The essence of the process is that the lexical item is used more frequently and in different contexts than it was used before when it carried only lexical meaning. The acquired new aspects of meaning (or new functions) further reinforce the more frequent usage.

[1] teenus.eki.ee/d-index

The DIC described here can provide information only about the increase in relative frequency. The implications of semantic change must be tackled in a separate module of a future lexicographic tool.

2. The Distribution Index and its calculation

Information about the relative frequency of word forms can be helpful when it comes to deciding whether a particular word form should be given the status of a headword in its own right. We have proposed an index of a statistical distribution of word forms (DI) as a heuristic for lexicographers (Vainik et al., 2021; Paulsen et al., 2021).

The idea behind the proposed DI lies in the assumption that proper forms of nouns tend to have constant distributions along with the case forms (combinations of number and case, e.g. plural elative and singular abessive) in the corpora. Based on the knowledge of normal distribution, it is possible to predict the frequencies of word forms on the basis of their lemma frequencies. The idea of the DI is to compare the actual (observed) frequency of a case form in a corpus with its expected frequency. The values of expected and observed frequency should be equal or close if the studied form follows the normal distribution. If there is a considerable difference between the values of expected and observed frequencies, one can conclude that there is an abnormal distribution.

2.1 Normal distribution of the case forms

The normal distribution of Estonian case forms was established in a previous study (Vainik et al., 2021). In that work, the distribution data of case forms from two annotated corpora — the balanced corpus of Estonian and the morphologically tagged corpus — were compared in order to control for the constancy of the proportions. The distribution of all of the case forms (i.e. 29 combinations of number and case) demonstrated very steady proportions in both of the corpora ($r = 0.999$; StDev 0.000). The mean values of the two corpora were established as the norms (see Table 1).

The norms were deduced relying on data on all types of declinable word classes: nouns, adjectives, numerals and pronouns. As such, the norms serve as generalised benchmarks for comparison.

Case	**DIC**	**Leipzig Glossing**	**Singular**	**Plural**
nominative	n	NOM	0.262	0.068
genitive	g	GEN	0.217	0.053
partitive	p	PART	0.102	0.037
additive	adt	ADT	0.011	
illative	ill	ILL	0.005	0.002
inessive	in	INE	0.042	0.007
elative	el	ELA	0.028	0.009
allative	all	ALL	0.028	0.008
adessive	ad	ADE	0.044	0.010

Case	DIC	Leipzig Glossing	Singular	Plural
ablative	abl	ABL	0.004	0.001
translative	tr	TRA	0.027	0.002
terminative	ter	TER	0.002	0.000
essive	es	ESS	0.004	0.001
abessive	ab	ABE	0.001	0.000
comitative	kom	COM	0.021	0.006

Table 1. Normal distribution of declinable words in Estonian

2.2 Formula for calculating the DI

In order to calculate the DI for an ambiform (i.e. a word form ambiguous in respect to its lexicographic status, also referred to as a wicked word form later in this paper), we need to guess which case form of which particular lemma it might be, i.e. the word form has to undergo tentative morphological analysis. For example, the word form *sõnul* would be interpreted tentatively to be the plural adessive case form of the lemma *sõna* 'word'.

To calculate the DI we need: 1) the observed frequency of the word form in a corpus (Z), 2) the norm of that particular case form (number + case) taken from a table of such norms (e.g. Table 1), and 3) the frequency of the lemma in a corpus (X). The DI is calculated according to the following formula:

$$DI = (Z - X \times Y) / X$$

2.3 The scale of DI values

The value of the DI can (theoretically) vary from nearly -1 to 1. Values near zero indicate normal distribution, and negative values indicate that the word form is under-represented as compared to its expected frequency. Values above zero indicate that the word form is used more often than expected by the norm. On a few occasions, a value can exceed 0.9, which indicates that the frequency of the lemma and the frequency of case forms are very close, i.e. the word occurs mostly in a certain case form. For example, *tikutulega* [match light-COM] '(search) diligently' occurs 2,547 times and the lemma *tikutuli* 'match light' 2,587 times in ENC2019. Lemmas of such case forms lack the normal paradigm, and their distribution is far from normal.

In an empirical study that compared the DIs of proper case forms to ambiforms, we were able to establish a tentative threshold value of DI (0.130). Values equal to or greater than the threshold are considered to show abnormal distribution (Vainik et al., 2021). Values higher than zero but lower than the threshold show moderate deviation from the normal distribution. The tentative scale of values and labels is presented in Table 2.

Values of DI	Label
< -0.05	*normist väiksem* 'under-represented'
-0.05 ... < 0.05	*normaalne* 'normal distribution'
0.05 ... < 0.130	*normist suurem* 'moderate over-representation'
0.130	*kriitiline* 'critical over-representation'

Table 2. Values and labels used in DIC

3. Description of the development of the calculator

3.1 The designed DIC functionalities

The DIC is a web-based application accessible to everyone. It takes an *ambiform* as input from the user and retrieves corpus data (frequencies of the word form and the suspected lemma), as well as the suspected morphological form. The tool calculates the distribution index of the input form and compares it to the ranked scale of word form emancipation. The DIC provides the outcome with a verbal label of the detected tendency of the distribution. The labels reflect the values determined in Table 2 (see the previous section): normist väiksem ('under-represented'), normaalne ('normal'), normist suurem ('moderate'), and kriitiline ('critical').

3.2 Prerequisites for building the DIC application

There are some inevitable prerequisites for creating the DIC application: 1) knowledge of the valid normal distribution of case forms (number + case, abstract), 2) the established scale of DI values, 3) the availability of an expeditious module for morphological analysis, and 4) the availability of a morphologically annotated corpus for retrieving the frequency data of forms and lemmas.

3.3 The main components of the application

The DIC application is written in the Python programming language and it uses the micro web framework Flask. Due to the specifics of the application, it is necessary to use two software components: one that performs a morphological analysis of the entered ambiform and another that requests statistical information about the frequency of the ambiform and its potential base forms from a representative corpus of texts.

The morphological analysis has to provide information about lemmas, parts of speech and the forms corresponding to the ambiform. In the current prototype, we use EstNLTK (version 1.6.7), which is a natural language toolkit for Estonian written in Python. It provides resources for basic NLP tasks: tokenisation, morphological analysis, lemmatisation, named entity recognition etc. (Orasmaa et al., 2016: 2460). Alternative tools for morphological analysis, such as R-package UDPipe[2], are not available yet[3]. The EstNLTK toolkit also seems natural because its tagging system coincides with that used by Sketch Engine: the platform that lexicographers are most familiar with. From a practical viewpoint, it is preferable to avoid discrepancies in tagging, e.g. it would be helpful to find similar long-tags when it comes to looking into the concordances of the particular ambiform in SketchEngine.

The second component of the DIC makes automated HTTP requests to the Estonian National Corpus 2019 (ENC2019), which is available on the SketchEngine platform. The requests are performed by using the Sketch Engine API[4]. ENC2019 is currently the newest and largest automatically annotated corpus of the Estonian language (approx. 1.5 billion words). The corpus is annotated with the EstNLK toolkit (version 1.6.7). The precision of the annotation is not yet known. Some problems with the compilation and annotation processes of Estonian corpora are discussed by Koppel (2020).

3.4 The DIC algorithm

The DIC algorithm performs a sequence of activities when calculating the D-index. The sequence is provided by an eight-line pseudocode, as follows (and explained below):

```
1: word ⟵ user entered ambiform
2: norm_freq ⟵ read_from_file
3: lemmas, postags, forms ⟵ estnltk_morf_anal(word)
4: for i ⟵ [1, …|lemmas|]:
5:        X, Z ⟵ query_from_SkE(word, lemmas[j], postags[j])
6:        Y ⟵ norm_freq[forms[j]]
7:        D_index ⟵(Z - X * Y) / X
8:        DI_label ⟵ find_di_label(D_index)
```

Rows 1 and 2: A user enters the input data —a word— and the norm_freq is read from a file. The norm_freq is the normal (expected) distribution of word forms, and it is previously specified based on the balanced corpus of Estonian and the morphologically disambiguated corpus (see section 2.1 above).

Row 3: All of the possible lemmas, postags, and forms are found for the entered word using the Estonian morphological analyser estnltk_morf_anal (EstNLTK is the Python library for Estonian language processing and analysis; see section 3.3 above).

Row 4: Repeat the sentences in rows 6 and 9 as many times as there are elements in the

[2] See more https://www.rdocumentation.org/packages/udpipe/versions/0.8.5, https://www.r-bloggers.com/2018/02/a-comparison-between-spacy-and-udpipe-for-natural-language-processing-for-r-users/, https://universaldependencies.org/

[3] Kairit Sirts, personal communication.

[4] See more at https://www.sketchengine.eu/documentation/api-documentation/

lemmas list (`|lemmas|`).

Row 5: The `query_from_SkE` method queries SketchEngine based on the `word`, from which we separate the information about the frequency of occurrence of the word (`Z`) and the frequency of occurrence of the `lemmas[j]` at the `postags[j]` (`X`).

Row 6: The program finds the norm proportion `Y` for the `forms[j]` in the dictionary `norm_freq`.

Row 7: Based on `X`, `Y` and `Z, the D_index` is calculated.

Row 8: Using the predefined scale, the `find_di_label` method is used to find the rating label (`DI_label`) corresponding to the `D_index`.

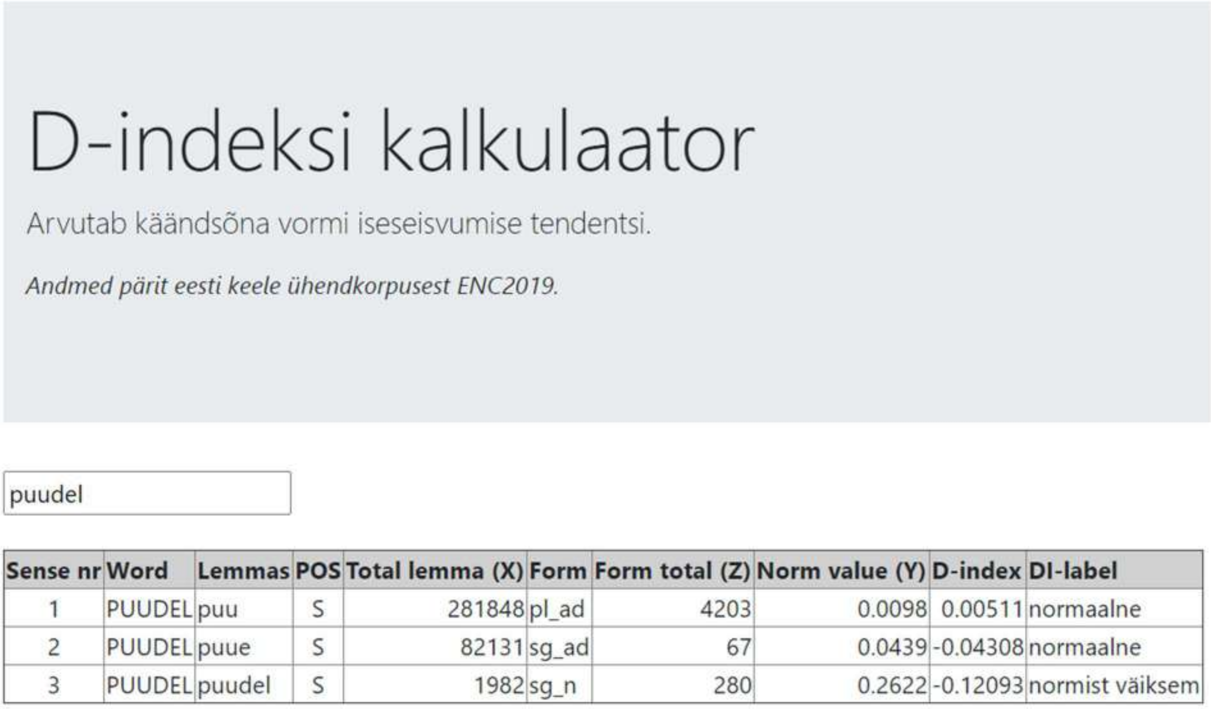
The number of D-indices of a single word (nominal) depends on how many initial lemma-postag forms morphological analysis and SkE query yield.

4. The DIC at work

The DIC works on the web. It can be opened in a separate window of a web browser while working in Ekilex or checking corpus data via SketchEngine. It is supported by the most common browsers (it has been tested on Microsoft Edge, Mozilla Firefox, Chrome, Vivaldi, and Brave).

Figure 1 presents the user interface of the DIC. The title translates as “A calculator of D-index” and the subtitle as “It calculates an autonomy tendency for case forms of declinable words” and “The data is retrieved from the corpus ENC2019”. There is a search box below the title and further below are situated tabular fields for the results of a query. There will be as many rows presented as there are different interpretations provided by the morphological analysis.

The form entered, *puudel,* has three homographic readings as different case forms (plural adessive, singular nominative and singular adessive, respectively) of three different lemmas: *puu* ‘tree’, *puudel* ‘poodle’, and *puue* ‘disability. The distribution rates and labels of these interpretations are presented in the last two columns. It can be concluded that the frequency of the form *puudel* is normal or below, no matter for which case and lemma it stands.



D-indeksi kalkulaator

Arvutab käändsõna vormi iseseisvumise tendentsi.

Andmed pärit eesti keele ühendkorpusest ENC2019.

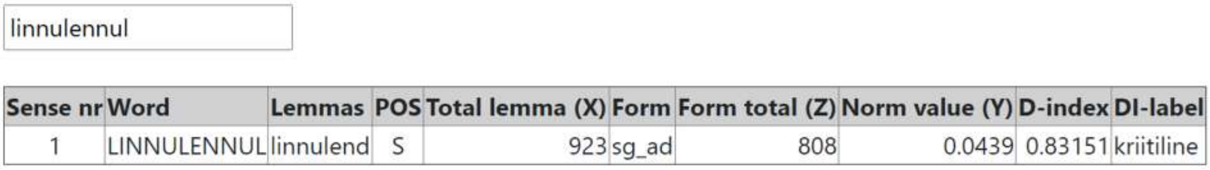
puudel

Sense nr	Word	Lemmas	POS	Total lemma (X)	Form	Form total (Z)	Norm value (Y)	D-index	DI-label
1	PUUDEL	puu	S	281848	pl_ad	4203	0.0098	0.00511	normaalne
2	PUUDEL	puue	S	82131	sg_ad	67	0.0439	-0.04308	normaalne
3	PUUDEL	puudel	S	1982	sg_n	280	0.2622	-0.12093	normist väiksem

Figure 1. The user interface of DIC.

4.1 Illustrations

In the following, we present some examples of how the DIC works. Here is a short list of word forms: *kombel, lahus, nõusolekul, linnulennul, peensusteni, alguses, habemega, lehes* and *sõlmes*. The results of the analysis are presented in Figure 2 (a—k). We have omitted the title sections to save space. The illustrations are grouped in descending order according to their DI values (and labels).

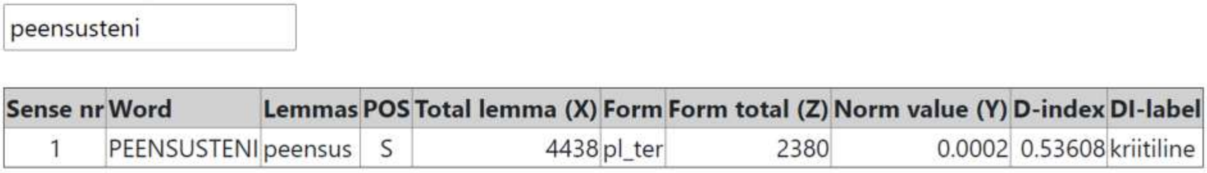


linnulennul

Sense nr	Word	Lemmas	POS	Total lemma (X)	Form	Form total (Z)	Norm value (Y)	D-index	DI-label
1	LINNULENNUL	linnulend	S	923	sg_ad	808	0.0439	0.83151	kriitiline

a) *linnulennul* [bird.fly-ADE] 'very fast'

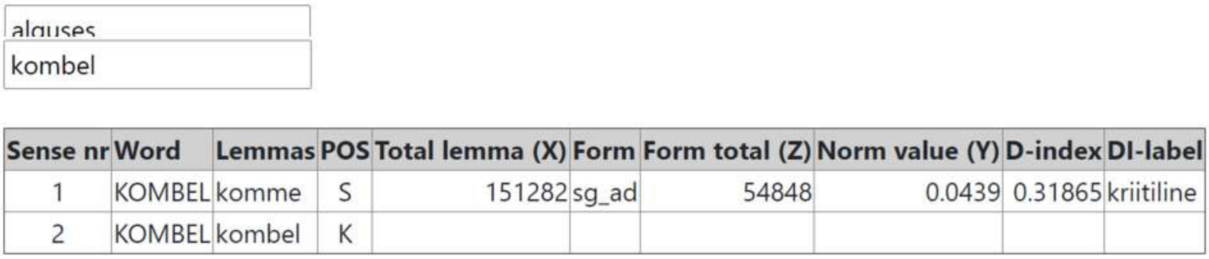
b) *peensusteni* [detail-PL.TER] 'scrupulously'



peensusteni

Sense nr	Word	Lemmas	POS	Total lemma (X)	Form	Form total (Z)	Norm value (Y)	D-index	DI-label
1	PEENSUSTENI	peensus	S	4438	pl_ter	2380	0.0002	0.53608	kriitiline

c) *alguses* [beginning-INE] 'at the beginning'



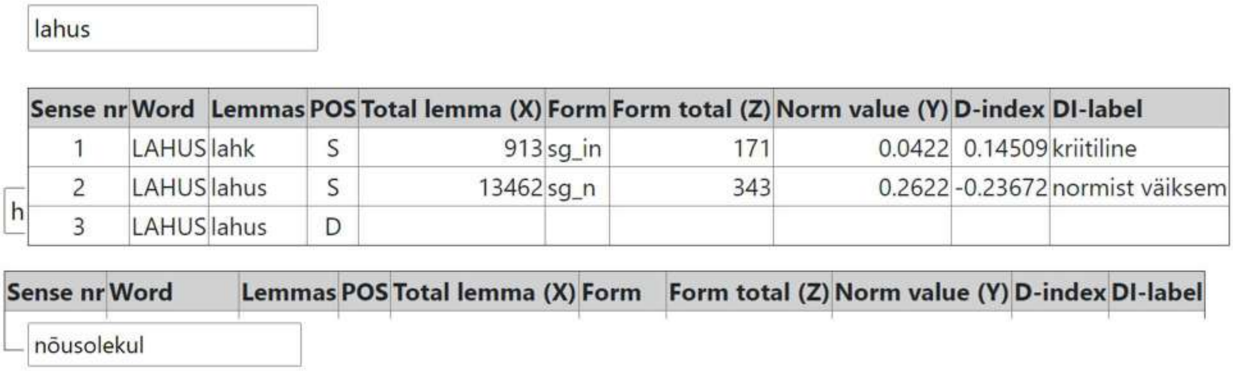
alguses

kombel

Sense nr	Word	Lemmas	POS	Total lemma (X)	Form	Form total (Z)	Norm value (Y)	D-index	DI-label
1	KOMBEL	komme	S	151282	sg_ad	54848	0.0439	0.31865	kriitiline
2	KOMBEL	kombel	K						

d) *kombel* [manner-ADE] ‘in a way’

e) *habemega* [beard-COM] ‘outdated’

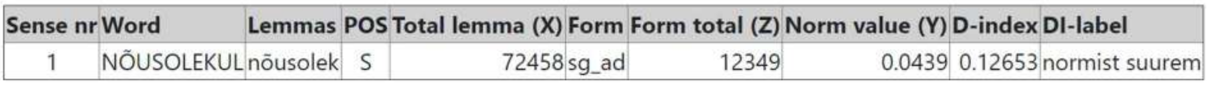


lahus

Sense nr	Word	Lemmas	POS	Total lemma (X)	Form	Form total (Z)	Norm value (Y)	D-index	DI-label
1	LAHUS	lahk	S	913	sg_in	171	0.0422	0.14509	kriitiline
2	LAHUS	lahus	S	13462	sg_n	343	0.2622	-0.23672	normist väiksem
3	LAHUS	lahus	D						

Sense nr	Word	Lemmas	POS	Total lemma (X)	Form	Form total (Z)	Norm value (Y)	D-index	DI-label

nõusolekul



Sense nr	Word	Lemmas	POS	Total lemma (X)	Form	Form total (Z)	Norm value (Y)	D-index	DI-label
1	NÕUSOLEKUL	nõusolek	S	72458	sg_ad	12349	0.0439	0.12653	normist suurem

f) *lahus* [division-INE] ‘separated (from)’

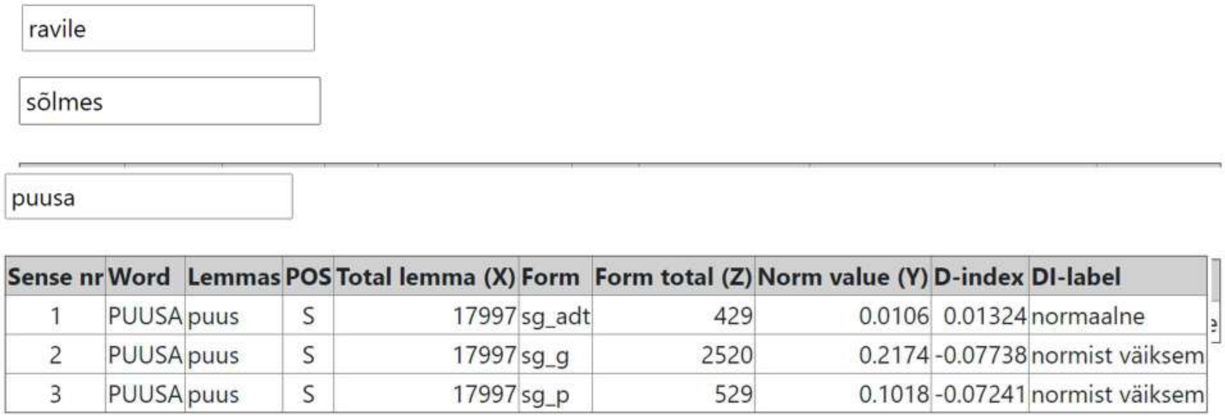
g) *nõusolekul* [agreement-ADE] ‘with the agreement of’

h) *ravile* [cure-ADE] ‘to a treatment’

i) *sõlmes* [knot-INE] ‘tangled’

j) *lehes* [leaf-INE] ‘covered with fresh leaves’; *lehes* [newspaper-INE] ‘in a newspaper’

k) *puusa* [hip-ADT] ‘(to) akimbo’



ravile

sõlmes

puusa

Sense nr	Word	Lemmas	POS	Total lemma (X)	Form	Form total (Z)	Norm value (Y)	D-index	DI-label
1	PUUSA	puus	S	17997	sg_adt	429	0.0106	0.01324	normaalne
2	PUUSA	puus	S	17997	sg_g	2520	0.2174	-0.07738	normist väiksem
3	PUUSA	puus	S	17997	sg_p	529	0.1018	-0.07241	normist väiksem

Figure 2. Illustrations of the DIC at work

It appears that the critical values of the DI vary considerably (from 0.8 down to the threshold value of 0.130). High D-indices can characterise forms with high, moderate and low lemma frequencies in absolute terms (compare c, b and a in Figure 2, for example). This is also the case with a normal distribution (compare i and j in Figure 2, for example). The comparability of the distributions, independent of the frequencies of forms or lemmas in absolute terms, is considered to be the advantage of the DI as a statistic (see Paulsen et al, 2019; Vainik et al., 2021). The examples d, f, i and k in Figure 2 illustrate the case when a form has more than one interpretation according to the corpus tagging. In some cases, there are homographic readings of the ambiform (e.g. f in Figure 2, where the form *lahus* can be interpreted as belonging to two alternative lemmas: *lahk* ‘division’ and *lahus* ‘dilution’). Another kind of multiplicity of interpretations originates in the decategorisation of certain case forms of nouns (e.g. d, f, i and k in Figure 2; see also Paulsen et al., 2019) and interpreting those as indeclinable words (adverbs — D, adpositions — K). Decategorisation may or may not diminish the DI value as a case form (as in i and d in Figure 2, respectively). The effects of decategorisation and the accuracy of corpus tagging are discussed in more detail in another paper (Paulsen et al., 2019). The case in j where the word *leht* is polysemous is the most complicated. The calculator is unable to distinguish the meanings and sums up all of the occurrences of both the form and its lemma. Thus, the potential over-representation of a form in one particular meaning, e.g. 'covered with fresh leaves', will go unnoticed on a purely statistical basis.

4.2 Evaluation of the DIC and its results

4. 2.1 Quantitative parameters

A single query by DIC took 1-1.5 seconds on average during the test period of the prototype. We noticed delays, occasionally, at times when Sketch Engine was slow anyway (for unknown reasons). The speed of the DIC is related to the smoothness of queries by Sketch Engine because the DIC retrieves its frequency data via the Sketch Engine API (see Section 3.3).

4.2.2 Quality of the results

The quality of the DIC can be estimated by comparing its output with some kind of approved standard. It is reasonable to assume that the decisions made by lexicographers so far can be used as a standard in this respect. As the problem to be solved by the assistance of the DIC is whether to include a particular word form in a dictionary as a stand-alone headword or not, we can use the DI level of the case-form-like approved headwords as a standard.

In the following, we describe the experiment of calculating DIs for a set of not yet established word forms and comparing their DI levels with similar case forms of nouns that have been approved as headwords in the CombiDic. We chose headwords from the CombiDic that are analysed as case forms only, in corpus texts by Vabamorf, and whose DIs thus purely represent their distribution as nouns and are not distorted by occasional decategorisation (see Paulsen et al., 2019 for discussion).

Table 3 presents the 30 ambiforms with their DIs based on the data of ENC2019[5]. The rows are arranged so that shared forms (number + case) are presented together. The groups are accompanied by data on their number in our database and average DI levels, as well as by examples of some headwords from the CombiDic, with the maximum and minimum value of the DI in each subcategory. The DI values exceeding the tentative threshold (0.130) set by the previous research (Vainik et al., 2021) are boldfaced in Table 3.

Eleven ambiforms out of 30 appeared to demonstrate critical over-representation (0.130), eight demonstrated moderate over-representation (0.05 ... 0.129) and eleven ambiforms

[5] An excerpt from our database of such ambiforms; see Paulsen et al. (2019).

demonstrated normal distribution (-0.04 ... 0.04).

A comparison with the DI values of the approved case-form-like headwords shows that their average well exceeds the threshold, which indicates that the approved forms generally tend to be distributed abnormally. There is remarkable variation, however, in each subcategory: the items with maximum DI values tend to be rather high (close to 0.95 occasionally) while the minimum DI values demonstrate perfectly normal distribution.

This observation — that word forms with only moderate or normal salience in a corpus (as measured by their DI) are approved as autonomous headwords in the CombiDic – can be explained in many ways. Firstly, the statistical distribution has not been the (main) concern in deciding lexicon membership. The CombiDic is an aggregated super dictionary by nature, and has inherited its content from many dictionaries compiled independently (Koppel et al., 2019, and Tavast et al., 2020). Secondly, the semantics of the word forms has naturally been the main concern in lexicography. The headwords with minimum DI levels in Table 3 are very special in terms of composition and meaning, mostly reflecting a kind of rural or robust undercurrent in the Estonian lexicon, which originates in the lifestyle of peasants. The word forms have been considered worth including in the dictionary because dictionaries are expected to assist in understanding literary and historical texts, too, and cannot be pure reflections of the newest corpora. Thirdly, the variance in the DI levels of dictionary headwords is great because not all language changes are traceable in the corpus data. The consistency of the corpus affects the statistical results obtained from it. Some case forms of nouns in our database of ambiforms just represent colloquial changes of usage that are not yet directly detectable using a corpus of written language. For example, in Table 3 the forms with normal DI levels, *VIGADETA* [mistake-PL.ABE] 'errorless' and *PÕHJUSENA* [põhjus-ESS] 'as caused', are in no way different from their approved analogues with normal DI levels: *takistusteta* [obstacle-PL.ABE] 'without obstacles' and *tulemusena* [result-ESS] 'as a result', respectively.

Table 3. Distribution indices of 30 ambiforms not present in the CombiDic compared to similar case forms present in the CombiDIc.

Ambiforms not included in the CombiDic			**Ambiforms approved as headwords in the CombiDic**			
			Average of the group			**Extremes of the group**
Ambiforms	**DI**	**Label**	**Form**	**N**	**Ave**	**Ambiforms with Max and Min values**
KOOSKÕLAS [harmony-INE] ‘in accord’	**0.756**	**critical**	sg in	67	**0.275**	*otseloodis* [stright.plummet-INE] 'vertically straight'
LÄHEDUSES [contiguity-INE] ‘nearby’	**0.547**	**critical**				
STIILIS [style-INE] 'à la mode'	**0.357**	**critical**				
KODUS [home-INE] 'at home'	**0.314**	**critical**				
LAPSEPÕLVES [childhood-INE] 'in childhood'	**0.279**	**critical**				
HÄDAS [trouble-INE] 'in trouble'	**0.187**	**critical**				
PAANIKAS [panic-INE] 'in a panic'	0.114	moderate				
RONGKÄIGUS [procession-INE] 'in procession'	0.112	moderate				
VARJUS [shadow-INE] 'in the lee of'	0.056	moderate				
MURES [worry-INE] 'worried'	0.054	moderate				*köies* [rope-INE] 'belayed'
RAAMES [frame-PL.IN] 'in the context of (smth)’	0.091	moderate	pl in	7	**0.326**	*üldjoontes* [general.line-PL.INE] ‘in general terms’
LEEKIDES [flame-PL.INE] 'in flame'	0.084	moderate				
PIIRES [border-PL.INE] 'within'	0.036	normal				
KORDADES [time-PL.INE] '(many) times'	0.008	normal				*litsides* [whore-PL.INE] 'sleep around'
VAHELDUSEKS [variance-TRA] 'for a change'	**0.447**	**critical**	sg tr	6	**0.282**	*tarbeks* [need-TRA] 'for'
VÕRDLUSEKS [comparison-TRA] 'for comparison'	**0.224**	**critical**				
PROOVIKS [try-TRA] 'on approval'	0.021	normal				
TANTSUKS [tants-TRA] 'for a dance'/'into a dance'	0.007	normal				*saateks* [accompany-TRA] 'for background'

Table 3 *(continued)*

Ambiforms not included in the CombiDic			**Ambiforms approved as headwords in the CombiDic**			
			Average of the group			**Extremes of the group**
Ambiforms	**DI**	**Label**	**Form**	**N**	**Ave**	**Ambiforms with Max and Min values**
JÕUGA [force-COM] 'by force'	0.059	moderate	sg kom	16	**0.365**	*kamaluga* [hand-COM] 'handful'
HINGEGA [soul-COM] 'passionately'	0.031	normal				
ÜLLATUSEGA [surprise-COM] 'with surprise'	0.006	normal				*kapaga* [cup-COM] 'in quantities'
PENSIONILE [pension-ALL] 'pension off'	**0.151**	**critical**	sg all	7	**0.172**	*tagaplaanile* [back.ground-ALL] 'to the background'
MINEKULE [leaving-ADE] 'to be leaving'	-0.008	normal				*verele* [blood-ALL] 'into bleeding'
HINNANGUL [estimate-ADE] ‘as estimated’	**0.528**	**critical**	sg ad	51	**0.346**	*esmapilgul* [first.glance-ADE] ‘at first glance’
VÕIMUL [power-ADE] 'in power'	0.028	normal				*pasal* [shit-ADE] 'diarrhea'
RÕÕMUST [joy-ELA] 'because of joy'	0.013	normal	sg el	6	**0.170**	*surmasuust* [death.mouth-ELA] 'escape death' *esirinnast* [forefront-ELA] 'from the forefront'
PÕHJUSENA [põhjus-ESS] 'as caused'	0.006	normal	sg es	4	**0.42**1	*kulutulena* [wildfire-ESS] 'extensively' *tulemusena* [result-ESS] 'as a result'
TÜKKIDEKS [piece-PL.TRA] 'into pieces'	0.074	moderate	pl tr	1	**0.267**	*ribadeks* [strip-TRA] ‘into strips’
ANDMETEL [data-PL.ADE] 'based on data'	**0.197**	**critical**	pl ade	7	**0.563**	*savijalgadel* [clay.foot-PL.ADE] 'shaky' *sulgpatjadel* [feather.pillow-PL.ADE] 'treasured'
VIGADETA [mistake-PL.ABE] 'errorless'	0.007	normal	pl ab	2	**0.206**	*viperusteta* [glitch-PL.ABE] 'without a glitch' *takistusteta* [obstacle-PL.ABE] 'without obstacles'

The results of the experiment suggest that the lexicographers could include the eleven ambiforms with critical DI values in Table 3 in a dictionary without hesitation while with the others additional — preferably semantic — consideration is needed. On the other hand, the status of word forms already included in the CombiDic can be validated — to some degree — automatically, based on their higher than threshold DI values.

The overall quality rating of the DIC can be formulated in this way: a higher than critical level of DI can be trusted as an indicator of the relative autonomy of a word form, while a lower than critical level of DI does not preclude such autonomy. The DIC thus provides relative heuristics rather than absolute ratings or true-value decisions.

5. Conclusion and discussion

There is a need for a measurable criterion when deciding the lexicographic status of some wicked case forms of nouns in Estonian that can take the meaning and function of indeclinable function words. We have proposed a distribution index (DI) as such a measure. The DI can be used as an indicator of the correspondence of a particular form's actual frequency with its predicted — in the normal distribution of case forms — elicitation degree.

We have described the steps taken to develop an application — the Distribution Index Calculator (DIC) — which can be used by lexicographers when working with wicked word forms (called ambiforms in this paper and elsewhere (e.g. Vainik et al., 2020; Paulsen et al., 2019)). The purpose of such an application is to provide the lexicographer with more elaborate statistical information than absolute frequencies and to process further annotated corpus data with the aim of developing a more specific indicator of the degree of grammaticalisation. We have described the prerequisites and the main components of the application, as well as having provided the algorithm.

As a result, the DIC is a web-based application accessible to everyone. It takes an *ambiform* as an input from the user and retrieves corpus data (frequencies of the word form and the suspected lemma), as well as the suspected morphological form. The tool calculates the distribution index of the input form and compares it to the ranked scale of word form autonomy. The DIC provides the outcome with a verbal label about the detected tendency of the distribution.

A substantial part of the paper was devoted to providing examples of the DIC at work and to comparing the results of the DIC with the decisions made by lexicographers when approving such forms for the CombiDic of Estonian. The conclusion was that the DIC provides relative heuristics rather than absolute ratings or true-value decisions. This is because a higher than critical level of DI can be trusted as an indicator of the relative autonomy of a word form, while a lower than critical level of DI does not preclude such autonomy, and additional inspection of the case forms is needed.

The idea of the DI and the calculator providing indices as a measurable statistic is based on the assumption that the case forms generally follow a constant proportion (i.e. their normal distribution) in corpus texts. It has also been stated by Koppel (2020) that "[...] patterns of

Estonian words are well established and rarely debated among lexicographers [...]”. However, the existence and categorisation of wicked case forms has been quite a problem for lexicographers (Paulsen et al., 2019; Karelson, 2005). The question of upgrading lexical items that traditionally were sub-headwords in dictionaries to headwords has arisen in the context of aggregating autonomous dictionaries into the unified CombiDic (and its underlying database, Ekilex) (Koppel et al., 2019; Tavast et al., 2020).

One can argue that the DIC does a task similar to the Sketch Engine’s function “frequent constructions”, i.e. revealing the relative prominence of certain forms. However, as the DI is based on a comparison with the normative distribution of case forms, our tool provides an instant comparison with the norm and is thus more informative about possible deviations. We believe that the DIC can be useful for lexicographers as it provides the results of the calculation, as well as information about the existence of alternative interpretations due to homonymy. No lexicographer has tried to work with the DIC yet, as it is still in development.

Since the setup of the DIC is generic, it can also be used to test the tendencies of morphological distribution in other languages with rich morphology. The language-specific normal distribution rates (number + case) need to be available and the scales have to be established beforehand. Finnish might be a good candidate for a trial[6], as there are similar grammaticalisation processes of nominal case forms (see e.g. the analysis of the grammaticalisation of body-part nouns into adpositions in Ojutkangas, 2001). “Most Finnic adpositions display elements of productive noun inflection and frequently apply one of the local case sets” (Grünthal, 2003: 47).

6. Limitations of the application and suggestions for future research and development

Some limitations of this work should be noted. The first and foremost is that the results provided by the calculator depend on the accuracy of the corpus tagging. The DIC cannot go beyond the existing annotation yet. Both the corpus tagging system and morphological analysis are based on the Vabamorf (OÜ Filosoft) software, using the EstNLTK 1.6 (Python) library. This is open-source software with broad functionality created specifically to analyse the morphology of the Estonian language (Orasmaa et al., 2016: 2461). However, the wicked case forms described in this paper also cause problems for morphological analysis. This is because there is no good procedure for their disambiguation when it comes to choosing between multiple available interpretations. If a word form has been approved as an indeclinable word for the lexicon of Vabamorf, this results in a tendency for the analysis of this particular word form to be split between different interpretations with questionable accuracy. Such examples appear in illustrations d, f and i in Figure 2. Split interpretations can result in a decrease in the DI level of that form from heightened value to normal.

[6] Data regarding the distribution of case forms in Finnish is available online: https://kaino.kotus.fi/visk/sisallys.php?p=1227&fbclid=IwAR1v5oF4UqIySTckF50KwZK11VBmR8RJdHa6UNATYF9O241B1LYJ4DsbtnI and https://kaino.kotus.fi/visk/sisallys.php?p=1228

Therefore, the results of the forms with multiple interpretations cannot be fully trusted.

Another limitation is that the meanings of a polysemous word cannot be separated yet. The DIC calculates the indices of word forms as if there were only one form deductible to one particular lemma. This is shown in illustration j in Figure 2.

The DIC is in the process of ongoing development. Multiple paths forward are available in this respect: one involves improving the current prototype, e.g. by refining and fine-tuning the norms, the scale and the threshold to meet the more specific needs of lexicographers. Adding statistical information about interpretations other than case forms is one option. Another way to improve the current prototype is by extending its coverage to multiple corpora, which would enable it to follow changes in the relative salience of wicked forms in different styles, e.g. colloquial vs general usage, or by tracking diachronic changes. It is also possible to make the interface of the DIC more attractive, e.g. showing its output using visualisations.

One of the directions of future work is to try to overcome deficiencies due to the current morphological annotation of the corpus. We have thought about testing a "zero hypothesis", i.e. ignoring the PoS definitions of morphological coding and retrieving data as "wild" word forms, summing up the numbers of the forms independently of their PoS tagging. We believe that such an approach would result in higher DI values of ambiforms with split interpretations. On the other hand, the information about their decategorisation would be lost. We are also open to trying some alternative systems of morphological tagging if available (e.g. Universal Dependencies PoS Tagger, TreeTagger and/or RFTagger).

The ultimate goal of future work is to incorporate the DIC into a more complex multi-search application, which would help lexicographers to attach POS tags to lexical units in a more systematic way. The multi-search application has to give a more comprehensive picture of a word form's behaviour in texts. A measure of statistical distribution will be combined with measures of morphosyntactic behaviour and semantic similarity to a prototype of the suspected word class.

7. Abbreviations

Glossing: ABE – abessive case; ABL — ablative case; ADE – adessive case; ADT – additive case; ALL – allative case; COM — comitative case; ELA – elative case; ESS — essive case; GEN – genitive case; ILL – illative case; INE — inessive case; PART – partitive case; PL – plural; SG – singular; TER — terminative case; TRA – translative case

8. Acknowledgements

This work was supported by Estonian Research Council grant PSG227.

9. References

Blensenius, K. & von Martens, M. (2019). Improving Dictionaries by Measuring Atypical Relative Word-form Frequencies. In I. Kosem et al. (eds.) *Proceedings of eLex 2019* *conference. 1–3 October 2019. Sintra, Portugal.* Brno: Lexical Computing CZ, s.r.o., pp. 660–675.

CombiDic = The EKI Combined Dictionary [EKI ühendsõnastik]. (2020). I. Hein, J. Kallas, O. Kiisla, K. Koppel, M. Langemets, T. Leemets, M. Melts, S. Mäearu, T. Paet, P. Päll, M. Raadik, M. Tiits, K. Tsepelina, M. Tuulik, U. Uibo, T. Valdre, Ü. Viks, P. Voll. Institute of the Estonian Language. Accessed at: Sõnaveeb 2020. https://sonaveeb.ee. [25.2.2021].

Ekilex. Accessed at: https://ekilex.eki.ee/ (20 March 2021)

Feltgen, Q., Fagard, B. & Nadal, R. (2017). Frequency patterns of semantic change: Corpus-based evidence of a near-critical dynamics in language change. *Royal Society Open Science 4*. DOI: 10.1098/rsos.170830.

Grünthal, R. (2003). Finnic Adpositions and Cases in Change. *Suomalais-Ugrilaisen Seuran toimituksia 244*. Helsinki: Finno-Ugrian Society.

Habicht, K., Penjam, P. & Prillop, K. (2011). Sõnaliik kui rakenduslik ja lingvistiline probleem: sõnaliikide märgendamine vana kirjakeele korpuses. [Parts of speech as a functional and linguistic problem: Annotation of parts of speech in the corpus of Old Written Estonian] *Estonian Papers in Applied Linguistics* 7, pp. 19–41.

Heine, B. & Kuteva, T. (2007). *The genesis of grammar. A reconstruction.* Oxford: Oxford University Press.

Hopper, P. J. & Traugott, E. C. (2003). *Grammaticalization.* 2nd ed. Cambridge: Cambridge University Press.

Kallas, J., Langemets, M. & Tender, T. (2019). Opening up Estonian dictionaries to European communities and language technology. In: Tender, T.; Eichinger, L. M. (Eds). *Language and Economy. Language industries in a multilingual Europe: EFNIL Conference 2019, Tallinn 2019.* Research Institute for Linguistics, Hungarian Academy of Sciences, pp. 149—154.

Karelson, R. (2005). Taas probleemidest sõnaliigi määramisel [The problems of PoS tagging revisited] *Eesti Rakenduslingvistika Ühingu aastaraamat, 1*(2004), pp. 53—70.

Koppel, K. (2020). *Näitelausete korpuspõhine automaattuvastus eesti keele õppesõnastikele* [Corpus-Based Automatic Detection of Example Sentences for Dictionaries for Estonian Learners]. PhD thesis. Tartu: Tartu Ülikooli Kirjastus.

Koppel, K., Tavast, A., Langemets, M. & Kallas, J. (2019). Aggregating dictionaries into the language portal Sõnaveeb: Issues with and without a solution. In I. Kosem, T. Zingano Kuhn, M. Correia, J. P. Ferreria, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek, & C. Tiberius (eds.) *Proceedings of the eLex 2019 conference. 1-3 October 2019, Sintra, Portugal.* Brno: Lexical Computing CZ, s.r.o., pp. 434—452.

Milintsevich, K. & Sirts, K. (2020). Lexicon-Enhanced Neural Lemmatization for Estonian. In: *Human Language Technologies – The Baltic Perspective.* IOS Press. (Frontiers in Artificial Intelligence and Applications), pp. 158—165. DOI: 10.3233/FAIA200618.

Ojutkangas, K. (2001). *Ruumiinosannimien kieliopillistuminen suomessa ja virossa. Suomalaisen Kirjallisuuden Seuran toimituksia* 845, Helsinki.

Orasmaa, S., Petmanson, T., Tkatšenko, A., Laur, S. & Kaalep, H.-J. (2016). EstNLTK – NLP Toolkit for Estonian. In N. Calzolari, K. Choukri, T. Declerck, M. Grobelnik, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis (eds.) *Proceedings of the*

Tenth International Conference on Language Resources and Evaluation (LREC 2016): The International Conference on Language Resources and Evaluation; Portorož, Slovenia; 2016. Portorož, Slovenia: ELRA, pp. 2460−2466. Available at: http://www.lrec-conf.org/proceedings/lrec2016/pdf/332_Paper.pdf

Paulsen, G., Vainik, E., Tuulik, M. & Lohk, A. (2019). The Lexicographer's Voice: Word Classes in the Digital Era. In I. Kosem, T. Zingano Kuhn, M. Correia, J. P. Ferreria, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek, & C. Tiberius (eds.) *Proceedings of the eLex 2019 conference. 1-3 October 2019, Sintra, Portugal.* Brno: Lexical Computing CZ, s.r.o., pp. 319−337. Available at: https://elex.link/elex2019/wp-content/uploads/2019/09/eLex_2019_18.pdf

Tavast, A., Koppel, K., Langemets, M. & Kallas J. (2020). Towards the Superdictionary: Layers, Tools and Unidirectional Meaning Relations. In Z. Gavriilidou, M. Mitsiaki & A. Fliatouras (eds.) *Proceedings of XIX EURALEX Congress: Lexicography for Inclusion, Vol. 1.* Alexandroupolis, Greece: Democritus University of Thrace, pp. 215−223.

Vainik, E., Paulsen, G. & Lohk, A. (2021). Käändevormist sõnaks: mida näitab sagedus? [From inflected form to a word: the role of frequency]. Accepted by *Estonian Papers in Applied Linguistics, 17.*

Vainik, E., Paulsen, G. & Lohk, A. (2020). A typology of lexical ambiforms in Estonian. In Z. Gavriilidou, M. Mitsiaki & A. Fliatouras (eds.) *Proceedings of XIX EURALEX Congress: Lexicography for Inclusion, Vol. 1.* Alexandroupolis, Greece: Democritus University of Thrace, pp. 119−130.

Multiword-term bracketing and representation in terminological knowledge bases

Pilar León-Araúz[1], Melania Cabezas-García[1], Pamela Faber[1]

[1] University of Granada, Granada, Spain
E-mail: pleon@ugr.es, melaniacabezas@ugr.es, pfaber@ugr.es

Abstract

Multiword terms (MWTs) are frequently consulted in terminological resources due to their structural, cognitive, and conceptual complexity. However, in most terminological resources they are not always well described, since they are often included as independent term entries with no information on how their constituents are related. An accurate management of MWTs of three or more constituents requires, as a first step, their structural disambiguation, also called bracketing. This paper examines MWT bracketing in order to enhance MWT representation by describing their structural dependencies. Based on NLP advances in bracketing, a protocol has been designed through corpus queries and evaluated according to the reliability of corpora and rules as well as the causes underlying failure. Automatising bracketing can help enhance the representation of MWTs in terminological knowledge bases, assisting both the terminologist and the final user, since making their relational structure explicit can favour knowledge acquisition.

Keywords: multiword term; bracketing; terminological knowledge base; terminology

1. Introduction

Multiword terms (MWTs) are frequently consulted in terminological resources due to their structural, cognitive, and conceptual complexity. However, in most terminological resources they are not always well described (Cabezas-García & Faber, 2017) or even well related to their heads and/or modifiers, since they are often included as independent term entries or unanalysed text strings, with no other information about their underlying relational structure. An accurate management and description of these terms requires an initial step that traditionally has not been among the main interests in terminology or specialised lexicography. This is bracketing, or structural disambiguation (Nakov & Hearst, 2005; Barrière & Ménard, 2014), which is necessary for the right interpretation of MWTs having three or more constituents, as in [*reactive power*] *consumption.* Knowledge of these dependencies facilitates MWT comprehension (i.e. reactive power is consumed instead of power consumption is reactive) and, consequently, translation. In Spanish, *consumo de potencia reactiva* would be the right choice instead of **consumo energético reactivo* or **consumo reactivo energético*, which would be the result of a misunderstood bracketing. The inclusion of MWTs in knowledge-based resources can benefit from their prior structural disambiguation, whose automatisation can assist both terminologists and final users. For instance, their representation can be enhanced by placing them in relation to other concepts' entries based on their dependencies, such as their hypernyms (consumption), thus facilitating knowledge acquisition.

Cabezas-García and León-Araúz (2019) proposed a series of manual steps for the bracketing of MWTs based on their linguistic properties and advances from NLP. At a later stage, León-Araúz and Cabezas-García (in press) added new steps in the form of a bracketing protocol and designed queries in Sketch Engine (Kilgarriff et al., 2004) with a view to automatising bracketing and analysing the reliability of every rule in two different English corpora: (i) a wind power corpus (since the set of MWTs belonged to this domain); and (ii) the Open Access Journal (DOAJ) corpus. The Sketch Engine's API was used to automatically query the corpora. Based on the results of the queries, rules were collectively applied to provide the bracketing of a 103 three-term MWT set. Although the automatic protocol worked in 83% of the cases, the bracketing failed in both corpora for 13 MWTs, thus suggesting a more qualitative study of the results, by analysing those MWTs and looking for possible causes.

This paper examines MWT bracketing in order to enhance MWT representation by describing their structural dependencies. The bracketing errors in León-Araúz and Cabezas-García (in press) were analysed and our results showed that an in-depth analysis of bracketing errors can be used to enhance the protocol. In turn, using an automatised bracketing protocol can result in a more accurate representation of MWTs in terminological resources. In particular, a specific module for MWT representation (Cabezas-García, 2019; 2020) has been designed in the terminological knowledge base EcoLexicon (https://ecolexicon.ugr.es/), which will include bracketing-related information. The remainder of this paper is structured as follows: Section 2 describes the procedure followed in order to automatise bracketing and evaluate its output; Section 3 proposes a new module for the description of MWTs in a terminological knowledge base; and Section 4 draws some conclusions and future lines of research.

2. Multiword-term bracketing

NLP has particularly focused on the structural disambiguation of MWTs, given their difficulties for NLP systems (Lauer, 1995; Girju et al., 2005; Nakov, 2007; Barrière & Ménard, 2014). Likewise, their difficulties in translation (i.e. one of the ultimate purposes of term bases) have been widely acknowledged. However, to the best of our knowledge, none of these findings have been applied in the design of MWT entries in terminological knowledge bases. In Section 2.1 the main bracketing models found in the literature are briefly described. In Section 2.2 the protocol applied in this research, based on the latter, is explained and evaluated.

2.1 Bracketing models

NLP has proposed two main models for the bracketing of three-term MWTs: the adjacency and dependency models. The adjacency model (Marcus, 1980; Pustejovsky et al., 1993) takes an MWT p1 p2 p3 and compares if p2 is more related to p1 or p3. For that purpose, the number of occurrences of p1 p2 and p2 p3 are compared. For instance, in *renewable energy technology* there are more occurrences of *renewable energy* than of *energy technology.* Thus, a left-bracketing structure is adopted ([*renewable energy*] *technology*). The dependency model (Lauer, 1995) compares whether p1 is more strongly associated with p2 or p3. Therefore, the analysis does not start from the central term, as in the adjacency model, but rather from the first one to the left. When p1 is more strongly associated with p2 than to p3, there is a left bracketing ([*tip speed*] *ratio*). In contrast, when p1 is dependent on p3, there is a right bracketing (*mean* [*wind speed*]).

Along the same lines, Grefenstette (1994) states that dependency structures govern how MWTs can be shortened: "*civil rights activist* can be bracketed as [*civil rights*] *activist*, which can be shortened to *rights activist* but not to *civil activist.* On the other hand, *Yale medical library* is properly bracketed as *Yale* [*medical library*] which can then be reduced to *Yale library* or *medical library*, but not to *Yale medical*" (Grefenstette, 1994, p. 65). Based on Grefenstette's approach, for a right bracketing, both p2 p3 (*medical library*) and p1 p3 (*Yale library*) should be more frequent than p1 p2 (*Yale medical*), whereas for a left bracketing p1 p2 (*civil rights*) should be more frequent than p1 p3 (*civil activist*), the latter actually being the same rule as the one proposed by the dependency model.

Apart from these models, Nakov and Hearst (2005) propose a series of surface patterns (i.e. hyphens and slashes, possessive genitive, internal capitalisation, brackets, concatenation, internal inflection, etc.) as signs indicating an internal grouping. For example, *brain's stem cell* would suggest a right bracketing (*brain* [*stem cell*]) because of the possessive genitive, whereas *tyrosine kinases activation* would indicate a left bracketing ([*tyrosine kinase*] *activation*) because of the internal inflection. They also suggest that paraphrases are useful for identifying internal dependencies in MWTs. For instance, *health care reform* is left-bracketed because paraphrases separating those groups can be found, as in "reform *in* health care". Paraphrases can be either verbal or prepositional.

2.2 Bracketing automatisation

Based on the models and patterns above, a set of queries was designed and sent to Sketch Engine's API in order to retrieve and compare the frequencies of all the possible groupings contained in a list of 103 MWTs selected from the wind energy specialised domain (Section 2.2.1). As mentioned above, two corpora were used to compare whether corpus size and/or domain specificity had an influence on the output: (i) a wind power corpus (WPC) specifically compiled for this research; and (ii) the Open Access Journal (DOAJ) corpus. The first consisted of wind energy specialised texts (i.e. scientific articles and PhD dissertations originally written in English) and had approximately three million words, whereas the latter covered all areas of science, technology, medicine, social science, and humanities and had approximately two billion words.

After that, the results were compared with the baseline (manually disambiguated by three annotators) and the protocol was evaluated in terms of rule and corpus reliability (Section 2.2.2). Since the protocol failed in both corpora for 13 MWTs, a more in-depth analysis was performed in order to discover the causes of protocol failure (Section 2.2.3) and improve it accordingly.

2.2.1 Preparing the dataset: queries and rules

The list of 103 MWTs, manually bracketed as a baseline, is included in Table 1.

offshore [wind farm]	installed [wind power]	[permanent magnet] generator
[tip speed] ratio	[wind turbine] design	[wind farm] project
[wind power] plant	[wind penetration] level	[wind speed] distribution
[wind power] generation	[wind speed] datum	[wind energy] production
[wind power] capacity	novel [wind turbine]	extreme [wind speed]
mean [wind speed]	domestic [hot water]	[wind tunnel] test
[wind power] production	[power generation] system	[wind energy] penetration
average [wind speed]	offshore [wind market]	offshore [wind park]
offshore [wind turbine]	[renewable energy] technology	[renewable energy] system
[renewable energy] source	[wind power] penetration	[wind speed] measurement
offshore [wind power]	[wind power] forecast	shrouded [wind turbine]
offshore [wind energy]	[wind power] development	[wind turbine] control
[wind energy] system	total [installed capacity]	micro [hydropower plant]
small [wind turbine]	conventional [power plant]	hybrid [wind farm]
high [wind turbine]	[power system] reliability	[blade element] theory
rated [wind speed]	offshore [wind project]	[reactive power] consumption
large [wind farm]	[wind turbine] model	[wind energy] potential
onshore [wind farm]	power [electronic converter]	installed [wind generation]
[wind turbine] blade	[wind turbine] generator	offshore [wind resource]
[wind power] output	[sound pressure] level	[wind turbine] application
low [wind speed]	[wind turbine] manufacturer	power [spectral density]
[wind turbine] rotor	[wind energy] project	[wind speed] forecasting
large [wind turbine]	[wind power] fluctuation	[wind power] integration
[control system] design	[heat transfer] medium	[transmission system] operator
average [capacity factor]	[wind power] project	[thermal power] plant
[wind energy] sector	[hydroelectric power] station	[time domain] simulation
unity [power factor]	urban [wind turbine]	[reactive power] control
[full load] hour	[hydro power] plant	[grid connection] cost

[wind turbine] component	[wind energy] capacity	[wind energy] application
[power system] operation	[hydroelectric power] plant	[voltage source] converter
net [capacity factor]	[wind resource] assessment	[sound power] level
[mass flow] rate	[wind farm] development	net [present value]
[wind energy] converter	[wind energy] density	conventional [wind turbine]
[wind turbine] system	[reactive power] compensation	[renewable energy] resource
[wind turbine] technology		

Table 1: List of MWTs manually bracketed

Based on bracketing models (2.1), the terms in Table 1 were decomposed in all possible groupings and/or searched for within different structures, as pointed out by the following 12 indicators:

1. MWTs decomposed in all possible groupings according to adjacency, dependency and shortening models (p1 p2, p2 p3, p1 p3) (for *offshore wind farm*, *offshore wind, wind farm, offshore farm*);

2. Insertions within the MWTs (p1 * p2 p3 and p1 p2 * p3) (*offshore* [*wind farm*] because *offshore* ***shrouded*** *wind farm*);

3. Longer MWTs where adjacent groupings act as modifiers (p1 p2 *, p2 p3 *), head (* p1 p2, * p2 p3) or middle modifiers (* p1 p2 *, * p2 p3 *) (*offshore* [*wind farm*] because ***onshore*** *wind farm*);

4. MWTs with a hyphen between adjacent groupings (p1-p2 p3, p1 p2-p3) ([*cell cycle*] *analysis* because *cell-cycle analysis*);

5. MWTs with the possessive genitive between adjacent groupings (p1's p2 p3, p1 p2's p3) (*brain* [*stem cell*] because *brain*___'s___ *stem cell*);

6. MWTs showing brackets around a single element (p1 p2 (p3), p1 (p2) p3, (p1) p2 p3) or a grouping ((p1 p2) p3, p1 (p2 p3)) ([*cell cycle*] *analysis* because **(***cell cycle***)** *analysis*);

7. MWTs where one of the adjacent groupings forms a monolexical compound (p1p2 p3, p1 p2p3) ([*gear box*] *manufacturer* because ***gearbox*** *manufacturer*);

8. MWTs where one of the first two elements is inflected for number (p1 p2s p3, p1s p2 p3) ([*tyrosine kinase*] *activation* because *tyrosine kinase*___s___ *activation*);

9. MWTs showing a different word order of the first two elements (p2 p1 p3) (*mean* [*total consumption*] because *total mean consumption*);

10. MWTs decomposed in all possible groupings having a prepositional paraphrase in between (p3 PREP p1 p2, p2 p3 PREP p1, p1 p3 PREP p2) ([*permanent magnet*] *generator* because *generator* ***with*** *permanent magnets*; [*mean wind*]

speed because *mean speed **of** wind*);

11. MWTs decomposed in all possible groupings having a verbal paraphrase in between (p3 V p1 p2, p1 p2 V p3, p2 p3 V p1, p1 V p2 p3) ([*permanent magnet*] *generator* because *generator **has** permanent magnets*);
12. MTWs where one of the adjacent groupings is followed by two capital letters (expecting an acronym) in brackets (p1 p2 (AA) p3, p1 p2 p3 (AA)) ([*direct current*] *generator* because *direct current **(DC)** generator*).

Consequently, 34 specific CQL (Corpus Query Language) queries were designed for the extraction of occurrences of each of the above structures (Table 2).

Bracketing indicators	**Structure retrieved**	**CQL queries**
Decomposed MWTs	p1 p2	[tag!="JJ.*|N.*"][lemma="**p1**"][lemma="**p2**"][tag!="N.*|JJ.*"]
	p2 p3	[tag!="JJ.*|N.*"][lemma="**p2**"][lemma="**p3**"][tag!="N.*|JJ.*"]
	p1 p3	[tag!="JJ.*|N.*"][lemma="**p1**"][lemma="**p3**"][tag!="N.*|JJ.*"]
Insertions	p1 * p2 p3	[lemma="**p1**"][tag="N.*|JJ.*|RB.*|VVN.*|VVG.*"]+ [lemma="**p2**"][lemma="**p3**"]
	p1 p2 * p3	[lemma="**p1**"][lemma="**p2**"][tag="N.*|JJ.*|RB.*|VVN.*|VVG.*"]+ [lemma="**p3**"]
Longer MWTs	p1 p2 *	[tag!="N.*|JJ.*"][lemma="**p1**"][lemma="**p2**"] [tag="JJ.*|N.*|RB.*|VVG.*|VVN.*" & lemma!= "**p3**"]* [tag="N.*" & lemma!= "**p3**"]
	* p1 p2	[tag="N.*|JJ.*"]+[lemma="**p1**"][lemma="**p2**"] [tag!="N.*|JJ.*"]
	p2 p3 *	[tag!="N.*|JJ.*"] [lemma="**p2**"][lemma="**p3**"] [tag="JJ.*|N.*|RB.*|VVG.*|VVN.*"]* [tag="N.*"]
	* p2 p3	[tag="N.*|JJ.*" & lemma!= "**p1**"]+ [lemma="**p2**"] [lemma="**p3**"] [tag!="N.*|JJ.*"]
	* p1 p2 *	[tag="N.*|JJ.*"]+ [lemma="**p1**"][lemma="**p2**"] [tag="JJ.*|N.*|RB.*|VVG.*|VVN.*" & lemma!="**p3**"]*[tag="N.*" & lemma!= "**p3**"]
	* p2 p3 *	[tag="N.*|JJ.*" & lemma!="**p1**"]+[lemma="**p2**"][lemma="**p3**"]

		[tag="JJ.*|N.*|RB.*|VVG.*|VVN.*" & lemma!="**p3**"]*[tag="N.*"]
Hyphen	p1-p2 p3	[lemma="**p1-p2**"][lemma="**p3**"]
	p1 p2-p3	[lemma="**p1**"][lemma="**p2-p3**"]
Possessive genitive	p1 p2's p3	[lemma="**p1**"][word="**p2**'s"][lemma="**p3**"]
	p1's p2 p3	[word="**p1**'s"][lemma="**p2**"][lemma="**p3**"]
Brackets	p1 p2 (p3)	[lemma="**p1**"][lemma="**p2**"][word="\("][lemma="**p3**"][word="\)"]
	(p1) p2 p3	[word="\("][lemma="**p1**"][word="\)"][lemma="**p2**"][lemma="**p3**"]
	p1 (p2) p3	[lemma="**p1**"][word="\("][lemma="**p2**"][word="\)"][lemma="**p3**"]
	(p1 p2) p3	[word="\("][lemma="**p1**"][lemma="**p2**"] [word="\)"] [lemma="**p3**"]
	p1 (p2 p3)	[lemma="**p1**"][word="\("][lemma="**p2**"] [lemma="**p3**"] [word="\)"]
Monolexical compound	p1p2 p3	[lemma="**p1p2**"][lemma="**p3**"]
	p1 p2p3	[lemma="**p1**"][lemma="**p2p3**"]
Inflection	p1 p2s p3	[lemma="**p1**"][lemma="**p2**" & tag="NNS"][lemma="**p3**"]
	p1s p2 p3	[lemma="**p1**" & tag="NNS"][lemma="**p2**"][lemma="**p3**"]
Word order	p2 p1 p3	[lemma="**p2**"][lemma="**p1**"][lemma="**p3**"]
Prepositional paraphrases	p3 PREP p1 p2	[lemma="**p3**"][]{0,2}[tag="IN" & lemma!="like"] []{0,2}[lemma="**p1**"][lemma="**p2**"][lemma!="**p3**"]
	p2 p3 PREP p1	[lemma!="**p1**"][lemma="**p2**"][lemma="**p3**"][]{0,2}[tag="IN" & lemma!="like"][]{0,2}[lemma="**p1**"]
	p1 p3 PREP p2	[tag!="JJ.*|N.*"][lemma="**p1**"][lemma="**p3**"][]{0,2}[tag="IN" & lemma!="like"][]{0,2}[lemma="**p2**"][tag!="JJ.*|N.*"]
Verbal paraphrases	p3 V p1 p2	[lemma="**p3**"][]{0,2}[tag="VV.*"][]{0,2}[lemma="**p1**"] [lemma="**p2**"][lemma!="**p3**"]

	p1 p2 V p3	[lemma="**p1**"][lemma="**p2**"][lemma!="**p3**"]{0,2}[tag="VV.*"] []{0,2}[lemma="**p3**"]
	p2 p3 V p1	[lemma!="**p1**"][lemma=" **p2**"][lemma="**p3**"][]{0,2}[tag="VV.*"] []{0,2}[lemma="**p1**"]
	p1 V p2 p3	[lemma="**p1**"][lemma!="**p2**"]{0,2}[tag="VV.*"][lemma!="**p1**"]{0,2} [lemma=" **p2**"][lemma="**p3**"]
Acronyms	p1 p2 (AA) p3	[lemma="**p1**"][lemma="**p2**"][word="\("][word="[A-Z]{2}(s)?"][word="\)"][lemma= "**p3**"]
	p1 p2 p3 (AA)	[lemma="**p1**"][lemma="**p2**"][lemma="**p3**"][word="\("][word="[A-Z]{2}(s)?"][word ="\)"]

Table 2: CQL queries

To retrieve all the data, each constituent of the 103 MWTs was automatically filled in the placeholders of p1, p2 and p3 and queries were sent to both corpora through Sketch Engine's API, which means that a total of 7,004 queries were performed. In order to avoid noise, all queries were applied to a single sentence (within <s/>) and sub-hits (lazy results causing a multiplying effect) were filtered out. For the same reason, some of the queries need to exclude certain elements. For example, when looking for the MWTs decomposed in three independent terms (p1 p2, p2 p3, p1 p3), the queries exclude any adjective or noun before and after them ([tag!="N.*|JJ.*"]) to avoid structures where the groupings are only part of longer MWTs.

Based on the figures retrieved through the Sketch Engine's API, the following 16 rules were developed in order to automatically compute the bracketing of each MWT (Table 3).

Adjacency	1. If p1 p2 > p2 p3 then (p1 p2) p3; If p1 p2< p2 p3 then p1 (p2 p3); Else, N/A
Dependency	2. If p2 p3 > p1 p3, then (p1 p2) p3; If p2 p3 < p1 p3, then p1 (p2 p3); Else, N/A
Shortening	3. If p1 p2 > p1 p3, then (p1 p2) p3 If p1 p3 & p2 p3 > p1 p2, then p1 (p2p3) Else, N/A

Insertions	4. If p1 p2 * p3 > p1 * p2 p3, then (p1 p2) p3; If p1 p2 * p3 < p1 * p2 p3, then p1 (p2 p3); Else, N/A
Longer MWTs	5. If p1 p2 * + * p1 p2 + * p1 p2 * > p2 p3 * + * p2 p3 + * p2 p3 *, then (p1 p2) p3; If p1 p2 * + * p1 p2 + * p1 p2 * < p2 p3 * + * p2 p3 + * p2 p3 *, then p1 (p2 p3); Else, N/A
Hyphen	6. If p1-p2 p3 > p1 p2-p3, then (p1 p2) p3; If p1-p2 p3 < p1 p2-p3, then p1 (p2 p3); Else p1 N/A
Possessive genitive	7. If p1 p2's p3 > p1's p2 p3, then (p1 p2) p3; If p1 p2's p3 < p1's p2 p3, then p1 (p2 p3); Else N/A
Brackets	8. If p1 p2 (p3) > (p1) p2 p3 + p1 (p2) p3, then (p1 p2) p3; If p1 p2 (p3) < (p1) p2 p3 + p1 (p2) p3, then p1 (p2 p3); Else N/A
	9. If (p1 p2) p3 > p1 (p2 p3), then (p1 p2) p3; If (p1 p2) p3 < p1 (p2 p3), then p1 (p2 p3); Else N/A
Monolexical compound	10. If p1 p2 p3 > p1 p2p3, then (p1 p2) p3; If p1 p2 p3 < p1 p2p3, then p1 (p2 p3); Else N/A
Internal inflection	11. If p1 p2s p3 > p1s p2 p3, then (p1 p2) p3; If p1 p2s p3 < p1s p2 p3, then p1 (p2 p3); Else N/A
Word order	12. If p2 p1 p3 > 0, then p1 (p2 p3); Else N/A
Prepositional paraphrases	13. If p3 PREP p1 p2 > p2 p3 PREP p1, then (p1 p2) p3; If p3 PREP p1 p2 < p2 p3 PREP p1, then p1 (p2 p3); Else N/A

	14. If p1 p3 PREP p2 > 0, then p1 (p2 p3) Else, N/A
Verbal paraphrases	15. If p3 V p1 p2 + p1 p2 V p3 > p2 p3 V p1 + p1 V p2 p3, then (p1 p2) p3; If p3 V p1 p2 + p1 p2 V p3 < p2 p3 V p1 + p1 V p2 p3, then p1 (p2 p3); Else N/A
Acronyms	16. If p1 p2 (AA) > p1 p2 p3 (AA), then (p1 p2) p3; If p1 p2 (AA) < p1 p2 p3 (AA), then p1 (p2 p3); Else N/A

Table 3: Bracketing rules

Most of the rules lead to either left or right bracketing (or N/A if no results or equal results are obtained), but two of them are only indicative of one. If rules 12 and 14 apply, they will indicate a left or right bracketing, respectively, but if they do not, that does not mean that the opposite bracketing applies. For instance, when applying rule 12 to *micro hydropower plant*, the word order *hydropower micro plant* is not found. However, this does not mean that it has a left bracketing. Furthermore, most of the rules compare the figures of two queries, but some others include the addition of several from different queries (5, 8 and 15). For instance, when rule 5 is applied to *wind power fluctuation*, longer MWTs formed by each of the possible groupings are compared and added (e.g. for *wind power*, longer MWTs, such as *wind power system*, *onshore wind power*, and *offshore wind power consumption*, are added and compared to the figures associated with *power fluctuation*). Finally, except for rules 12 and 14, all the rules but one (3) are composed of two opposing conditions. Rule 3 is a mixture of the left-bracketing condition of the dependency model and two nested conditions (p1 p3 > p1 p2 & p2 p3 > p1 p2).

In sum, the protocol is composed of 12 indicators formulated in 34 queries, whose results are compared in 16 bracketing rules. Once the rules were applied and the bracketing candidates obtained (based on the agreement of most rules, which all have the same weight), the results were compared to the baseline.

2.2.2 Evaluating the protocol: rules and corpus reliability

Our results showed that the protocol allows for the correct bracketing of MWTs in more than 83% of the cases as the average in both corpora, but some of the rules are more productive and/or reliable than others, certain differences between the corpora can also be found, and the confidence level of all rules (i.e. the probability to match with the baseline based on the number of rules agreeing on the same result) shows differences among the MWTs in the dataset.

The performance of the rules for disambiguating purposes is based on their likelihood to retrieve results from corpora and their ability to actually solve MWT bracketing as compared to the baseline. The balance between frequency and reliability is what constitutes the basis for a weighted protocol. This means that there are rules that do not retrieve any result very often, but they are highly reliable when they do. For instance, the possessive rule had a 100% matching rate but could only be used with seven MWTs. In contrast, there are rules that are always likely to retrieve results but do not always deliver an output matching the baseline.

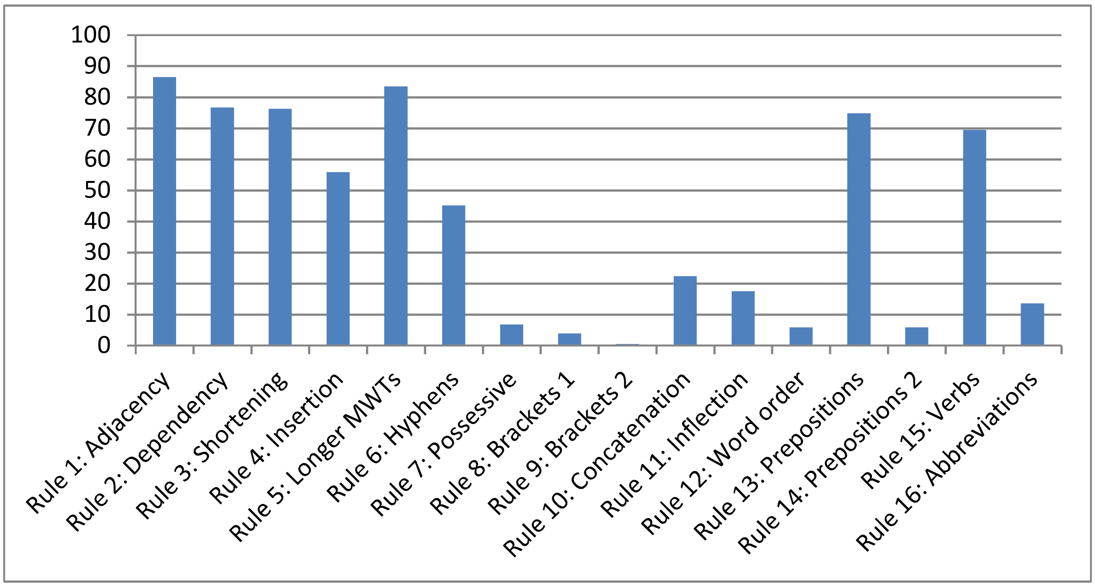


Figure 1: Performance of bracketing rules

Figure 1 shows the performance of each of the rules considering both factors. Adjacency (86.4%), longer MWTs (83.5%), dependency (76.7%) and shortening (76.2%) are, collectively, the most useful rules.

As for the corpora, the agreement with the baseline based on the queries on the WPC outperformed that of the DOAJ. Another difference is the varying performance of the protocol on left or right bracketing. Generally speaking, left bracketing is better identified in both corpora, but the difference is even more noticeable in the WPC.

Corpus size and type were thus found to have an influence on the results. The WPC, although smaller in size, provided better bracketing results for the MWT dataset (86.4% vs. 79.6%), as it belongs to the wind power domain. Domain-specificity is thus a key factor for the performance of the protocol over size.

When looking at the rules individually, differences can also be found when comparing corpora (Figure 2) from both quantitative and qualitative points of view.

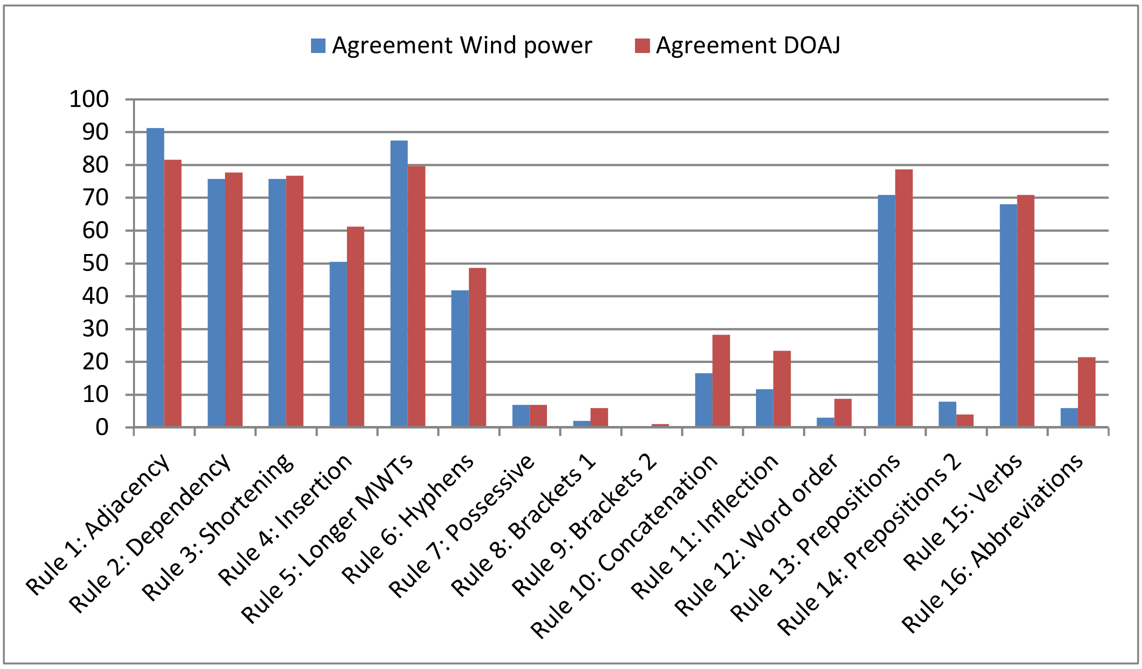


Figure 2: Quantitative and qualitative performance of bracketing rules in both corpora

As previously mentioned, the most reliable rules in both corpora were those related to adjacency, dependency, or the capacity to form new longer MWTs, followed by prepositional or verbal paraphrases and insertions. However, the DOAJ provided better results for certain indicators related to the "surface patterns" reported by Nakov (2007) (e.g. hyphens, concatenation, inflection, abbreviations, etc.), since such patterns will be more likely found in larger corpora. Among the most reliable rules, adjacency and longer MWTs performed better in the WPC, whereas dependency, shortening and insertion performed better in the DOAJ, which might indicate that the former are domain-dependent and the latter size-dependent. This can be verified when looking at the figures (Figure 3) from a purely qualitative way (i.e. not taking into account when no results are retrieved from the corpora and bracketing cannot be computed). In that case most of the rules except for brackets, prepositional paraphrases (and only that of p1 p3 PREP p2) and abbreviations were more reliable in the WPC.

The fact that right bracketing has a lower matching rate with the baseline, especially in the DOAJ, opens a new line of inquiry regarding the nature of these MWTs and their syntactic structure, since the choice of the dataset, based on frequency, was not balanced in terms of left/right bracketing or syntactic structures. The main differences between the corpora are the following: adjacency is equally reliable for left and right bracketing in the WPC as opposed to the DOAJ, where right bracketing reliability scores higher; the insertion and longer MWTs rules work in opposing directions; the inflection rule in the DOAJ only shows reliability for left bracketing.

In the WPC, 100% reliability is shown for hyphens and possessives in the case of left bracketing and for word order for right bracketing. In the DOAJ, 100% reliability is found for bracketed groupings in the case of right bracketing. In both of them, 100% reliability is found for bracketed single words, word order, type 2 prepositional paraphrases and abbreviations in the case of right bracketing.

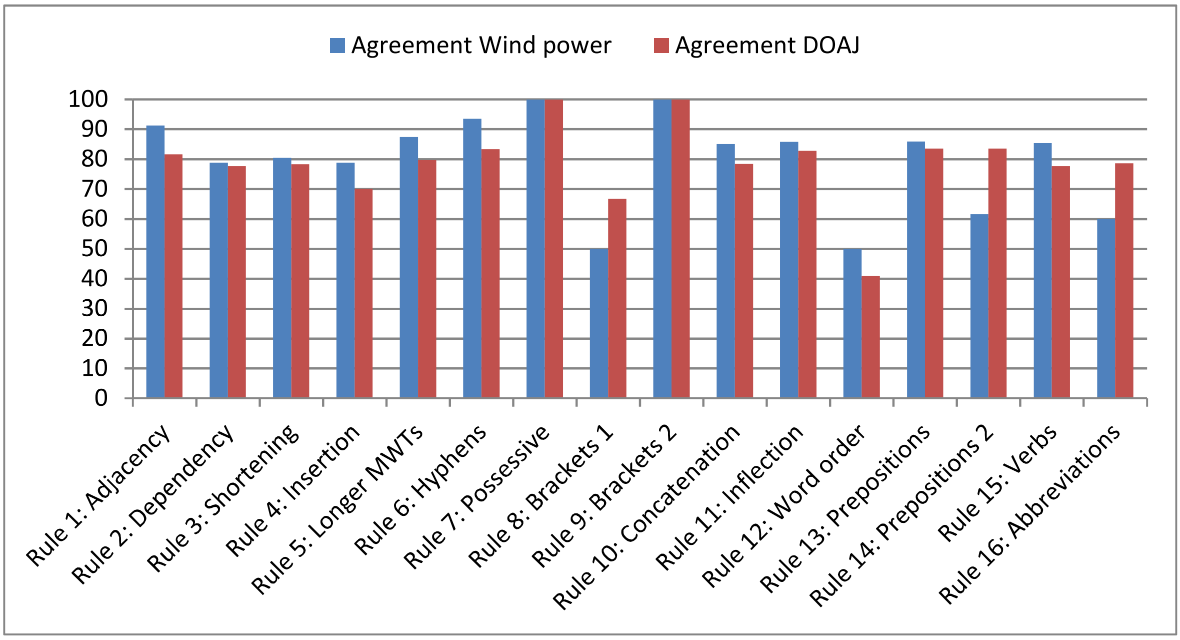


Figure 3. Qualitative performance of bracketing rules in both corpora

Regarding the overall evaluation of the protocol, the output was analysed based on the following: (1) whether the resulting bracketing agreed with the baseline; (2) whether the candidate bracketing was the same in both corpora: and (3) the confidence of each bracketing based on the number of rules pointing in the same direction without considering N/A results (no results from the queries). For instance, for [*wind turbine*] *blade*, even if only 68.75% of the rules could be applied, 100% of them pointed to a left bracketing.

In half the cases, the rules showed a 100% confidence, 51.45% for the WPC and 41.74% for the DOAJ, from which 96.22% and 95.34%, respectively, agreed with the baseline. The only failed bracketings with a 100% confidence were *offshore wind project* (in both corpora), *sound power level* in the WPC, and *offshore wind park* in the DOAJ. From the bracketings showing 80 to 99% confidence (20.38% in both corpora), 90.47% and 85.71% agreed with the baseline. From 50 to 79% confidence (28.15% and 37.86%), 65.51% and 58.97% agreed with the baseline.

In the WPC alone, erroneous bracketing only occurred for *hydroelectric power station* (and only because the application of all rules gave a N/A output), whereas in the DOAJ failures included *wind power plant*, *wind power generation*, *wind power output*, *power electronic converter*, *sound pressure level*, *wind energy density*, *wind energy production*, and *reactive power consumption.* The fact that more erroneous bracketings were found through the DOAJ might indicate again that domain-specificity is what matters the most, since in this corpus many different domains converge and the constituents of these MWTs might accept very different combinations outside the wind power domain.

In both corpora the bracketing failed for the following 13 MWTs: *offshore wind power*, *offshore wind energy*, *wind penetration level*, *offshore wind project*, *hydroelectric power plant*, *hydro power plant*, *micro hydropower plant*, *installed wind generation*, *offshore wind resource*, *thermal power plant*, *sound power level*, *mass flow rate* and *offshore wind park*. We have thus selected this list to perform a more in-depth analysis of possible causes.

2.2.3 Understanding the causes of protocol failure

The 13 MWTs where the protocol failed in both corpora are shown in Table 4 with both outputs and confidence levels.

Baseline	**WPC output**	**Confidence**	**DOAJ output**	**Confidence**
offshore [wind power]	N/A	50%	[offshore wind] power	62.5%
offshore [wind energy]	[offshore wind] energy	62.5%	[offshore wind] energy	55.5%
[wind penetration] level	N/A	50%	wind [penetration level]	66.6%
offshore [wind project]	[offshore wind] project	100%	[offshore wind] project	100%
[hydroelectric power] plant	N/A	50%	hydroelectric [power plant]	70%
[hydro power] plant	hydro [power plant]	55.5%	hydro [power plant]	55.5%
micro [hydropower plant]	N/A	50%	[micro hydropower] plant	66.6%
installed [wind generation]	[installed wind] generation	71.4%	[installed wind] generation	66.6%
offshore [wind resource]	[offshore wind] resource	85.7%	[offshore wind] resource	85.7%
[thermal power] plant	N/A	50%	thermal [power plant]	75%
[sound power] level	sound [power level]	100%	sound [power level]	83.3%
[mass flow] rate	mass [flow rate]	66.6%	mass [flow rate]	83.3%
offshore [wind park]	[offshore wind] park	80%	[offshore wind] park	100%

Table 4: 13 MWTs where the bracketing protocol failed

In most cases, the system delivered the baseline's opposite bracketing, but in five cases the results retrieved by the WPC were N/A, since the results pointed to a 50% confidence, which indicates again that a domain-specific corpus outperforms a large one. The results of these MWTs were analysed based on the following possible causes: (i) the nature of the MWTs (e.g. the left/right bracketing, omission of constituents, their syntactic structure, exceptions to the rule); (ii) the formulation of the corpus queries; and (iii) the rules' confidence level; and (iv) the fact that some rules might be noisier than helpful, thus biasing the results.

Based on their syntactic structures, most of the MWTs (9) follow the structure A+N+N; only one MWT shows the Participle+N+N structure, which could be subsumed under the latter; and three N+N+N structures are found. Considering that A+N+N structures only amount to 30% of the initial 103 MWT dataset, this could point to a degree of bracketing difficulty for such structures, although this should be confirmed by replicating the study with a more balanced dataset in terms of syntactic structure.

In terms of left or right bracketing, the set of failed MWTs is really balanced (six and seven respectively) as compared to their proportion in the original 103 MWT set (34 right-bracketed and 69 left-bracketed MWTs), which suggest that this factor does not necessarily influence the success of the protocol.

There seems to be a trend in failure for MWTs having *plant* or *level* as their head and *offshore wind* as modifiers. In some of these cases, a variable bracketing could occur even in human scenarios, which is often the result of multidimensionality and could explain why the rules did not solve the bracketing of most MWTs in this 13-element set, since 11 of them contain the above mentioned heads or modifiers. For instance, the constituents *offshore wind* could be bracketed together indicating the wind type (e.g. [*offshore wind*] *power* would refer to the energy produced from this type of wind).

Alternatively, the opposite grouping would instead highlight the location relation between *offshore* and the head. For example, *offshore* [*wind power*] would allude to a type of energy produced in that specific location. Furthermore, constituents such as *turbine* or *farm* could have been elicited between *wind* and *power* (the true term being *offshore wind turbine/farm power*), in which case *offshore* would refer to the place where those devices are located. The same concept can thus be seen from different angles, so both human and automatic procedures could be likely to provide contradictory bracketed structures. In this sense, the cases of *offshore wind project* and *offshore wind resource* might be a case of human bracketing failure (despite inter-annotator agreement), since confidence figures are particularly striking. These are the only two MWTs, together with *sound power level* and *offshore wind park*, where confidence level scored so high in the wrong direction as compared to the baseline. In contrast, most failed bracketings showed a confidence level of 50-60%, which points to the possibility of setting a threshold above 60%.

Something similar could happen with *hydro power plant*, where [*hydro power*] *plant* would be a plant that uses water power (in a more general sense of energy) and *hydro* [*power plant*] would imply a plant generating power (in the sense of electricity) that uses water. *Hydroelectric power plant* would fall under the same hypothesis, however, with its synonym *hydroelectric power station* the protocol did not fail. The same happened with sound power level, which got a failed bracketing while a very similar term (*sound pressure level*) got it right. This reinforces the hypothesis that *power*, due to polysemy, is especially prone to multidimensionality.

Regarding the formulation of corpus queries, no errors possibly influencing the results were found. The last step was to wonder whether there were certain rules that might be more misleading than helpful in these MWTs, opening the possibility of constraining the protocol for the rest of the MWTs in the set. Table 5 shows the performance of each rule for each MWT in the WPC/DOAJ. However, no significant patterns were found, which means that if each rule were to have a different weight, weights cannot be inferred by analysing erroneous bracketings.

	offshore [wind power]	offshore [wind energy]	[wind penetration] level	offshore [wind project]	[hydroelectric power] plant	[hydro power] plant	micro [hydropower plant]
Rule 1: Adjacency	Agree / Agree	Agree / Agree	Agree / Fail	Fail / Fail	Fail / Fail	Fail / Fail	Agree / Agree
Rule 2: Dependency	Fail / Fail	Fail / Fail	Agree / Agree	Fail / Fail	Agree / Agree	Agree / Agree	Fail / Fail
Rule 3: Shortening	Fail / Fail	Fail / Fail	Agree/Agree	Fail / Fail	Agree / Agree	Agree / Agree	Fail / Fail
Rule 4: Insertion	N/A/Fail	N/A / Fail	Fail / Fail	Fail / Fail	N/A / Fail	Fail/ Agree	N/A/ N/A
Rule 5: Longer MWTs	Agree / Agree	Fail / Agree	Fail / Fail	Fail / Fail	Fail / Fail	Fail / Fail	Agree / Agree
Rule 6: Hyphens	Agree / N/A	Agree / N/A	Agree / N/A	N/A / N/A	N/A / Agree	Agree / Fail	N/A / Fail
Rule 7: Possessive	N/A / N/A	N/A / N/A	N/A / N/A	N/A / N/A	N/A / N/A	N/A / N/A	N/A / N/A
Rule 8: Brackets 1	N/A /	N/A /	N/A / N/A	N/A /	N/A / N/A	N/A / N/A	N/A / N/A

	Agree	N/A		N/A			
Rule 9: Brackets 2	N/A / N/A	N/A / N/A	N/A / N/A	N/A / N/A	N/A / N/A	N/A / N/A	N/A / N/A
Rule 10: Concatenation	Agree / N/A	N/A / Agree	N/A / N/A	Fail / N/A	N/A / Agree	Agree / Fail	N/A / Fail
Rule 11: Inflection	N/A / N/A	N/A / N/A	N/A / N/A	N/A / N/A	N/A / N/A	N/A / N/A	N/A / N/A
Rule 12: Word order	N/A / N/A	N/A / Agree	N/A / N/A	N/A / N/A	N/A / N/A	Fail / N/A	N/A / N/A
Rule 13: Prepositions	Fail / Fail	Fail/ Fail	Fail/ Fail	Fail/ Fail	N/A/ Fail	Fail/ Fail	N/A / N/A
Rule 14: Prepositions 2	N/A / N/A	Agree / N/A	N/A / N/A	N/A / N/A	N/A / N/A	N/A / Fail	N/A / N/A
Rule 15: Verbs	Fail / Fail	Fail / Fail	Fail / N/A	Fail / Fail	N/A / Fail	N/A / Fail	N/A / N/A
Rule 16: Abbreviations	N/A / N/A	N/A/ N/A	N/A/ N/A	N/A/ N/A	N/A/ N/A	N/A/ N/A	N/A/ N/A

Table 5: Rules' performance on 13 failed bracketings in the WPC

	installed [wind generation]	offshore [wind resource]	[thermal power] plant	[sound power] level	[mass flow] rate	offshore [wind park]
Rule 1: Adjacency	Agree / Agree	Fail / Agree	Fail / Fail	Fail / Fail	Fail / Fail	Fail / Fail
Rule 2: Dependency	Fail / Fail	Fail / Fail	Agree / Agree	Fail / Fail	Agree / Agree	Fail / Fail
Rule 3: Shortening	Fail / Fail	Fail / Fail	Agree / Agree	Fail / Fail	Agree/Agree	Fail / Fail
Rule 4: Insertion	Fail / Fail	Fail / Fail	N/A / Fail	N/A / N/A	N/A / Fail	N/A / Fail

Rule 5: Longer MWTs	Fail / Agree	Fail / Fail	Fail / Fail	Fail / Fail	Fail / Fail	Fail / Fail
Rule 6: Hyphens	N/A / N/A	N/A / N/A	N/A / Agree	N/A / N/A	N/A / Fail	N/A / Agree
Rule 7: Possessive	N/A / N/A	N/A / N/A	N/A / N/A	N/A / N/A	N/A / N/A	N/A / N/A
Rule 8: Brackets 1	N/A / N/A	N/A / N/A	N/A / N/A	N/A / N/A	N/A / N/A	N/A / N/A
Rule 9: Brackets 2	N/A / N/A	N/A / N/A	N/A / N/A	N/A / N/A	N/A / N/A	N/A / N/A
Rule 10: Concatenation	N/A / N/A	N/A / N/A	N/A / Fail	N/A / N/A	Fail / Fail	Agree / N/A
Rule 11: Inflection	N/A / N/A	N/A / N/A	N/A / Fail	N/A / N/A	N/A / N/A	N/A / N/A
Rule 12: Word order	N/A / N/A	N/A / N/A	N/A / Fail	N/A / N/A	Fail / Fail	N/A / N/A
Rule 13: Prepositions	Agree / N/A	Fail/ Fail	Fail/ Fail	N/A / Agree	N/A / Fail	N/A / N/A
Rule 14: Prepositions 2	N/A / N/A	N/A / N/A	N/A / N/A	N/A / Fail	N/A / Fail	N/A / N/A
Rule 15: Verbs	Fail / Fail	Agree / Fail	Agree / Fail	N/A / N/A	N/A / Fail	N/A /N/A
Rule 16: Abbreviations	N/A/ N/A	N/A/ N/A	N/A / Fail	N/A/ N/A	N/A / Fail	N/A/ N/A

Table 5: Rules' performance on 13 failed bracketings in the WPC II

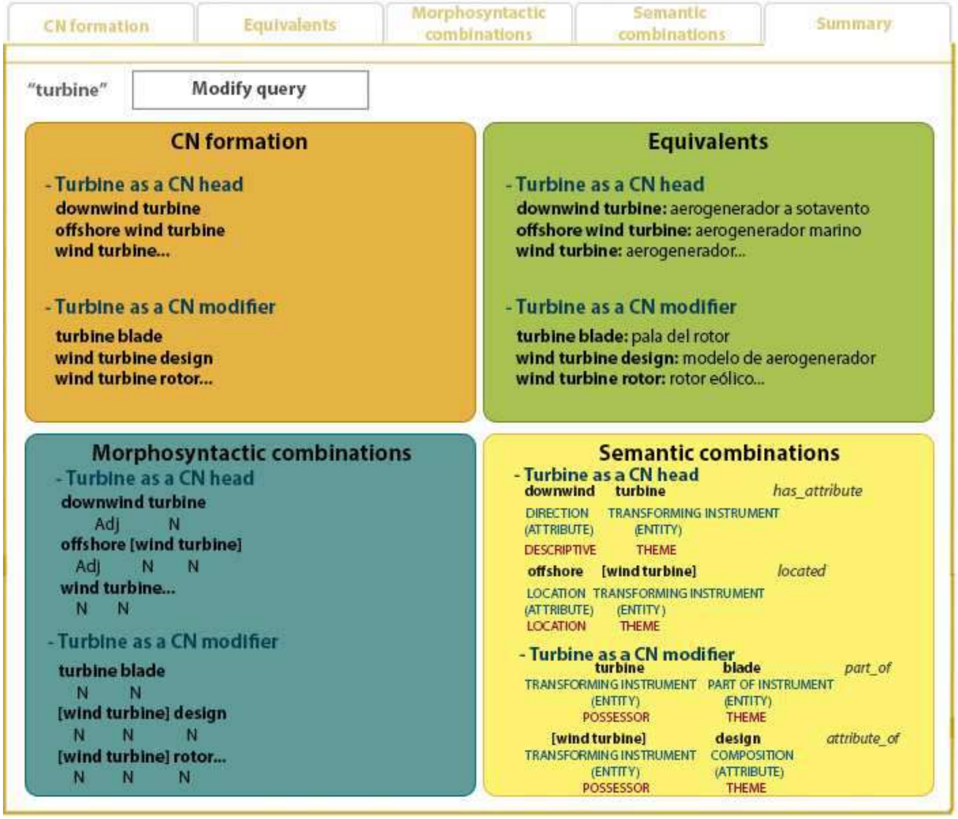
In any case, the protocol delivered promising results that could be applied in any terminology management scenario needing a thorough description of MWTs.

3. Multiword-term representation in terminological knowledge bases

An accurate representation of MWTs in terminological knowledge bases involves providing users with access to the implicit information codified in such specialised units, namely their structural dependencies and the semantic relations encoded among the constituents. Since the second depends on the first, automatising bracketing facilitates the inclusion of such information. Furthermore, establishing equivalence and performing cross-lingual comparisons are only possible through the semantics implied.

In EcoLexicon, a new module for the description of MWTs has been designed. When users query a monolexical term, they can access all of the MWTs where the search term appears as a constituent, whether it is the head or a modifier. Figure 4 shows the summary view of four different tabs where different types of information are provided, in this case regarding the search term *turbine*.

The results of this view are a summary of what is obtained in the specific views that will be described below, namely (i) MWT formation, (ii) Equivalents, (iii) Morphosyntactic combinations, and (iv) Semantic combinations. As can be observed in Figure 4, the CN formation bubble shows some of the MWTs that include the term turbine. These examples are also shown in the Equivalents bubble along with their main Spanish equivalents. The Morphosyntactic combinations bubble focuses on bracketing and part-of-speech tagging. Finally, the Semantic combinations bubble also shows bracketing, as well as annotation with semantic categories (blue), semantic roles (red), and the internal semantic relation (grey, on the right).



CN formation

- Turbine as a CN head
downwind turbine
offshore wind turbine
wind turbine...

- Turbine as a CN modifier
turbine blade
wind turbine design
wind turbine rotor...

Equivalents

- Turbine as a CN head
downwind turbine: aerogenerador a sotavento
offshore wind turbine: aerogenerador marino
wind turbine: aerogenerador...

- Turbine as a CN modifier
turbine blade: pala del rotor
wind turbine design: modelo de aerogenerador
wind turbine rotor: rotor eólico...

Morphosyntactic combinations

- Turbine as a CN head
downwind turbine
Adj N
offshore [wind turbine]
Adj N N
wind turbine...
N N

- Turbine as a CN modifier
turbine blade
N N
[wind turbine] design
N N N
[wind turbine] rotor...
N N N

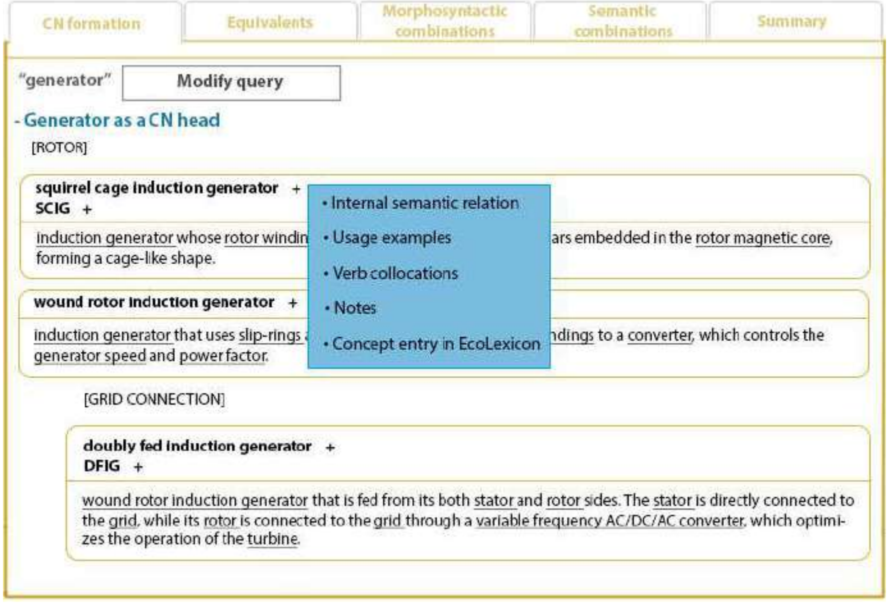
Semantic combinations

- Turbine as a CN head
downwind turbine *has_attribute*
DIRECTION TRANSFORMING INSTRUMENT
(ATTRIBUTE) (ENTITY)
DESCRIPTIVE THEME
offshore [wind turbine] *located*
LOCATION TRANSFORMING INSTRUMENT
(ATTRIBUTE) (ENTITY)
LOCATION THEME

- Turbine as a CN modifier
turbine blade *part_of*
TRANSFORMING INSTRUMENT PART OF INSTRUMENT
(ENTITY) (ENTITY)
POSSESSOR THEME
[wind turbine] design *attribute_of*
TRANSFORMING INSTRUMENT COMPOSITION
(ENTITY) (ATTRIBUTE)
POSSESSOR THEME

Figure 4: Summary view of the MWT module in EcoLexicon

Figure 5 shows an extract from the MWT formation tab, where the term *generator* is shown as the head of three terms hierarchically organised, linked to their definitions and highlighted conceptual dimensions (i.e. rotor or grid connection) as well as related term variants (i.e. *SCIG*, *DFIG*). MWTs whose modifier is generator (e.g. *generator torque control*) can also be obtained.



- Generator as a CN head

[ROTOR]

squirrel cage induction generator +
SCIG +

induction generator whose rotor windin ars embedded in the rotor magnetic core, forming a cage-like shape.

wound rotor induction generator +

induction generator that uses slip-rings ndings to a converter, which controls the generator speed and power factor.

[GRID CONNECTION]

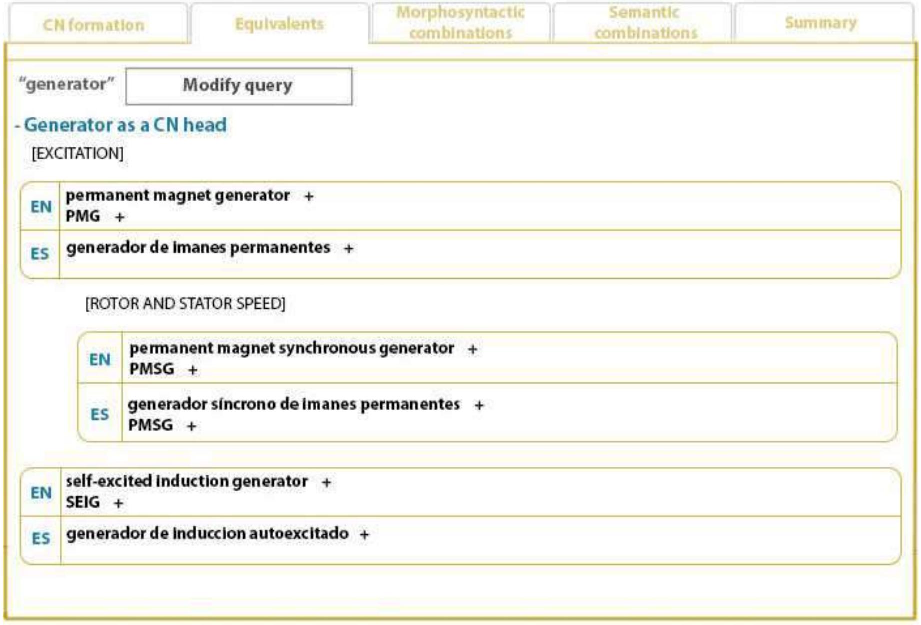
doubly fed induction generator +
DFIG +

wound rotor induction generator that is fed from its both stator and rotor sides. The stator is directly connected to the grid, while its rotor is connected to the grid through a variable frequency AC/DC/AC converter, which optimizes the operation of the turbine.

Figure 5: Extract from the MWT formation tab for *generator*

By clicking on the plus sign next to each term, users can access additional information: (i) internal semantic relations between the constituents of the MWT, (ii) usage examples, (iii) verb collocations; (iv) notes, (v) and the main term entry in the knowledge base. The *internal semantic relation* option shows the MWT head and modifier, as well as the semantic relation that links them. In MWTs formed by more than two constituents, bracketing facilitates this distinction between head and modifier, and is thus included in this view (e.g. *wound rotor induction generator* > [wound rotor] *part_of* [induction generator]).

Figure 6 shows an extract from the MWT equivalents tab, where the MWTs with *generator* as their head are now related to their corresponding terms in Spanish. Additional languages, such as French, are planned to be included in the near future. The same secondary options are offered as in the previous view, except for the definition, which is included here as a secondary option.



"generator"

- Generator as a CN head

[EXCITATION]

EN	permanent magnet generator + PMG +
ES	generador de imanes permanentes +

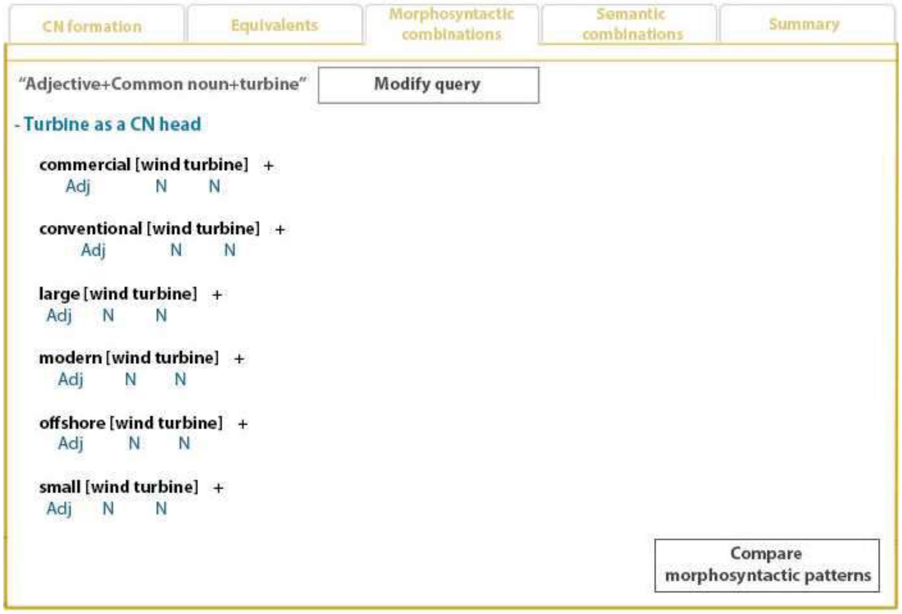
[ROTOR AND STATOR SPEED]

EN	permanent magnet synchronous generator + PMSG +
ES	generador síncrono de imanes permanentes + PMSG +

EN	self-excited induction generator + SEIG +
ES	generador de induccion autoexcitado +

Figure 6: Extract from the MWT equivalents tab for *generator*

Figure 7 shows an extract from the morphosyntactic combinations tab, where the MWTs with *turbine* as their head are presented according to their morphosyntactic structure and bracketing.



"Adjective+Common noun+turbine"

- Turbine as a CN head

commercial [wind turbine] +
Adj N N

conventional [wind turbine] +
Adj N N

large [wind turbine] +
Adj N N

modern [wind turbine] +
Adj N N

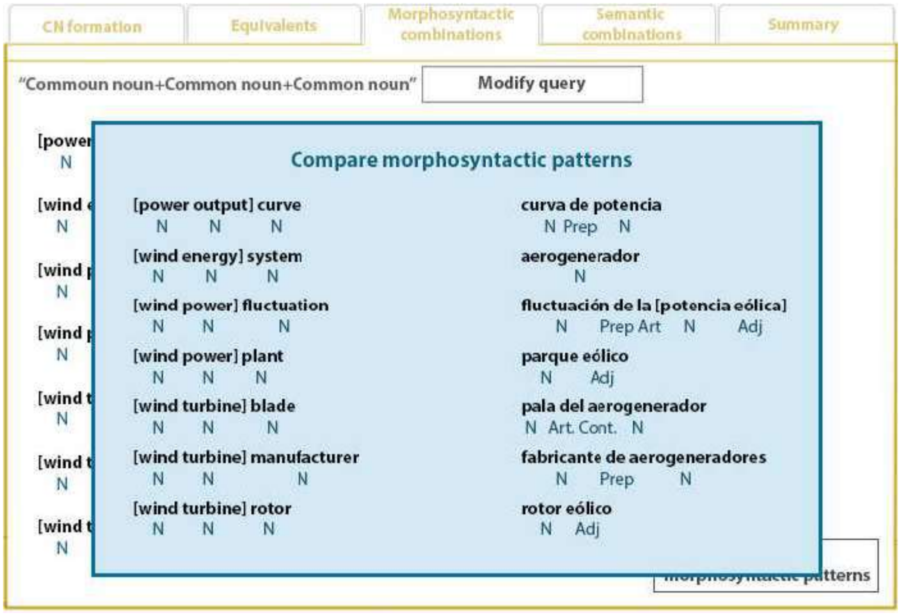
offshore [wind turbine] +
Adj N N

small [wind turbine] +
Adj N N

Figure 7: Extract from the Morphosyntactic combinations tab for *turbine*

By clicking on the plus sign next to each term, users can access additional information. In this view, the semantic relation is not provided since such semantic information is not relevant in this section. However, bracketing plays a central role, as it facilitates morphosyntactic analysis and MWT management.

Furthermore, when clicking in Compare morphosyntactic patterns, a bilingual view will be displayed (Figure 8). The results that meet the search criteria will be shown, together with their main variants in the target language. These are annotated with the part-of-speech of each constituent, so that the morphosyntactic patterns of term formation in both languages can be compared. Users can also observe that bracketing does not always correspond in the two languages (e.g. when the equivalent has fewer constituents, as in *power output curve* and *curva de potencia*).


CN formation
Equivalents
Morphosyntactic combinations
Semantic combinations
Summary
"Commoun noun+Common noun+Common noun"
Modify query

Compare morphosyntactic patterns

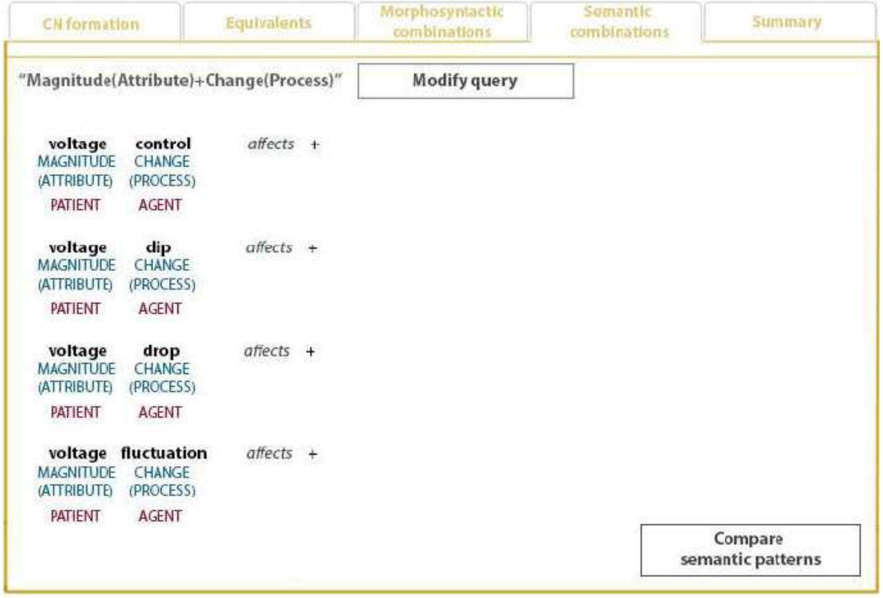
[power output] curve N N N	curva de potencia N Prep N
[wind energy] system N N N	aerogenerador N
[wind power] fluctuation N N N	fluctuación de la [potencia eólica] N Prep Art N Adj
[wind power] plant N N N	parque eólico N Adj
[wind turbine] blade N N N	pala del aerogenerador N Art. Cont. N
[wind turbine] manufacturer N N N	fabricante de aerogeneradores N Prep N
[wind turbine] rotor N N N	rotor eólico N Adj

Figure 8: Extract from the Compare morphosyntactic combinations tab for *turbine*

Figure 9 shows an extract from the semantic combinations tab, where the semantic categories MAGNITUDE(ATTRIBUTE) and CHANGE(PROCESS) are queried to obtain the MWTs that include them. The MWTs retrieved are tagged with their bracketing structure (if they have three or more constituents), and their semantic categories and roles. For instance, in *voltage control*, *voltage* belongs to the category of MAGNITUDE(ATTRIBUTE) and *control* belongs to the category of CHANGE(PROCESS). In this MWT, *control* is the agent since it affects *voltage*, the patient. Next to each MWT, its internal semantic relation is also shown.

By clicking on the plus sign next to each term, users can access additional information. Unlike the previous views, an *additional semantic information* option is provided, which displays more specific data for users interested in further conceptual characterisation.

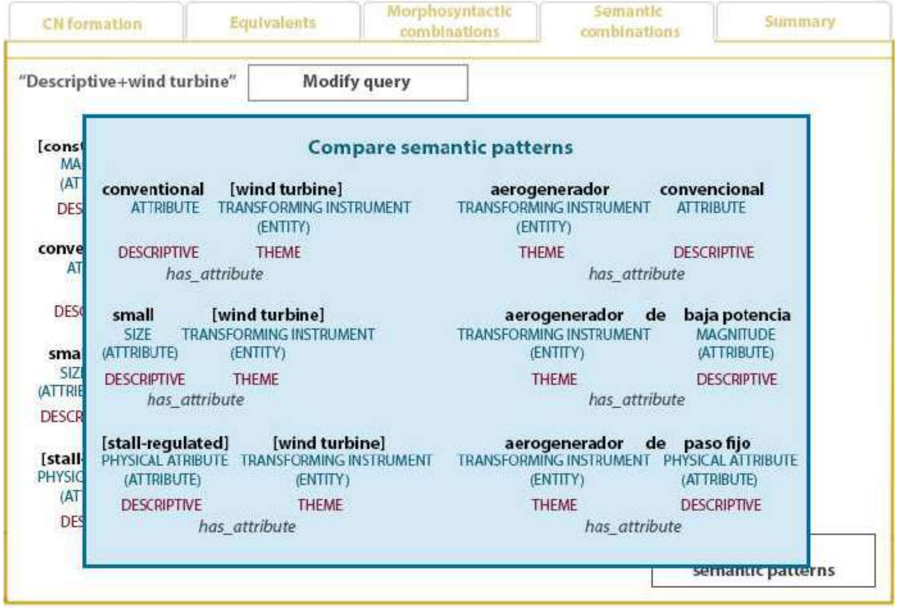
The *Compare semantic patterns* option is also provided (Figure 10). This section can be used to compare the semantic pattern of our results with that of their translation equivalents. A cross-linguistic approach to common phenomena such as variation or multidimensionality can thus be obtained, and the semantic annotation of MWTs in both languages can be contrasted (e.g. *small wind turbine*, based on size, vs its equivalent *aerogenerador de baja potencia*, based on power). Not surprisingly, bracketing is the key to ascertaining the basic parts of MWTs and facilitate their understanding.



"Magnitude(Attribute)+Change(Process)"

voltage	control	affects +
MAGNITUDE (ATTRIBUTE)	CHANGE (PROCESS)	
PATIENT	AGENT	
voltage	**dip**	*affects* +
MAGNITUDE (ATTRIBUTE)	CHANGE (PROCESS)	
PATIENT	AGENT	
voltage	**drop**	*affects* +
MAGNITUDE (ATTRIBUTE)	CHANGE (PROCESS)	
PATIENT	AGENT	
voltage	**fluctuation**	*affects* +
MAGNITUDE (ATTRIBUTE)	CHANGE (PROCESS)	
PATIENT	AGENT	

Figure 9: Extract from the Semantic combinations tab



"Descriptive+wind turbine"

Compare semantic patterns

conventional	**[wind turbine]**	**aerogenerador**	**convencional**
ATTRIBUTE	TRANSFORMING INSTRUMENT (ENTITY)	TRANSFORMING INSTRUMENT (ENTITY)	ATTRIBUTE
DESCRIPTIVE	THEME	THEME	DESCRIPTIVE
has_attribute		*has_attribute*	
small	**[wind turbine]**	**aerogenerador de**	**baja potencia**
SIZE (ATTRIBUTE)	TRANSFORMING INSTRUMENT (ENTITY)	TRANSFORMING INSTRUMENT (ENTITY)	MAGNITUDE (ATTRIBUTE)
DESCRIPTIVE	THEME	THEME	DESCRIPTIVE
has_attribute		*has_attribute*	
[stall-regulated]	**[wind turbine]**	**aerogenerador de**	**paso fijo**
PHYSICAL ATRIBUTE (ATTRIBUTE)	TRANSFORMING INSTRUMENT (ENTITY)	TRANSFORMING INSTRUMENT (ENTITY)	PHYSICAL ATTRIBUTE (ATTRIBUTE)
DESCRIPTIVE	THEME	THEME	DESCRIPTIVE
has_attribute		*has_attribute*	

Figure 10: Extract from the Compare semantic patterns tab

4. Conclusions

In this paper, a bracketing protocol has been presented together with its practical application in the design and compilation of a MWT module in a terminological knowledge base. Regarding the protocol, we concluded that the most productive rules are adjacency, longer MWTs, dependency, shortening, and paraphrases.

It is also advisable to perform the queries in domain-specific corpora, and not necessarily large ones. When large corpora are available, other surface patterns might prove more useful in terms of precision.

As for the MWT module described in this paper, it is intended to be useful for a wide variety of users, ranging from translators and interpreters, terminologists and technical writers, to students and environmental specialists. This resource includes different types of information that assists in both comprehension and production tasks. A systematic approach was adopted with a view to enhancing the heterogeneous description of MWTs in language resources, as well as specific problems such as the lack of consideration of internal dependencies or bracketing.

5. Acknowledgements

This research was carried out as part of project *Translation-oriented Terminology Tools for Environmental Texts* (FFI2017-89127-P), funded by the Spanish Ministry of Economy and Competitiveness.

6. References

Barrière, C. & Ménard, P. A. (2014). Multiword noun compound bracketing using Wikipedia. In *Proceedings of the First Workshop on Computational Approaches to Compound Analysis.* ACL and Dublin City University, pp. 72–80.

Cabezas-García, M. & Faber, P. (2017). A Semantic Approach to the Inclusion of Complex Nominals in English Terminographic Resources. In R. Mitkov (ed.) *Computational and Corpus-Based Phraseology*, Lecture Notes in Computer Science, 10596. Cham: Springer, pp. 145-159.

Cabezas-García, M. (2019). *Los compuestos nominales en terminología: formación, traducción y representación.* PhD dissertation. Granada, Universidad de Granada.

Cabezas-García, M. & León-Araúz, P. (2019). On the Structural Disambiguation of Multi-word Terms. In G. Corpas Pastor & R. Mitkov (eds.) *Computational and Corpus-Based Phraseology*, Lecture Notes in Computer Science, 11755. Cham: Springer, pp. 46-60.

Cabezas-García, M. (2020). *Los términos compuestos desde la Terminología y la Traducción.* Berlin: Peter Lang.

EcoLexicon. Accessed at: https://ecolexicon.ugr.es/. (1 February 2021)

Girju, R., Moldovan, D., Tatu, M. & Antohe, D. (2005). On the semantics of noun compounds. *Computer Speech & Language*, 19(4), pp. 479-496.

Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery.* Kluwer Academic Press.

Kilgarriff, A., Rychly, P., Smrz, P. & Tugwell, D. (2004). The Sketch Engine. In G.

Williams & S. Vessier (eds.) *Proceedings of the Eleventh EURALEX International Congress*. EURALEX, pp. 105-116.

Lauer, M. (1995). *Designing Statistical Language Learners: Experiments on Noun Compounds*. PhD dissertation. Australia, Macquarie University.

León-Araúz, P. & Cabezas-García, M. (in press). Evaluating a bracketing protocol for multiword terms. In *Recent Advances in Multiword Units in Machine Translation and Translation Technology.* John Benjamins.

Marcus, M. (1980). *A Theory of Syntactic Recognition for Natural Language.* MIT Press.

Nakov, P. (2007). *Using the Web as an Implicit Training Set: Application to Noun Compound Syntax and Semantics.* PhD dissertation. Berkeley, University of California at Berkeley.

Nakov, P. & Hearst, M. (2005). Search engine statistics beyond the n-gram: application to noun compound bracketing. In *Proceedings of the Ninth Conference on Computational Natural Language Learning, CoNLL 2005.* ACL, pp. 17–24.

Pustejovsky, J., Anick, P. & Bergler, S. (1993). Lexical semantic techniques for corpus analysis. *Computational Linguistics,* 19(2), pp. 331–358.

Frame-based terminography: a multi-modal knowledge base for karstology

Špela Vintar[1], Vid Podpečan[2], Vid Ribič[3]

[1] University of Ljubljana, Aškerčeva 2, SI – 1000 Ljubljana
[2] Jožef Stefan Institute, Jamova 39, SI – 1000 Ljubljana
[3] Kofein dizajn, Beethovnova 9, SI – 1000 Ljubljana
E-mail: spela.vintar@ff.uni-lj.si , vid.podpecan@ijs.si, vid@kofein.si

Abstract

We present an innovative approach to the representation of domain-specific knowledge which combines traditional concept-oriented terminography with knowledge frames and augments linguistic data with images, videos, interactive graphs and maps. The interface is simple and intuitive, prompting the user to enter a query term in any of the three languages (English, Croatian and Slovene). If the term is found it is described through textual definitions from various sources, its frame derived from annotated data, a graph depicting the neighbourhood of the concept and – if feasible – a map of geolocations for the queried term. The frame represents aggregated and structured knowledge as it describes the concept through a set of semantic relations. Graphs enable the user to browse through related concepts and explore the domain in a visually represented network. The underlying knowledge base of karstology was created within the TermFrame project and is based on an implementation and extension of the frame-based approach to terminology.

Keywords: frame-based terminography; karstology; knowledge base; visualisation

1. Introduction

The notion of frames as templates of knowledge structures (Faber, 2009; Faber et al., 2011) has found great resonance in the field of terminology as it efficiently combines the textual, contextual and cognitive layers of knowledge into a comprehensive theoretical and practical framework. In the TermFrame project we approach the domain of karstology from an interdisciplinary perspective to create a multilingual and multi-modal interactive knowledge base tailored to different types of users: domain experts, students, and researchers, but also non-experts interested in karst.

Karstology itself is an interdisciplinary field studying karst, a special type of landscape which develops on soluble rocks such as limestone or gypsum. Typical karst landmarks include caves, sinkholes, various rock formations and complex water systems with streams which may sink and continue their flow subterraneously. Apart from being a field of interest for geography, hydrology, speleology and geology, karst systems – especially caves – are popular tourist destinations and important areas of environmental protection, which is why we envisage interested non-experts as potential users of our knowledge base.

The web user interface to the knowledge base is designed in line with the principles of usability as defined by Jakob Nielsen through the following five key features (Nielsen, 1996): *learnability* (how simple the interface is for a first-time visitor), *efficiency* (how quickly the user can complete their task), *memorability* (how well does the user master the interface after a period of non-use), *errors* (the number of errors the user makes during use, their gravity and difficulty of correction), and *satisfaction* (how pleasing the interface design is).

The remainder of this article is structured as follows: After a brief overview of related work in Section 2 we dedicate Section 3 to the various sources of information for our knowledge base. We describe the resources, processing steps and tools used to create each of the layers presented to the user. Section 4 focuses on the mode of presentation itself and the rationale of designing the search interface so that it can be accessible and usable for all of our potential target groups. We conclude with a brief discussion and plans for future work.

2. Related work

Frame-based approaches to terminology (FBT; Faber, 2012) have become mainstream in the past decade. While the EcoLexicon as the first of its kind continues to improve and expand (Faber et al., 2016; León-Araúz et al., 2019), other authors and projects integrate frames or conceptual templates into their knowledge representations (Roche et al., 2019; Bihua et al., 2020; Giacomini, 2018).

Since specialised knowledge is often conceptualised as a network, numerous examples of knowledge visualisations in the form of graphs can be listed, such as multilingual databases of colexification patterns CLICS [1] (Mayer et al., 2014), Wikipedia visualisation (WikiGalaxy[2]) or biological domain knowledge exploration software such as Biomine Explorer (Podpečan et al., 2019). The latter implements a rich network visualisation and manipulation interface which sits on top of the Biomine search engine serving the relevant parts of the enormous Biomine network according to the user's query. Cytoscape (Shannon et al., 2003) is one of the most important examples of feature-complete network analysis software. While it was originally developed for biological research, it has since grown into a general, extensible platform for complex network analysis and visualisation. Gephi (Bastian et al., 2009) implements very efficient algorithms for the visualisation of extremely large networks, but does not implement many data integration options and is thus limited to visualisation and basic analysis of general networks. OmicsNet (Zhou and Xia, 2018) implements a visual analytics platform for multi-omics integration and features 3D visualisation in the browser.

[1] http: //clics.lingpy.org

[2] http://wiki.polyfra.me/

3. Resources for the TermFrame knowledge base

3.1 Concepts and textual definitions

The creation of our trilingual knowledge base for karstology was performed in stages (cf. Vintar et al., 2019). First, specialised corpora in English, Croatian and Slovene were compiled, ensuring optimal coverage of the domain. The corpora are comparable and contain relevant contemporary works on karstology, including books, articles, doctoral and master's theses, glossaries and encyclopaedia. The composition of the corpus is described in more detail in Vintar and Stepišnik (2020). The English subcorpus contains just under two million words, while the Slovene and the Croatian subcorpora are smaller and together consist of around one million words.

Some of the corpus texts were available only in printed format, so that a full digitisation procedure was required, including scanning, OCR and manual proofreading; others were obtained directly from publishers, authors and internet sources. For some of the texts copyright issues remain unresolved, and such texts were used only as a source of definitions and their digitised versions have been discarded. The cleared part of the comparable corpus will be released through the Clarin.si repository[3].

In the second stage, definitions of karst concepts were collected from the TermFrame corpora using the ClowdFlows definition extraction tool (Pollak et al., 2012). The final data set consists of 725 annotated definitions for English, 786 for Slovene and 661 for Croatian. All definitions were manually annotated in accordance with our domain model specifying the semantic categories and relations relevant for karstology (cf. Vintar et al., 2019). An example of an annotated definition can be seen in Figure 1.

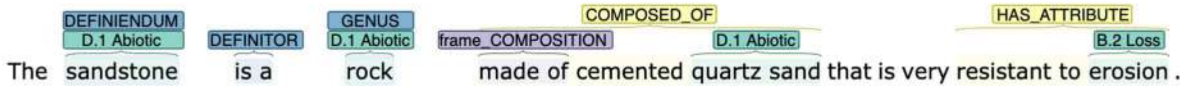

The sandstone is a rock made of cemented quartz sand that is very resistant to erosion .

Figure 1: Annotated definition in WebAnno

The domain model specifies five top-level categories dividing karst terms into Landforms, Processes, Geomes, Entities/Properties and Instruments/Methods. Each category is associated with a set of semantic relations used to define or describe it; these combinations can also be referred to as definition templates and help organise and represent knowledge in a systematic manner. In addition to the categories and relations, each definition is also analysed for definition elements, so that we annotate the DEFINIENDUM, GENUS and SPECIES (the latter is relevant for extensional definitions).

[3] https://www.clarin.si/repository/xmlui/

The definitions in all three languages are contained in a common database where each definiendum – which can be in English, Slovene or Croatian – is assigned a concept ID, thus linking equivalents to a specific and unique meaning. A concept may have several definitions in one language (most notably karst, for which there are as many as 13 English definitions) and several terms designating it, or it may not have an equivalent in all three languages.

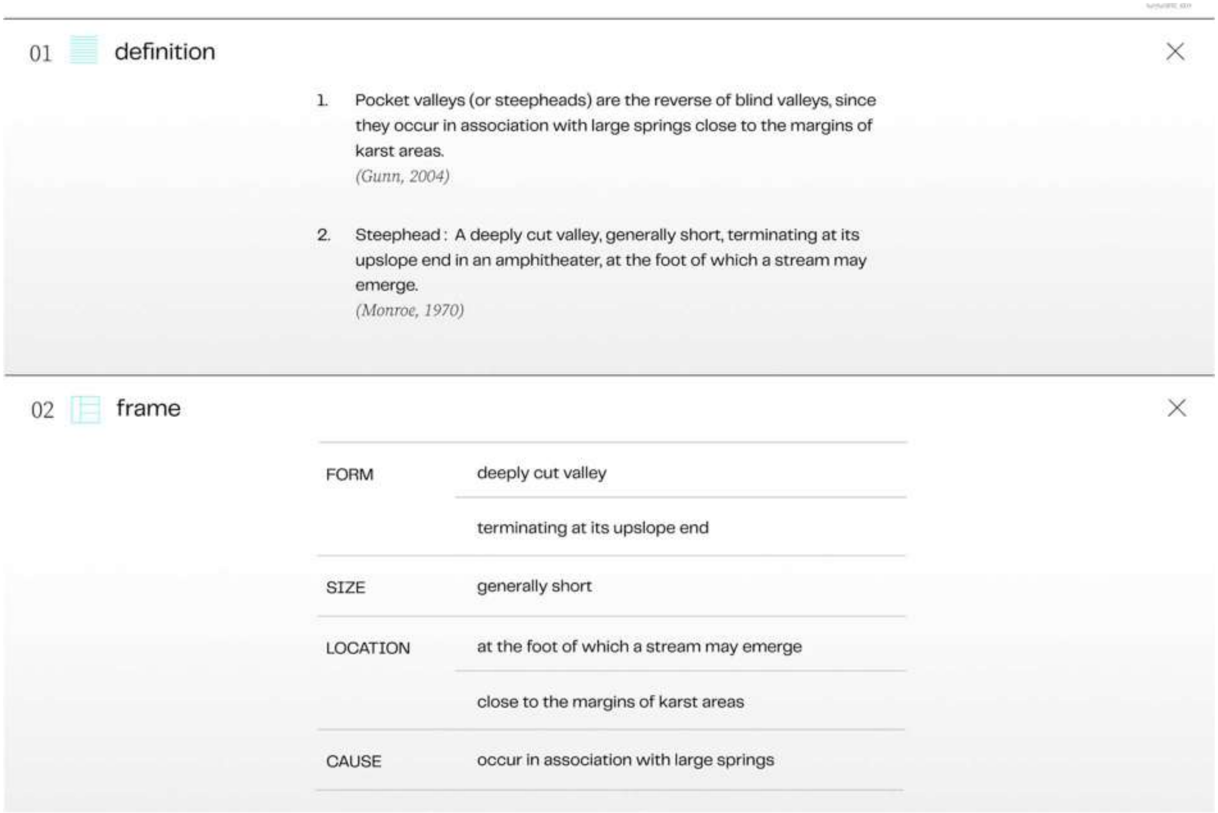
3.2 Representing frames

In order to allow further processing of annotated definitions in all three currently supported languages they have to be exported from the WebAnno annotation software (Eckart de Castilho et al., 2016). We use the common .tsv format which is one of the available outputs of WebAnno. Due to the complexity of the annotated data, any simple text format (including .csv) is ill-suited for this task. The following issues need to be handled by the parser in order to extract the correct and complete data.

- The annotation of a text is composed of annotation blocks which contain annotations of sentences. These blocks are separated by empty lines and comments.
- Single cells may contain additional inner separation characters.
- An annotation can span any number of cells in the same column, either in a contiguous block or possibly separated with other annotations.
- Annotations spanning multiple tokens are characterised by annotation serial numbers (counters) in square brackets following the annotation name. However, serial numbers are not present in annotations spanning single tokens.

We implemented the parser using the popular Pandas framework,[4] which offers several data manipulation and selection features which made our task easier. First of all, the csv parser is configured so that the .tsv export of WebAnno is stored correctly into an internal data structure (Pandas' DataFrame). Then, the complete annotation data is split into sentence annotation blocks using sentence ID as the grouping key. The possibility of intra-cell separation is handled next by duplicating the row for each such value while assigning a new, unique index. This is followed by extracting the actual tokens belonging to each annotation. Pandas' powerful data selection functions are used to simplify this task. Finally, the complete annotation data is stored in an internal format and ready to be converted into a format suitable for the representation of frames in a table or visualisation in a graph.

[4] https://pandas.pydata.org/



1. Pocket valleys (or steepheads) are the reverse of blind valleys, since they occur in association with large springs close to the margins of karst areas.
(Gunn, 2004)

2. Steephead : A deeply cut valley, generally short, terminating at its upslope end in an amphitheater, at the foot of which a stream may emerge.
(Monroe, 1970)

FORM	deeply cut valley
	terminating at its upslope end
SIZE	generally short
LOCATION	at the foot of which a stream may emerge
	close to the margins of karst areas
CAUSE	occur in association with large springs

Figure 2: Definitions and frame

The data presented as the frame of the query term collects all annotated semantic relations from different definitions and displays them in the order of the "ideal" definition. Thus, if Surface landforms are typically defined through their FORM, SIZE, LOCATION and CAUSE, the frame tab will list all strings from the definitions that had been annotated as either of these relations. The main added value of the frame-based approach is that the information about the term is aggregated from different textual sources, and that it is structured in a manner which reflects the cognitive template surrounding the Surface landform concept category.

3.3 Visualisation

There are several possibilities how to define a graph structure using the extracted annotations. Currently, graphs are created according to the following rules applied to each sentence annotation block.

1. For every "definiendum" definition element create:
   a. a node from its tokens,
   b. a node from its category, and
   c. a directed edge named *has_category* from the token node to the category node.

2. For every "genus" node create a node from its tokens.
3. For every "definiendum" token node and every "genus" token node create a directed edge named *is_a* from the first node to the second node.
4. For every "relation" definition element create:
    a. a node from its tokens,
    b. a directed edge from the "definiendum" token node to the "relation" token node, and give it the name of the relation.

The visualisation backend software stack consists of the following components. First, the data loader provides fast loading from serialised data structures containing the graphs with the topology as described above. Second, the graph extraction component performs subgraph extraction according to input parameters. Currently, one or more nodes can be used as the input query. The extractor performs neighbourhood search from the specified nodes using the currently default depth limit of 2 and returns the resulting subgraph. Finally, the exporter serialises the extracted subgraph into a selected format. We use JSON to pass the subgraph data to the frontend, but several other formats are supported and can be used for server-side processing or for download.

The visualisation of the graph corresponding to the user query is implemented using the open source vis.js library[5] which is a dynamic, browser based visualisation library. It enables interactive and efficient visualisation of reasonably large graphs (up to a few thousand nodes). In our case, however, the size of graphs is limited to only few dozens of nodes because of the neighbourhood search depth limit of 2.

When the JSON containing the graph data is received from the backend, a vis.js DataSet structure is created first. It contains information about nodes and edges and any additional node and edge data that is required by the graph visualisation user interface. Then, a visualisation canvas is created and populated with the contents of the DataSet. Several visualisation parameters are set to values which enable clear visualisation of small knowledge graphs.

The graph displayed alongside the query is interactive in the sense that each node which corresponds to a term in our knowledge base can be clicked by the user. This action runs a new query so that the entire results window is refreshed and a new set of definitions, frame, graph, etc. is displayed.

3.4 Images and videos

Since most of our karst concepts pertain to tangible landscape entities, we obtained a collection of images and videos depicting karst phenomena. Images are labelled with concept IDs and integrated into the search interface. Images and aerial photographs of

[5] https://visjs.org/

karst forms and processes were obtained during systematic field surveys and morphographic mapping for documentation and field research of karst conducted by Dr. Uroš Stepišnik and colleagues from 2006 to 2021. Apart from karst documentation and research, the visual materials are also used for didactic purposes in teaching the physical geography of karst at the Department of Geography at the University of Ljubljana (Stepišnik, 2020). Classical photographic equipment and unmanned aerial vehicles were used for photographic documentation. The image and video material is available for the purposes of the TermFrame project under the CC-BY-NC-ND license.

3.5 Maps

For the most central and frequent karst landforms which are described in our corpus through actual geolocations, we created maps displaying these locations. Place names were automatically extracted using the GeoNames.org database as a source of global geographical names and REZI[6], a publicly available registry of geographical names for Slovenia and Croatia. The extracted names were supplemented with GPS coordinates and imported into Google MyMaps to create maps of documented locations of the relevant landform.

4. Designing the interface

The search interface is designed to be as simple and user-friendly as possible, focusing primarily on usability for non-linguists. The user can enter a karst term in any of the three languages and the results will be displayed in tabs. After the image or video, the user can read all the definitions for the concept from different sources, then view the "framed" definition, browse a clickable graph of related terms and, if available, see the locations of the concept on the map.

The web interface is a WordPress installation with some custom modifications tailored to the needs of our project. We have developed a database importer in order to easily import new terms into the website. The importer processes the entries from a csv file and maps them to the corresponding posts in a WordPress database. On top of that we have a cron job which obtains the data for the graph visualisation via API. A cron job is a simple software utility that schedules tasks to run at certain time intervals in the future. This API is specifically developed for this project and returns information about nodes and edges for any karst concept we have in the database. This data is subsequently processed so it can be used with a vis.js library to display the graph in the frontend.

[6] https://egp.gu.gov.si/egp/?lang=en

4.1 Target audience

The first step in designing the user experience is to define the target users of the interface. In our case the interface addresses several target groups of experts and non-experts. Experts from the domains of geography, karstology and speleology will be able to consult the karst knowledge base during their work, compare definitions by different authors and browse for similar concepts. Linguists and terminologists will explore mainly the linguistic aspects of the terms and their definitions, and both groups will benefit from the equivalents in other languages, related concepts and graphs thereof.

The more general target group of non-experts will explore karst phenomena through images, videos, textual descriptions and maps.

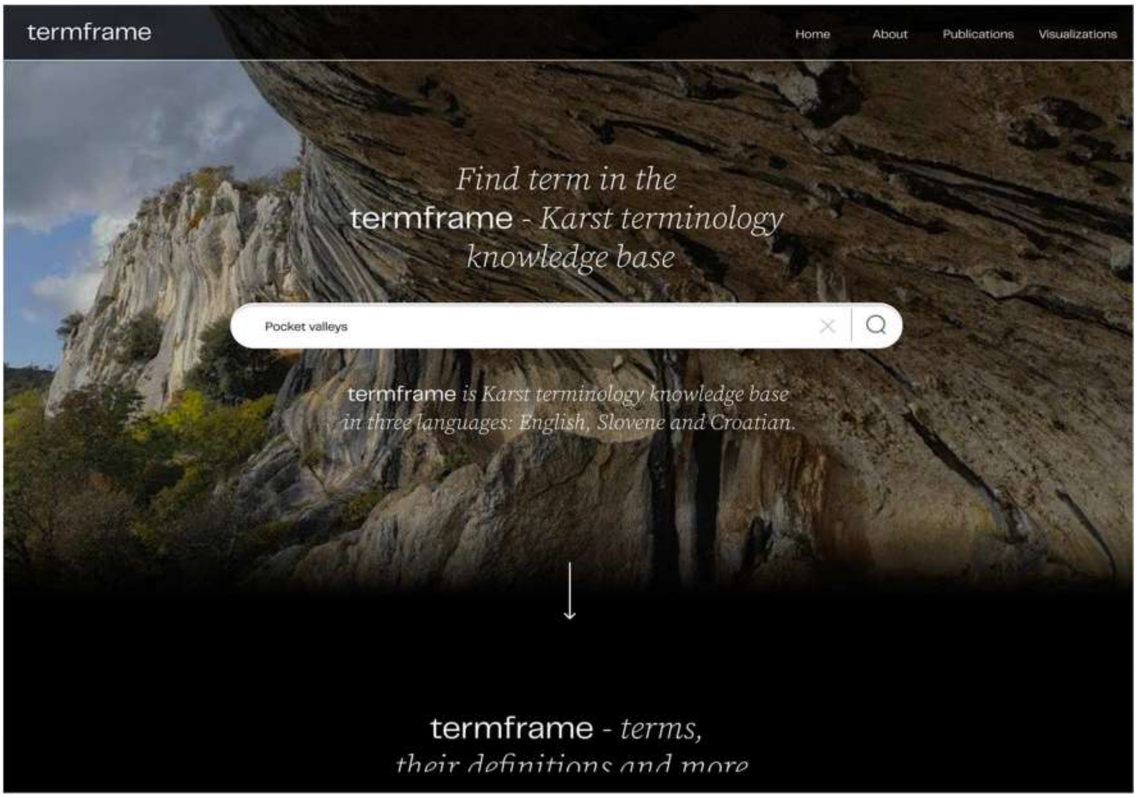
4.2 Browse vs. search

While designing the user interface we first needed to resolve the question of how to represent the knowledge base to facilitate user access and satisfy the five Nielsen criteria of usability mentioned above. The choice was between two user scenarios, browse or search, whereby each has its advantages and disadvantages. Browsing allows the user to search through a list or hierarchy. If the list is unordered, the search time increases linearly with the number of items to choose from. Since our knowledge base contains over 1,700 terms, such browsing would be extremely inefficient.

4.3 User journey

All target groups share the same mode of access to the knowledge base, but upon receiving a response to the query the user may select the most relevant type of content presentation. First, the user enters a query into the search field (for example "pocket valleys", see Figure 3).

We therefore selected searching as the access scenario. The user enters a query term and immediately receives hits – provided the knowledge base contains the query term. Browsing can be resumed via the graph of related terms where the number of items to choose from is considerably smaller, while still allowing the user to explore without knowing exactly what to search for. The search engine will first display the results in the language of the query term, but the user may switch languages if the same concept is described in the other two languages.

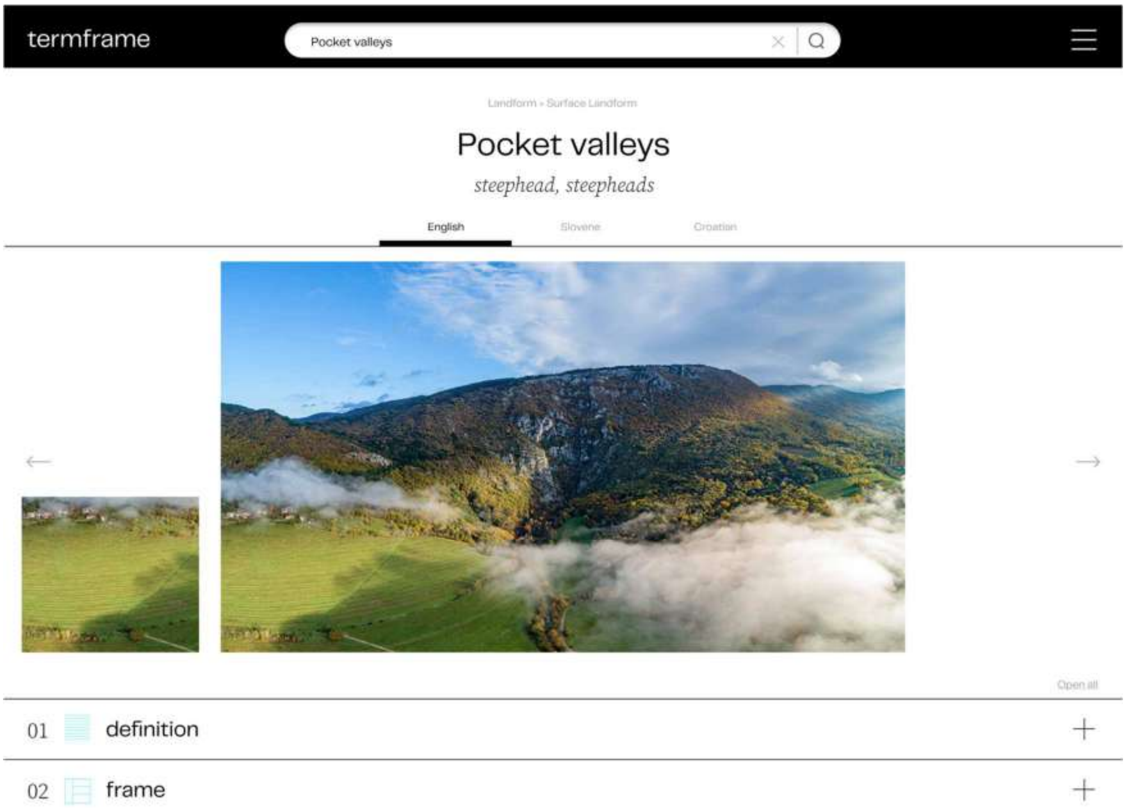


Find term in the
termframe - *Karst terminology knowledge base*

termframe *is Karst terminology knowledge base in three languages: English, Slovene and Croatian.*

termframe - *terms, their definitions and more*

Figure 3: Main search field



Pocket valleys

steephead, steepheads

01 definition +

02 frame +

Figure 4: Displaying the found term

The query term is displayed in the header of the page, together with the semantic category and subcategory above the term and its synonyms below it (Figure 4). Under the image or video is a list of four expandable tabs for the user to choose from. A domain expert will presumably focus on the definitions and the frame (Figure 2), a linguist might explore the graph (Figure 5), and a non-expert user might open the map (Figure 6) and look for locations of the karst phenomenon.

The web site contains two additional tabs. Under Visualisations, several versions of the entire knowledge network are presented displaying selected layers of information (e.g. terms and categories, terms and geni, terms and relations). The Publications tab lists the complete bibliography of project-related articles.

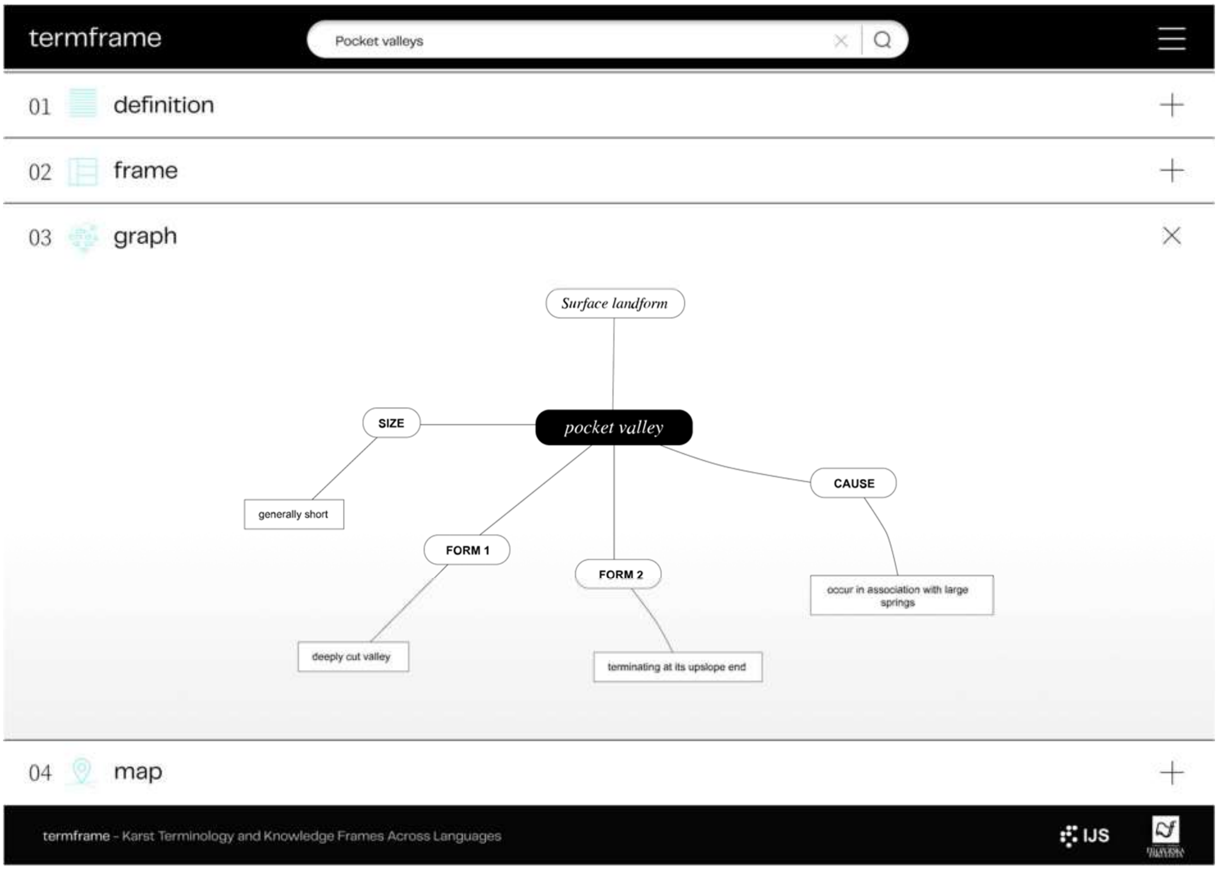


Figure 5: Graph

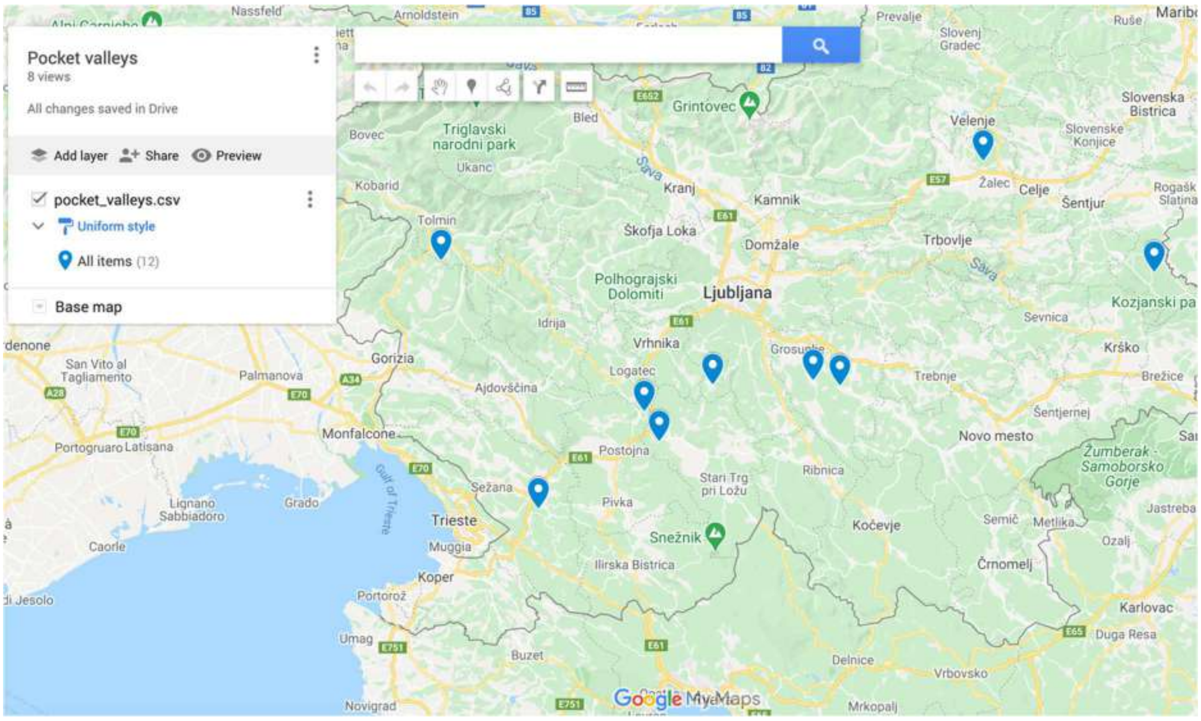


Figure 6: Map of pocket valleys in Slovenia

5. Conclusion

We have described a new resource for karstology which presents structured knowledge in an attractive and innovative manner. The rationale of the design is that even highly specialised knowledge which has partly been obtained using complex text mining techniques can still be accessible and visually compelling. The frame-based restructuring of definitions seems a promising approach which links the textual level of knowledge with the cognitive, spatial and visual spheres.

Since the user interface is still being completed at the time of writing, no usability studies have been performed yet. An evaluation of the web interface by different target groups remains one of our goals for the future. Since karst phenomena in Slovenia and Croatia, but also elsewhere, attract large numbers of visitors who may be interested to explore the karst knowledge base on a hand-held device, we envisage the development of an app which would incorporate location data to the display of maps and images.

Upon project completion (by the end of 2021), several datasets will be made available through the Clarin.si repository for English, Slovenian and Croatian: 1. the TermFrame corpora (except for the works for which distribution was explicitly denied), 2. the extracted and semantically annotated definitions, and 3. the parsed annotations in table format which can be used for visualisation or other form of analysis. The online knowledge base described above is available online without registration.

6. Acknowledgments

This work is supported by the Slovenian Research Agency grant J6-9372 TermFrame: Terminology and knowledge frames across languages, 2018-2021, research programme grant P6-0215 Slovene Language - Basic, Contrastive, and Applied Studies (2019-2024) and by European Union's Horizon 2020 research and innovation programme under grant agreement No. 825153, project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media).

7. References

Bastian, M., Heiman, S. & Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. International AAAI Conference on Weblogs and Social Media.

Bihua, Q. I. U. (2020). A Frame-based Version of NATO Glossaries. China Terminology, 22(3), p. 33.

Eckart de Castilho, R., Mújdricza-Maydt, É., Yimam, S. M., Hartmann, S., Gurevych, I., Frank, A. & Biemann, C. (2016). A Web-based Tool for the Integrated Annotation of Semantic and Syntactic Structures. In Proceedings of the LT4DH workshop at COLING 2016, Osaka, Japan.

Faber, P. (2009). The Cognitive Shift in Terminology and Specialized Translation. MonTI. Monografías de Traducción e Interpretación 1, pp. 107-134. https://doi.org/10.6035/MonTI.2009.1.5

Faber, P., León-Araúz, P. & Reimerink, A. (2011). Knowledge representation in EcoLexicon. *Technological innovation in the teaching and processing of LSPs: proceedings of TISLID* 10, pp. 367-386.

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

Faber, P., León-Araúz, P., & Reimerink, A. (2016). EcoLexicon: new features and challenges. GLOBALEX, pp. 73-80.

Giacomini, L. (2018). Frame-based Lexicography: Presenting Multiword Terms in a Technical E-dictionary. In *Proceedings of the XVIII EURALEX International Congress.*

Hick, W.E. (1952). On the rate of gain of information. Quarterly Journal of Experimental Psychology. 4 (4:1), pp. 11–26. doi:10.1080/17470215208416600

León-Araúz, P., Reimerink, A. & Faber, P. (2019). EcoLexicon and by-products: Integrating and reusing terminological resources. Terminology, 25 (2), pp. 222-258.

Mayer T., Terhall, A. & Urban, M. (2014). An Interactive Visualization of Crosslinguistic Colexification Patterns. In Proceedings of VisLR: Visualization as Added Value in the Development, Use and Evaluation of Language Resources, pp. 1-8.

Nielsen, J. (1996). Usability metrics: tracking interface improvements. In IEEE

Software, vol. 13, no. 6, pp. 1-2, Nov. 1996, doi: 10.1109/MS.1996.8740869.

Podpečan, V., Ramšak, Ž., Gruden, K., Toivonen, H. & Lavrač, N. (2019). Interactive exploration of heterogeneous biological networks with Biomine Explorer. Bioinformatics, 24 June 2019, pii: btz509, doi: 10.1093/bioinformatics/btz509.

Pollak, S., Vavpetič, A., Kranjc, J., Lavrač, N. & Vintar, Š. (2012). NLP workflow for on-line definition extraction from English and Slovene text corpora. Vienna: KONVENS 2012, pp. 53-60.

Pollak, S., Podpečan, V., Miljkovic, D., Stepišnik, U. & Vintar, Š. (2020). The NetViz terminology visualization tool and the use cases in karstology domain modeling. In Proceedings of the 6th International Workshop on Computational Terminology (COMPUTERM 2020), pp. 55–61.

Roche, C., Costa, R., Carvalho, S. & Almeida, B. (2019). Knowledge-based terminological e-dictionaries: The EndoTerm and al-Andalus Pottery projects. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, *25*(2), pp. 259-290.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research.13(11), pp. 2498–504.

Stepišnik, U. (2020). Fizična geografija krasa. Ljubljana: Znanstvena založba Filozofske fakultete.

Vintar, Š., Saksida, A., Vrtovec, K. & Stepišnik, U. (2019). Modelling specialized knowledge with conceptual frames: The TermFrame approach to a structured visual domain representation. In I. Kosem et al. (eds.) Proceedings of eLex 2019, pp. 305-318.

Vintar, Š. & Stepišnik, U. (2020). TermFrame: A Systematic Approach to Karst Terminology. Dela, (54), pp. 149-167. https://doi.org/10.4312/dela.54.149-167

Zhou, G. & Xia, J. (2018). OmicsNet - a web-based tool for creation and visual analysis of biological networks in 3D space. Nucleic Acids Research (doi:10.1093/nar/gky510).

A cognitive perspective on the representation of MWEs in electronic learner's dictionaries

Thomai Dalpanagioti

School of English Language and Literature, Aristotle University of Thessaloniki, 54124
Thessaloniki, Greece
E-mail: thomdalp@enl.auth.gr

Abstract

One of the main pending methodological issues in lexicography is the representation of multiword expressions (MWEs). Their heterogeneous and fuzzy nature has given rise to diverse typologies in linguistic theory and to a variable and inconsistent treatment in lexicographic practice. Addressing this issue in the context of pedagogical lexicography is of vital importance because, due to a complex interplay of features of form, meaning and use, MWEs present major difficulties for learners as regards reception, production and retention. This paper thus examines the representation of different types of MWEs in online versions of English monolingual learner's dictionaries and points out the need for a more rational, motivated and systematic lexicographic treatment. We argue for a cognitively oriented approach to MWEs that draws on Frame Semantics and the Conceptual Metaphor and Metonymy Theory. The proposal is illustrated through two case studies, which demonstrate how MWEs are integrated in a motivated semantic network of the motion verbs *crawl* and *dash*. The flexibility of the electronic medium can make it feasible to design cognitively informed features of the dictionary microstructure to improve the representation of MWEs.

Keywords: multiword expressions; monolingual learner's dictionaries; Frame Semantics; Conceptual Metaphor and Metonymy Theory; motion verbs

1. Introduction

This paper is motivated by the elusive nature of multiword expressions (MWEs) which are notoriously difficult to handle in lexicography. Although dictionary practices continuously develop, it remains unclear how MWEs should be represented in dictionaries. By overcoming space constraints and making new search paths feasible, the potential of the electronic medium has been widely recognised (de Schryver, 2003; Atkins & Rundell, 2008). MWEs have received much attention from a lexicographic and natural language processing perspective (for an overview see Gantar et al., 2019). However, challenges still remain at both macro- and microstructural levels, and the lack of "a comprehensive theoretical approach to the treatment of all types of MWEs in lexicography" is noted (ibid.: 143).

Focusing on English monolingual learner's dictionaries (MLDs) as representatives of the most recent developments in lexicography, several studies have observed considerable variation in the treatment of MWEs (e.g. Atkins & Rundell, 2008: 394-397; Walker, 2009: 289-291). For example, the same MWEs have been recorded under different entries and in a different manner, e.g. as fixed expressions needing an explanation or as simple examples, highlighted within "focus boxes" or indicated by special labels (e.g. "idiom", "phrasal verb", "phrase"). The lack of consistency in the selection and wording of MWEs seems to result from differences in what each dictionary regards as collocation, idiom, etc., and from the large number of variant forms observed in corpora. At the level of the macrostructure the consultation process may have become easier due to access flexibility in electronic dictionaries; MWEs can be retrieved automatically wherever entered as long as they have received "lemma-sign status" (de Schryver, 2003: 178; Atkins & Rundell, 2008: 253). However, at the level of the microstructure no major change has been made in the description or arrangement of MWEs; they are usually presented as a list of hyperlinks at the end of an entry with no clear indication of how they are connected to the lemma's semantic network (Wojciechowska, 2020).

Against this background and on account of the user perspective in MLDs, we propose that cognitive semantic theories, namely Frame Semantics (Fillmore, 2006 [1982]) and the Conceptual Metaphor and Metonymy Theory (Lakoff & Johnson, 1980), can help us improve the lexicographic treatment of most MWEs since they are – at least to some extent – motivated. Considering Lakoff's (1987: 346) claim that "it is easier to learn something that is motivated than something that is arbitrary", the paper draws examples from a small-scale corpus-based and cognitively-oriented pre-lexicographic database for motion verbs to outline an informed and more user-friendly treatment of MWEs.

To set the scene, section 2 discusses MWEs from a typological and lexicographic perspective, while section 3 considers what cognitive semantic theories can contribute to the ongoing question of the representation of MWEs in dictionaries. Section 4 demonstrates the practical solutions proposed through two case studies focusing on the manner-of-motion verbs *crawl* and *dash*. By reviewing the treatment of the *crawl-* and *dash-* MWEs in online versions of MLDs and reconstructing the microstructures of the entries, we illustrate a cognitively informed treatment of MWEs – complementary to the preliminary corpus-based extraction of typical word combinations.

2. MWEs and lexicographic issues

MWEs have long been a focus of great interest in the field of lexicology and lexicography due to their pervasive but also fuzzy nature. From a theoretical perspective, numerous attempts have been made to capture the complex interaction of idiomaticity and flexibility, giving rise to terminological diversity. From a lexicographic perspective, however, the representation of MWEs in dictionaries has not been extensively researched, and "the status of MWEs in lexicography still remains unsettled" (Wojciechowska, 2020: 584). This study does not aim to offer one more classification of MWEs; rather it uses Gantar et al.'s (2019) integrative typology as a point of reference with a view to discussing the lexicographic treatment of MWEs in two case studies.

Bringing together three classifications (i.e. Atkins & Rundell's, 2008: 164, Bergenholtz & Gouws's, 2014, and Baldwin & Kim's, 2010), Gantar et al. (2019) present a lexicographically relevant typology consisting of seven types of MWEs: collocations (e.g. *severe criticism*), fixed phrases and idioms (e.g. *to have a heart of gold*), compounds (e.g. *lame duck*), proverbs (e.g. *half a loaf is better than no bread*), phrasal verbs (e.g. *take off*), light-verb constructions (e.g. *take a walk*), and prepositional phrases (e.g. *with regard to*). This typology is built on gradable criteria such as collocability, contiguity, idiomaticity, compositionality, figuration and fixedness (ibid.: 141-142). In fact, despite variation in terminology it is generally agreed that there is a scalar relationship between types of MWEs exhibiting gradability of one or more of the following broad dimensions: (a) semantic/pragmatic specialisation and metaphoricity, (b) lexico-grammatical fixedness/variation, and (c) frequency of occurrence (for an overview see e.g. Dalpanagioti, 2018: 425-427). However, not only are there fuzzy borders between different types of MWEs, but also between co-occurrence patterns in the broad sense of typical contextual environment and the narrower sense of MWEs (ibid.). As Fellbaum (2016: 412) points out, “there are no hard rules to distinguish between merely preferred co-occurrences and more or less fixed collocations that arguably have lexical status”.

The interplay of features of form, meaning and use makes the representation of MWEs in dictionaries a challenge. Decisions regarding “what”, “where” and “how” are not easy to take, and thus there is a lack of consistency in the lexicographic treatment of MWEs. For example, Oppentocht and Schutz (2003: 218) observed that phraseological entities “can often be found under more than one entry, in different forms, and even with different explanations”, while more recently Gantar et al. (2019: 156) underlined the need for standardisation in categorising and tagging MWEs in dictionary databases and identifying their canonical forms and variants. Relevant in this respect is Bergenholtz and Gouws’s (2014) call for differential treatment of MWEs in light of users’ needs (reception vs. production) and dictionary function (communicative vs. cognitive). Learner’s dictionaries in particular should rise to the challenge of representing both their meaning and full range of usage (Fellbaum, 2016: 424).

Corpus data and the electronic medium have opened exciting possibilities for learner’s dictionaries. As regards phraseological information, developments mainly concern its coverage and access (Lew, 2012: 349-351; Paquot, 2015: 469; Dziemianko, 2017: 669; Wojciechowska, 2020). An increasing number of word combinations seems to be channelled into electronic dictionaries though various microstructural components (e.g. definitions, examples, subentries, boxes), while more effective search options are also offered (e.g. fuzzy matching, type-ahead search, menus, signposts, hyperlinks). However, the potential of the electronic medium has not yet been fully realised, and suggestions to further this include developing user-friendly customisation options and blending electronic dictionaries with learning environments (Lew, 2012: 353, 361), systematically specifying word combinations in terms of genre and register (Paquot, 2015: 470), integrating corpus-query tools into dictionary platforms (Paquot, 2015: 476), and reflecting the semantic relations between MWEs (Wojciechowska, 2020). Elaborating on the last research direction, this study argues for a cognitively oriented approach to MWEs.

3. The potential contribution of cognitive semantic theories

There seems to be a growing trend to advocate the application of cognitive linguistics in lexicography (see e.g. Geeraerts, 1990; Fillmore & Atkins, 1992; Van der Meer, 1999; Moon, 2004; Molina, 2008; Wojciechowska, 2012; Kövecses & Csábi, 2014; Jiang & Chen, 2015; Ostermann, 2015; Xu & Lou, 2015; Wiliński, 2016; Dalpanagioti, 2019). As Geeraerts (2007: 1168) explains, what cognitive linguistics can contribute to lexicography is a more realistic conception of semantic structure. While corpus linguistics has revolutionised lexicography by providing access to vast amounts of authentic language data and foregrounding the role of context, cognitive linguistics can make dictionary entries more reasonable and streamlined. Relevant studies mainly propose ways of ordering and defining senses to make semantic relations more transparent; however, MWEs have not received much attention. In this context, the present study aims to demonstrate how the combined use of Frame Semantics and the Conceptual Metaphor and Metonymy Theory can help improve the treatment of MWEs in electronic dictionaries.

The main assumption of Frame Semantics is that words must be grouped and explained in relation to a "(semantic) frame", i.e. a structured background of experience which constitutes a kind of prerequisite for understanding the meaning of a word (Fillmore, 1985: 224). Every semantic frame consists of specific "frame elements" (FEs), i.e. the "various participants, props, and other conceptual roles" involved in the schematic representation of a situation (Fillmore & Petruck, 2003: 359). Frame semantics links these situation-specific semantic roles to their syntactic realisations (grammatical functions and phrase types), thus specifying valence in both semantic and syntactic terms.

Targets of annotation in the Berkeley FrameNet project are typically single words but can also be MWEs such as phrasal verbs (e.g. *give in* in the frame [Giving_in]) or idioms (e.g. *kick the bucket* in the frame [Death]) (Ruppenhofer et al., 2016: 21). Focusing on predicates with a clear syntax-semantics mapping, FrameNet marks MWEs only with a Target label with no FE/grammatical function/phrase type annotation (ibid.: 59). However, MWEs receive special attention in the context of another frame semantic project for German, the SALSA (SAarbrücken Lexical Semantics Annotation and Analysis) project, which addresses the issue of metaphor representation. What is proposed for single-word and multi-word metaphors is a double annotation scheme with “a source frame representing the literal meaning, and a target frame representing the figurative meaning” (Burchardt et al., 2009: 216); by contrast a single frame annotation is assigned to (pure) idioms. Since the strategy of double frame semantic annotation allows for capturing both the overall meaning (target frame) and the internal structure (source frame) of metaphorical MWEs, it could be a useful starting point for a motivated lexicographic treatment.

Conceptual motivation has been discussed in relation to idiomatic expressions within the framework of Conceptual Metaphor and Metonymy Theory (as laid out by Lakoff & Johnson, 1980) and its application in language learning. For instance, Gibbs (1993) argues that there are thousands of idioms which, without being predictable, seem to be motivated partially by metaphorical/metonymic schemes of thought very much alive in everyday reasoning. Similarly, Dobrovol'skij (2011: 56) defines motivation as "transparency of conceptual links between source and target" and posits that "there are many idioms which are not semantically analyzable, and yet they are motivated". Applied cognitive linguistic studies point out the pedagogical benefits of raising learners' awareness of motivated meaning and semantic networks; for example, Boers and Lindstromberg (2006) and Kövecses (2012) make special reference to the usefulness of conceptual metaphor in the comprehension and retention of figurative idioms.

The implications of Conceptual Metaphor and Metonymy Theory for pedagogical lexicography are mostly discussed in relation to ordering and defining senses. For instance, Van der Meer (1999) argues that making learners aware of the extensions of words, by ordering senses in the dictionary from literal to figurative, can facilitate vocabulary learning. Similarly, it is important to show the relation between senses in the wording of definitions; as Lew (2013: 299) explains, "foregrounding the links between different shades of meaning may help repair some of the damage done by artificially chopping semantic space into separate dictionary senses". Lexicographic applications of the Conceptual Metaphor and Metonymy Theory to the treatment of MWEs can be traced in specialised dictionaries for phrasal verbs or idioms, which seek to express the underlying conceptual motivation (for an overview see Kövecses & Csábi, 2014: 129-130), and in the "metaphor boxes" of the MEDAL (print and electronic) dictionaries (for an overview see Moon, 2004). Metaphor boxes provide an explanation of a metaphorical concept in terms of the mapping between source and target domains, and group together illustrative examples for words and phrases that realise the mapping; they were developed for about 60 concepts and have been placed in the macrostructure near the relevant target domain headword to facilitate encoding in L2.

Within the context of corpus-based, electronic, pedagogical lexicography, we use two case studies as a framework for making suggestions that move beyond reference to one MWE type (e.g. idioms) or customisable macrostructural arrangement (e.g. metaphor-based). We proceed to demonstrate how insights from Frame Semantics and Conceptual Metaphor and Metonymy Theory can be systematically combined to improve the treatment of MWEs.

4. Case studies: *to crawl* and *to dash*

Whereas metalexicographic studies can be selective about the MWEs examined for the purposes of illustration, in practical lexicographic work an exhaustive analysis of the polysemy and phraseology of words is required. To discuss the role and (actual and proposed) treatment of MWEs within the framework of a holistic lexicographic portrait, we present two case studies that draw data from a pre-lexicographic database for motion verbs; for a short description of the corpus-based and cognitively oriented features of the database see Dalpanagioti (2018: 422-423). Examining the entries for the verbs *crawl* and *dash*, we focus on the microstructural representation of MWEs of various types; in terms of Gantar et al.'s (2019) typology, they can be classified as collocations (*crawl the Net/web*, *dash someone's hopes*), idioms (*crawl out of the woodwork*, *make your skin/flesh crawl*), proverbs: routine/situational formulas (*I must dash*, *dash it all*), and phrasal verbs (*crawl with*, *dash off*). We thus proceed to first compare the "Big Five" MLDs with regard to their representation of MWEs (section 4.1), and then to present an alternative cognitively informed treatment (section 4.2).

4.1 The treatment of MWEs in the "Big Five" MLDs

Aspects of form, meaning and presentation of MWEs are examined in the *crawl* (v) and *dash* (v) entries of the online editions of OALD, LDOCE, COBUILD, CALD and MEDAL. To facilitate the comparative analysis of the data, we have collected the relevant information for the MWEs accessed through the *crawl* (v) and *dash* (v) entries in Table 1 and Table 2, respectively.

With regard to coverage, we do not expect to find great differences, since all these dictionaries are corpus-informed. Striking instances, nevertheless, are *crawl the Net/web* and *crawl back to*, which are recorded in only one dictionary, i.e. LDOCE and CALD respectively.[1] Variant forms, such as *make your skin/flesh crawl*, *come/crawl out of the woodwork*, *dash it/dash it all*, seem to be consistently recorded with only slight differences. Similarly, there is agreement on the semantic and pragmatic information reflected in definitions and labels; in particular, corpus-derived information on implications and register restrictions seem to be systematically provided.

However, variation can be observed with regard to the arrangement of MWEs. Although hyperlinking MWEs to a separate entry seems to be the most common practice among the five MLDs, there are various positions in which hyperlinks are placed. More precisely, MWE hyperlinks may appear as separate senses (e.g. *dash somebody's hopes* in LDOCE, *make your skin/flesh crawl* in COBUILD), in an “idioms” or “phrasal verbs” box (e.g. *dash off* in OALD and CALD), in a right-hand panel with more results (e.g. *crawl/come out of the woodwork* in LDOCE and CALD), or in both a box and a right-hand panel (e.g. *make your skin/flesh crawl* and *dash it (all)* in MEDAL). When MWEs are not hyperlinked they are defined and illustrated in the main entry as a separate sense (a typical practice in COBUILD) or in a sub-entry in a box (a strategy preferred by OALD), or, less often, they are located among illustrative examples without being highlighted (e.g. *I must dash* in COBUILD and CALD).

[1] In fact, the Word Sketches for *crawl* (v) in two web corpora available through Sketch Engine (i.e. ukWaC and enTenTen18) confirm the high frequency of its occurrence with nouns denoting a Web location such as *Web*, *Internet*, *website*, *net*, etc. (semantic preference). In contrast, there is not enough evidence to support the recording of *crawl back to* as an idiom.

OALD	**LDOCE**	**COBUILD**	**CALD**	**MEDAL**
be crawling with	*be crawling with something*	*be crawling with*	*be crawling with sb/sth*	*crawl with* (usually progressive)
(informal) to be full of or completely covered with people, insects or animals, in a way that is unpleasant	to be completely covered with insects, people etc.	If you say that a place is crawling with people or animals, you are emphasizing that it is full of them. [informal, emphasis]	to be full of insects or people in a way that is unpleasant	1. to be full of people in a way that is unpleasant 2. to be covered in insects
label: phrasal verb hyperlink in a box	sense 6 hyperlink	sense 4	sense signpost: ‘Fill’	label: phrasal verb hyperlink in a box
make your skin crawl	*make somebody’s skin crawl*	*to make your skin crawl* or *make sb's flesh crawl*	*make sb's skin crawl*	*make your skin/flesh crawl*
to make you feel afraid or full of horror	(informal) to make someone feel very uncomfortable or slightly afraid	If something makes your skin crawl or makes your flesh crawl, it makes you feel shocked or disgusted.	If someone or something makes your skin crawl, you think they are very unpleasant or frightening	to give you a very unpleasant and slightly frightened feeling

label: idiom sub-entry in a box	hyperlink in the “More results” panel	sense 6 hyperlink	label: idiom hyperlink in the “More meanings” panel	label: phrase hyperlink in a box & in the “Other entries for this word” panel
come/crawl out of the woodwork	*crawl/come out of the woodwork*	- *(come out of the woodwork)*	*come/crawl out of the woodwork*	*come/crawl out of the woodwork*
(informal, disapproving) if you say that somebody comes/crawls out of the woodwork, you mean that they have suddenly appeared in order to express an opinion or to take advantage of a situation	if someone crawls out of the woodwork, they suddenly and unexpectedly appear in order to take advantage of a situation, express their opinion etc. – used to show disapproval		(mainly disapproving) to appear after having been hidden or not active for a long time	to suddenly appear after a long time, especially for unpleasant reasons
label: idiom sub-entry in a box	hyperlink in the “More results” panel		label: idiom hyperlink in the “More meanings” panel	label: phrase hyperlink in the “Other entries for this word” panel
-	*crawl the Net/web*	-	-	-

	if a computer program crawls the Net, it quickly searches the Internet to find the particular information you need			
	sense 7 hyperlink			
-	-	-	*crawl back (to sb)*	-
			to admit that you were wrong and ask someone to forgive you or ask them for something that you were offered and refused in the past	
			label: idiom hyperlink in a box	

Table 1: *Crawl* MWEs in the "Big Five" MLDs

OALD	**LDOCE**	**COBUILD**	**CALD**	**MEDAL**
I must dash	*(I) must dash/(I) have to dash*	*dash* (*I have to dash/ must dash* in examples; not highlighted)	*I must dash*	*I must dash/I have to dash*
I must dash (= leave quickly), *I'm late.*	*(British English, spoken)* used to tell someone that you must leave quickly	If you say that you have to dash, you mean that you are in a hurry and have to leave immediately. [informal]	UK *I must dash - I've got to be home by seven.*	used for saying that you must leave quickly because you are in a hurry
example under sense 1	sense 3 hyperlink	sense 2	example under sense ‘Move quickly’	label: phrase spoken hyperlink in a box & in the “Other entries for this word” panel
dash somebody’s hopes	*dash somebody’s hopes*	*dash* (*dash hopes* in examples)	*dash sb's hopes*	*dash someone’s hopes*
to destroy somebody’s hopes by making what they were hoping for	to disappoint someone by telling them that what they	If an event or person dashes someone's hopes or expectations, it destroys	to destroy someone's hopes	to make it impossible for someone to do what

impossible	want is not possible	them by making it impossible that the thing that is hoped for or expected will ever happen. [journalism, literary]		they hoped to do
label: idiom sub-entry in a box	sense 2 hyperlink	sense 6	label: idiom hyperlink in a box	label: phrase hyperlink in a box & in the “Other entries for this word” panel
dash (it)! / dash it all!	*dash it (all)!*	*dash/ dash it/ dash it all*	*dash*	*dash it (all)*
(old-fashioned, British English) used to show that you are annoyed about something	*(British English, old-fashioned)* used to show that you are slightly annoyed or angry about something	You can say dash or dash it or dash it all when you are rather annoyed about something. [British, informal, old-fashioned, feelings]	(*UK, old-fashioned, informal*) used to express anger	used when you are annoyed about something
label: idiom sub-entry in a box	sense 5 hyperlink	label: exclamation sense 10	label: exclamation separate entry: *dash* (*Oh dash (it)!* as an example)	label: phrase informal old-fashioned hyperlink in a box & in the “Other entries for this word” panel

dash something off	*dash off*	*dash off*	*dash sth off*	*dash off*
to write or draw something very quickly	1. to leave somewhere very quickly 2. *dash something off* to write or draw something very quickly	1. If you dash off to a place, you go there very quickly. 2. If you dash off a piece of writing, you write or compose it very quickly, without thinking about it very much.	to write something quickly, putting little effort into it	1. [intransitive] to leave quickly or suddenly because you are in a hurry 2. [transitive] to write or draw something quickly because you are in a hurry
label: phrasal verb hyperlink in a box	label: phrasal verb hyperlink	label: phrasal verb hyperlink	label: phrasal verb hyperlink in a box	label: phrasal verb hyperlink in a box & in the “Other entries for this word” panel

Table 2: *Dash* MWEs in the "Big Five" MLDs

Besides dictionary-specific preferences, it is important to notice how the same MWEs are classified across the dictionaries and whether the same MWE types are treated consistently. As regards classification, in Table 1 and Table 2 we can find clear-cut cases like *dash off*, which is labelled as "phrasal verb" and accessed through a hyperlink in all dictionaries, but also more challenging cases like *be crawling with* and *dash somebody's hopes*, which are tagged as fixed phrases ("phrasal verb", "idiom") in some dictionaries and as contextual realizations of a sense in others. As regards the question of consistency, there does not seem to be an identifiable type-specific treatment. Irrespective of whether MWEs are collocations, idioms, phrasal verbs or situational formulas, the general tendency is to present them separately from the main entry (in separate hyperlinked entries or in separate boxes in the entry) and even when they appear among numbered senses there is no indication of their relation.

To sum up, based on the examination of the sample entries we can conclude that corpus analysis has led to a high degree of consistency in the representation of MWE variant forms, meanings, implications and illustrative examples. However, corpus analysis cannot address the issue of linking semantically related units into a coherent network unless combined with an appropriate theoretical model. Focusing thus on the "where" and "how", rather than on the "what", we outline a cognitively oriented representation of MWEs in the two case studies.

4.2 A cognitively informed treatment of MWEs

Instead of detaching MWEs from the main entry, we propose incorporating them in the network of lexical units (LUs). Drawing information from a database that has applied a corpus-based and cognitively oriented methodology to establishing LUs (Dalpanagioti, 2013; 2018), we reconstruct the skeletal structure of the entries *crawl* (v) and *dash* (v). The semantic networks of the verbs appear in Table 3 and Table 4, and demonstrate the links between single-word and multi-word LUs.[2]

Since separate senses generally correspond to different semantic frames and assign different FEs (Atkins, Rundell & Sato, 2003: 335-337), we cluster corpus uses and distinguish LUs (single-word and multi-word ones) based on FrameNet's frames.[3] To lend further support to the frame-based sense

[2] Corpus examples are not included in Table 3 and Table 4 because the study focuses on arranging and presenting LUs rather than establishing them based on corpus uses; besides, there seems to be considerable agreement in the senses and uses provided in the MLD entries examined above. Variant forms of MWEs have been clustered together under the same LU (see e.g. the idiom schema *make someone's skin/flesh/scalp crawl*).

[3] Descriptions of all FrameNet frames mentioned in Table 3 and Table 4 are available online at https://framenet.icsi.berkeley.edu/fndrupal. The only exception is the [Self_motion]$_{\text{figurative}}$ frame (*crawl*, LU4) which has been introduced and described in Dalpanagioti (2013: 17-19).

distinctions, we consider how they are motivated by the cognitive mechanisms of metaphor and metonymy. Promoting a cognitive-based rather than a frequency-based approach to the ordering of LUs (Van der Meer, 1999: 203-4; Lew, 2013: 293), we proceed from literal to metonymic to metaphorical extensions and organise LUs into a tiered structure with two main clusters of related senses in each table.

While in Table 3 all LUs correspond to discrete frames, in Table 4 we notice that the frames [Departing] and [Cause_impact] are mentioned twice. This is due to our decision to distinguish between LUs that evoke the same frame, when corpus uses exhibit distinct semantic-pragmatic nuances not reflected in frame distinctions (e.g. *dash it (all)* is separated from the other [Cause_impact] uses because it serves a special discoursal function). However, in combining semantic and contextual criteria for determining LUs, we pay particular attention not to elevate mere contextual variations to the status of an LU, because it is easy to lose sight of the semantic integrity of words by means of excessive splitting (Atkins & Rundell, 2008: 313). Relevant in this respect is the collocation *dash someone's hopes* (Table 4, LU7), which is treated as a usage pattern rather than as a stand-alone LU.

The (pre-lexicographic) cognitive semantic analysis presented in Table 3 and Table 4 has practical implications for the representation of MWEs in online MLDs. First of all, it is evident that all instances of the various MWE types examined are motivated, i.e. they have clear conceptual links with other LUs. However, these are not reflected in current dictionary practices, which create distance between semantically related LUs, for instance, by hyperlinking MWEs to separate entries or listing them in separate boxes. What is suggested instead

Clusters of senses	**LU**		**Frame**	**Conceptual motivation**
I. Motion	1	move along with the body close to the ground	[Self-motion]	literal sense; natural locomotion of insects/ reptiles with legs and literal extension to the motion of human beings (toddlers) on the basis of similarity of posture
	2	Phrasal verb: *crawl with something/ someone* (progressive colligation) be covered/ crowded with movers (creatures or people)	[Abounding_with]	CONTAINER FOR CONTENT metonymy from LU1; shift of emphasis from the SELF-MOVERS to the LOCATION where motion takes place
	3	Idiom: *make someone's skin/ flesh/ scalp crawl* make someone feel fear or revulsion	[Stimulate_emotion]	motivated by the [Abounding_with] LU and the metonymies PHYSIOLOGICAL EFFECT FOR EMOTION and BODY PART (skin, flesh, scalp) FOR PERSON/ EXPERIENCER experiential basis: when we feel horrified or revolted we have the sensation that insects are moving over our skin; i.e. we feel as if crawling with insects
	4	move forward slowly	[Self_motion]$_{\text{figurative}}$	extension from LU1 (collocate type: human); experiential grounding: when you crawl, your speed is reduced - metonymy: shift of emphasis from the manner of motion of humans (i.e. on hands and knees) to their speed of motion (i.e. slow)

				- metaphor: further extension to the slow speed of any kind of activity SELF_MOVER: human, vehicle, plant, substance, path, process, time, fear
	5	Idiom: *crawl/come out of the woodwork* appear for unpleasant reasons	[Coming_to_be]	extension from LU1 (collocate type: human) via the metaphor LACK OF VIRTUE IS DOWN (weak/ dishonest people are characterised as “worms”, i.e. underground movers) it implies contempt
II. Action	6	behave in a servile manner; try hard to please someone in authority in order to get an advantage Colligation: *crawl to someone*	[Subordinates_and_superiors]	extension from LU1 (collocate type: human) via the metaphor BEING SUBJECT TO CONTROL IS DOWN; experiential grounding: lowering the body to the ground is a gesture of submission it implies disapproval of the behaviour and of the people involved
	7	search the Internet for information Collocation: *crawl the Web*	[Scouring]	extension from LU1 (collocate type: insect) on the basis of the Computing sense of spider, and the metaphors ACTION (i.e. searching) IS MOTION (i.e. path traversing) and ABSTRACT STRUCTURE OF A COMPLEX SYSTEM (i.e. information database) IS PHYSICAL STRUCTURE (i.e. spider web) SEARCHER: computer program (e.g. *web spider*, *search*

				engine, *software*) GROUND: Internet (e.g. *Web*, *(web)site*, *net*) it implies that the software carries out the search quickly and lists the results

Table 3: Integrating *crawl* MWEs in a motivated semantic network

Clusters of senses	**LU**		**Frame**	**Conceptual motivation**
I. Motion	1	run towards a goal very quickly or hastily	[Self_motion]	literal sense; violent manner of motion
	2	Proverb (routine formula): *(I) must/ have (got) to dash* used for saying that you must leave quickly because you are in a hurry	[Departing]	spoken expression related to LU1 (collocate type: human); it serves a special discoursal function, i.e. to excuse yourself for leaving
	3a	Phrasal verb: *dash off* leave a place quickly because you are in a hurry	[Departing]	special case of LU1; the particle *off* contributes the SOURCE FE to the [Self_motion] of LU1
	3b	Phrasal verb: *dash off something* write or draw something quickly	[Text_creation]	extension from LU3a via the EVENT STRUCTURE metaphor: MANNER OF ACTION IS MANNER OF MOTION mapping: leaving in a hurry (literal source: place) → writing

				something in a hurry (metaphorical source: mind) it implies that you are not thinking very much or trying very hard
II. Impact	4	hit against a surface with great force	[Impact]	literal extension from LU1 by adding the element [contact by impact] to self-motion: SELF-MOVER = IMPACTOR
	5	make something move violently against a surface, usually so that it breaks	[Cause_impact]	causative extension from LU4: the action of the verb has an effect on an entity (IMPACTOR) so that it will move forcibly/ violently and hit against another entity, the IMPACTEE (ACTION FOR RESULT metonymy); it implies (physical) damage
	6	Proverb (routine formula): *dash it (all)!* exclamation used to express annoyance	[Cause_impact]	spoken expression motivated by LU5; it serves a special discoursal function, i.e. its sole meaning is its implication (the speaker is annoyed about something)
	7	destroy someone's hopes, dreams, plans, etc., thus disappointing them Collocation: *dash someone's hopes*	[Destroying]	extension from LU5 via the metaphors THOUGHTS/ FEELINGS ARE OBJECTS and BAD IS DOWN; it implies cruelty and emotional damage (frustration) UNDERGOER: *hope, expectation, dream, effort, prospect, spirits* (restricted set of collocates)

Table 4: Integrating *dash* MWEs in a motivated semantic network

is to take advantage of the flexibility of the electronic medium to translate cognitively oriented information into (microstructural) dictionary features.

Adding frame-based signposts as guidewords and using a tiered structure with clusters of senses ordered in a logical manner can be applied to whole entries to make connections more transparent. We should note in this respect that only CALD uses guidewords in the entries examined, yet without rational arrangement of sense divisions (e.g. the “Fill” MWE *be crawling with* appears far from the “Move” sense after the “Try to please” section), and only MEDAL uses a tiered structure, yet without incorporating MWEs in it. Besides these general techniques, MWEs in particular could be recorded (with frame-based signposts) in alphabetical order in a menu at the top of the entry to facilitate access, but placed within the related sense division in the entry text to indicate semantic motivation. For example, the idiom variants *make someone’s skin/flesh/scalp crawl* could be placed under the “Motion” cluster after the *be crawling with* motivating LU. In this way, the entry could draw users’ attention to both the overall meaning (target frame) through the frame-based signpost and the internal structure (source frame) of metaphorical MWE through its position.

The descriptions of conceptual motivation in Table 3 and Table 4 can be used not only to position MWEs inside the entries, but also to systematically incorporate a new type of information in electronic entries. What is proposed in this respect is to create short and simplified notes about how MWEs are connected to the motivating meaning and include them in definitions and/or in awareness-raising notes. For example, the use of the word “movers” in the definition of *crawl with something/someone* (Table 3, LU2) is a clue to its link to the literal motion LU1; similar is the function of the parallel use of adverbs (“quickly”, “violently”) in the definitions in Table 4. As regards awareness-raising notes, they can have the form of hyperlinked notes that explain the underlying motivation of MWEs. The relevant information in Table 3 and Table 4 should be expressed in a simplified manner following the example set by MEDAL’s metaphor boxes. For instance, complementing MLDs’ quite similar definitions of *make someone’s skin/flesh/scalp crawl* (see Table 1) with a note on the experiential grounding of the idiom schema (see Table 3) would facilitate learners’ understanding and recall. Enriching learners’ dictionaries with cognitive information is expected to have positive effects on L2 vocabulary learning (see e.g. Yang & Wei 2015), but more user studies are needed to firmly support this.

5. Conclusion

Situated within the framework of "cognitive lexicography" (Ostermann, 2015), this paper has explored the relevance of cognitive approaches, namely Frame Semantics and the Conceptual Metaphor and Metonymy Theory, to the lexicographic treatment of MWEs. A review of two entries in the online versions of the "Big Five" MLDs has revealed the need for a rational organising framework that could help users (learners) make sense of the rich corpus-derived information on MWEs (including variant forms, illustrative examples, implications and usage constraints). In reconstructing the skeletal structure of the sample entries, we have demonstrated the motivation of different types of MWEs and their link to the rest of the LUs. This conceptual information can be reflected in various elements of the microstructure – frame-based signposts, tiered structure, points of access through menus and related sense divisions, clues in definitions and notes – to show the relation between the unit (meaning) and its components (form). This suggestion can complement cognitively oriented macrostructural practices like MEDAL's "metaphor boxes" and make a step towards treating MWEs holistically within motivated semantic networks.

In "post-editing lexicography", where lexicographically relevant information can automatically be extracted from corpora and drafted in preliminary entries, organising single-word and multi-word LUs in a coherent and principled manner still seems to be one of the most challenging tasks. As relevant lexical databases like Framenet and MetaNet develop further, they could be integrated into the editorial workflow and more ways to channel cognitive linguistic insights into MLDs could be devised.

6. References

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

Atkins, S. B. T., Rundell, M. & Sato, H. (2003). The contribution of FrameNet to practical lexicography. *International Journal of Lexicography*, 16(3), pp. 333–357.

Baldwin, T. & Kim, S. N. (2010). Multiword expressions. In N. Indurkhya & F. J. Damerau (eds.) *Handbook of Natural Language Processing.* 2nd Edition. Boca Raton, USA: CRC Press, pp. 267-292.

Bergenholtz, H. & R. Gouws. (2014). A Lexicographical Perspective on the Classification of Multiword Combinations. *International Journal of Lexicography*, 27(1), pp. 1–24.

Boers, F. & Lindstromberg, S. (2006). Cognitive linguistic applications in second or foreign language instruction: Rationale, proposals, and evaluation. In G. Kristiansen, M. Achard, R. Dirven & F.J. Ruiz de Mendoza Ibáñez (eds.) *Cognitive Linguistics: Current Applications and Future Perspectives.* Berlin/New York: Mouton de Gruyter, pp. 305-355.

Burchardt, A., Erk, K., Frank, A., Kowalski, A., Padó, S. & Pinkal, M. (2009). FrameNet for the semantic analysis of German: Annotation, representation and automation. In H. C. Boas (ed.) *Multilingual FrameNets in Computational Lexicography: Methods and Applications.* Berlin/New York: Mouton de Gruyter, pp. 209-244.

Dalpanagioti, Th. (2013). Frame-semantic Issues in Building a Bilingual Lexicographic Resource: A Case Study of Greek and English Motion Verbs. *Constructions and Frames*, 5(1), pp. 5–38.

Dalpanagioti, Th. (2018). A Frame-semantic Approach to Co-occurrence Patterns: A Lexicographic Study of English and Greek Motion Verbs. *International Journal* *of Lexicography*, 31(4), pp. 420–451.

De Schryver, G.-M. (2003). Lexicographers' Dreams in the Electronic-dictionary Age. *International Journal of Lexicography*, 16(2), pp. 143-199.

Dobrovol'skij, D. (2011). The structure of metaphor and idiom semantics (a cognitive approach). In S. Handl & H.-J. Schmid (eds.) Windows to the Mind: Metaphor, Metonymy and Conceptual Blending. Berlin/New York: Mouton de Gruyter, pp. 41-62.

Dziemianko, A. (2017). Electronic dictionaries. In P. A. Fuertes-Olivera (ed.) *Routledge Handbook of Lexicography.* London & New York: Routledge, pp. 663-683.

Fellbaum, C. (2016). The treatment of multi-word units in lexicography. In Ph. Durkin (ed.) *The Oxford Handbook of Lexicography.* Oxford: Oxford University Press, pp. 411-424.

Fillmore, C. J. (1982). Frame semantics. In *Linguistics in the Morning Calm.* Seoul: Hanshin Publishing, pp. 111–137. Reprinted in D. Geeraerts (ed.) (2006) *Cognitive Linguistics: Basic Readings.* Berlin: Mouton de Gruyter, pp. 373-400.

Fillmore, C. J. (1985). Frames and the Semantics of Understanding. In *Quaderni di Semantica*, 6(2), pp. 222-254.

Fillmore, C. J. & Atkins, B. T. S. (1992). Toward a frame-based lexicon: The semantics of risk and its neighbors. In A. Lehrer & E. F. Kittay (eds.) *Frames, Fields, and Contrasts.* Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 75–102.

Fillmore, C. & Petruck, M. (2003). FrameNet Glossary. In *International Journal of Lexicography*, 16(3), pp. 359-361.

FrameNet. Accessed at: https://framenet.icsi.berkeley.edu/fndrupal. (5 April 2021)

Gantar, P., Colman, L., Parra Escartín, C. & Martínez Alonso, H. (2019). Multiword Expressions: Between Lexicography and NLP. *International Journal of Lexicography* 32.2: 138–162.

Geeraerts, D. (1990). The lexicographical treatment of prototypical polysemy. In S. L. Tsohatzidis (ed.) *Meanings and Prototypes. Studies in Linguistic Categorization.* London and New York: Routledge, pp. 195–210.

Geeraerts, D. (2007). Lexicography. In D. Geeraerts & H. Cuyckens (eds.) *The Oxford Handbook of Cognitive Linguistics.* Oxford: Oxford University Press, pp. 1160-1174.

Gibbs, R. W. Jr. (1993). Why idioms are not dead metaphors. In Cr. Cacciari & P. Tabossi (eds.) *Idioms: Processing, Structure, and Interpretation.* New Jersey: Lawrence Erlbaum Associate Publishers, pp. 57-77.

Hoey, M. (2005). *Lexical Priming: A New Theory of Words and Language.* London: Routledge.

Jiang, G. & Qiaoyun C. (2015). A Micro Exploration into Learner's Dictionaries: A prototype theoretical perspective. *International Journal of Lexicography*, 30(1), pp. 108-139.

Kövecses, Z. (2012). A cognitive linguistic view of learning idioms in an FLT context. In D. Geeraerts, R. Dirven, J. Taylor & R. Langacker (eds.) *Applied Cognitive Linguistics, II, Language Pedagogy.* Berlin, Boston: De Gruyter Mouton, pp. 87-

116.

Kövecses, Z. & Csábi, S. (2014). Lexicography and Cognitive Linguistics. *Revista Española de Lingüís-tica Aplicada/Spanish Journal of Applied Linguistics*, 27(1), pp. 118-139.

Lakoff, G. & Johnson, M. (1980). *Metaphors We Live By.* Chicago and London: The University of Chicago Press.

Lew, R. (2012). How can we make electronic dictionaries more effective? In S. Granger & M. Paquot (eds.) *Electronic Lexicography.* Oxford: Oxford University Press, pp. 343-361.

Lew, R. (2013). Identifying, ordering and defining senses. In H. Jackson (ed.) *The Bloomsbury Companion to Lexicography.* London: Bloomsbury Publishing, pp. 284-302.

Molina, C. (2008). Historical Dictionary Definitions Revisited from a Prototype Theoretical Standpoint. *Annual Review of Cognitive Linguistics* 6, pp. 1-22.

Moon, R. (2004). On Specifying Metaphor: An Idea and its Implementation. *International Journal of Lexicography* 17(2), pp. 195-222.

Oppentocht, L. & Schutz, R. (2003). Developments in electronic dictionary design. In P. Van Sterken-burg (ed.) *A Practical Guide to Lexicography.* Amsterdam/Philadelphia: John Benjamins, pp. 215-227.

Ostermann, C. (2015). *Cognitive Lexicography: A New Approach to Lexicography Making Use of Cognitive Semantics.* Lexicographica. Series Maior 149. Berlin: Mouton de Gruyter.

Paquot, M. (2015). Lexicography and phraseology. In D. Biber & R. Reppen (eds.) *Cambridge Handbook of English Corpus Linguistics.* Cambridge: Cambridge University Press, pp 460-477.

Ruppenhofer, J., Ellsworth, M., Petruck, M., Johnson, C. & Scheffczyk, J. (2016). *FrameNet II: Extended Theory and Practice.* Accessed at: https://framenet2.icsi.berkeley.edu/docs/r1.7/book.pdf. (5 April 2021)

van der Meer, G. (1999). Metaphors and dictionaries: The morass of meaning, or how to get two ideas for one. *International Journal of Lexicography*, 12(3), pp. 195-208.

Walker, C. (2009). The Treatment of Collocation by Learners' Dictionaries, Collocational Dictionaries and Dictionaries of Business English. *International Journal of Lexicography*, 22(3), pp. 281-299.

Wiliński, J. (2016). Frame, metaphor and metonymy in onomasiological lexicography. In M. Fabiszak, K. Krawczak & K. Rokoszewska (eds.) *Categorization in Discourse and Grammar.* Berlin: Peter Lang, pp. 65-82.

Wojciechowska, S. (2020). Access Routes to BODY PART Multiword Expressions in the 'Big Five' MELDs: Use of Hyperlinks. *Lexikos*, 30, pp. 583-608.

Xu, H. & Lou, Y. (2015). Treatment of the Preposition *to* in English Learners' Dictionaries: A Cognitive Approach. *International Journal of Lexicography*, 28(2), pp. 207–231.

Yang, N. & Wie, X. (2015) Metaphor information in Macmillan English Dictionary for

Advanced Learners: Presentation & Effectiveness. *International Journal of Lexicography*, 29(4), pp. 424-451.

Dictionaries

CALD: *Cambridge Advanced Learner's Dictionary.* Accessed at: http://www.dictionary.cambridge.org. (5 April 2021)

COBUILD: COBUILD Advanced English Dictionary. Accessed at: http://www.collinsdictionary.com. (5 April 2021)

LDOCE: *Longman Dictionary of Contemporary English.* Accessed at: http://www.ldoceonline.com. (5 April 2021)

MEDAL: *Macmillan English Dictionary for Advanced Learners.* Accessed at: http://www.macmillandictionary.com. (5 April 2021)

OALD: *Oxford Advanced Learner's Dictionary.* Accessed at: http://www.oxfordlearnersdictionaries.com. (5 April 2021)

The structure of a dictionary entry and grammatical properties of multi-word units

Monika Czerepowicka

University of Warmia and Mazury in Olsztyn (Poland)
E-mail: monika.czerepowicka@uwm.edu.pl

Abstract

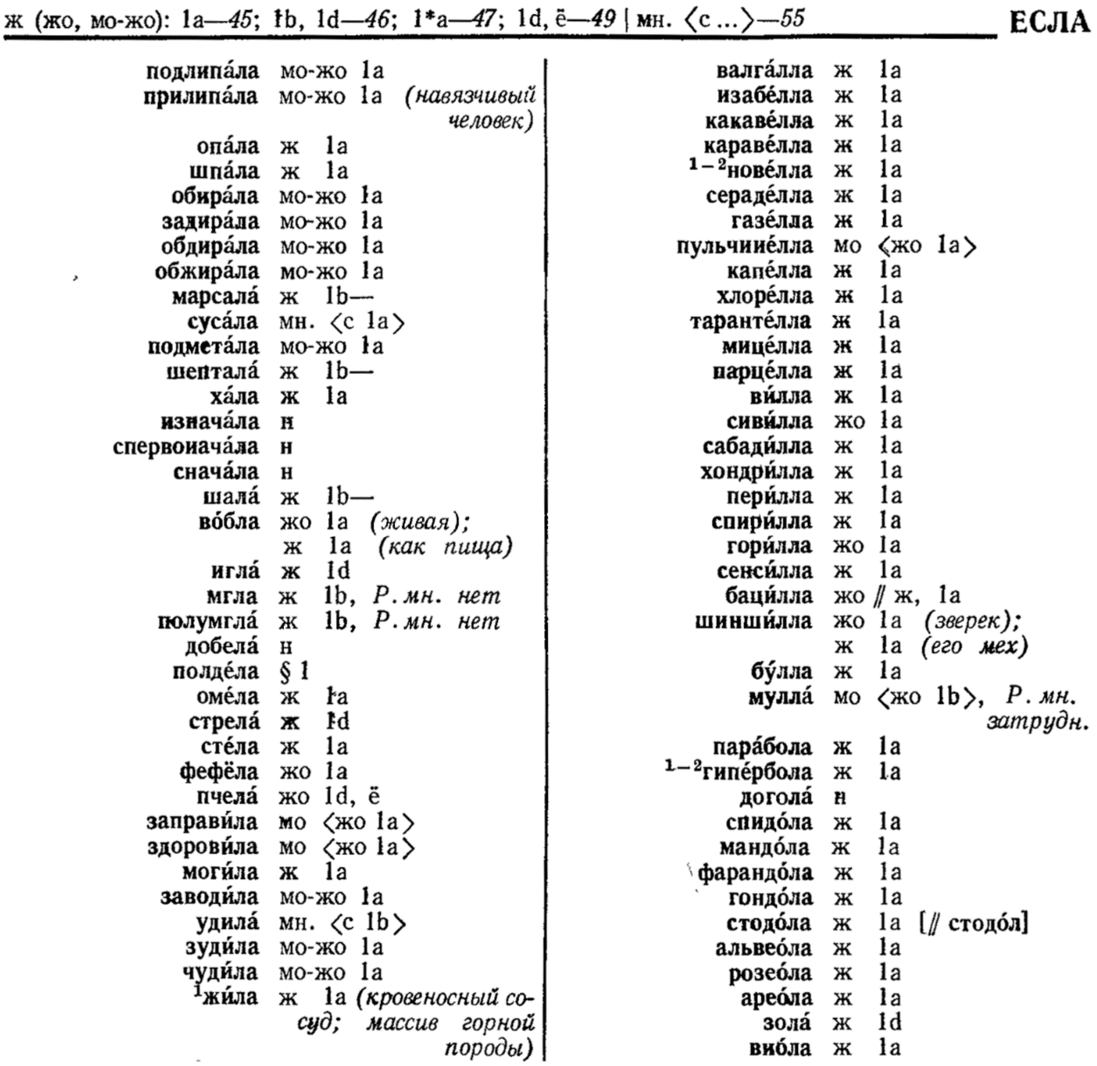
Users of advanced inflectional languages expect dictionaries to provide clear inflectional information so that the creation or use of a given form does not generate additional problems. The development of technologies and tools for machine language processing has naturally made contemporary inflectional dictionaries advanced electronic works that contain tools for the individualisation of their content in line with users' needs. The main concern of this article is the influence of the grammatical properties of language units on lexicographic description, in particular the structure of a dictionary entry. This issue will be discussed with reference to *Verbel. The Inflectional Dictionary of Polish Verbal Phrases*, which is an electronic dictionary listing over 5,000 multi-word units, giving all their paradigmatic forms directly. Although it is a specialist study providing a formal description of units, thanks to the proper structure of entries it is possible to be used also by non-specialists. The opportunity of choosing the scope of lexicographic information in *The Verbel Dictionary* is guaranteed by a two-stage scheme of the entry which consists of a general and detailed description of units.

Keywords: multi-word units; inflection; dictionary; e-lexicography

1. Introduction

The subject under scrutiny is the part of lexicographic description which reports on the grammatical, mainly inflectional, information about a unit. It is assumed that language units are differentiated on the basis of their semantic and grammatical features, thus they can also be discontinuous (cf. Baldwin and Kim, 2010; Bogusławski, 1976; Mel'čuk, 2006; Mel'čuk & Zholkovsky, 1984; Sag et al., 2002). Regardless of their formal structure, however, they should be uniformly described. The position advocated in this study is that multi-word units of language should be accompanied by an equally detailed, precise and consistent inflectional description as lexemes. For this reason, a rigorous, algorithmic model will be applied to provide such a description in *Verbel. The Inflectional Dictionary of Polish Verbal Phrases* (Kosek et al., 2020).

The idea of grammatical dictionaries providing all paradigmatic forms of a unit is particularly important and useful for inflected languages, such as Polish and other Slavic languages. One can find a few methodological models that make it possible to describe units of language in an adequately detailed and precise manner and have been used in dictionaries. *The Grammatical Dictionary of Russian* (Rus. *Грамматический словарь русского языка*) by Andriey Zaliznyak (1977) is one of the first dictionaries of this type. Zaliznyak’s approach was highly innovative in dispensing with the construct of the morpheme and putting the notion of the paradigm in the spotlight, heralding the rise of ‘word-and-paradigm’ and other realisational theories in morphology (Iosad and others 2018: 176). Zaliznyak's morphological model consists in constructing paradigm forms from an abstract lexeme representation using rewrite rules. The dictionary contains about 100 000 units of language with their grammatical characteristics presented by symbols and listed in *a tergo* order (Fig. 1).

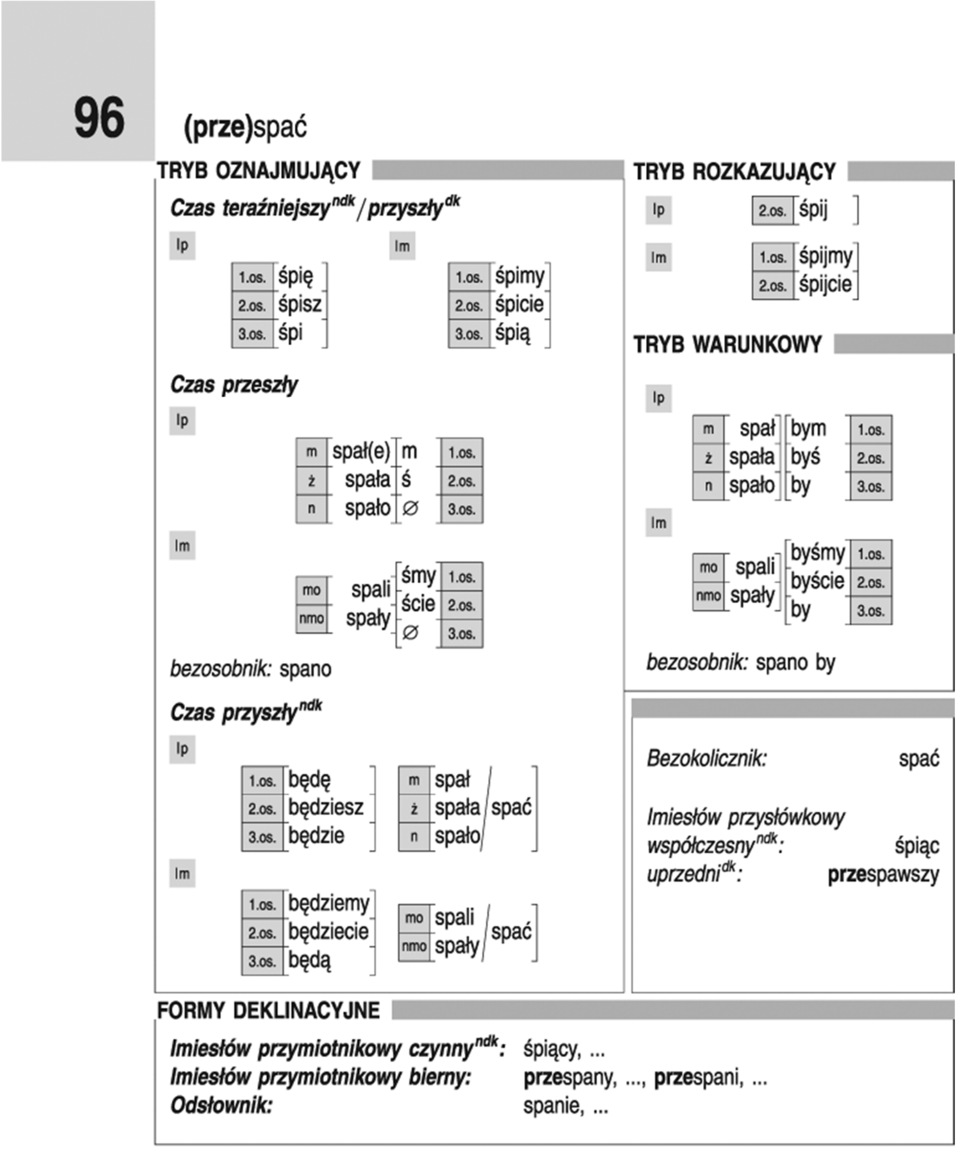


ж (жо, мо-жо): 1a—*45*; 1b, 1d—*46*; 1*a—*47*; 1d, ё—*49* | мн. ⟨с ...⟩—*55* **ЕСЛА**

подлипа́ла	мо-жо	1a
прилипа́ла	мо-жо	1a *(навязчивый человек)*
опа́ла	ж	1a
шпа́ла	ж	1a
обира́ла	мо-жо	1a
задира́ла	мо-жо	1a
обдира́ла	мо-жо	1a
обжира́ла	мо-жо	1a
марсала́	ж	1b—
суса́ла	мн.	⟨с 1a⟩
подмета́ла	мо-жо	1a
шептала́	ж	1b—
ха́ла	ж	1a
изнача́ла	н	
спервоиача́ла	н	
снача́ла	н	
шала́	ж	1b—
во́бла	жо	1a *(живая);*
	ж	1a *(как пища)*
игла́	ж	1d
мгла	ж	1b, *Р. мн. нет*
полумгла́	ж	1b, *Р. мн. нет*
добела́	н	
полде́ла	§ 1	
оме́ла	ж	1a
стрела́	ж	1d
сте́ла	ж	1a
фефёла	жо	1a
пчела́	жо	1d, ё
заправи́ла	мо	⟨жо 1a⟩
здорови́ла	мо	⟨жо 1a⟩
моги́ла	ж	1a
заводи́ла	мо-жо	1a
удила́	мн.	⟨с 1b⟩
зуди́ла	мо-жо	1a
чуди́ла	мо-жо	1a
^{1}**жи́ла**	ж	1a *(кровеносный сосуд; массив горной породы)*
валга́лла	ж	1a
изабе́лла	ж	1a
какаве́лла	ж	1a
карavéлла	ж	1a
$^{1-2}$**нове́лла**	ж	1a
сераде́лла	ж	1a
газе́лла	ж	1a
пульчинне́лла	мо	⟨жо 1a⟩
капе́лла	ж	1a
хлоре́лла	ж	1a
таранте́лла	ж	1a
мице́лла	ж	1a
парце́лла	ж	1a
ви́лла	ж	1a
сиви́лла	жо	1a
сабади́лла	ж	1a
хондри́лла	ж	1a
пери́лла	ж	1a
спири́лла	ж	1a
гори́лла	жо	1a
сенси́лла	ж	1a
баци́лла	жо // ж,	1a
шинши́лла	жо	1a *(зверек);*
	ж	1a *(его мех)*
бу́лла	ж	1a
мулла́	мо	⟨жо 1b⟩, *Р. мн. затрудн.*
пара́бола	ж	1a
$^{1-2}$**гипе́рбола**	ж	1a
догола́	н	
спидо́ла	ж	1a
мандо́ла	ж	1a
фарандо́ла	ж	1a
гондо́ла	ж	1a
стодо́ла	ж	1a [// **стодо́л**]
альвео́ла	ж	1a
розео́ла	ж	1a
арео́ла	ж	1a
зола́	ж	1d
вио́ла	ж	1a

Figure 1. Entries from *The Grammatical Dictionary of Russian* (Rus. *Грамматический словарь русского языка*) by A. Zaliznyak (1977)

Paradigms can also be shown precisely by illustrative tables. This way of data presentation is used in grammatical dictionaries of verbal units of French (Bescherelle, 1978) or Polish (Saloni, 2007). All verbs are arranged in groups distinguished on the basis of their morphological structure and inflectional properties. A total of 106 patterns were identified following in-depth and detailed analyses of the Polish conjugation. The paradigm of each pattern is presented with an example verb in a table. Thanks to proper presentation of the formal structure of a given verb the user can inflect other verbs belonging to the same group and noted in the dictionary (Fig. 2).



96 **(prze)**spać

TRYB OZNAJMUJĄCY

Czas teraźniejszy^{ndk}/ przyszły^{dk}

lp		lm	
1.os.	śpię	1.os.	śpimy
2.os.	śpisz	2.os.	śpicie
3.os.	śpi	3.os.	śpią

Czas przeszły

lp

m	spał(e)	m	1.os.
ż	spała	ś	2.os.
n	spało	∅	3.os.

lm

mo	spali	śmy	1.os.
nmo	spały	ście	2.os.
		∅	3.os.

bezosobnik: spano

Czas przyszły^{ndk}

lp

1.os.	będę	m	spał	spać
2.os.	będziesz	ż	spała	
3.os.	będzie	n	spało	

lm

1.os.	będziemy	mo	spali	spać
2.os.	będziecie	nmo	spały	
3.os.	będą			

TRYB ROZKAZUJĄCY

lp	
2.os.	śpij
lm	
1.os.	śpijmy
2.os.	śpijcie

TRYB WARUNKOWY

lp

m	spał	bym	1.os.
ż	spała	byś	2.os.
n	spało	by	3.os.

lm

mo	spali	byśmy	1.os.
nmo	spały	byście	2.os.
		by	3.os.

bezosobnik: spano by

Bezokolicznik:	spać
Imiesłów przysłówkowy	
współczesny^{ndk}:	śpiąc
uprzedni^{dk}:	**prze**spawszy

FORMY DEKLINACYJNE

Imiesłów przymiotnikowy czynny^{ndk}:	śpiący, ...
Imiesłów przymiotnikowy bierny:	**prze**spany, ..., **prze**spani, ...
Odsłownik:	spanie, ...

Figure 2. A table from *The Polish Verb. Inflection, Dictionary of 12,000 lexemes* (Pl. *Czasownik polski.* Odmiana, słownik 12 000 czasowników) of Z. Saloni (2001)

The development of electronic dictionaries gives an obvious opportunity to note the paradigms of all units directly (*in extenso*). However, the lexicographic description should present the nature of each unit in all its inflectional complexity. Tools to construct an appropriately precise scheme to provide the inflectional information of Polish units are included in the concept of a morphological description, proposed by Janusz S. Bień and Zygmunt Saloni (1982). The methodological perspective adopted by the authors proved to be effective in the case of lexemes, which was confirmed by *The Grammatical Dictionary of Polish* (Pl. *SGJP*; Saloni et al., 2015), which contains descriptions of over 300,000 Polish units. This theoretical model has proven successful in machine processing as well, being used in the morphosyntactic marking of the National Corpus of Polish (Pl. NKJP; Przepiórkowski et al., 2012). However, it should be emphasised that it concerns lexemes. Since multi-word units require an equally detailed and rigorous description, as mentioned above, it has been decided to implement the model in the inflectional dictionary of Polish verbal phrases.

In this paper, terms such as "phrase", "phraseologism" and "multi-word units" are applied to refer to discontinuous units of language. Verbal units of this type can be defined as connections of at least two words that perform the function of the centre of the sentence, similarly to verbal lexemes. Because of the degree of unification of unit components, the possibility of replacing some of them and resultant changes of meaning, one can distinguish idioms, light verb constructions, and collocations among them. However, in this study we do not consider differences between the mentioned semantic types of verbal multi-word units but focus on their inflected and morphosyntactic features and their influence on the structure of a dictionary entry. Still, we discuss morphological types of Polish multi-word units as well as the basis of the theoretical model used in *The Verbel Dictionary.* The key terms it comprises are *a morphological word*, *a paradigmatic word*, *a flexeme* and *a vocabula*, and they reflect the multi-step procedure of a comprehensive grammatical description of language units.

2. Description model

A *morphological word* is defined here as a sequence of letters (graphemic shape; *signifiant*) interpreted grammatically and semantically (*signifié*). It is a complete linguistic sign. Its grammatical properties are determined on the basis of morphological features of each type of word – nouns, verbs, adjectives etc. Apart from traditional morphological categories, such as case, number, gender, and person, the register of morphological words also includes non-traditional categories, resulting from detailed inflectional description. These include, *inter alia*, such categories as agglutination and vocalism, both connected to each other and with inflection by person. Agglutination is a grammatical feature noted in the past tense inflection. The person-number morpheme (of the 1st or 2nd person) usually appears immediately after a verbal stem, forming one word in textual form: *robił-**em*** ('I did' masc.), *robił-**eś*** ('you did' masc.), *robili-**śmy*** ('we did' masculine personal), *robili-**ście*** ('you did' pl. masculine personal). Still, these morphemes can be torn off the verb stem and glued to another word in a sentence, e.g.: *Blat miał**eś*** - *Blat**eś** miał.* ('You had a tabletop'). There are syntactic constructions where this kind of operation is required, such as in some dependency sentence phrases: *Jan chciał, żeby**śmy poszli** na spacer.* - **Jan chciał, żeby **poszliśmy** na spacer* ('Jan wanted us to go for a walk').

The vocalism category is also observed in the past tense. The shape of the agglutination morpheme depends on the ending of the verb stem. If the stem ends in a consonant, like in masculine forms, e.g. *robił* ('he did'), the agglutination morpheme becomes vocal:

*robił**em*** (‘I did’ masc; the first-person is created by adding the agglutination morpheme to the past verb stem). If the verb stem ends in a vowel (as it is in non-masculine forms), the agglutination morpheme becomes non-vocal, e.g. *robiła**m*** (‘I did’ fem.), *robiły**śmy*** (‘we did’ non-masculine). When generating Polish verb forms, all such subtle morphological features must be taken into account.

Polish verbal morphological words are heterogeneous. They cannot be classified by the same morphological categories. Apart from formal signs, they differ in semantic identification. The full paradigm of the verb includes verbal adjectives (participles) and nouns, i.e. forms that are inflected by cases, in addition to forms inflected by person and number. Furthermore, conjugation forms are subject to morphological categories to varying degrees – for example, the category of genus is manifested in the past tense (*robił* ‘he did’, *robiła* ‘she did’) and certain complex future forms (*będę robił* ‘I will be doing’ masc., *będę robiła* ‘I will be doing’ fem.). In the case of other verb forms, gender neutralisation can be noted. Among verbal paradigmatic forms there are also those that cannot be assigned any other grammatical category than aspect, these are: infinitive, adverbial participles, impersonal forms (Pl. *bezosobnik*, forms with -*no*, -*to*), e.g. *robić* ‘to do’, *robiąc* ‘doing’, *robiono* ‘it was done.’ This prompts us to classify them into groups that fall under the same morphological categories. We call sets of forms differentiated on the basis of the same morphological categories *flexemes* or *paradigmatic words.*

The level of complexity of the Polish conjugation system calls for a special treatment. The paradigm of verbal units consists of various types of paradigmatic words (non-past and past forms, participles, conditional forms, imperative, etc.), which form separate sub-paradigms. For example, flexemes of the past tense are inflected by person, number, gender, while non-past forms (present and future simple tense) and imperative – by person and number. The theoretical problem related to consistent morphosyntactic description of verbal units is closely related to the number of flexemes belonging to a given unit, and thus to the multitude and variety of inflectional forms. In the case of verbal units, it is a systemic phenomenon, which ultimately determines the architecture of a dictionary entry which becomes a super-class – a *vocabula.* A vocabula, i.e. a dictionary entry, groups paradigmatic words with the same semantic root, so it consists of various types of flexemes characterised by different morphological features. Verbs with regular inflection patterns (full paradigm) encompass 8 or 10 flexemes (depending on the aspect).

This multistage procedure provides the basis for both the description of abstract language units and their textual realisations, at the same time providing tools for separate levels of linguistic description. This type of research perspective seems to be particularly helpful in machine language processing.

3. The specificity of verbal multi-word units

It should be noted that the paradigm of verbal multi-word units depends on their morphosyntactic properties as well as morphological structure. They differ in both internal syntax (mutual relations of multi-word unit components) and external syntax (matching and requirements with regard to other sentence elements; cf. Lewicki, 1986). Based on their morphosyntactic features, one can distinguish three types of Polish multi-word units. Thus, there are phrases (type 1) which are characterised by an open position for the subject in the nominative, {*ktoś*} *zbija bąki* (lit. {*someone*$_{\text{Nom}}$} *is shooting herons*, 'someone is getting lazy'), {*ktoś*} *przypina komuś łatkę*, (lit. {*someone*$_{\text{Nom}}$} *is sticking a patch on someone else*, 'someone is attributing a negative feature or behaviour to someone else'). In contrast, other phrases do not open up a position for the nominative argument (type 2). This position is permanently filled in lexically, for example, *oczy*$_{\text{Nom}}$ *wychodzą* {*komuś*} *na wierzch* (lit. {*someone's*} *eyes*$_{\text{Nom}}$ *go to the surface*, 'someone is really surprised'), *włos*$_{\text{Nom}}$ {*komuś*} *z głowy nie spadnie* (lit. *not a hair*$_{\text{Nom}}$ *will fall off* {*someone's*} *head*, 'someone will be safe'). The third group of phrases does not show any collocability with the nominative argument, which is why it is characterised by a very limited paradigm, e.g.: {*komuś*$_{\text{Dat}}$} *pada na mózg* (lit. *it falls on* {*someone's*$_{\text{Dat}}$} *brain*, 'someone acts irrationally'), {*komuś*$_{\text{Dat}}$} *przybywa na wadze* (lit. {*someone*$_{\text{Dat}}$} *has more on the scales*, 'someone is putting on weight'), {*komuś*$_{\text{Dat}}$} *brak piątej klepki* (lit. *someone*$_{\text{Dat}}$ *lacks the fifth plank*, 'someone is crazy'). A unit's belonging to a given type determines its inflectional paradigm. Vocabulas of the first type can potentially have a full inflectional paradigm, with any limitations resulting from their semantic features (meaning). The second and third type units show numerous limitations in terms of variation by categories of person and number.

From an essentially morphological point of view verbal multi-word units can be divided into two groups: verbs and predicates. Both types differ in their formal structure and the scope of inflectional forms. The VERB class includes mainly phrases based on the inflective verb with a potentially regular inflection paradigm, such as: {*ktoś*} *dzwoni zębami* (lit. {*someone*} *rings their teeth*, 'someone feels cold'), {*ktoś*} *pada komuś do nóg* (lit. {*someone*} *falls down to someone else's feet*, 'someone shows their respect towards someone else'), as well as {*komuś*$_{\text{Dat}}$} *dzwoni w uszach* (lit. *it rings in* {*someone's*$_{\text{Dat}}$} *ears*, 'someone has tinnitus'), {*komuś*$_{\text{Dat}}$} *pada na mózg* (lit. *it falls on* {*someone's*$_{\text{Dat}}$} *brain*, 'someone acts irrationally'[1]).

The PRED class consists of units whose verbal component belongs to (primarily) defective verbs, which do not inflect by person and number, only by mode and tense

[1] This type of property is a characteristic feature of inflectional languages, such as Polish, as can be seen in the provided translations. Syntactic complexity, as a result of which the logical subject of the sentence is not expressed in the nominative case but in the dependent case, can only be rendered using literal translation. Equivalent units in English retain the typical canonical syntactic structure in which the subject, performer, or person affected by the state is expressed grammatically in the nominative form.

categories, for example *można*, *należy*, *trzeba* (‘can,’ ‘should,’ ‘need to’). In grammar studies they are called "modal verbs." As in the case of lexemes, the share of predicative multi-word units in the total number of phrases recorded in the dictionary is little. Among over 5,000, only 12 are PRED entries, e.g. {*komuś*/*czemuś*} *można wszystkie żebra policzyć* (lit. *one can count* {*someone's*/*something's*} *ribs*, ‘someone/something is thin’), {*komuś*} *brak słów* (lit. {*someone*$_{\mathrm{Dat}}$} *lacks words*, ‘someone does not know what to say’).

All detailed information about units' paradigms and their limitations are marked at the formal level in graphs.

4. Verbel. The Inflection Dictionary of Verbal Phraseological Units

The theoretical model mentioned in section 2 shows particular steps of language description: from the level of text realisation and interpretation (morphological words), through grouping forms according to their morphosyntactic features (flexemes), to the mental abstractive level in the form of units of language (vocabulas). Therefore, it was decided to implement it in electronic inflectional dictionary of verbal multi-word units. *Verbel* is a digitally born dictionary. Its purpose is to give a full paradigmatic description of multi-word units. It is not the only dictionary of this kind. *Verbel* originates from works related to the description of Polish multi-word units for the purpose of an in-depth analysis of Polish texts and is a continuation of the *SEJF* dictionary (*The Grammatical Lexicon of Polish Multi-Word Expressions*), which is a lexical resource of Polish nominal, adjectival and adverbial multi-word expressions, consisting of about 4,700 multi-word units (Czerepowicka, 2014; Czerepowicka & Savary, 2018). However, the level of complexity of the Polish conjugation system calls for special treatment of verbal words. It turns out to be incompatible with the model used in the *SEJF*, which is simpler and the entry's structure is flat. Consequently, a lexicographic description required a significant reconstruction, which determined the final hierarchical structure of the entry in the *Verbel* dictionary. Since the lexicographic information reflects levels of linguistic description, the dictionary can be applicable in NLP of Polish, such as deep mechanisms of language processing or multi-word units' identification in text. Although it has been compiled with machine processing in mind, it can be useful also for human users.

The dictionary contains over 5,000 verbal multi-word different units, both syntactically and morphologically. The distribution of dictionary entries, including their types mentioned in Section 3, is shown in Table 1. The complexity of the unit's paradigm depends, *inter alia*, on which group the unit belongs to.

	Units
1st type	4770
2nd type	289
3rd type	55
Total	5,126

Table 1. Distribution of verbal types in the *Verbel* dictionary

4.1 The structure of a unit

The basic unit in the dictionary is a vocabula, i.e. a unit from the highest level of morphological description. Phraseologisms are recorded in the 3rd person singular in the non-past tense if it exists, such as: {*ktoś/coś*} *dolewa oliwy do ognia*, (lit. {someone/something} is *adding oil to the fire*, 'someone/something is adding fuel to the fire'; {*coś*} *bierze w łeb* (lit. *{something} is taking to the head*, 'something, like a plan, is unsuccessful') – imperf; {*ktoś/coś*} *doleje oliwy do ognia* lit. *{someone/something} will add oil to the fire*, 'someone/something will add fuel to the fire'); {*coś*} *weźmie w łeb* (lit. *{something} will take to the head*, 'something, like a plan, will become unsuccessful') – perf.

In line with the Polish lexicographic tradition verbal units should be recorded in the infinitive form. However, there are important reasons to deviate from the known path. The 3rd person form shows the unit in its natural syntactic and semantic context. It also helps to identify a conjugation group, which can be especially useful for human users of the dictionary. This method of lemmatisation was postulated in specialised descriptions (cf. Tokarski, 1973) and has been used in a few Polish dictionaries (cf. Bogusławski & Garnysz-Kozłowska, 1979; Bogusławski-Wawrzyńczyk, 1993; Bogusławski & Danielewiczowa, 2005; Dunaj, 1996). Still, there is not one way of recording units in contemporary dictionaries that is dominant.

Beside the type of the unit, each entry gives general and detailed information on the unit. The entry is characterised by an appropriate structure comprised of stages of each units' description, see Fig. 3.

General and detailed information is grouped into particular tabs in the application: general information about the unit on the vocabula level (OPIS OGÓLNY HASŁA, lit. 'general description of an entry'), detailed inflectional information understood as pointing to the main flexeme and a list of all of them (OPIS JEDNOSTKI, lit. 'description of a unit'), a formal description of sub-paradigms in the form of graphs (OPIS ODMIANY FLEKSEMU, lit. 'description of the flexeme's inflection'), forms of particular flexemes (FORMY FLEKSEMU, lit. 'forms of flexeme') and paradigms of individual units

(Wszystkie formy jednostki, lit. ‘all forms of the unit’).

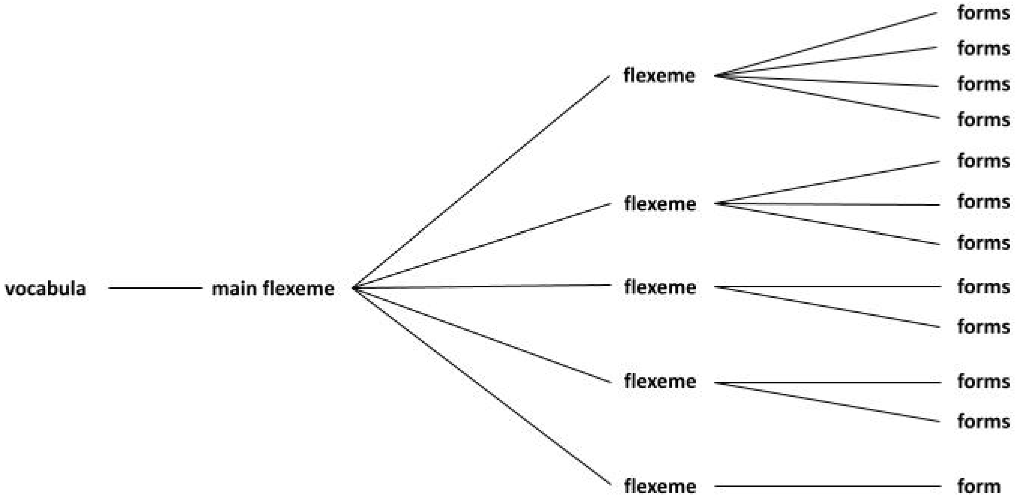


Figure 3. A scheme for the structure of an entry in the *Verbel* dictionary

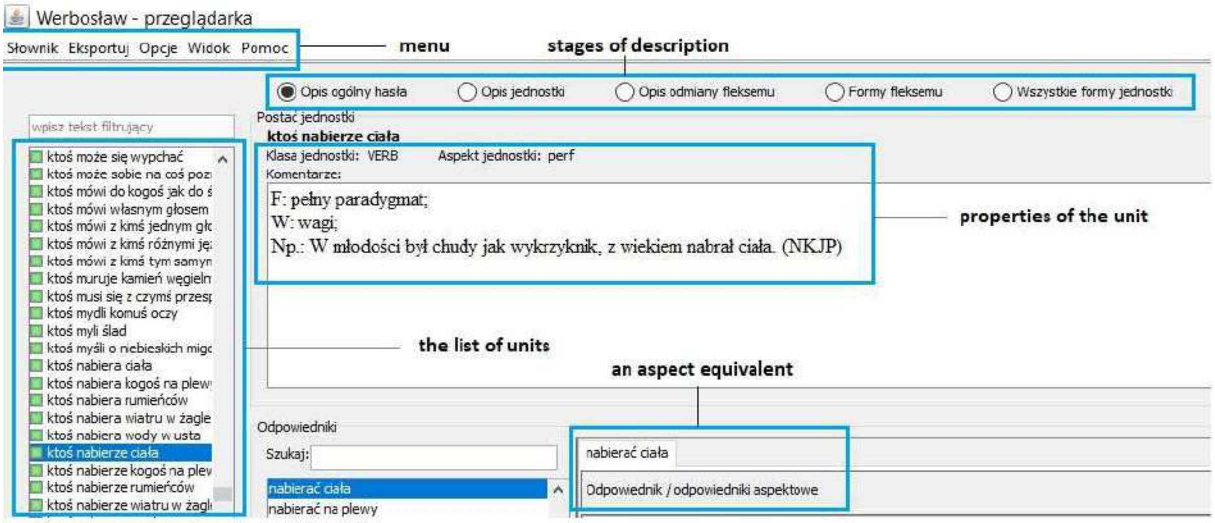
4.2 A general description of the unit

At the initial stage of description, each multi-word verbal unit is assigned to one of two morphological types of vocabulas: verbs (VERB) or predicates (PRED).

In addition to assigning units to a grammatical class, the value of the aspect of phraseologisms is noted – perfective (perf) or imperfective (imperf). The unit’s aspect equivalents, if any were determined, are also included here. What is more, this element of an entry presents general descriptive information about the paradigm (F), e.g. full, in the case of defective paradigm, and the types of excluded inflections are provided. It includes other general data about the unit, e.g. possible non-verb variants (W), pragmatic information (P), normative information (N), supplementary grammatical information (G), and examples (Np.) and selectively the meaning of the described units:

{*ktoś*} *nabiera rumieńców*	*somebody blushes*
F: pełny paradygmat	F: full paradigm
W: cery, kolorów	W: *lit.* complexion, colours
P: tryb rozkazujący w funkcjach wtórnych	P: imperative in secondary functions (a wish, a threat)

Examples included in entries come from original texts – from NKJP and the resources of the Polish Internet. The shape of a typical entry is shown in Figure 4:



ktoś może się wypchać
ktoś może sobie na coś poz
ktoś mówi do kogoś jak do ś
ktoś mówi własnym głosem
ktoś mówi z kimś jednym gło
ktoś mówi z kimś różnymi ję
ktoś mówi z kimś tym samyn
ktoś muruje kamień węgieln
ktoś musi się z czymś przesp
ktoś mydli komuś oczy
ktoś myli ślad
ktoś myśli o niebieskich migd
ktoś nabiera ciała
ktoś nabiera kogoś na plew
ktoś nabiera rumieńców
ktoś nabiera wiatru w żagle
ktoś nabiera wody w usta
ktoś nabierze ciała
ktoś nabierze kogoś na plev
ktoś nabierze rumieńców
ktoś nabierze wiatru w żagl

Postać jednostki
ktoś nabierze ciała

Klasa jednostki: VERB   Aspekt jednostki: perf

Komentarze:

F: pełny paradygmat;
W: wagi;
Np.: W młodości był chudy jak wykrzyknik, z wiekiem nabrał ciała. (NKJP)

nabierać ciała
nabierać na plewy

nabierać ciała
Odpowiednik / odpowiedniki aspektowe

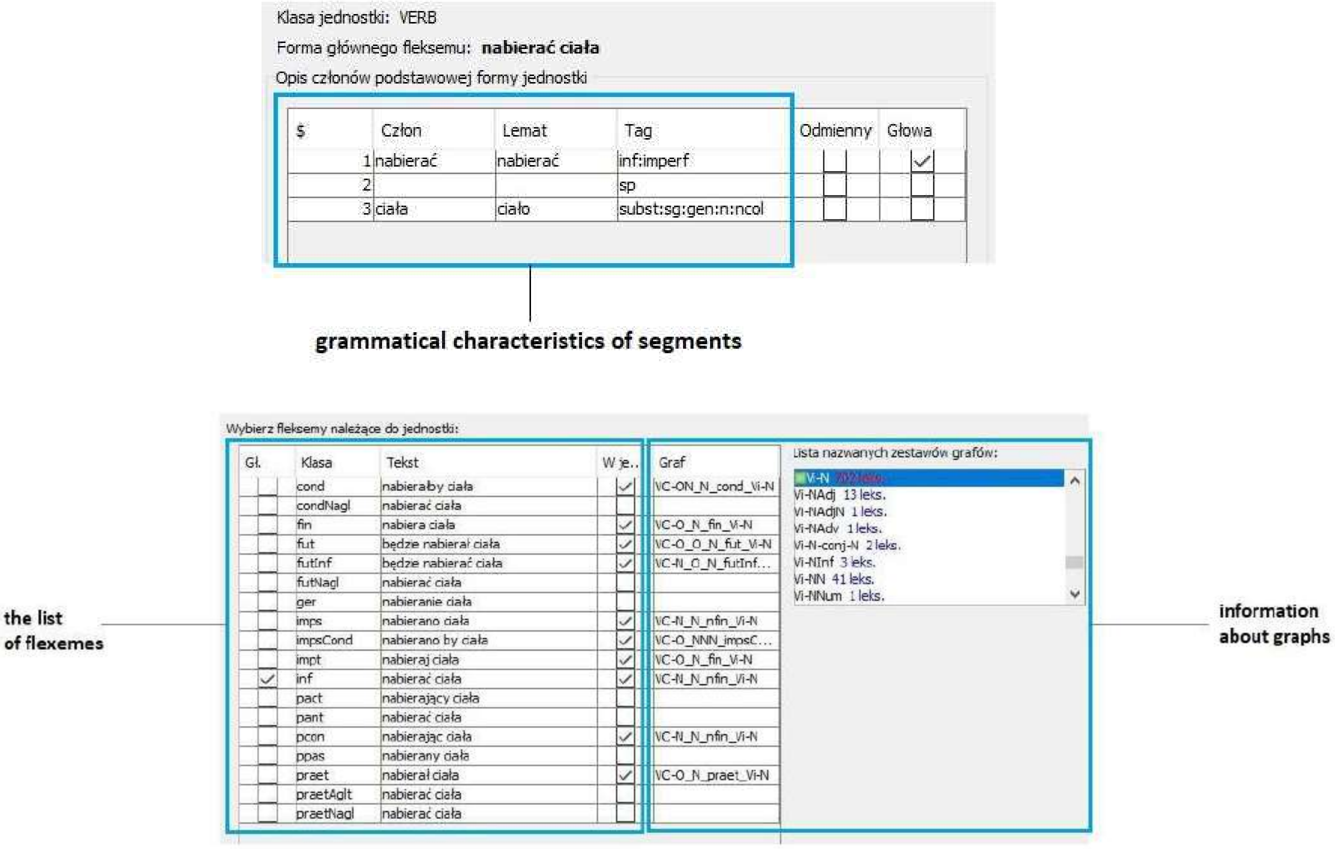
Figure 4. The general description of an entry in the *Verbel* dictionary

Descriptive information about the unit gives an idea of its properties and meaning, additionally illustrating its use in a sentence. This part of the application roughly coincides with the traditional lexicographic description and is advantageous for the human user.

4.3 Inflectional information

The following tabs contain more formal inflectional and paradigmatic description of the unit. The next step is to indicate the form of the main flexeme and grammatical characteristics of each component of the multi-word unit. The main flexeme is provided in the infinitive form of the verbal component along with all the lexical parts of the unit, excluding open positions marked with the pronouns *someone*, *something*, e.g. *mieć ręce pełne roboty* (lit. *to have hands full of work* 'to be busy'), *dolać oliwy do ognia* (lit. *to add oil to the fire*, 'to add fuel to the fire'), *pomóc jak umarłemu kadzidło* (lit. *to help like the incense helps the dead*, 'to be of no help at all'). The choice of the infinitive for the base form (main flexeme) was determined by the way forms are created in the dictionary application. They are obtained on the basis of the infinitive form in the morphological generator Morfeusz used in the dictionary (see Woliński, 2014).

Grammatical description of individual components consists of lemmatization and indicating an appropriate morphosyntactic tag. The dictionary provides rudimentary information on the internal syntax of the unit, e.g. by pointing out its main segment – head (Głowa), see Figure 5.



Klasa jednostki: VERB

Forma głównego fleksemu: **nabierać ciała**

Opis członów podstawowej formy jednostki

$	Człon	Lemat	Tag	Odmienny	Głowa
1	nabierać	nabierać	inf:imperf	□	☑
2			sp	□	□
3	ciała	ciało	subst:sg:gen:n:ncol	□	□

Wybierz fleksemy należące do jednostki:

Gł.	Klasa	Tekst	W je..	Graf
□	cond	nabierałby ciała	☑	VC-ON_N_cond_Vi-N
□	condNagl	nabierać ciała	□	
□	fin	nabiera ciała	☑	VC-O_N_fin_Vi-N
□	fut	będzie nabierał ciała	☑	VC-O_O_N_fut_Vi-N
□	futInf	będzie nabierać ciała	☑	VC-N_O_N_futInf...
□	futNagl	nabierać ciała	□	
□	ger	nabieranie ciała	□	
□	imps	nabierano ciała	☑	VC-N_N_nfin_Vi-N
□	impsCond	nabierano by ciała	☑	VC-O_NNN_impsC...
□	impt	nabieraj ciała	☑	VC-O_N_fin_Vi-N
☑	inf	nabierać ciała	☑	VC-N_N_nfin_Vi-N
□	pact	nabierający ciała	□	
□	pant	nabierać ciała	□	
□	pcon	nabierając ciała	☑	VC-N_N_nfin_Vi-N
□	ppas	nabierany ciała	□	
□	praet	nabierał ciała	☑	VC-O_N_praet_Vi-N
□	praetAglt	nabierać ciała	□	
□	praetNagl	nabierać ciała	□	

Lista nazwanych zestawów grafów:

- Vi-N [illegible]
- Vi-NAdj 13 leks.
- Vi-NAdjN 1 leks.
- Vi-NAdv 1 leks.
- Vi-N-conj-N 2 leks.
- Vi-NInf 3 leks.
- Vi-NN 41 leks.
- Vi-NNum 1 leks.

Figure 5. A part of the description of a unit in the *Verbel* dictionary

Then, specific types of flexemes are assigned to each entry. Their number is determined by the value of aspect and specific inflectional features of the unit. For instance, full paradigm imperfective units contain 10 types of paradigmatic words, perfective ones – 8. This tab is crucial for the structure of the entry in the dictionary, as it contains a list of all the flexemes belonging to a given unit, see Figure 5.

4.4 Formal description

Generation of the forms of individual flexemes in the dictionary is based on graphs (cf. Marciniak et al. 2011). The relation of a graph to a flexeme is one-sided: each flexeme, regardless of the complexity of its forms, is attached to exactly one graph, but one graph can be assigned to many flexemes which consist of the same number of segments and have the same set of forms. The invariance of units (especially visible in the forms of the past tense, future compound tense, and the conditional mood) is recorded in the form of successive paths in the graph (see Fig. 6). For the purpose of describing over 5,000 phraseological units, 818 individual graphs were created. They contain information about inflectional categories and aspects. The markers of grammatical categories and their values follow the tagset of NKJP.

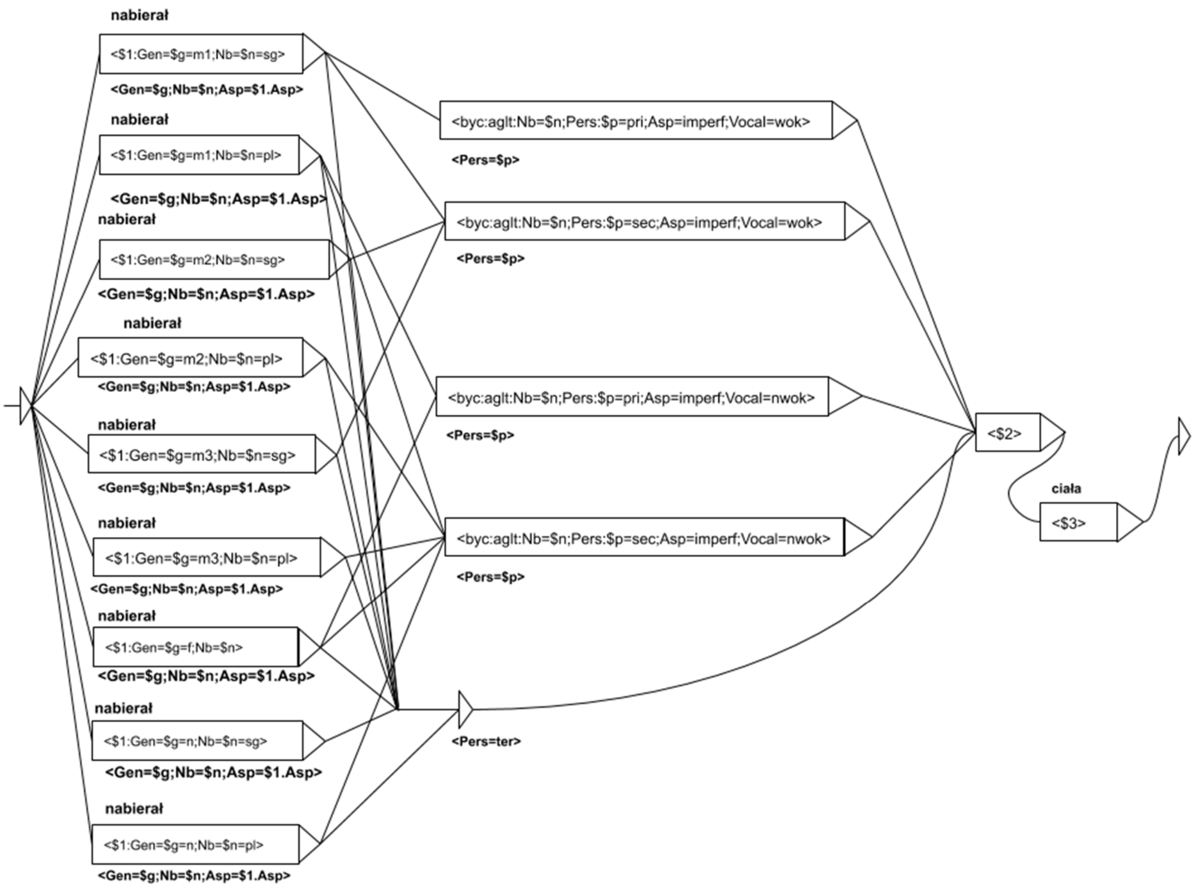
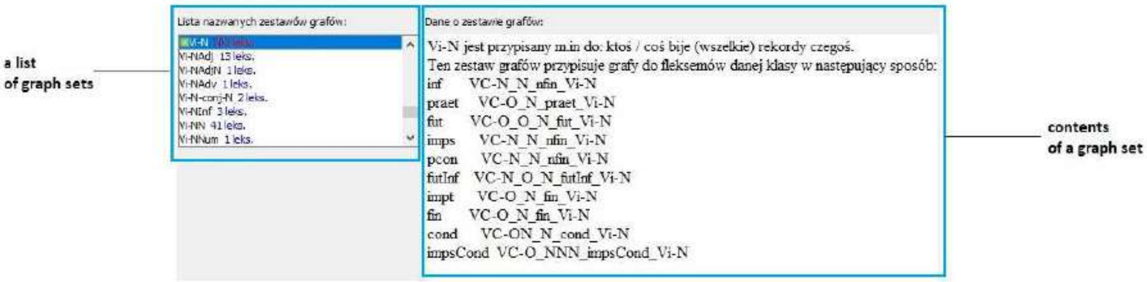

nabierał
<$1:Gen=$g=m1;Nb=$n=sg>
<Gen=$g;Nb=$n;Asp=$1.Asp>
nabierał
<$1:Gen=$g=m1;Nb=$n=pl>
<Gen=$g;Nb=$n;Asp=$1.Asp>
nabierał
<$1:Gen=$g=m2;Nb=$n=sg>
<Gen=$g;Nb=$n;Asp=$1.Asp>
nabierał
<$1:Gen=$g=m2;Nb=$n=pl>
<Gen=$g;Nb=$n;Asp=$1.Asp>
nabierał
<$1:Gen=$g=m3;Nb=$n=sg>
<Gen=$g;Nb=$n;Asp=$1.Asp>
nabierał
<$1:Gen=$g=m3;Nb=$n=pl>
<Gen=$g;Nb=$n;Asp=$1.Asp>
nabierał
<$1:Gen=$g=f;Nb=$n>
<Gen=$g;Nb=$n;Asp=$1.Asp>
nabierał
<$1:Gen=$g=n;Nb=$n=sg>
<Gen=$g;Nb=$n;Asp=$1.Asp>
nabierał
<$1:Gen=$g=n;Nb=$n=pl>
<Gen=$g;Nb=$n;Asp=$1.Asp>
<byc:aglt:Nb=$n;Pers:$p=pri;Asp=imperf;Vocal=wok>
<Pers=$p>
<byc:aglt:Nb=$n;Pers:$p=sec;Asp=imperf;Vocal=wok>
<Pers=$p>
<byc:aglt:Nb=$n;Pers:$p=pri;Asp=imperf;Vocal=nwok>
<Pers=$p>
<byc:aglt:Nb=$n;Pers:$p=sec;Asp=imperf;Vocal=nwok>
<Pers=$p>
<Pers=ter>
<$2>
ciała
<$3>

Figure 6. An example of the graph of a past tense flexeme

The graph above presents one of the most complex flexemes – of the past tense. The multitude of paths in the graph is dictated by the complex morphological structure of this type of form. When generating them, several morphosyntactic parameters should be taken into account at the same time, such as person, gender, and vocalism.

Graphs can be grouped into sets on the basis of their morphological and syntactic features. The same set of graphs is assigned to phraseologisms with a similar formal structure and with exactly the same inflectional paradigm. Each set contains a list of flexemes belonging to the unit along with graphs assigned to individual flexemes (Fig.7).


a list
of graph sets

Lista nazwanych zestawów grafów:

Vi-N [illegible]
Vi-NAdj 13 leks.
Vi-NAdjN 1 leks.
Vi-NAdv 1 leks.
Vi-N-conj-N 2 leks.
Vi-NInf 3 leks.
Vi-NN 41 leks.
Vi-NNum 1 leks.

Dane o zestawie grafów:

Vi-N jest przypisany m.in do: ktoś / coś bije (wszelkie) rekordy czegoś.
Ten zestaw grafów przypisuje grafy do fleksemów danej klasy w następujący sposób:

inf	VC-N_N_nfin_Vi-N
praet	VC-O_N_praet_Vi-N
fut	VC-O_O_N_fut_Vi-N
imps	VC-N_N_nfin_Vi-N
pcon	VC-N_N_nfin_Vi-N
futInf	VC-N_O_N_futInf_Vi-N
impt	VC-O_N_fin_Vi-N
fin	VC-O_N_fin_Vi-N
cond	VC-ON_N_cond_Vi-N
impsCond	VC-O_NNN_impsCond_Vi-N

contents
of a graph set

Figure 7. A list of graphs belonging to a Vp-N set

There is a total of 504 graph sets in *Verbel.* Grouping sets allows one to draw conclusions regarding the number of particular syntactic-morphological types of Polish multi-word units. Almost 30% of graph sets concern regular paradigms with a complete set of forms, a vast majority of which belong to imperfective units. Sets that support the greatest number of units have one nominal complement, both imperfective and perfective verbal component. Respectively, they are attributed to 722 and 595 verbal forms from over 5000 units. However, a significant part of the sets is needed to create the forms of incomplete, defective paradigms. There are more than 100 sets belonging to single, individual units, such as: *{ktoś} przewraca się w grobie* (*{somebody} turns* (*over*) *in* (*one's*) *grave*), *{ktoś} zjadłby konia z kopytami* (*{somebody} could eat a horse*), *{ktoś} nie dałby za {coś} złamanego grosza* ({somebody} *doesn't give / won't give single penny for {something}*).

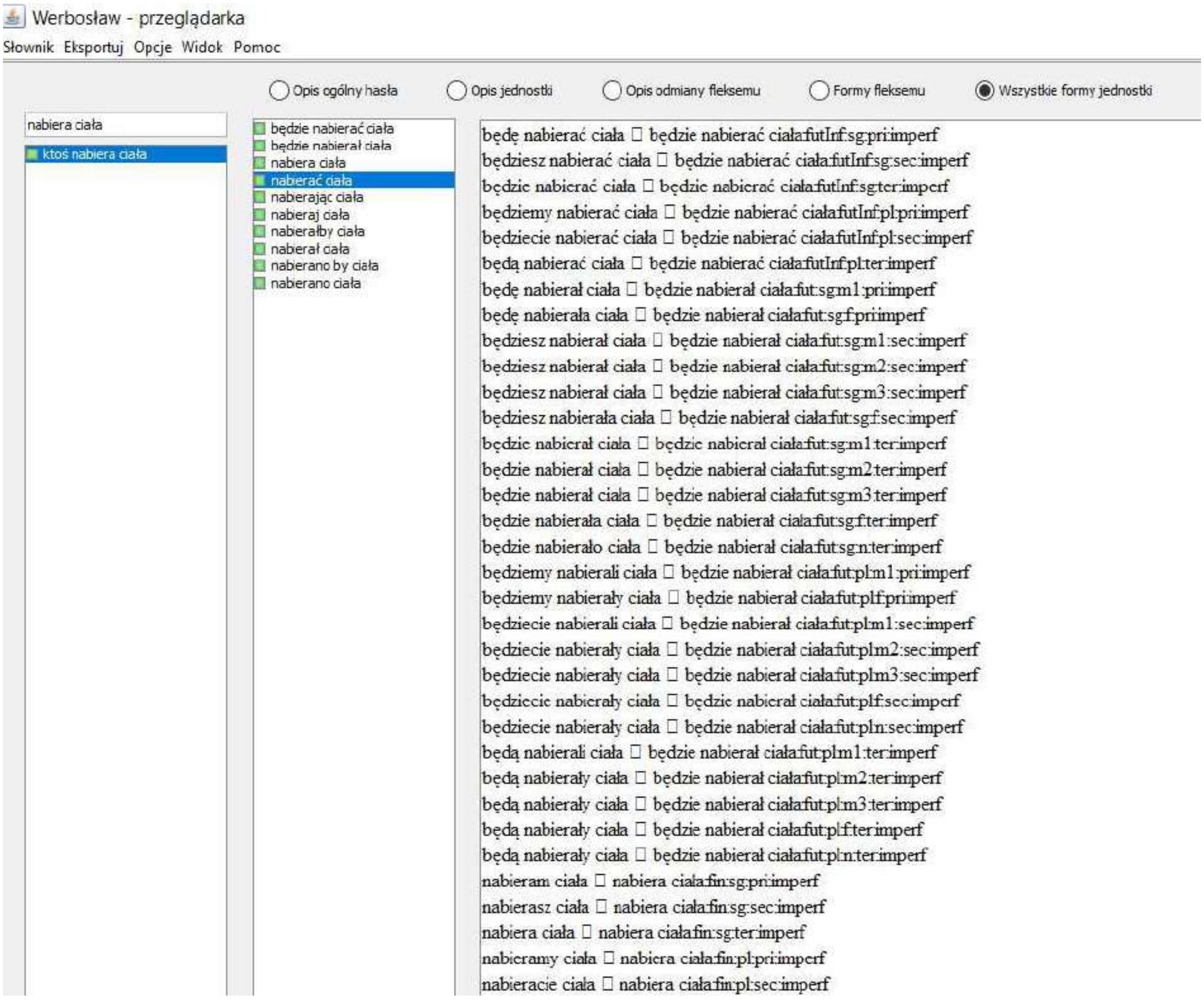
4.5 A full paradigm

On the basis of graphs, the application generates forms of individual flexemes which constitute separate sub-paradigms. In turn, a set of all sub-paradigms constitutes a complete paradigm of the unit. It is a list of all forms of the unit together with a morphosyntactic tag (Fig. 8).

5. Conclusions

Obtaining a full paradigm of a given multi-word unit in the *Verbel* dictionary takes place gradually according to the principle from general to particular, i.e. from general descriptive information about the unit to a list of all its forms (morphological words) with the inflectional characteristics assigned to them. It seems that the data provided at the initial stage (general information about the variant, data about the value of aspect, presence of an aspect equivalent, the meaning and examples) and the final stage (all inflectional forms of the unit) constitute a sufficient lexicographic description.

Placing individual types of information in separate dictionary tabs gives the user the freedom to apply it. It is very likely that an average user will be satisfied with the general description, and perhaps they will also look at the list of forms or flexemes. On the other hand, a specialist – a linguist, lexicographer, computer scientist – will be curious about various stages of the description and technical ways of their presentation. Although the dictionary contains detailed information described according to a specialised linguistic model, its basic use does not require extensive specialist knowledge. Tabs containing details of the formal description can be omitted without losing the functionality of the dictionary.



nabiera ciała

🟩 ktoś nabiera ciała

🟩 będzie nabierać ciała
🟩 będzie nabierał ciała
🟩 nabiera ciała
🟩 nabierać ciała
🟩 nabierając ciała
🟩 nabieraj ciała
🟩 nabierałby ciała
🟩 nabierał ciała
🟩 nabierano by ciała
🟩 nabierano ciała

```
będę nabierać ciała □ będzie nabierać ciała:futInf:sg:pri:imperf
będziesz nabierać ciała □ będzie nabierać ciała:futInf:sg:sec:imperf
będzie nabierać ciała □ będzie nabierać ciała:futInf:sg:ter:imperf
będziemy nabierać ciała □ będzie nabierać ciała:futInf:pl:pri:imperf
będziecie nabierać ciała □ będzie nabierać ciała:futInf:pl:sec:imperf
będą nabierać ciała □ będzie nabierać ciała:futInf:pl:ter:imperf
będę nabierał ciała □ będzie nabierał ciała:fut:sg:m1:pri:imperf
będę nabierała ciała □ będzie nabierał ciała:fut:sg:f:pri:imperf
będziesz nabierał ciała □ będzie nabierał ciała:fut:sg:m1:sec:imperf
będziesz nabierał ciała □ będzie nabierał ciała:fut:sg:m2:sec:imperf
będziesz nabierał ciała □ będzie nabierał ciała:fut:sg:m3:sec:imperf
będziesz nabierała ciała □ będzie nabierał ciała:fut:sg:f:sec:imperf
będzie nabierał ciała □ będzie nabierał ciała:fut:sg:m1:ter:imperf
będzie nabierał ciała □ będzie nabierał ciała:fut:sg:m2:ter:imperf
będzie nabierał ciała □ będzie nabierał ciała:fut:sg:m3:ter:imperf
będzie nabierała ciała □ będzie nabierał ciała:fut:sg:f:ter:imperf
będzie nabierało ciała □ będzie nabierał ciała:fut:sg:n:ter:imperf
będziemy nabierali ciała □ będzie nabierał ciała:fut:pl:m1:pri:imperf
będziemy nabierały ciała □ będzie nabierał ciała:fut:pl:f:pri:imperf
będziecie nabierali ciała □ będzie nabierał ciała:fut:pl:m1:sec:imperf
będziecie nabierały ciała □ będzie nabierał ciała:fut:pl:m2:sec:imperf
będziecie nabierały ciała □ będzie nabierał ciała:fut:pl:m3:sec:imperf
będziecie nabierały ciała □ będzie nabierał ciała:fut:pl:f:sec:imperf
będziecie nabierały ciała □ będzie nabierał ciała:fut:pl:n:sec:imperf
będą nabierali ciała □ będzie nabierał ciała:fut:pl:m1:ter:imperf
będą nabierały ciała □ będzie nabierał ciała:fut:pl:m2:ter:imperf
będą nabierały ciała □ będzie nabierał ciała:fut:pl:m3:ter:imperf
będą nabierały ciała □ będzie nabierał ciała:fut:pl:f:ter:imperf
będą nabierały ciała □ będzie nabierał ciała:fut:pl:n:ter:imperf
nabieram ciała □ nabiera ciała:fin:sg:pri:imperf
nabierasz ciała □ nabiera ciała:fin:sg:sec:imperf
nabiera ciała □ nabiera ciała:fin:sg:ter:imperf
nabieramy ciała □ nabiera ciała:fin:pl:pri:imperf
nabieracie ciała □ nabiera ciała:fin:pl:sec:imperf
```

Figure 8. A full paradigm of the unit {*ktoś*} *nabiera ciała* (lit. {*someone*} *gets a body*, 'someone puts on weight')

In this regard, the dictionary may be useful for an average user, not a specialist, especially since the transition from a general description to a full paradigm does not require going through all the description steps in a sequence. Intermediate stages may prove interesting for researchers who focus on describing natural language in a formal manner. In this sense, the dictionary can reach a wide audience. Perhaps it is far from a particularly user-friendly dictionary, and to become one it needs an appropriate interface. Currently, it is an offline dictionary, and taking into account the expectations of users and technological development the web version would be of greater value. However, these are purely technical conditions. When it comes to the lexicographic layer, the dictionary contains data that is appropriately organised to be a resource for a wide audience.

Since each single form carries a label that includes the name of the base flexeme and its grammatical characteristics (see Figure 8), the data contained in the *Verbel* dictionary can be useful in marking multi-word units in other linguistic tools: text corpora and treebanks, especially because the morphosyntactic marker system used in the dictionary is compatible with the tagset of Polish National Corpus. That is why the dictionary can also be applied in further research on multi-word units in texts.

6. References

Baldwin, T. & Kim, S. N. (2010). Multiword Expressions. In: N. Indurhya, F.J. Damerau (eds.) *Handbook of Natural Language Processing.* Boca Raton, USA: CRC Press, pp. 267–292.

Bień, J. S. & Saloni, Z. (1982). Pojęcie wyrazu morfologicznego i jego zastosowanie do opisu fleksji polskiej (wersja wstępna). *Prace Filologiczne*, XXXI, pp. 31–45.

Bogusławski, A. (1976). O zasadach rejestracji jednostek języka. *Poradnik Językowy*, 8(342), pp.356–364.

Czerepowicka, M. (2014). SEJF – Słownik elektroniczny jednostek frazeologicznych, *Język Polski*, XCIV, vol. 2, pp. 116-129.

Czerepowicka, M., Savary, A. (2018). SEJF – a Grammatical Lexicon of Polish Multi-Word Expressions. In: Z. Vetulani, J. Mariani (eds.) *Human Language Technologies as a Challenge for Computer Science and Linguistics: 7th Language and Technology Conference, LTC 2015, Poznań, Poland, November 27–29 2015, Revised Selected Papers, In memoriam Adam Kilgarriff.* Lecture Notes in Artificial Intelligence, vol. 10930. Berlin: Springer-Verlag, pp. 59–73.

Iosad, P., Koptjevskaja-Tamm, M., Piperski, A. & Sitchinava, D. (2018). Depth, brilliance, clarity: Andrey Anatolyevich Zaliznyak (1935–2017). *Linguistic Typology*, 22(1), pp. 175–184.

Kosek, I., Czerepowicka, M., Przybyszewski, S. (2020). *Verbel. Elektroniczny słownik paradygmatów polskich frazeologizmów czasownikowych. Teoria, problemy, prezentacja.* Olsztyn: University of Warmia and Mazury.

Lewicki, A. M. (1986). Składnia związków frazeologicznych. *Bulletin de la Société Polonaise de Linguistique*, XL, pp. 75–83.

Marciniak, M., Savary, A., Sikora, P. & Woliński, M. (2011). Toposław – a lexicographic framework of multi-word units. In: Z. Vetulani (ed.) *Human Language Technology. Challenges for Computer Science and Linguistics: 4th Language and Technology Conference, LTC 2009, Poznań, Poland, November 6-8, 2009, Revised Selected Papers.* Lecture Notes in Artificial Intelligence, vol. 6562. Berlin: Springer-Verlag, pp. 139–150.

Mel'čuk, I. (2006). Explanatory Combinatorial Dictionary. In: G. Sica (ed.) *Open Problems in Linguistics and Lexicography.* Monza: Polimetrica, pp. 225–355.

Przepiórkowski, A., Bańko, M., Górski, R. L. & Lewandowska-Tomaszczyk, B. (eds.). (2012). *Narodowy Korpus Języka Polskiego.* Warsaw: Wydawnictwo Naukowe PWN.

Sag I.A., Baldwin T., Bond F., Copestake A., Flickinger D. (2002). Multiword Expressions: A Pain in the Neck for NLP. In: Gelbukh A. (eds.) *Computational Linguistics and Intelligent Text Processing.* CICLing 2002. Lecture Notes in Computer Science, vol. 2276. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-45715-1_1.

Saloni, Z., Świdziński, M. (1998). *Składnia współczesnego języka polskiego*, Edition IV, changed. Warsaw: Wydawnictwo Naukowe PWN.

Woliński, M. (2014). Morfeusz reloaded. 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 2014, pp. 1106–1111, Reykjavík, Iceland, 2014. ELRA.

Tokarski, J. (1973). *Fleksja polska.* Warsaw: Państwowe Wydawnictwo Naukowe.

Dictionaries & Websites:

Bogusławski, A. & Danielewiczowa, M. (2005). *Verba polona abscondita. Sonda słownikowa III.* Warsaw: Elma Books.

Bogusławski, A. & Garnysz-Kozłowska, T. (1979). *Addendum to Polish phraseology. An introductory issue.* Edmonton: Linguistic Research.

Bogusławski, A. & Wawrzyńczyk, J. (1993). *Polszczyzna, jaką znamy. Nowa sonda słownikowa.* Warsaw: University of Warsaw.

Dunaj, S. (ed.). (1996). *Słownik współczesnego języka polskiego.*Warsaw: Wilga.

Mel'čuk, I. & Zholkovsky, A. (1984). *Explanatory Combinatorial Dictionary of Modern Russian. Semantico-syntactic Studies of Russian Vocabulary.* Viena: Wiener Slawistischer Almanach.

NKJP: Narodowy Korpus Języka Polskiego. Accessed at: http://nkjp.pl/ (20 May 2021).

Saloni, Z. (2007). *Czasownik polski. Odmiana. Słownik 12 000 czasowników.* Edition III, changed. Warsaw: Wiedza Powszechna. Accessed at: https://depot.ceon.pl/handle/123456789/20067 (27 May 2021).

SEJF: *Słownik elektroniczny jednostek frazeologicznych.* Accessed at: http://zil.ipipan.waw.pl/SEJF (20 May 2021).

SGJP: *Słownik gramatyczny języka polskiego.* Accessed at: http://sgjp.pl/ (20 May 2021)

Verbel: *Verbel. Elektroniczny słownik paradygmatów polskich frazeologizmów czasownikowych.* Accessed at: http://uwm.edu.pl/verbel/ (20 May 2021).

Zaliznyak, A. A. (1977). *Grammatičeskij slovar' russkogo jazyka. Slovoizmenenie* [*A grammatical dictionary of Russian: Inflection*]. Moscow: Russkij jazyk.

Dictionaries as collections of lexical data stories: an alternative post-editing model for historical corpus lexicography

Ligeia Lugli[1]

[1] SOAS, room 339, 10 Thornhaugh St, London WC1H 0XG, UK
E-mail: ll34@soas.ac.uk

Abstract

This paper proposes a model of dictionary post-editing inspired by data-journalism. It starts by problematising the parallel, drawn in the description of this year's eLex conference theme, between lexicographic and machine-translation post-editing. It then proceeds to outline data-journalism workflows and to illustrate how these may offer a suitable blueprint for automating and post-editing corpus-driven historical dictionaries of low-resource languages. In particular, the paper highlights the usefulness of adopting an iterative development model, whereby minimal automated entries are incrementally augmented with curated information, and of switching to data-visualisations as the main medium of communication.

Data-journalists concentrate much of their post-editing efforts in plotting the data into highly customised visualisations capable of narrating their interpretation of a story while also allowing multiple lines of inquiry. This paper suggests that historical lexicographers would benefit from similarly directing their post-editing efforts into weaving data into customised, lemma-specific, visualisations capable of guiding users towards further exploration.

The paper concludes with practical examples drawn from two ongoing historical dictionary projects, *A Visual Dictionary and Thesaurus of Buddhist Sanskrit* and *A Visual Dictionary of Tibetan Verb Valency*, which are adopting data-journalism workflows to post-edit automatically generated entries and data-visualisations into 'lexical data stories'.

Keywords: historical lexicography; data-journalism; post-editing; Sanskrit; Tibetan

1. Machine-translation post-editing for lexicography: a critique

For decades lexicography has been on a path of increasing automation. The late 90s and early 2000s vision of machines taking up the bulk of lexicographic work is now coalescing into reality (Grefenstette, 1998; Rundell, 2002). Hypothetical notions regarding the role of humans in a largely automated workflow are quickly being replaced by practical strategies for post-editing automated dictionary drafts. It is therefore a good moment to look at industries that already possess well established post-editing workflows and consider which could be most profitably adapted to which lexicographic endeavour.

The description of this year's eLex conference theme conceptualises lexicographic post-editing as akin to the post-editing practices honed in the field of machine-translation,[1] a parallel already drawn by Jakubíček a few years ago (Jakubíček, 2017). While machine-translation post-editing workflows may be profitably adapted to some

[1] '...This technological progress leads to new methodological approaches where most editorial work consists of post-editing of automatically created content – similarly to post-editing of machine-translated texts.' (eLex 2021 introductory paragraph, https://elex.link/elex2021/)

lexicographic projects (e.g. Baisa et al., 2019), they are not likely to constitute an optimal model for lexicography in general, and especially not for historical lexicography of low-resource languages, which is the focus of this paper. This is mostly due a fundamental difference in the nature and goals of translation and historical lexicography.

In machine-translation projects, computers generate a draft translation from an input text and humans refine it. The degree of manual refinement (i.e. post-editing) varies depending on how similar to a human-made translation the final product should be. ‘Light’ post-editing is often sufficient to ensure that the message of source text is rendered accurately, while more labour intensive 'full' post-editing may be required to achieve a perfectly smooth reading experience in the target language, akin to a human translation (Nitzke et. al., 2019). In other words, machine-translation post-editing practices are articulated along two axes, accuracy, intended as faithfulness to the source text, and readability of the output text.

The relevance of these axes to historical lexicography is doubtful. While basic readability is indeed important, dictionary entries need not be specimen of great prose. Given their standardised wording and rigorously structured format, text generation templates should be capable of producing perfectly readable, if perhaps not enjoyable, dictionary entries (see Section 2 below). Post-editing for readability is therefore not likely to constitute a priority for many historical dictionary projects. Accuracy, by contrast, is a very likely priority. However, what constitutes accuracy in translation and in lexicography is entirely different. As such, machine-translation post-editing practices may well not be the best route to lexicographic accuracy.

The reason for this lies in a fundamental difference in the relationship between input data and output text in translation and lexicography. Translation aims at transforming its source data (by transposing it into another language), whereas lexicography aims at illustrating trends in its source data and deriving conclusions from them. This impacts the efficacy of text post-editing for accuracy in the two fields. In translation, manipulating the wording of the machine-generated draft directly affect its accuracy. Post-editing is thus an efficient path to improving the quality of computer-generated translations. While changing the text of automated dictionary drafts may also improve the overall dictionary quality, this is not an efficient path to increased accuracy. Lexicographic accuracy resides not so much in the wording of the entries as in the quality of sample, analysis and interpretation of the corpus data. Lexicographic accuracy is thus more directly impacted by addressing the representativeness of the corpus used, the level of detail of the linguistic annotation recorded in the corpus and the relevance of the statistical information automatically derived from it (Frankenberg-Garcia et al., 2020; Baisa et al., 2019). As it will be discussed in section 3, post-editing may not be the most efficient way to address these matters in historical lexicography of low resource languages, where the efforts could rather be concentrated in enriching a small corpus with detailed linguistic information.

Moreover, what constitutes an accurate representation of the input data is much more subjective in historical lexicography than it is in machine-translation. Interpretation is typically straightforward in automatically translated texts—literary works, puns and ambiguous prose lying still largely beyond the scope of machine-translation, and best translated from scratch by humans (Nitzke et. al., 2019). This means that machine-translation post-editing can realistically aim to achieve an uncontroversial version of the translated text; a version that is going to be equally useful to all its readers.

The situation is more complex in lexicography. Much of what goes into a dictionary entry, from sense categorisation and sense descriptions up to example selection, is highly interpretive. In the case of historical lexicography matters of philological uncertainty, disputed dating and difficulties of interpretation further complicate the picture. Adopting a machine-translation post-editing model in historical lexicography hardly does justice to this complexity, or to dictionary users. It implies a conceptualisation of dictionary entries as a definitive top-down account of a word's semantics and usage, which risks misrepresenting interpretation and subjective choices as purely descriptive accounts. This vastly limits the usefulness of historical dictionaries as tools for research. Post-editing models that allow users to pursue different interpretations of the data and provide a transparent record of lexicographers' editing choices may yield more versatile and useful resources.

Finally, a post-editing model inspired by machine-translation raises concerns of sustainability for historical dictionary projects that depend on public funding. Public funding cycles for humanities projects are relatively short, covering typically a period of three years in the UK and USA (e.g. schemes funded by the Arts and Humanities Research Council and National Endowment for the Humanities). As a result, historical dictionary projects often need to produce a minimally viable product very quickly in order to showcase their outputs early and secure follow-up funding for further work. If dealing with low-resource languages or specialised domains, they also often need to create and process corpora from scratch and thus invest a significant portion of their first funded period into developing the source data necessary for their dictionaries. Under these circumstances, it is advisable to develop dictionaries iteratively, by first publishing automated entries based on corpus data and then gradually refining and augmenting them through further iterations (Lugli, 2019). This makes it possible to align lexicographic outputs with funding cycles, but it is important to note that this model is efficient only in so far as there is no overlap in the work required for each iteration. It is doubtful that this is best achieved through the adoption of post-editing practices inspired by machine-translation.

In machine-translation contexts, the choice between different levels of post-editing (bare machine output, light post-editing or full post-editing) occurs early on in a project. The literature on automated translation construes the relationship between light post-editing and full post-editing as one of alternative editorial strategies, rather than as a progression between different editorial stages, since arguably both involve much of the same tasks (see the post-editing decision tree in Nitzke et al. 2019, 246). While the practices developed for machine-translation can surely be adapted to the needs of lexicography (as accomplished, for example in the project described in Baisa et al., 2019), in light of the limitations outlined above it seems useful to expand the pool of reference models available for dictionary post-editing. I propose that we consider one model that is remarkably close to historical lexicography in several respects: data-journalism.

2. Text automation and post-editing in journalism

Data-journalism is a branch of journalism that focusses on deriving news stories from datasets and typically conveys much of the information through data-visualisations.[2] The complexity of data-journalism pieces ranges from relatively simple graphs and narratives, such as those charting the spread of COVID-19, ubiquitous in newspapers these days, to the more nuanced and interpretive pieces published in dedicated data-journalism outlets, such as *The Economist*'s *Graphic Detail.*

Like translation, journalism has undergone considerable levels of automation in recent years. As with machine-translation, drafts of news pieces are now routinely generated automatically and then refined through human curation (Marconi, 2020; Diakolpoulos, 2019; Graefe, 2016). The processes of text generation and post-editing, however, differ between the translation and news industries. The difference is, again, rooted in the relationship between input data and automatically generated output. While translations transpose the input data into a new language, news pieces elaborate on the input data, typically producing entirely new text from and about numeric inputs.

An output text's relationship with the input data varies depending on the type of news. Reports on sport matches or election results summarise the input data; financial news may highlight trends and changes in assets' value; in-depth analyses may draw conclusions from the input data, or use them to support a specific argument. While all kinds of data-based news can be (and indeed are) automated, the degree and quality of the automation, as well as the post-editing strategies required to reach a publishable product vary.

There is consensus in the literature on automated journalism that the best automated output is achieved with types of news that have a relatively rigid format, a predictable vocabulary, rely on highly structured data and describe (rather than interpret) the input data. These types of news include market and weather reports as well as sports

[2] My use of the term data-visualisation is close to the definition provided by Bakakis: 'Data visualization is the presentation of data in a pictorial or graphical format, and a data visualization tool is the software that generates this presentation. Data visualization provides users with intuitive means to interactively explore and analyze data, enabling them to effectively identify interesting patterns, infer correlations and causalities, and supports sense-making activities. ' (Bakakis, 2018), but I extend my application of the term to cover cases of static (i.e. non-interactive) data-visualisation as well.

and election results—all of which have been routinely automated for years (Carlson, 2015; Diakopoulos, 2019). For these types of news, automated text is published with minimal or no human post-editing (Graefe et al., 2018; Diakopoulos, 2019).[3] This is not to say that these news pieces do not require any human labour at all. Rather, the labour is concentrated in pre-processing. Before any automated news writing can take place, humans need to prepare the data and templates that will be used to generate the text of news pieces. Data preparation includes the usual steps of cleaning, wrangling and verification, and needs to be performed on any new data used. This appears to be the weakest link in run-of-the-mill news automation, as the errors discussed in the literature are all due to poorly pre-processed data (e.g. Diakopoulos, 2019: 133; Marconi, 2020: 69). Template preparation is more robust, but rapidly evolving. Traditionally, templates for automatic text-generation are 'hard coded'. News editors prepare set templates for each type of news, detailing the order in which the information is to be presented, as well as alternative sentence structures to be used convey each piece of information and pools of synonyms to choose from to ensure some variation in the automated texts. The results of this procedure are consistently good and often indistinguishable from human writing (Diakopoulos, 2019: 126). In recent years, the creation of templates has been partially automated and machines are now able to structure a piece and concatenate (and in some cases craft) sentences on the basis of rules and/or statistical models derived from news corpora (Diakopoulos, 2019: 98 ff.; Leppänen et al., 2017). This obviously leads to faster pre-processing by drastically reducing the need for detailing domain-specific templates. The overall time and labour required to achieve a publishable product, however, is not reduced. Dynamically created templates tend to introduce problems of readability and thus require more post-editing efforts. Unsurprisingly, the news industry prefers to invest resources in labour-intensive template creation and dispense with (or minimise) post-editing, rather than opt for the reverse (Diakopoulos, 2019). This is an efficient choice as even though they may not generalise well across different types of news, detailed templates are still re-usable for all news within a given category. Post-editing by contrast is piece-specific; it is not re-usable at all, at least for now.[4]

The opposite is true for news stories that are based on data but require investigation, interpretation and are best conveyed through original narratives. That is, the type of news stories that is most typically referred to as 'data-journalism'.[5] Even though data

[3] RADAR, a leading news project, only manually checks the output of one in ten automated news pieces (Diakopoulos 2019, 134).

[4] See Diakolopoulos's brief discussion of 'distant editing' as a prominent desideratum in the news industry (Diakopoulos 2019, 134 and 247-248).

[5] Several definitions of data-journalism and discussions of its relative position within the field journalism vis-à-vis other computer-enhanced forms of news-making have been put forward (see Coddington 2018 for a comprehensive review). For the purposes of this paper, the generic characterisation of data-journalism as an approach to crafting news stories that is centred on the acquisition, analysis, interpretation and publication of data will suffice (cf. Usher, 2016: 90; Howard, 2014: 2-5; both cited in Coddington, 2018: 17).

play a central role in these stories, they cannot simply be plugged in a text template. The narrative is too unique to be amenable to templates; no matter how sophisticatedly constructed they might be (Stray, 2019; Caswell and Doerr, 2018). The efficient choice for these stories is to switch from a paradigm of pure automation to one of augmentation, whereby machines generate a minimal description from the data and leave it to journalists to investigate and flesh out the narrative of the story (Diakopoulos, 2019: 46ff; Graefe, 2016: 29). While the journalism literature refers to this process as 'augmentation' or 'human-machine interaction' (Marconi, 2020: 69-71; Diakopoulos, 2019: 247-248), it is a form of post-editing, in so far as it amounts to the manual curation of an automatically generated draft. Still, it differs from machine-translation post-editing in two important respects: it is iterative and not centred on text.

In data-journalism, the initial automated summary of data can constitute a minimal viable product (or 'minimally viable story', Marconi, 2020). This product may not be fit for publication in a newspaper, but it is usually good enough to be immediately released in the form of a blog post or as a news alert (Young and Hermida, 2015). The automated summary can then be enriched with more information and interpretation in successive stages—possibly depending on the amount of interest that each iteration of the story generates among the public (Marconi, 2020).[6] Besides being efficient for news production, this iterative story development is also empowering for the reader. It provides early and comprehensive access to granular data that would otherwise not be available, such as real time information on local crime or a detailed breakdown of minor election results, which journalists would rarely have the time to report manually (Young and Hermida, 2015; Leppänen et al., 2017; Marconi 2020).

Unredacted automatic reports may not make for a very enjoyable read, though. Fortunately, the dullness of automated text can be entirely bypassed by presenting the automated data summary in the form of data-visualisations. Reliance on data-visualisations is one of the most salient features of data-journalism (Coddington, 2018; Kennedy et al., 2019).[7] Tools for the automatic identification of potentially newsworthy leads typically supply journalists with visual analytics (Diakopoulos, 2019: 57, 48ff; Wiedmann, 2018; Stray 2019), and systems are in place to automatically generate publication-ready data-visualisations to accompany data-driven news (Alhalaseh et al., 2018). The initial automatically generated minimally viable story could thus take the form of a graph or data-visualisation dashboard (e.g. Diakopoulos, 2019: 49 fig 2.1).

Post-editing also focusses on visualisations. Much of the educational literature on how to craft data-journalism stories stresses the importance of editing the visualisations

[6] See Marconi 2020, chapter 3 for a detailed explanation of iterative journalism.

[7] Data-visualisations are perceived by some as having replaced writing as the "main semiotic mode" of journalistic storytelling (Kennedy et al., 2019).

accompanying the story so that they communicate the main points of the narrative, highlight the author's interpretation of the data and guide the user towards specific insights (Thudt et al., 2018; Stopler et al., 2018; Kennedy et al., 2019). Given the interpretative nature of data-journalism, another topic that is emphasised in this literature is the role of interactive data-visualisation in encouraging users to explore multiple lines of inquiry, reach different interpretations and reveal bias (Thudt et al., 2018; Diakopouls, 2018, 246). By curating data-visualisations and letting users explore the dataset used for a story, journalists increase transparency and civic engagement, two cornerstones of data-journalism ethics (Coddington, 2018; cf. Kennedy et al., 2019). These practices may also help historical lexicographers meet the needs of their audiences.

3. A data-journalism post-editing model for lexicography

The post-editing practices developed for data-based news pieces could be profitably transferred to historical corpus lexicography of low-resource languages. This subset of lexicography possesses some characteristics that make it an especially good fit for the newsroom's approach to post-editing.

First of all, its low-resource aspect. Limited budget and manpower make it necessary to prioritise efforts very carefully, and dependence on public funding makes iterative dictionary development especially suitable for this type of lexicography. Under these circumstances, the newsroom practice of shifting labour from post-editing single-purpose texts to preparing data and templates for the automatic generation of multiple texts is appealing.

This model of labour allocation may even work better in lexicography than in news production, for two reasons. As mentioned earlier, poorly prepared data and complex narratives are the two main obstacles to post-editing-free news automation. Neither of these apply to lexicography. Data preparation is challenging in journalism because news data change continuously and thus require constant monitoring and checking. By contrast, the data used for historical dictionaries typically amounts to a language corpus that only needs be prepared once. Moreover, while only a fraction of news stories fit the requirements for template-based text generation, dictionary entries, with their fixed structure, formulaic phraseology and well-ordered integration of corpus data, are perfectly amenable to simple templates, which can easily be enriched with dynamic data-visualizations to allow users to actively engage with the data behind the entries. Indeed, the dictionary post-editing model inspired by machine-translation also leverages this characteristic of lexicographic entries by slotting automatically extracted and sorted corpus data in specified fields within an entry (e.g. Měchura, 2017). The difference in the data-journalism model is that an automated minimally viable entry can be published without any post-editing and still be highly engaging thanks to reliance on interactive data-visualisations and highly curated corpus data (cf. Baisa et al., 2019).

This, again, works best for low-resource historical languages. For three reasons. First, the corpora available for these languages are typically rather small by contemporary standards and are often created for specific lexicographic purposes. This allows for more fine-grained annotations to be encoded in the corpus than is typically possible for larger corpora. It also allows for manual curation of the annotations, which as a result may be more accurate and detailed than the automated tagging typical of large corpora (Lugli, 2019). Such accurate and fine-grained annotations in turn allow for a wider range of information to be automatically derived from the corpus and plugged into entry templates, thus enabling the creation of fairly rich automated entries (see the next section for examples).

Second, new historical dictionaries of low-resource languages typically bring to the public lexical data that would not otherwise be available (e.g. data from newly created corpora or newly discovered manuscripts). Hence their audiences are likely to benefit from early access to new lexicographic material, even if it is in the minimal form of an automated entry.

Finally, historical dictionaries of low-resource languages tend to be used for research purposes, often by highly trained academics. Some of the work typically required in dictionary post-editing, such as checking the automated selection of examples, can therefore be offloaded to users, who may even prefer to filter through examples themselves, using custom parameters, rather than be given a fixed set of sentences pre-selected by lexicographers.[8] Given the uncertainty surrounding much historical material, especially for low-resource languages, these users are also likely to prefer having the option of engaging directly with the data rather than being given solely a top-down interpretation of the meaning and evolution of a given lemma. A purely automated entry presenting annotated corpus data could thus serve this user pool, especially if it offers the possibility to explore the data interactively.

To this end, the data-journalism practice of publishing automated news stories as or with data-visualisations is, again, better suited for historical lexicography than the machine-translations model of a text-centred dictionary entry. Since dictionaries have been moving away from prescriptivist definitions and towards descriptions of words' use, conveying the content of lexicographic entries through data-visualisation has become easier. An automated description of corpus information is easier to render graphically than verbally. Easily programmable data-visualisations can efficiently represent data that would require complex sentences and elaborate text-generation templates to be described in text (see next section for examples).

Overall, data-visualisations require less post-editing than text. Problems of syntax, infelicitous wording or clumsy sentence concatenation do not apply to charts. Still,

[8] This is the feedback we received from prospective users of both the Tibetan and Sanskrit dictionaries.

post-editing data-visualisation is advisable. Automatically generated charts may fail to highlight the most interesting aspect of the data, or obfuscate important patterns in a sea of data points. Especially so, if the same set of visualisations is applied to all lemmata in the dictionary, regardless of their semantic characteristics or distributional patterns. Some charts will inevitably fit better one lemma than another. Hand picking the best chart for each lemma and selecting the most effective colour scheme or interactive options for each type of information is thus an important task.

Data-visualisation post-editing can be the focus of a second iteration of the dictionary. Here the automated entries generated in the first iteration can be augmented with a view to guide users through the data. A third iteration can further augment the entries with the addition of a text narrative that explains the lexicographer's interpretation of the data. This final iteration would combine high-level lexicographic curation with interactive data-exploration, thus balancing guided and self-directed use of the resource and allowing multiple interpretations.

Given that post-editing is both labour intensive and single-purpose, it may be expedient to limit it to a subset of entries (cf. Baisa et al., 2019). Following the data-journalism model, lexicographers could concentrate their manual efforts on lemmata that are deemed especially interesting, either because they attract the most views from users or because they satisfy some predefined statistical test. For example, polysemic words that display dramatic diachronic changes could be the focus of detailed entries that explain their development and semantic plasticity, while monosemic words that are homogeneously distributed across periods may be satisfactorily represented by automated minimal entries.

4. Examples from Tibetan and Sanskrit lexicography

A post-editing model inspired by data-journalism has been applied to two historical dictionaries of low-resource languages currently under development. Both dictionaries are still undergoing their first iteration and are presently best characterised as working prototypes. Both are highly specialised lexical resources, one is a dictionary and thesaurus of Buddhist Sanskrit aimed at translators of Buddhist literature (*A Visual Dictionary and Thesaurus of Buddhist Sanskrit*), the other is a diachronic valency lexicon of Tibetan verbs (*A Visual Dictionary of Tibetan Verb Valency*).

The two resources differ completely in content and aim, but have been developed following the same 'deferred post-editing' iterative model, whereby a completely automated version of the dictionary is released as proof of concept and is then followed by incrementally augmented iterations (Lugli, 2019). While both dictionaries have adopted an approach to post-editing that closely resembles that of automated journalism (especially the one described in Marconi, 2020), it should be noted that this connection is made *a posteriori* and the dictionary development model was not originally inspired by data-journalism. By contrast, new editorial directions regarding data-visualisation post-editing, which will be introduced shortly, have been explicitly modelled after data-journalism best practices.

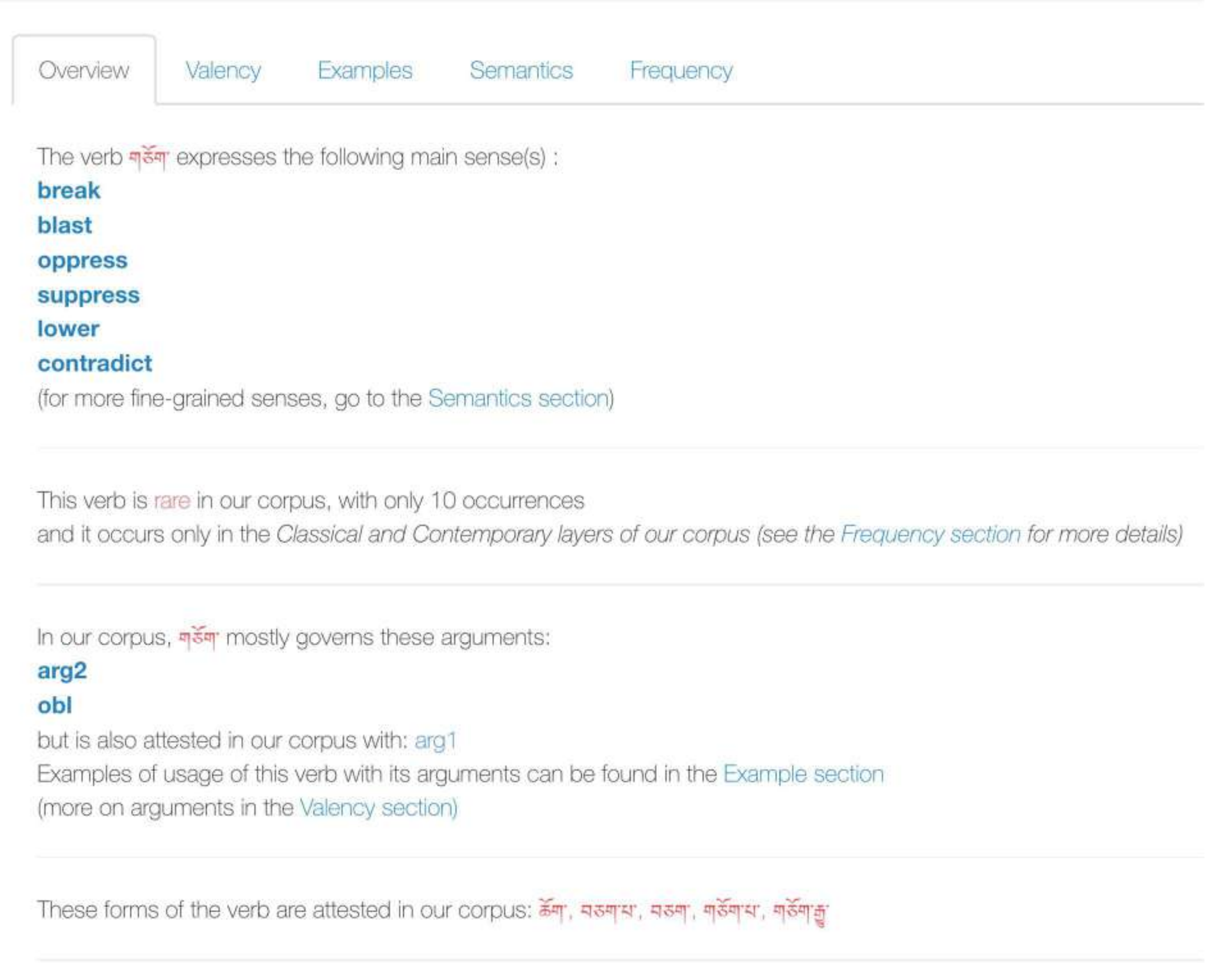
As with automated news, the dictionary projects discussed here shifted the bulk of lexicographic work from post-editing to data and template preparation. For the Tibetan project, most of the lexicographic work consisted of annotating a small diachronic corpus with verb argument structure, which is the primary focus of this resource. Additionally, a sample of about five thousand sentences instantiating the top hundred most frequent verbs are also being annotated semantically, by specifying the meaning that the headword verb takes in each sentence. Since other good dictionaries of Tibetan verbs exist, in this project we have opted for using the sense categorisation provided in pre-existent resources (specifically Hill, 2010). By contrast, the Buddhist Sanskrit project focusses on fine-grained semantic analysis, with lexicographers annotating a small corpus of sampled sentences with original information regarding word senses, semantic prosody, as well as conceptual and syntactic relations. The annotation process in both projects has been time consuming, but has resulted in a re-usable, multi-purpose dataset that makes lexicographic analysis not only time efficient but also completely transparent by clearly associating each data point with the corresponding interpretation provided by the lexicographers (Lugli, 2019).[9]

For both dictionaries, the annotated corpus data is plugged into a programmatic template that for each headword generates three main types of outputs: 1) a short text summary, 2) a variety of interactive data-visualisations and 3) a dynamic list of examples that can be filtered according to various parameters. At present, the automated text is minimal. In the Sanskrit dictionary it only provides a breakdown of the headword's senses, whereas in the Tibetan valency lexicon it also adds a summary of frequency, diachronic distribution and valency structure of each headword and, where applicable, the light verb constructions (a type of multiword expression) in which it participates (Figure 1). Following feedback from peers and users we will expand the automated text summaries to include more descriptive prose and some examples.

Examples are currently displayed in a separate tab in both resources. The Sanskrit dictionary has a 'quick examples' tab that displays examples that have been manually selected by lexicographers during corpus annotations and a 'more examples' tab that allows users to access all the sentences that have been annotated for a lemma and filter them by genre, sense, semantic prosody and grammatical features. The Tibetan lexicon does not offer hand-picked examples. Instead, it sorts the annotated sentences according to a 'good example' score inspired by the Sketch Engine's Gdex paradigm (Kilgariff et al., 2008). Soon, it will offer users the possibility to manipulate this score according to their own preferred parameters. After all, what constitutes a good example largely depends on what a user wishes to see exemplified. The default score aims at prioritising sentences that express a complete thought, are relatively short, contain no anaphoric references and only minimal 'noise', such as long lists of verbs

[9] Lugli 2019 discusses in detail the efficiency of this workflow.

and strings of modifiers (Lugli, 2019, 208-209). Yet, some users may wish to see examples of the interplay between valency patterns and anaphoric markers, or see how the headword verb is strung together with other verbs in formulaic lists.



The verb གཅོག expresses the following main sense(s) :
break
blast
oppress
suppress
lower
contradict
(for more fine-grained senses, go to the Semantics section)

This verb is rare in our corpus, with only 10 occurrences
and it occurs only in the *Classical and Contemporary layers of our corpus (see the Frequency section for more details)*

In our corpus, གཅོག mostly governs these arguments:
arg2
obl
but is also attested in our corpus with: arg1
Examples of usage of this verb with its arguments can be found in the Example section
(more on arguments in the Valency section)

These forms of the verb are attested in our corpus: ཆོག, བཅག་པ, བཅག, གཅོག་པ, གཅོག་རྒྱུ

Figure 1. Automatically generated lemma overview in the Visual Dictionary of Tibetan Verb Valency (mangalamresearch. shinyapps.io/VisualDictionaryOfTibetanVerbValency/, accessed on 8/4/2021)

Examples are not the only area where the display preferred by lexicographers and users may differ. Both dictionaries allow users to interact with the graphs to view their own preferred combination of variables, switch between different types of charts, or change between normalised and absolute frequencies. More importantly, the Sanskrit dictionary allows users to customise periodisation and other metadata and adjust all data-visualisations accordingly. This is crucial when dealing with Sanskrit literature, where dating of texts is uncertain and often hotly disputed (Lugli, 2018).

In sum, the first iteration of both dictionaries offers users a wealth of manually annotated data and the possibility to explore it interactively and, potentially, to reach their own conclusions. One important limitation of these first iterations is that they do not make explicit the conclusions reached by the lexicographic teams. While the interpretation of each sentence is granularly recorded in the source data in the form of linguistic annotations, the automated entries do not provide an overall interpretation of the semantic or syntactic history of headwords. Such interpretation is the object of our post-editing phase and constitutes the focus of further iterations.

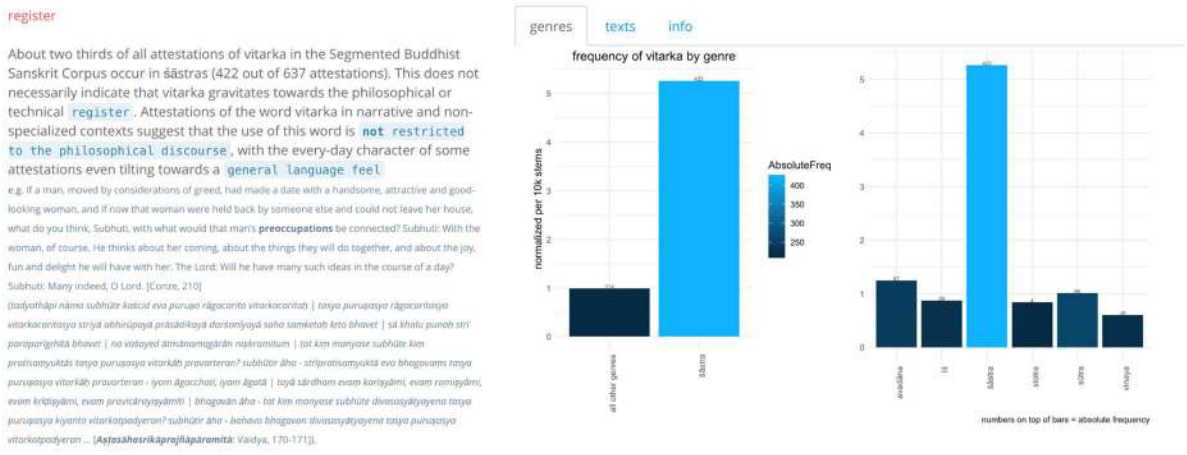
A second iteration is currently being planned for the Buddhist Sanskrit dictionary. It will centre around the creation of 'lexical portraits', curated interpretations of the data that mix narrative and edited data-visualisations. While the details of our post-editing pipeline are still being tried and tested, the general principles are clear. They are inspired by data-journalism workflows in that they aim to lead lexicographers and users alike through a progression from a minimal automated summary to an interpretive explanation that blends human-written text with purpose-specific graphs. The process starts with lexicographers receiving automated summaries for each lemmata. These summaries take the form of visual analytics and touch upon four main areas, 1-2) lemma and sense distribution over subcorpora, 3) lexical context in which each word-sense tends to occur and 4) distribution of the semantic prosody of each sense over the subcorpora. The automated dashboard also highlights the cross-section of subcorpora where the most change is detected (e.g. periods or genre or philosophical tradition). Lexicographers create a text narrative that explicitly interprets the information provided in the automated summary and relates it to areas of interest for translators (the primary target audience of this dictionary), such as register, level of technical specialisation, connotation and comparison with near-synonyms. While drafting the narrative, lexicographers are asked to 1) refer to specific example sentences (examples are taken from the first iteration of the dictionary), 2) select the appropriate chart to illustrate each aspect of the data that is discussed in the narrative and 3) edit the charts to maximise their communicative power. This last point is probably the least practiced in historical lexicography, so a few examples are in order.[10]

The following examples are taken from an entry prototype that we are developing for the second iteration of the Buddhist Sanskrit dictionary. Since we are still working on this prototype, only a single proof-of-concept entry is currently available online in this new format, the lexical portrait of the word *vitarka*. The prototype is accessible from the dictionary entry on *vitarka*, but this is presently not yet integrated in the dictionary application, but hosted separately at mangalamresearch.shinyapps.io/LexicalPortrait_Vitarka/.

Most of the charts in this prototype are post-edited versions of the automated charts included in the automated summary. An automated graph showing the frequency of the lemma compared to its near synonyms, for example, has been edited by manually trimming the pool of near-synonyms shown to enhance the readability of the graph.

[10] This is not to say that no efforts have been made in the direction of data-visualisation post-editing within historical lexicography, but these efforts still seem very rare (e.g. Hoenen, 2018).

Other charts have been edited to facilitate interpretation. For example, the automated summary contained a barchart illustrating the normalised frequency of the lemma in each text type. It was clear from the chart that the lemma is dramatically more frequent in the genre *śāstra* (treatise) than in the other genres, but the fragmentation into several bars obfuscated the focal contrast between the frequency of *vitarka* in *śāstras* and in the rest of Buddhist literature, which is explicitly referred to in the narrative accompanying the graph. To make the comparison clearer, we added a second chart that displays the cumulative frequency of the headword in all other genres (figure 2). Finally, we edited a wordcloud to highlight the link between lexical context and semantics. The automated version of this wordcloud highlighted the words that surround *vitarka* according to their collocational strength. Some rather obscure words that happen to co-occur with *vitarka* with statistically significant frequency were prominently displayed. The resulting data-visualisation was not very informative, as the highlighted words scarcely contributed to the interpretation of the headword's semantics. To improve on this, we manually experimented with different parameters and eventually changed them to highlight words according to the number of texts in which they co-occur with the headword. This produced a more informative picture where the most prominent items clearly point to the two different senses of the headwords.



register

About two thirds of all attestations of vitarka in the Segmented Buddhist Sanskrit Corpus occur in śāstras (422 out of 637 attestations). This does not necessarily indicate that vitarka gravitates towards the philosophical or technical `register`. Attestations of the word vitarka in narrative and non-specialized contexts suggest that the use of this word is **`not restricted to the philosophical discourse`**, with the every-day character of some attestations even tilting towards a `general language feel`

e.g. If a man, moved by considerations of greed, had made a date with a handsome, attractive and good-looking woman, and if now that woman were held back by someone else and could not leave her house, what do you think, Subhuti, with what would that man's **preoccupations** be connected? Subhuti: With the woman, of course. He thinks about her coming, about the things they will do together, and about the joy, fun and delight he will have with her. The Lord: Will he have many such ideas in the course of a day? Subhuti: Many indeed, O Lord. [Conze, 210]

(tadyathāpi nāma subhūte kaścid eva puruṣo rāgacarito vitarkacaritaḥ | tasya puruṣasya rāgacaritasya vitarkacaritasya striyā abhirūpayā prāsādikayā darśanīyayā saha saṃketaḥ kṛto bhavet | sā khalu punaḥ strī paraparigṛhītā bhavet | na vaśayed ātmānam agārān niṣkramitum | tat kiṃ manyase subhūte kiṃ pratisaṃyuktās tasya puruṣasya vitarkāḥ pravarteran? subhūtir āha - strīpratisaṃyuktā eva bhagavaṃs tasya puruṣasya vitarkāḥ pravarteran - iyam āgacchati, iyam āgatā | tayā sārdhaṃ evaṃ kariṣyāmi, evaṃ ramiṣyāmi, evaṃ krīḍiṣyāmi, evaṃ pravicārayiṣyāmīti | bhagavān āha - tat kiṃ manyase subhūte divasasyātyayena tasya puruṣasya kiyanto vitarkotpadyeran? subhūtir āha - bahavo bhagavan divasasyātyayena tasya puruṣasya vitarkotpadyeran ... [***Aṣṭasāhasrikāprajñāpāramitā***: Vaidya, 170-171]).

Figure 2. A portion of the prototype lexical portrait (mangalamresearch.shinyapps.io/LexicalPortrait_Vitarka/, accessed on 8/4/2021).

The parameters used to generate each graph are detailed in the 'info' tab accompanying each graph. The text slotted in the 'info' sections is automatically generated via a template that describes the default parameters used to generate the graph and is edited whenever the parameters are manually changed. The text of the narrative, by contrast, is unlikely to be amenable to automation. While we may experiment with generating an automatic draft for the lexicographers to post-edit in a machine-translation fashion, it seems that the interpretive and original content of the narrative is better suited to the augmentation model of data-journalism, whereby only the visual analytics are automatically generated and the storytelling is left to the human author.

5. Acknowledgements

Part of the work presented in this paper has been funded by UKRI as part of the project *Lexicography in Motion: A History of the Tibetan* Verb (AH/P004644/1). The Sanskrit dictionary is funded by the Mangalam Research Center for Buddhist Languages.

6. References

A Visual Dictionary and Thesaurus of Buddhist Sanskrit. Accessed at: mangalamresearch.shinyapps.io/VisualDictionaryOfBuddhistSanskrit/ (30 March 2021)

A Visual Dictionary of Tibetan Verb Valency. Accessed at: mangalamresearch.shinyapps.io/VisualDictionaryOfTibetanVerbValency/ (30 March 2021)

Alhalaseh, R., Munezero,M., Leinonen, L. & Leppänen, L. (2018). Towards Data-Driven Generation of Visualizations for Automatically Generated News Articles. *Proceedings of the 22nd International Academic Mindtrek Conference.*

Baisa, V. et al. (2019). Automating Dictionary Production: a Tagalog-English-Korean Dictionary from Scratch. In I. Kosem, T. Zingano Kuhn, M. Correia, J. P. Ferreria, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek & C. Tiberius C. (eds.) *Smart Lexicography: eLex 2019*, pp. 805–818.

Bakakis, N. (2018). Big data visualization tools. *arXiv* 1801.08336.

Caswell, D. & Doerr, K. (2018). Automated Journalism 2.0: Event-Driven Narratives, from Simple Descriptions to Real Stories. *Journalism Practice*, 12(4), pp. 477–496.

Coddington M. (2018). Defining and Mapping Data Journalism and Computational Journalism: A Review of Typologies and Themes. In S. Eldridge II (ed.) *The Routledge Handbook of Developments in Digital Journalism Studies.* New York: Routledge.

Diakopoulos, N. (2018). Ethics in Data-Driven Storytelling. In N. Henry Riche et al. (eds.) *Data-Driven Storytelling.* London: CRC, pp. 233–247.

Diakopoulos, N. (2019). *Automating the News: How Algorithms Are Rewriting the Media.* Harvard University Press.

Frankenberg-Garcia, A., Rees, G. P. & Lew, R. (2020). Slipping through the Cracks in e-Lexicography. *International Journal of Lexicography*, doi: 10.1093/ijl/ecaa022.

Graefe, A. (2016). *Guide to Automated Journalism.* Columbia University, Tow Center for Digital Journalism.

Graefe, A., Haim, M. & Brosius, H. B. (2018). Perception of Automated Computer-Generated News: Credibility, Expertise and Readability. *Journalism*, 19(5), pp. 95–610.

Graphic Detail. Accessed at: https://www.economist.com/graphic-detail (28 May 2021)

Grefenstette, G. (1998). The Future of Linguistics and Lexicographers: will there be Lexicographers in the Year 3000? In T. Fontenelle et al. (eds.) *Proceedings of the Eighth Euralex Conference*, Liège.

Hill, N. (2010). *A Lexicon of Tibetan Verb Stems as Reported by the Grammatical Tradition.* Munich: Bayerische Akademie der Wissenschaften.

Hoenen, A. (2018). Annotated Timelines and Stacked Area Plots for Visualization in Lexicography. *Elexis Workshop*, Galway 2018.

Jakubíček, M. (2017). The advent of post-editing lexicography. *Kernerman Dictionary News*, July 2017.

Kennedy, H. et al. (2019). Data Visualisations: Newsroom Trends and Everyday Engagements. In J. Gray & L. Bounegru (eds.) *The Data Journalism Handbook 2: Towards a Critical Data Practice.* Amsterdam: Amsterdam University Press.

Kilgarriff, A., Husak, M., McAdam, K., Rundell, M. & Rychly, P. (2008). GDEX: Automatically Finding Good Dictionary Examples in a Corpus. In E. Bernal & J. DeCesaris (eds.) *Proceedings of the Thirteenth EURALEX International Congress.* Barcelona: Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra, pp. 425–432.

Leppänen, L., Munezero, M., Granroth-Wilding, M., Toivonen, H. (2017). Data-Driven News Generation for Automated Journalism. In *Proceedings of the 10th International Conference on Natural Language Generation*, pp. 188–197.

Lugli, L. (2018). Drifting in timeless polysemy: Problems of chronology in Sanskrit lexicography. *Dictionaries: Journal of the Dictionary Society of North America.* Vol. 39 (1), pp. 105–129.

Lugli, L. (2019). Smart lexicography for under-resourced languages: lessons learned from Sanskrit and Tibetan. In I. Kosem, T. Zingano Kuhn, M. Correia, J. P. Ferreria, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek & C. Tiberius C. (eds.) *Smart Lexicography: eLex 2019*, pp. 198–212.

Marconi, F. (2020). *Newsmakers: Artificial Intelligence and the Future of Journalism.* New York: Columbia University Press.

Měchura, M. B. (2017). Introducing Lexonomy: an open-source dictionary writing and publishing system. In I. Kosem et al. (eds.) Electronic Lexicography in the 21st Century: Lexicography from Scratch. *Proceedings of the eLex 2017 conference*, Leiden.

Nitzke, J., Hansen-Schirra, S., Canfora, C. (2019). Risk management and post-editing competence. *The Journal of Specialised Translation*, 31, pp. 240–259.

Rundell, M. (2002). Good Old-fashioned Lexicography: Human Judgement and the Limits of Automation. In M.-H. Corréard (ed.) *Lexicography and Natural Language Processing. A Festschrift in Honour of B. T. S. Atkins.* Euralex.

Stopler, Ch.D., Lee, B., Henry Riche, N. & Statsko, J. (2018). Data-Driven Storytelling Techniques: Analysis of a Curated Collection of Visual Stories. In N. Henry Riche et al. (eds.), *Data-Driven Storytelling.* London: CRC, pp. 85–105.

Stray, J. (2019). Making Artificial Intelligence work for Investigative Journalism. *Digital Journalism*, 7(8), pp. 1076–1097.

Thudt A, Walny J., Gschwandtner, Th., Dykes, J. & Statsko, J. (2018). Exploration and Explanation in Data-Driven Storytelling. In Nathalie Henry Riche et al. (eds.), *Data-Driven Storytelling.* London: CRC, pp. 59–83.

Wiedmann, G., Yimam, S. M. & Biemann, Ch. (2018). A Multilingual Information Extraction Pipeline for Investigative Journalism. *arXiv* 1809.0022v.1

Young, M. L. & Hermida, A. (2015). From Mr. and Mrs. Outlier to Central Tendencies. *Digital Journalism* 3(3), pp. 381–397.

The Latvian WordNet and Word Sense Disambiguation: Challenges and Findings

Ilze Lokmane[1], Laura Rituma[2], Madara Stāde[3], Agute Klints[4]

[1]University of Latvia, Department of Latvian and Baltic Studies, Visvalža 4a, Riga, LV-1050
[2]Institute of Mathematics and Computer Science, University of Latvia, Raina bulvaris 29, Riga, LV-1050
[3]University of Latvia, Department of Latvian and Baltic Studies, Visvalža 4a, Riga, LV-1050
[4]University of Latvia, Department of Latvian and Baltic Studies, Visvalža 4a, Riga, LV-1050
E-mail: ilze.lokmane@lu.lv, laura.rituma@lumii.lv, stade.madara@gmail.com, agute.klints@gmail.com

Abstract

The article addresses the issues of word sense disambiguation within the process of developing an electronic lexical semantic resource, the Latvian WordNet. Apart from word senses, the resource also contains semantic paradigmatic relations between these senses, and therefore sense granularity must align with the need for creating synonymous, hyponymic, meronymic and antonymic links between Latvian words, as well as external links with the Princeton WordNet.

The development of the Latvian WordNet started in 2020 and it is based on two sources: a summarising electronic dictionary Tēzaurs.lv and available corpora. Because the word senses listed in Tēzaurs.lv are not directly usable for the needs of computer linguistics due to a number of reasons, the developers of the Latvian WordNet checked and revised the senses manually based on corpus data. Thus, the work on distinguishing word senses serves two purposes: 1) creating a Latvian WordNet, and 2) improving the structure of existing entries in the dictionary Tēzaurs.lv.

The article primarily focuses on the elaboration of common criteria for distinguishing word senses. The analysis concentrates on verbs as these are the most complex part of speech from the point of view of making sense distinctions. The authors conclude that the process is based on a set of criteria that form a certain hierarchy depending on the semantic group of verbs, namely, syntactic distribution, semantic distribution, as well as the interrelation between the two, and semantic decomposition of senses. Particular attention is paid to the interrelations of superordinate senses and subsenses, from which it is possible to conclude that an absolutely uniform and consistent subsense distinction is not likely to be possible, and, therefore, in cases of uncertainty, decisions are made in favour of what is needed to develop the Latvian WordNet.

Keywords: word sense disambiguation; sense distinction; electronic lexical semantic resource; syntactic and semantic distribution; lexical decomposition

1. Introduction

The article focuses on major challenges and some preliminary findings in the field of word sense disambiguation with respect to the development of a Latvian WordNet[1], i.e. structured, machine-readable wide coverage inventory of word senses and semantic

[1] Project "Latvian WordNet and word sense disambiguation" No. LZP-2019/1- 0464

relations, such as synonymy, hyponymy, meronymy, antonymy and similar between these senses in Latvian. By word sense disambiguation we understand the task of determining which sense of a word is being used in a particular context (Jurafsky & Martin, 2020: 1). Therefore, finding criteria for deciding when different uses of a word should be represented as discrete senses is crucial. The aim of the project is to determine the senses of 5,000 commonly used Latvian words and to establish semantic links between them, but at the present stage approximately 150 words have been processed, most of which have multiple senses. The work is carried out using a specifically developed tool, which is described in more detail in Section 3.

The development of the Latvian WordNet began in 2020 and it is based on two sources: digital versions of pre-existing monolingual general and specialist dictionaries and available corpora.

An important Latvian lexical resource maintained by the Institute of Mathematics and Computer Science of the University of Latvia (IMCS UL) is Tēzaurs.lv (Spektors et al., 2016), which is a large (~ 378,000 entries in the last release in March, 2021) digital compilation of legacy dictionaries[2]. Our experience indicates that the word senses listed in Tēzaurs.lv are not directly usable for the purpose of computational linguistics due to issues with sense granularity and boundaries, as well as the outdated nature of many of the senses. Therefore, the word senses available in this resource are checked and revised using a corpus-based approach to determine if the senses are still currently relevant, whether any new senses have appeared or whether specific uses of a word demonstrate the validity of word sense distinction (based on a similar revising of sense distinctions and definitions in Estonian WordNet see Kerner, Orav & Parm, 2010).

The main source data for the lexical analysis is The Balanced Corpus of Modern Latvian (10 million tokens), which is also maintained by IMCS UL but has become the de-facto reference corpus for Latvian linguistic research (Levāne-Petrova, 2019). However, not all word senses can be found in the corpus, therefore other corpora are employed for identifying and illustrating less common or colloquial word senses: Corpus of the Saeima (the Parliament of Latvia) (Darģis et al., 2018), Latvian Blog Corpus 2015 (Laizāns, 2015), Latvian Web Corpus 2007 (Dzerins & Dzonsons, 2007) and CommonCrawl of Latvian 2020.

A corpus-based approach results in a better set of word senses than the commonly used alternative of directly mapping Princeton WordNet concepts to translations in the target language, which implicitly transfers the English linguistic patterns of many concepts that are often not a good match for the target language. While a corpus based approach requires more effort, we have chosen this to ensure the linguistic validity of the resulting resource.

[2] The total number of Tēzaurs.lv sources is 329.

In addition, such an approach would meet the needs of both WordNet development and the improvement of word sense inventory of Tēzaurs.lv. Therefore, the development of the Latvian WordNet is primarily a linguistic (lexicographic) challenge, as the separation of senses is performed by manually aligning corpus evidence with lexicographic data.

2. Problematic Issues of Distinguishing Word Senses

Before describing the process of WordNet creation and the criteria for distinguishing senses, we would like to point out the main issues that arose in this process.

First of all, the process of word sense distinction is a complicated task in itself. We tend to agree with cognitive linguists that the question of how many senses the word has may not have a clear-cut answer. There is always a question whether two different uses of a word exemplify two separate senses, or contextual modulations of the same sense (Taylor, 2009: 144). Some linguists even claim that a word has just a single abstract meaning which is instantiated in a range of sometimes very different usage situations (Taylor, 2009: 147–148).

Therefore, the word sense system is not a stationary and entirely fixed one, and semantic derivation is an active and ongoing process. It could be said that the range of word meanings is continuous and diffuse, and the fixation of individual meanings is linked to a certain degree of schematisation. The concept of polysemy, on the other hand, is based on the idea of discreteness of lexical meanings and, as a consequence, researchers and lexicographers, in particular, try to discern strict boundaries around what is in fact an unclear grey area.

Therefore, lexicographic resources display a considerable variation in the number of word senses. Even though overall coverage of the senses is the same, dictionaries may have differently clustered senses and subsenses, with the same semantic space merged and split in various ways. For example, metaphoric and metonymic meaning extensions are not always set apart as distinct meanings. In addition, it is possible to use certain words creatively in new contexts, and it is not easy to determine whether it illustrates an already existing meaning or is considered an individual metaphorical or metonymic use and, hence, does not require including in the dictionary.

Thus, the question of what marks the point when a meaning should be regarded as a distinct sense or subsense and included in a dictionary is probably one of the most difficult issues of lexicographic work. As Allen (1999: 61) states, lexicographers can be divided into two broad categories - 'lumpers' and 'splitters': "The 'lumpers' like to lump meanings together and leave the extraction of the nuance of meaning that corresponds to a particular context to the user, whereas the 'splitters' prefer to enumerate differences of meaning in more detail; the distinction corresponds to that between summarizing and analyzing." Furthermore, Jackson (2002: 89) admits, that "most dictionaries tend to be of the 'splitting' type, though different dictionaries do not necessarily agree on where to make the splits between senses." This is also fully applicable to existing dictionaries of the Latvian language.

In our opinion and given the point of view of the user of the dictionary, it is better to list fewer senses, thus making the entry more transparent and reader-friendly. A lexicographer is able to discern between slight nuances of meaning, whereas an everyday user outside the realm of linguistics might find it difficult to grasp the difference between word senses, especially if they are accompanied by long and complex definitions. Initially, we planned to generalise the division of word senses in the dictionary and make it less detailed, but over the course of the work it became clear that a general division is not always entirely useful for WordNet purposes. In addition, the legacy of Tēzaurs.lv had to be treated with great care in order not to erase the dialectal, terminological and other word senses included there, even if modern language corpora do not contain examples of their use. Therefore, the corpus-based approach applies only to a certain part of the word senses.

Therefore, and **secondly**, in the revision of word senses a compromise was necessary between two extremes: an excessively generalised or fine-grained division of word senses. The need for a more detailed division arises in cases when synonymous, hyponymic and other semantic relations between senses are formed, as well as during the formation of external links with the Princeton WordNet. Our definitive solution for cases of ambiguity aligns with the needs of WordNet: word senses are identified in more detail when a sense and subsense form individual synonymic or other semantic links to a sense or subsense of another word.

Thirdly, despite the substantial semantic differences between various parts of speech and separate semantic groups within a part of speech, the selected approach to word sense distinction should be as consistent as possible. The defined criteria and their application are described in more detail in Section 4.

Fourthly, we encountered the problem of defining and dividing superordinate senses and subsenses. In such cases, it was noticeably more difficult to identify a consistent solution that would be equally applicable to words in all semantic groups, therefore defining subsenses is the most subjective step in the WordNet creation process and requires a more detailed explanation.

Latvian lexicographers have so far avoided studying the theoretical problems of word subsense, so the division found in the Latvian language dictionaries is inconsistent and intuitive. Semanticists, on the other hand, do not examine the problem of separating superordinate senses from subsenses and regard it as a topic pertaining more to lexicography. The basis for identifying a subsense is usually more detailed semantic differences attributable to the same sense, as well as grammatical and functional features of the word (LLVV, 1972: 11). They are as follows:

1) A subsense can differ from a superordinate sense by a certain semantic component. For example, the verb *uztvert* (*to catch*) has a sense 'to grasp' with a subsense 'to grasp and deflect'[3], therefore the semantic component 'to deflect' is added.

2) A subsense can differ from a superordinate sense by semantic distribution, namely, the semantic roles of the participants of the situation or the semantic groups they pertain to. For example, the verb *rakt* (*to dig*), has a sense 'to impale and move soil or dirt with a shovel', which indicates a person as the agent, whereas the subsense reveals other possible agents, such as equipment or animals. The semantics of the instrument is also different: humans dig with a shovel, while animals dig by using their muzzle or limbs.

3) A subsense can differ from a superordinate sense by syntactic distribution, for example, the superordinate sense can have transitive and subsense intransitive properties or vice versa. The creators of the Latvian WordNet believe that the use of a transitive verb without a direct object should not be considered as a subsense if the object can be understood from context or situation or if it is so general that it is not necessary to be named. For example, the word *dzert* (*to drink*) has a transitive superordinate sense 'to imbibe and swallow (a liquid)' and an intransitive subsense, e.g. *Dzert gribi?* 'Do you want a drink?'[4] Only if the sense of a verb that is being used in its intransitive use is joined by a new semantic component is there a basis for defining a subsense, as is demonstrated by the verb *lasīt* (*to read*), which has the transitive superordinate sense 'to take in a written text' and an intransitive subsense, which has the added semantic element of 'being able to'.

4) Cases of diathesis demonstrate the interrelation of semantic and syntactic distribution. Here, a situation is illustrated by the same verb from different points of view. The participants in the situation remain the same, but their syntactic status is changed. For example, the act of digging involves both the agent (*cilvēks rok* 'a person is digging') and the instrument (*rakt ar lāpstu* 'to dig with a shovel'), as well as patients of different kinds: that, which is moved (*rakt zemi* 'to dig soil') and that, which is created (*rakt bedri* 'to dig a hole'). Syntactically, only one of them can be realised at a time, but the situation as a whole does not change. The instrument can also be used as a subject (*lāpsta labi rok* 'this shovel digs well', *ekskavators rok* 'the excavator is digging'). Various cases of diathesis have been extensively examined in semantic studies (Paducheva, 2004: 51–79), as well as divided into types, which differ slightly in each respective language. In other semantic theories such extensions of a certain verb have been described as metonymic (Pustejovsky, 1998: 31–33), whereas in cognitive semantics this process is called profiling (Saeed, 2000: 328–330).

[3] All sense definitions referred to in this article are taken from Tēzaurs.lv.

[4] All examples of word usage are taken from Latvian language corpora.

Therefore, it can be concluded that to a certain extent subsenses can illustrate the continuity of lexical semantics of words and the gradual transition from one sense to another. It can be seen further in the paper that subsenses can be distinguished on the same principles as superordinate senses (see Section 4).

Fifthly, an optimal definition (sometimes called a gloss) of sense is necessary, as the definition method of a word sense can affect the entire system of word senses. For instance, a more general definition may lead to two or more senses being combined whereas specific definitions allow the contrary, i.e. splitting a sense into separate senses or subsenses.

Different forms of definition are appropriate to different types of words (Jackson, 2002: 94). Practical lexicography offers three main methods of defining sense: definition by synonym, definition by periphrasis and a scientific definition. Each of the listed methods has its advantages and disadvantages, which we will examine in more detail.

In the process of developing the Latvian WordNet, definition by synonym has been one of the most useful methods, as it facilitates finding synonym links between senses of various words, e.g. *domāt* (*to think*), the third sense of which is 'to care for'. However, taking into account the revelation of the lexical semantics of a word, this approach to definition also has notable disadvantages. Firstly, there is a risk of circularity (Jackson 2002: 94), secondly, by using a synonym, the meaning is essentially left unexplained, and thirdly, not all senses have synonyms. Moreover, the synonym used in the definition could have multiple senses as well.

Definition by periphrasis, unlike definition by synonymy, attempts to determine the semantic components that form the sense, e.g. *skriet* (*to run*) – 'to move steadily by springing steps, so that both feet occasionally leave the ground at the same time at each step'. For this method it is important to find the essential features, i.e. those that distinguish the realia from others, and not to include irrelevant information. The number of specific features should be sufficient (Zuicena, 2010: 370) and the words used in the definition should be simpler than the word that is being defined (Jackson, 2002: 93). Therefore, this method is similar to lexical decomposition. However, this approach has certain limitations: the first is that the proportion of words which lend themselves to this sort of analysis is relatively restricted; the second is that the analysis leaves much semantic knowledge unaccounted for (Cruse, 2004: 242). In practical lexicography the periphrastic definition method is often used intuitively, thus it is not always sufficiently accurate and is used mostly in cases when there are no synonyms.

A scientific approach or at least elements of it are sometimes used to define sense, such as the noun *bullis* (*a bull*) – 'a male representative of hollow-horned or antlered ruminants'. There are reasonable objections to this type of explanation, namely, that the definition of a scientific concept is not part of ordinary linguistic competence (Goddard, 1998: 28). However, it should be kept in mind that language users may have certain (albeit rudimentary) scientific knowledge of specific realia. Although explanatory dictionaries are not encyclopediae, there is no strict boundary between the meaning of a word and the knowledge of certain realia.

It should also be noted that the definition of a word sense often requires information on typical distribution. It is mostly used in verb definitions, e.g. *čivināt* (*to twitter*) – 'to make short, rhythmic chirping noises (about birds)'. When defining verbs of certain semantic groups, it is even impossible to do without this approach. For example, specific senses of sound verbs cannot be fully revealed either by synonymy or periphrasis. Definitions can also be supplemented by elements typical of the referent, introduced by the adverb *parasti* (*typically, usually*) (Jackson, 2003: 95), e.g. *glāze* (*a glass*) – 'a small (usually cylindrical) drinking container without a handle made of glass or other material'.

It should also be taken into account that there is no universal principle or method for defining the senses of words of all semantic groups and parts of speech. For example, distribution is more important for defining the semantics of verbs than it is for nouns. Polyvalent verbs are more effectively defined by describing their distribution (e.g. the meaning of the verb *īrēt* (*to rent*) can be revealed by listing *who, what, to whom, for how long and for what payment*), whereas in case of verbs with zero valency, e.g. *snigt* (*to snow*) the distribution analysis yields little information and other methods should be employed.

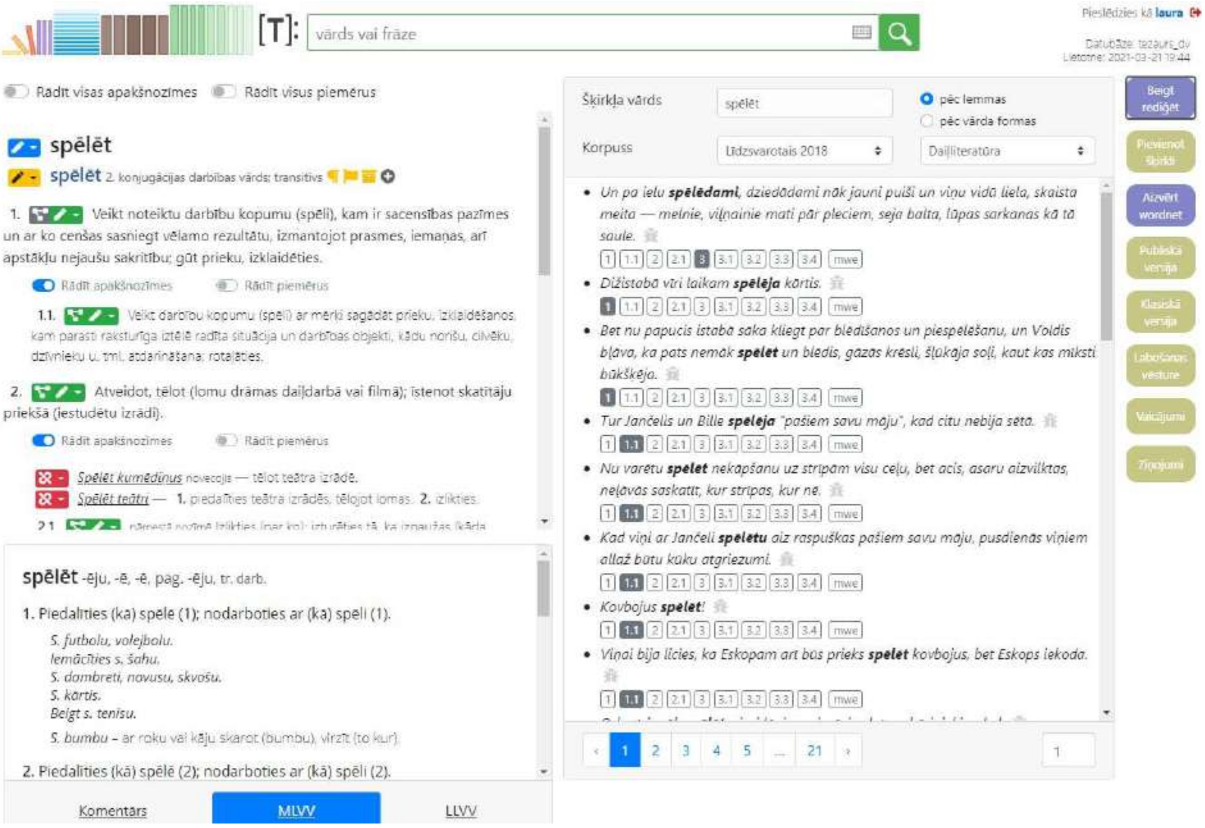
And lastly, certain problems are also caused by the separation of distinct word senses and multi-word expressions. However, this topic deserves separate research, therefore it is not examined in this article.

3. Lexicographic Infrastructure and Tools

The software infrastructure for this work is based on the existing tools for maintaining the Tēzaurs.lv lexicographic platform which was already used for maintenance of structured data for entries, glosses, word senses and usage examples. As we wanted to base the Latvian WordNet on the existing Tēzaurs.lv word sense data where possible, we chose to extend the Tēzaurs.lv editor tools with the required functionality instead of managing the WordNet data in a separate existing tool (for example, WordNet Loom and DebVisDic). This choice adds certain complexity due to need to balance the requirements (for example, for the word sense granularity) of the WordNet project with the expectations of generic dictionary users of Tēzaurs.lv, as they would see the same word senses, but it also has the potential to make the resulting resource more accessible to a wider general audience, which would be less likely to use separate tools for browsing WordNet data. The choice of integration also means that all work on improving word sense definitions and usage examples improves the general dictionary data.

The technical platform for the Tēzaurs.lv lexicographical database is built as JavaScript (Vue.js) web interface to a custom PostgreSQL database for the lexical data.

In order to manage WordNet data, we extended the Tēzaurs.lv database and tools with support for managing synsets and semantic links (including external links to the Princeton WordNet), as well as streamlining functionality for mapping corpus examples to specific word senses and subsenses (see Figure 1). The data is developed in an internal environment with quarterly releases of new data versions to the general public on the Tēzaurs.lv online platform. At project milestones, we plan to release the WordNet data along with the Tēzaurs.lv lexical database in machine-readable structured format.



spēlēt

spēlēt 2. konjugācijas darbības vārds; transitīvs

1. Veikt noteiktu darbību kopumu (spēli), kam ir sacensības pazīmes un ar ko cenšas sasniegt vēlamo rezultātu, izmantojot prasmes, iemaņas, arī apstākļu nejaušu sakritību; gūt prieku, izklaidēties.

1.1. Veikt darbību kopumu (spēli) ar mērķi sagādāt prieku, izklaidēšanos, kam parasti raksturīga iztēlē radīta situācija un darbības objekti, kādu norišu, cilvēku, dzīvnieku u. tml. atdarināšana; rotaļāties.

2. Atveidot, tēlot (lomu drāmas daiļdarbā vai filmā); īstenot skatītāju priekšā (iestudētu izrādi).

Spēlēt kumēdiņus novecojis — tēlot teātra izrādē.

Spēlēt teātri — 1. piedalīties teātra izrādēs, tēlojot lomas. 2. izlikties.

2.1. pārnestā nozīmē Izlikties (par ko); izturēties tā, ka izpaužas (kāda

spēlēt -ēju, -ē, -ē, pag. -ēju, tr. darb.

1. Piedalīties (kā) spēlē (1); nodarboties ar (kā) spēli (1).

S. futbolu, volejbolu.
Iemācīties s. šahu.
S. dambreti, novusu, skvošu.
S. kārtis.
Beigt s. tenisu.

S. bumbu – ar roku vai kāju skarot (bumbu), virzīt (to kur).

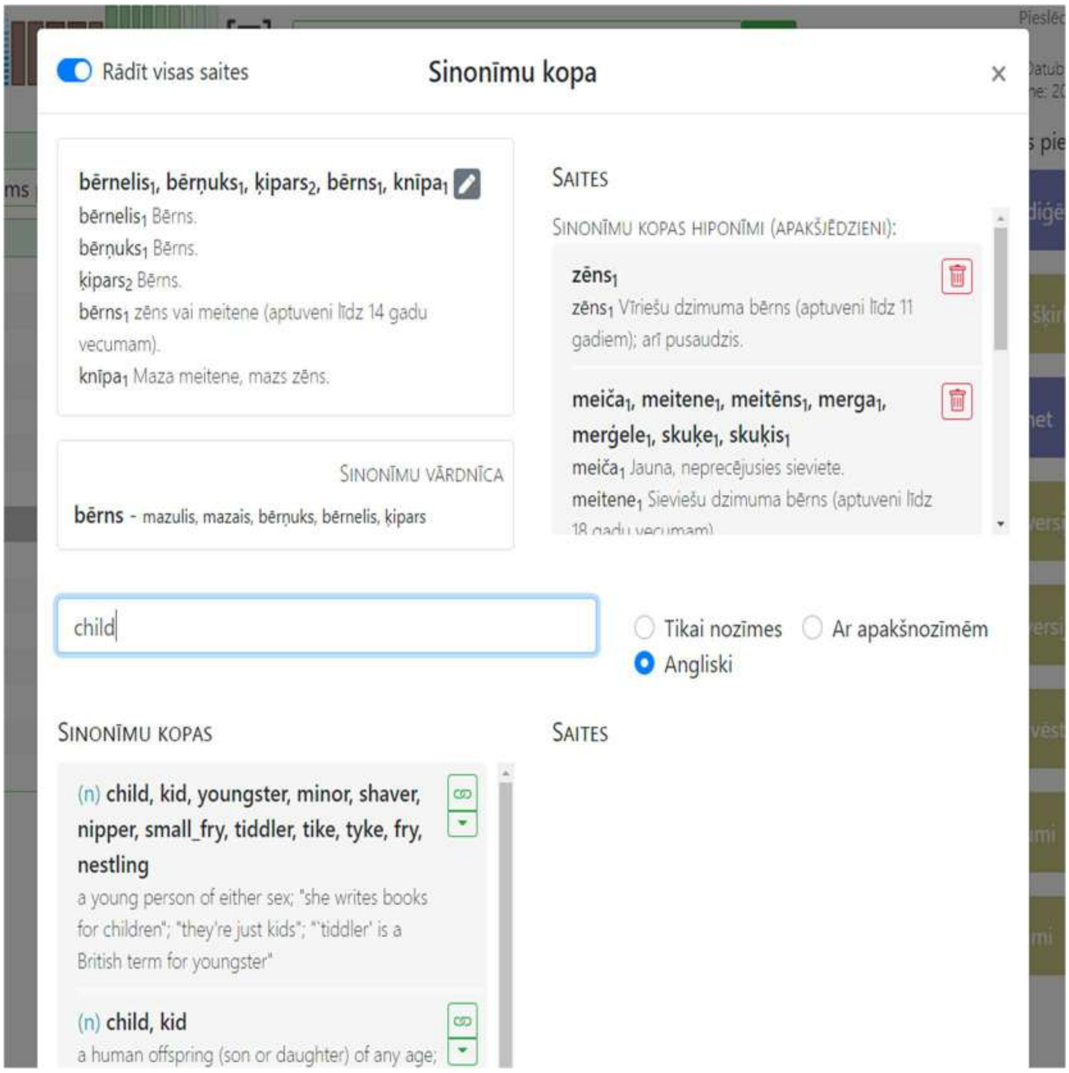
2. Piedalīties (kā) spēlē (2); nodarboties ar (kā) spēli (2).

Šķirkļa vārds: spēlēt
● pēc lemmas
○ pēc vārda formas
Korpuss: Līdzsvarotais 2018 | Daiļliteratūra

- *Un pa ielu **spēlēdami**, dziedādami nāk jauni puiši un viņu vidū liela, skaista meita — melnie, viļņainie mati pār pleciem, seja balta, lūpas sarkanas kā tā saule.*
- *Dižistabā vīri laikam **spēlēja** kārtis.*
- *Bet nu papucis istabā sāka kliegt par blēdīšanos un piespēlēšanu, un Voldis bļāva, ka pats nemāk **spēlēt** un blēdis, gāzās krēsli, šļūkāja soļi, kaut kas mīksti būkšķēja.*
- *Tur Jančelis un Bille **spēlēja** "pašiem savu māju", kad citu nebija sētā.*
- *Nu varētu **spēlēt** nekāpšanu uz strīpām visu ceļu, bet acis, asaru aizvilktas, neļāvās saskatīt, kur strīpas, kur ne.*
- *Kad viņi ar Jančeli **spēlētu** aiz raspuškas pašiem savu māju, pusdienās viņiem allaž būtu kūku atgriezumi.*
- *Kovbojus **spēlēt**!*
- *Viņai bija licies, ka Eskopam arī būs prieks **spēlēt** kovbojus, bet Eskops iekoda.*

Figure 1. The sense editing and example selection function view in the tool. The left side shows senses and subsenses listed in Tēzaurs.lv and two other dictionaries for comparing differences. The right side shows all the examples with the corresponding lemma in the selected corpus; each example can be marked with the matching word sense number.

The workflow consists of the following steps: 1) editing entries by modifying word senses, their order and definitions and adding new entries and senses, 2) browsing through various examples from different corpora and adding them to word senses or multi-word expressions in an entry (10–30 examples for each sense), 3) creating synsets between separate meanings of various words, 4) creating various types of links between synsets, 5) linking Latvian meanings/synsets with those of the Princeton WordNet (see Figure 2).



Sinonīmu kopa

bērnelis$_1$, bērņuks$_1$, ķipars$_2$, bērns$_1$, knīpa$_1$

bērnelis$_1$ Bērns.
bērņuks$_1$ Bērns.
ķipars$_2$ Bērns.
bērns$_1$ zēns vai meitene (aptuveni līdz 14 gadu vecumam).
knīpa$_1$ Maza meitene, mazs zēns.

SINONĪMU VĀRDNĪCA

bērns - mazulis, mazais, bērņuks, bērnelis, ķipars

SAITES

SINONĪMU KOPAS HIPONĪMI (APAKŠJĒDZIENI):

zēns$_1$
zēns$_1$ Vīriešu dzimuma bērns (aptuveni līdz 11 gadiem); arī pusaudzis.

meiča$_1$, meitene$_1$, meitēns$_1$, merga$_1$, merģele$_1$, skuķe$_1$, skuķis$_1$
meiča$_1$ Jauna, neprecējusies sieviete.
meitene$_1$ Sieviešu dzimuma bērns (aptuveni līdz 18 gadu vecumam)

child

○ Tikai nozīmes ○ Ar apakšnozīmēm
● Angliski

SINONĪMU KOPAS

(n) child, kid, youngster, minor, shaver, nipper, small_fry, tiddler, tike, tyke, fry, nestling
a young person of either sex; "she writes books for children"; "they're just kids"; "'tiddler' is a British term for youngster"

(n) child, kid
a human offspring (son or daughter) of any age;

SAITES

Figure 2. The window for creating synsets and semantic links; the process of reviewing word senses for “child”. The synset with all the synonyms included is shown at the upper part of the window. Below the synset there are synonym suggestions from the dictionary of synonyms. The search window is in the middle, where the developer can search for word senses on Tēzaurs.lv or corresponding English synsets on the Princeton WordNet. All links added to the synset are displayed on the right side.

From the WordNet perspective the main motivation of selecting a substantial quantity of examples from corpora is to use them as training data for supervised machine learning in developing a Word Sense Disambiguation system. As the usage examples are searched in corpus, the selected wordform/inflection is annotated with the manually chosen word sense identifier, forming a sense-annotated corpus. The review of examples also helps to ensure that the chosen word sense split is based on actual usage, and a manually chosen subset of most representative examples are also used in the public Tēzaurs.lv version to aid dictionary readers by illustrating the differences between specific word senses, in contrast to the earlier approach of Tēzaurs.lv which used automatically selected corpus examples for the whole entry, without explicit linking to word senses.

4. Word Sense and Subsense Distinction Criteria and their Applications

As mentioned before, the criteria of distinguishing word senses can differ depending on various parts of speech and even semantic groups (e.g. sound and directional verbs). The approach chosen in the development of the Latvian WordNet is based on word sense separation by a set of features. As verbs may be considered the most challenging part of speech with respect to deciding how many discrete senses a word has, we will examine this part of speech by concentrating on criteria which have proved to be useful.

Latvian is a highly inflected language, and thus the **syntactic distribution** of the verbs, namely, valency frame (arguments and their coding), has to be taken into account first of all (on the implementation of valency models of verbs in Polish WordNet see Dziob & Piasecki, 2018). The syntactic distribution shows what syntactic constructions a word is a part of, e.g. whether it has a direct or indirect object, certain adverbial modifiers, etc. Syntactic distribution can be particularly important when separating the senses of highly desemanticised and grammaticalised verbs. For example, the verb *būt* (*to be*) has a meaning ‘to be situated’, which becomes clear in a construction involving adverbials of place, e.g. *Visapkārt mājai ir priedes* ‘There are pine trees all around the house’, whereas the meaning ‘to belong’ can be understood in a construction containing the dative of possession: *Tev būs tieši tāda māja* ‘A house just like this will someday belong to you’.

The role of syntactic distribution in word sense distinction can also be illustrated by the verb of cognition *domāt* (*to think*). For example, the distribution of the sense ‘to consider’ is typically associated with an object clause introduced by conjunction *ka* (*that*) (*Domāju, ka tas nav godīgi pret auto izmantotājiem* ‘I think that it isn't fair to car users’) or deicitc adverbs *tā* (*thus, this way*) and *tāpat* (*in the same way, similarly*) (*Tā jau es domāju* ‘That's what I thought’), whereas the sense ‘to envisage, to get ready’ is demonstrated when combined with infinitive: *Ko tu domā darīt ar tiem?* ‘What are you thinking of doing with them?’.

Although syntactic distribution could be considered a fairly objective criterion in distinguishing word senses, it should be noted that sometimes two different senses can be used in the same syntactic construction. For example, the verb *domāt* (*to think*) in combination with a prepositional phrase can represent both the basic sense of ‘to think’ (*Es nezinu, par ko domāja viņš* ‘I don't know what he was thinking about’), as well as the secondary sense of ‘to care for’ (*Katrs īpašnieks sāktu domāt tikai par savu peļņu* ‘Each owner would start to think only of their own profit’). The latter sense can be identified based on the semantics of the object – the desirable things that are obtained through effort (e.g., *profit*). Therefore, it is not surprising that in some instances of word use there is ambiguity between these two senses, e.g. *Par to viņiem nav jādomā* ‘They don't have to think about it’.

Interestingly, in Latvian the verb *domāt* (*to think*) has two senses that mostly materialise in one grammatical form, namely, the past passive participle. The first one, meaning 'to be meant for a certain purpose' is used in combination with adverbials of purpose (*Bibliotēka domāta ne tikai lasīšanai, bet arī sarunām* 'The library is meant not only for reading but also for having talks'), whereas the second sense 'to understand by' is used with a prepositional phrase (*Ar meitām un dēliem ir domāti vecāku miesīgie pēcnācēji* 'One's direct descendants are understood by the terms 'daughters and sons'').

Secondly, **semantic distribution** including the **semantic roles** and **semantic features** of the arguments has proved to be useful. The semantic distribution of verbs includes the semantic roles of the participant (e.g. agent, patient, experiencer, beneficiary, addressee, instrument) and general or more specific semantic features (e.g. *animate / inanimate, abstract / concrete, countable / uncountable*).

The main problem associated with this method is that it is not clearly defined which semantic roles or characteristics are sufficiently important to be taken into account in the process of word sense distinction, e.g. whether the semantic opposition *human / other living beings* always enables one to fully differentiate between senses or not. Traditionally, in Latvian lexicography the verbs of motion, like *iet* (*to go*), *skriet* (*to run*) and so on, have different senses based on whether the action is performed by a human or animal, however, the developers of WordNet have chosen to overlook this in favour of a view that the nature of direction is not greatly changed by this. In this case, the animacy / inanimacy of the subject is a much more important characteristic. For example, in the basic sense of the verb *skriet* (*to run*) the subject is animate, whereas in derived senses it is an inanimate object (*Pa lāstekām uz leju skrien ūdens pilītes* 'Water droplets are running down the icicles'), physical phenomenon (*Uguns skrien uz priekšu* 'Fire is running forward') or phenomenon related to the subjective perception of humans (*Laiks skrēja nemanot* 'The time ran by unnoticed'; *Domas skrēja ātri* 'Thoughts ran through (one's) head'). In this case, the process of word sense distinction is based on the semantic groups of subjects, which can be viewed as a justified approach, given that significant features of the action directly depend on the subject: physical movement through space with or without legs, or movement through time or mental space. In contrast, the sense distinction process for the verb *mainīties* (*to change*) is not based on the animateness of the subject, even though it can relate to both animate subjects (*Nemaz neesi pa šiem gadiem mainījies* 'You haven't changed a bit over these years'), as well as inanimate ones (*Tomēr beidzamjā laikā situācija ir mainījusies* 'However, in recent times the situation has changed'). In our view, the process of change is a very general one and is not affected by the animateness of the subject.

A more interesting situation is presented by transitive verbs, where the semantic features and semantic roles of not only the subject but also the object can be crucial. Besides a direct object in the accusative, the verb *dot* (*to give*) takes an indirect object in the dative as well. It is also important to note that the direct object can have a wide spectrum of meaning, from a real object to abstract states, conditions etc. The position of the subject can be occupied not only by people or a group of people but also, for example, by circumstances. That is, everything that can serve as the basis for someone receiving something. So, the act of giving is interpreted very broadly as a causal relationship. Due to the previously examined semantic features, the verb *dot* (*to give*) is an often used one and has a wide distribution. This is also one of the verbs which tend to grammaticalise in many languages (Heine & Kuteva, 2002: 149–155), meaning that the semantics of the verb itself often play a fairly insignificant role in the semantics of phrases.

Word sense distinction for the verb *dot* (*to give*) is mainly based on the semantics of the object: it can be an inanimate object (*Nu tad dod to grozu un desmit santīmus šurp* 'Then give me the basket and 10 santims'), a state or a circumstance (*Nolēmām dot iespēju jaunam censonim* 'We decided to give the new contestant a chance'), information (*Norādes dot jau es varu* 'At least I can give directions'). At the same time, the structure of senses of this verb effectively demonstrates the interaction of grammatical and semantic criteria, for example, with the word sense 'to procure, to provide (conditions)', which has two subsenses. The first one, 'to have by birth', is usually realised through the passive participle in the past tense (*Viņam no dabas ir daudz dots* 'He was already given much from birth'), whereas the second subsense 'to let' is demonstrated through a syntactic construction with the infinitive (*Dodiet man arī pamēģināt!* 'Let me try!').

Thirdly, the differences in syntactic and / or semantic distribution are often combined with differences in **semantic components.** According to lexical decomposition theory, a word's sense may be broken down into smaller semantic components or features. As Cruse (2004: 235) states, "it is probably true to say that virtually every attempt to explicate a rich word meaning ends up by giving some sort of breakdown into simpler semantic components". In some cases, the semantic components that the meaning is composed of are the only criterion that delimits senses. For example, the verb *dot* (*to give*) has the sense of 'to allow to use (something) or take into possession', the semantic elements of which differ from the basic sense: instead of the physical act of giving, it describes the act of giving permission, even though the semantic type of the object is the same (*Ķeizars došot zemi* 'They say the Emperor will give land'). Semantic components influence, for example, the metaphorical subsense 'to pretend' of the verb *spēlēt* (*to play*): *Viņš spēlē gudrinieku* 'He's playing the smart guy'.

The method of semantic decomposition is more relevant in the analysis of monovalent or zero-valent verbs. However, it is also associated with the following problems.

1) It is problematic to define the semantic components, as they can have various degrees of generalisation. Semantic components can be identified best by comparing, for example, the senses of two words or the use of one word in different contexts.

2) The naming of semantic components can also be quite problematic, as words of natural language need to be used and the choice of words will affect the identification of semantic components as well. One attempt at solving this problem is by choosing a

limited number of words, which are used to explain the meaning of other words (see, for example, Wierzbicka, 1996; Goddard, 1998). However, there is no such inventory of semantic components fit for explaining all words of a language, and it is unlikely it could exist, or it would otherwise be too vast for convenient use.

3) The number of semantic components is not finite; in practice, each researcher puts forward a set of semantic components corresponding to the purpose of his research. However, in the work of a lexicographer and also in the development of electronic resources, such an approach would not present a solution, as the entire vocabulary of a language would have to be covered.

4) Even if a detailed decomposition or a word sense is possible, it is not possible to determine specifically how many and which semantic components must differ in order to register different word senses in a dictionary. In this case, a consistent solution is not possible, and the work of the lexicographer, as a rule, involves the use of intuition to determine which semantic components are sufficiently important for their change to create a new sense. If each case of a single differing semantic component was considered a new word sense, the resulting division of senses would be too exhaustive. Therefore, this criterion is usually applied in combination with the syntactic and semantic distribution, which was mentioned earlier.

And lastly, the difference in semantic components can be indicated by the possibility to replace one word with various **synonyms** in different contexts. As substitution with a synonym is a traditional and widely used method of explaining meaning, it can also be used in word sense or subsense distinction. For example, the word *spēlēt* (*to play*) can be substituted by verb *atskaņot* (*to perform*) in connection with music or a piece of music (*spēlēt / atskaņot skaņdarbu, mūziku* ‘to play / perform music, a piece of music’), but not in connection with a musical instrument (*spēlēt vijoli* ‘to play the violin’, but not *atskaņot vijoli* ‘to perform the violin’). That is a sufficient basis for a subsense ’to use (a musical instrument) to create sound’ to be established. This subsense is also the only one that forms hyponymic relationships with words *trinkšķināt* (*to fiddle*), *čīgāt* (*to saw*), as well as other words for playing musical instruments. The synonyms used in the definitions of word meanings can directly refer to synsets, but it should be noted that synonymy is essentially a relative concept, as the meanings of words can be more or less synonymous and they can have more or less in common.

5. Conclusions

The division of a word's lexical semantics into separate senses may vary depending on the purpose. The aim of word sense distinction in the context of development of WordNet is to obtain such a degree of word sense granularity that would allow to create synonymous, hyponymic, meronymic and antonymic links between word senses and subsenses and at the same time be transparent and easily perceived by any user of the Tēzaurs.lv electronic dictionary, including language learners. In cases of uncertainty, the decision is made in favour of what is needed to develop the Latvian WordNet.

The procedure of distinguishing word senses is based on a set of specific criteria, which are not equally substantial but jointly form a certain hierarchy. However, not all semantic groups demonstrate this hierarchy in the same way. In the sense distinction of polyvalent verbs syntactic distribution (syntactic functions of arguments and ways of coding) and semantic distribution (semantic roles of arguments and general or more specific semantic features) are more important, with semantic components and the possible replacement by a synonym playing a secondary role.

Although the concept of subsense has not been clearly defined yet, in the process of developing the Latvian WordNet the separation of senses and subsenses of verbs has proven necessary. Mostly, a subsense is a way of displaying metonymic (and less often metaphorical) shifts, which cannot be given the status of a separate sense. Regarding verbs, a subsense is most often distinguished by the semantic group of the subject or object. However, it should be emphasised that a consistent solution to subsense distinction is not likely, as it is not possible to determine exactly how large or significant the differences should be in order to consider them as a sign of a separate sense. The authors of the project have tried to formulate the superordinate sense in a sufficiently broad manner for it also to include subsenses. In cases when such an approach was not possible, a subsense was converted into an independent sense. In the formation of synsets and semantic links between word senses, the subsenses listed in the Latvian WordNet function in the same way as superordinate senses: they can form synsets or other semantic relations with other word senses.

Further work on the development of the Latvian WordNet will show whether the selected criteria for word sense distinction will prove useful for automatic word sense disambiguation and linking the Latvian WordNet with the Princeton WordNet. However, the authors are confident that the results of the chosen approach of manually processing the data are of a high quality and will serve as a valuable contribution to the development of lexicography and semantics of the Latvian language.

Acknowledgements

This research work was supported by the Latvian Council of Science, project “Latvian WordNet and word sense disambiguation”, project No. LZP-2019/1- 0464.

References

Allen, R. (1999). Lumping and splitting. *English Today*, 15 (4), pp. 61–63.

Cruse, A. (2004). *Meaning in language. An introduction to semantics and pragmatics.* Oxford: Oxford UP.

Darģis, R., Auziņa, I., Bojārs, U., Paikens, P., Znotiņš, A. (2018). Annotation of the Corpus of the Saeima with Multilingual Standards. *Proceedings of the 2018 ParlaCLARIN Workshop.*

Dzerins, J. & Dzonsons, K. (2007). Harvesting national language text corpora from the

Web. *Proceedings of the 3rd Baltic Conference on Human Language Technologies (Baltic HLT).*

Dziob, A. & Piasecki, M. (2018). Implementation of the Verb Model in plWordNet 4.0. Available at: https://www.aclweb.org/anthology/2018.gwc-1.14.pdf

Goddard, C. (1998). *Semantic analysis. A practical introduction.* Oxford & New York: Oxford UP.

Heine, B. & Kuteva, T. (2002). *World lexicon of grammaticalization.* Cambridge: Cambridge UP.

Jackson, H. (2002). *Lexicography. An introduction.* London & New York: Routledge.

Jurafsky, D. & Martin, J.H. (2020). Word Senses and WordNet. In: *Speeeh and Language Processing* (3rded.draft). Available at: https://web.stanford.edu/~jurafsky/slp3/

Kerner, K., Orav, H., & Parm, S. (2010). Growth and revision of Estonian wordnet. *Principles, Construction and Application of Multilingual Wordnets*, pp. 198–202.

Laizāns, M. (2015). *Latviešu valodas korpusa izveide no emuāru tekstiem. Bakalaura darbs.* (*Creation of a Latvian Language corpus from blog posts. Bachelor thesis.*) Rīga: Latvijas Universitāte.

Levane-Petrova, K. (2019). LVK2018: Līdzsvarotais mūsdienu latviešu valodas tekstu korpuss, tā nozīme gramatikas pētījumos (LVK2018: The Balanced Corpus of Modern Latvian and its role in grammar studies) *Language: Meaning and Form 10*, pp. 131–146.

LLVV: *Latviešu literārās valodas vārdnīca. (Dictionary of Standard Latvian)* 1.–8. (1972–1996). Rīga: Zinātne.

Paducheva, E. (2004). *Dinamicheskie modeli v semantike leksiki.* (*Dynamic models in lexical semantics*) Moskva: Yazyki slavyanskoj kul'tury.

Pustejovsky, J. (1998). *The generative lexicon.* Cambridge etc.: The MIT Press.

Saeed, J. I. (2000). *Semantics.* Oxford & Massachusetts: Blackwell publishers.

Skadiņa, I., Veisbergs, A., Vasiļjevs, A., Gornostaja, T., Keiša, I. & Rudzīte, A. (2012). *The Latvian language in the digital age.* Springer.

Spektors, A., Auziņa, I., Darģis, R., Gruzitis, N., Paikens, P., Pretkalniņa, L., ... & Saulīte, B. (2016). Tēzaurs. lv: the Largest Open Lexical Database for Latvian. *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC 2016), pp. 2568-2571.

Taylor, J. R. (2009). *Linguistic categorization.* Oxford: Oxford UP.

Wierzbicka, A. (1996). *Semantics. Primes and universals.* Oxford & New York: Oxford UP.

Zuicena, I. (2010). Vārda nozīmes skaidrojums "Mūsdienu latviešu valodas vārdnīcā". (Explanation of word senses in the "Dictionary of Modern Latvian"). *Vārds un tā pētīšanas aspekti.* 14(1). Liepāja: LiePA, pp. 369–374.

Finding gaps in semantic descriptions. Visualisation of the cross-reference network in a Swedish monolingual dictionary

Kristian Blensenius[1], Emma Sköldberg[2], Erik Bäckerud[3]

[1] University of Gothenburg, Gothenburg (Sweden)
[2] University of Gothenburg, Gothenburg (Sweden)
[3] The Swedish Academy Dictionary, Lund (Sweden)
E-mail: kristian.blensenius@gu.se, emma.skoldberg@svenska.gu.se,
erik.backerud@svenskaakademien.se

Abstract

Providing lexical information in dictionary entries by cross-referencing between semantically related headwords is very important, both from a reception-oriented and a production-oriented perspective. This study presents a survey of cross-references in a comprehensive monolingual dictionary of Swedish. It discusses cross-referencing in dictionaries in general as well as in the Swedish dictionary, focusing on the following four types of paradigmatic cross-references: SEE, COMPARE, SYNONYM, and OPPOSITE. By using data-visualisation software, the semantic network in the dictionary is overviewed in a new way. Furthermore, errors, gaps as well as other areas of improvement in the dictionary related to cross-referencing are discovered. Moreover, the relationships between the existing cross-references, how they are introduced in the dictionary and the dictionary's intended target groups are addressed. The study also reveals that the traditional lexicographic policies of the dictionary need to be adjusted to take advantage of the transition from paper to electronic publication.

Keywords: cross-references; paradigmatic relations; Swedish; lexicography; semantics

1. Introduction

According to Atkins and Rundell (2008: 238), ”every dictionary has its own palette of admissible ways of cross-referring from one entry to another”. The main aim of this study is to present a survey of cross-references in a comprehensive monolingual dictionary of modern Swedish, in this case the second edition of *Svensk ordbok utgiven av Svenska Akademien* (‘Contemporary Dictionary of the Swedish Academy’; henceforth SO). The study is conducted using a software for visualisation of graph data. We will show how the software can be used to find errors, gaps as well as other areas of improvement in the dictionary.

The outline is as follows: In section 2, the different functions of cross-references in general are discussed. Section 3 presents the different types of cross-references in SO. Section 4 focuses on the usage of four different paradigmatic cross-references placed in special fields in the microstructure of the dictionary and on how they can be visualised. These cross-references are indicated by the labels SE (‘see’, henceforth: SEE), JFR (abbr. of *jämför* ‘compare’, henceforth: COMPARE), SYN. (abbr. of *synonym* ‘synonym’, henceforth: SYNONYM), and MOTSATS (‘opposite’, henceforth: OPPOSITE). Section 5, finally, gives some final remarks.

2. Cross-references in dictionaries

Information on semantically related headwords in the dictionary entry is very important, especially from a reception-oriented perspective. Synonyms, antonyms, etc. serve to provide access to additional lexical information and lexical sets (e.g. relations between nouns such as *north*, *south*, *east*, and *west*) as well as delimiting the meaning of the headword (see e.g. Järborg, 1989: 20; Hult et al., 2010). For instance, the meaning of food-related English verbs like *chop*, *grind*, *mash*, and *shred* becomes clearer when comparing the definitions of the words. This kind of information can also serve to enhance users' knowledge of connotations, pragmatic characteristics, etc. For example, users can, by comparing different entries, be made aware of the different emotive meanings of adjectives denoting (degrees of) overweight, e.g. *chubby*, *corpulent*, *fat*, *plump*, and *stout.*

In her classical work *Words in the Mind. An Introduction to the Mental Lexicon*, Aitchison (2003) discusses four types of relationships between stimulus words and response words in association tests; *co-ordination* (e.g. *salt* - *pepper*), *collocation* (e.g. *bright* - *red*), *superordination* (e.g. *color* - *red/blue/green*), and *synonymy* (e.g. *hungry* - *starved*) (ibid. 2003: 84–91). She states that the consistent answers given in association tests seem to testify that meaning relations between different words also have psychological validity. Related words seem to be stored so that they form a system within which the associations take place.

Furthermore, information about semantically related headwords is, according to Malmgren (2009: 98), extremely important from a *production*-oriented perspective; providing synonyms can help users to write or speak with a varied vocabulary. Moreover, knowledge of antonyms and other classes of converse pairs of words can also be useful when it comes to paraphrasing (cf. *not nervous* and *calm*).

3. SO

3.1 The second edition of SO and its precursors

The second edition of SO, published in spring 2021, has been compiled at the Department of Swedish at the University of Gothenburg. The dictionary is primarily based on three previous (printed) dictionaries, including the first edition of SO, released in 2009. The dictionary includes approx. 65,000 headwords. In short, the monolingual definition dictionary is descriptive, and it deals with contemporary general language. SO is mainly intended as a reception dictionary, but it can also be used for production, and the target user groups are native speakers and advanced L2 learners.

The second edition of SO will (in contrast to the first edition from 2009) only be published digitally, as a dictionary app and at the Swedish Academy dictionary portal

Svenska.se. In preparing the second edition, the editorial team has made efforts in using the digital format as much as possible. For example, the content of the dictionary has become more accessible to the users than before, by having been made searchable in different ways. Furthermore, the SO lexicographers have tried to make it easier for users to both review the contents of the entries and go from one entry to another by adding more cross-references (these are indicated by labels, e.g. SYNONYM) and links.

At the same time, the current lexicographic team has updated the editorial guidelines with regards to cross-references between different entries. On many occasions, e.g. in the case of the verb ‘die’, there are plenty of synonyms or near-synonyms. However, when it comes to Swedish, there is a good selection of synonym dictionaries (see e.g. the website Synonymer.se, which is probably the most used Swedish dictionary of today). For this reason, and due to time constraints, inclusion of synonyms in SO has not been prioritised in the revision of the dictionary.

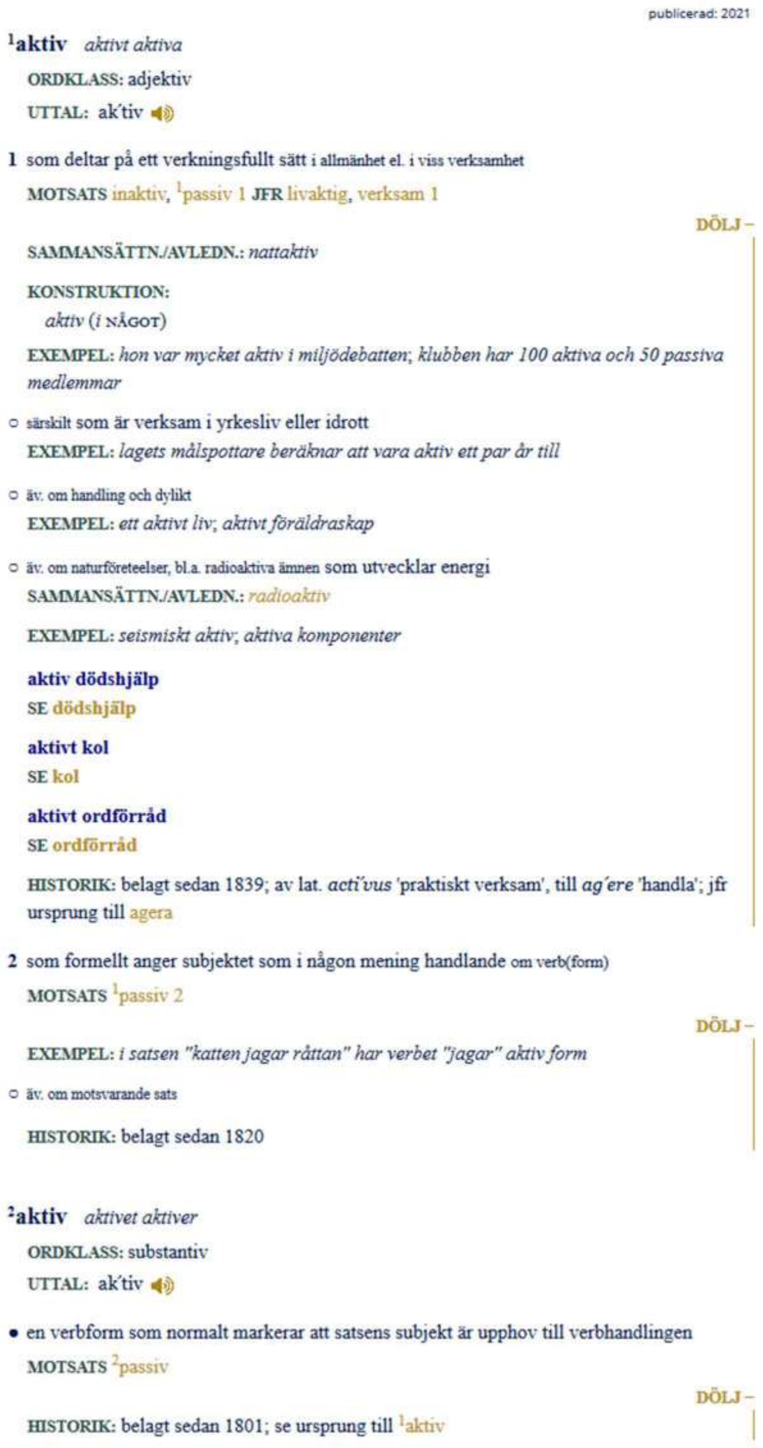
Furthermore, in connection with the practical lexicographic work with so-called “controversial” words, the SO lexicographers have aimed to include cross-references from offensive headwords to more neutral headwords in the dictionary, but not the other way around. For instance, in the second edition of SO, there is a link from the slightly archaic and derogatory adjective **homofil** (‘homo’) to the more neutral adjective **homosexuell** (‘homosexual’). However, there is no cross-reference from the headword **homosexual** to **homofil** (see Petersson & Sköldberg 2020 for more details on this work; also cf. "reciprocal cross-references" and "one-way cross-references" in Svensén 2009:389).[1]

3.2 Different types of microstructural cross-references in SO

In the following, we will discuss different types of cross-references included in the microstructure of the second edition of SO.

As a starting point, let us look at the entries ^{1}**aktiv** (‘active’) and ^{2}**aktiv** (‘active voice’) in the web version of SO in Figure 1.

[1] We use bold to indicate headwords. All translations into English were made by the authors.



publicerad: 2021

^{1}**aktiv** *aktivt aktiva*

ORDKLASS: adjektiv

UTTAL: ak´tiv

1 som deltar på ett verkningsfullt sätt i allmänhet el. i viss verksamhet

MOTSATS inaktiv, ^{1}passiv 1 JFR livaktig, verksam 1

DÖLJ –

SAMMANSÄTTN./AVLEDN.: *nattaktiv*

KONSTRUKTION:

aktiv (*i* NÅGOT)

EXEMPEL: *hon var mycket aktiv i miljödebatten; klubben har 100 aktiva och 50 passiva medlemmar*

○ särskilt som är verksam i yrkesliv eller idrott

EXEMPEL: *lagets målspottare beräknar att vara aktiv ett par år till*

○ äv. om handling och dylikt

EXEMPEL: *ett aktivt liv; aktivt föräldraskap*

○ äv. om naturföreteelser, bl.a. radioaktiva ämnen som utvecklar energi

SAMMANSÄTTN./AVLEDN.: *radioaktiv*

EXEMPEL: *seismiskt aktiv; aktiva komponenter*

aktiv dödshjälp

SE **dödshjälp**

aktivt kol

SE **kol**

aktivt ordförråd

SE **ordförråd**

HISTORIK: belagt sedan 1839; av lat. *acti´vus* 'praktiskt verksam', till *ag´ere* 'handla'; jfr ursprung till agera

2 som formellt anger subjektet som i någon mening handlande om verb(form)

MOTSATS ^{1}passiv 2

DÖLJ –

EXEMPEL: *i satsen "katten jagar råttan" har verbet "jagar" aktiv form*

○ äv. om motsvarande sats

HISTORIK: belagt sedan 1820

^{2}**aktiv** *aktivet aktiver*

ORDKLASS: substantiv

UTTAL: ak´tiv

- en verbform som normalt markerar att satsens subjekt är upphov till verbhandlingen

MOTSATS ^{2}passiv

DÖLJ –

HISTORIK: belagt sedan 1801; se ursprung till ^{1}aktiv

Figure 1: The entries ^{1}**aktiv** and ^{2}**aktiv** in the web version of the second edition of SO.

Starting from the top of the entry ^{1}**aktiv**, there are four different paradigmatic cross-references. The headwords **inaktiv** (‘inactive’) and ^{1}**passiv 1** (‘passive’) are introduced by the OPPOSITE label (i.e. MOTSATS in the figure). Furthermore, the users are encouraged to COMPARE the headword (JFR in the figure) with the adjective headwords **livaktig** (‘lively’) and **verksam 1** (‘energetic’). It should be noted that in the dictionary two other types of paradigmatically related words are also included, namely the ones introduced by the labels SEE and SYNONYM (see section 4.2 below).

Second, there are cross-references in the idiomatic-expressions section (and other kind of multiword expressions). During the preparation of the first edition of SO, new guidelines were set up for the lemmatisation of idioms. For instance, an idiom including an adjective and a noun (e.g. *den röda tråden* 'the common thread, theme') was placed and defined in the noun entry with a cross-reference to the expression in the adjective entry. In the case of **¹aktiv,** we find three cross-references of this type (see e.g. the multiword expression *aktiv dödshjälp* 'active euthanasia' with cross-reference to the entry **dödshjälp**).

As already indicated, the first edition of SO was primarily a printed dictionary, and the reason for using cross-references between idioms was limited space (cf. Rundell 2015). However, as the second edition of SO is only published electronically, it is at least possible to present the same information on an idiom in more than one entry, instead of using cross-references. A more radical change would be to make the idioms in SO more independent/visible/searchable and less dependent on their constituent parts. In this case, the traditional lexicographic policies of SO need to be adjusted to take advantage of the change in publication format.

Finally, as indicated by Figure 1, there are also links among the etymologies. In the case of **¹aktiv,** the dictionary users are (more or less) informed that they should compare the historical information of this word with the etymology of the verb **agera** ('act').

In **²aktiv** ('active voice'), there is one cross-reference and one link, to the OPPOSITE **²passiv** (also a verb form) and to the etymology of the adjective **¹aktiv**, respectively.

In addition to the cross-references and links found in the example in Figure 1, some words in the SO definitions are hyperlinked. One example is the noun entry **feminist** with the definition 'followers of feminism' where the word *feminism* is hyperlinked. These links are new in relation to the first edition from 2009, and they form also an important part of the semantic network in the dictionary. At present there are 13,000 links of this type in SO, so they are still relatively few. This can be related to the situation in English dictionaries. For example, Rundell stated already in 2015 that "In the digital editions of most (if not all) of the British learner's dictionaries, every word in an entry is hyperlinked" (Rundell, 2015:315). According to the author, it is advantageous that users can rapidly find the entry for a word if they are unsure about any word in a definition. However, according to the same author, it is never ideal for the users to have to go from one entry to another in order to get the full picture. If the meaning of certain keywords in the SO definitions (such as *feminism* in the example above) is made explicit in more than one entry (e.g. in the adjective entries **feminist** and **feministisk** 'feminist'), this problem can possibly be reduced.

To sum up, there are cross-references and links in slightly different places in the entries

[1]**aktiv** and [2]**aktiv**, and they partly serve different functions. Some are intended to facilitate for users to grasp the meaning of the headword by relating them to other relevant headwords in the dictionary. Other cross-references are only "space savers". The question is whether users understand this, if the labels are sufficiently informative, etc. (also see section 4.2 below).

4. The cross-reference network in SO

The tool we use to produce the graphics in this text is called Constellation (https://www.constellation-app.com/). It is a free open-source software for data visualisation and analytics originally developed at the Australian Signals Directorate in 2012.

To obtain the data describing all cross-references as a graph, the presentation-ready HTML files for the dictionary have been scanned by a Java program that generates a comma-separated text file with one line for every reference. This is one of the input formats accepted by the Constellation program. Each line contains identifiers and part of speech for the source and destination headwords as well as information about the kind of reference being made. An advantage of this simple data representation is that it makes it easy to select different subsets of the data for use in specific cases.

4.1 Visualisation of the cross-reference network in SO

The second edition of SO includes approx. 76,000 cross-references and links of the types discussed in section 3.2. The visualisation format is very space consuming, and some entries are connected to other entries by a great number of cross-references. For instance, **gå** ('go', verb) includes no less than 88 cross-references and links. Other examples from SO are **hand** ('hand', noun, 75 cross-references and links), **dag** ('day', noun, 71), **fot** ('foot', noun, 55), **stå** ('stand' verb, 54), and **dra** ('pull', verb, 54). The large number of cross-references and links in these entries is related to the fact that the words occur in a significant number of idioms.

Some examples from SO to illustrate how the software works and can be used in practical lexicography are presented below. The relationship between the entries [1]**aktiv** and [2]**aktiv** and (a subset of) other headwords in the second edition of SO can be visualised as in Figure 2.

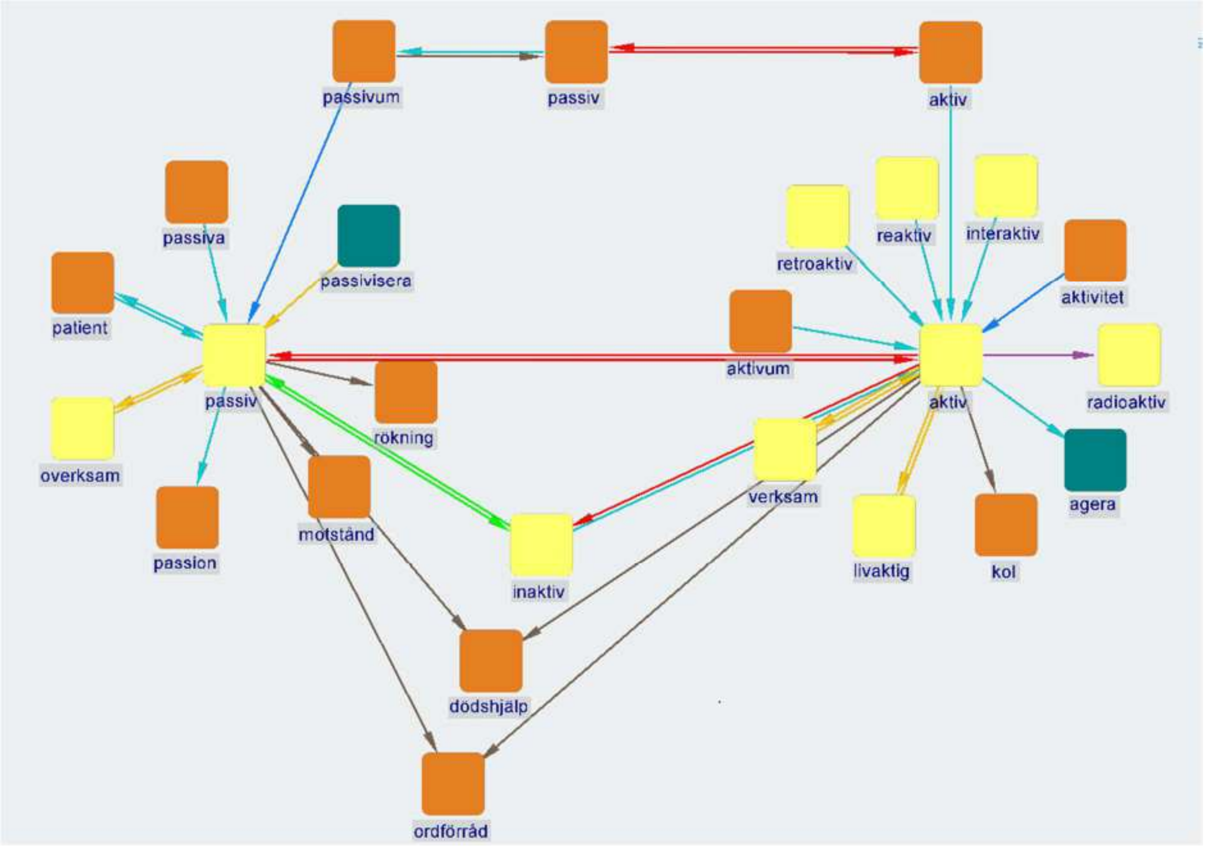


Figure 2: The semantic network including the headwords ¹**aktiv** and ²**aktiv** in the second edition of SO, illustrated by the Constellation software.

It can be seen that the figure includes boxes with different colours. Yellow boxes represent adjectives, orange boxes represent nouns, and green boxes represent verbs. The direction of the arrows indicates the direction of the cross-references. Furthermore, the colours of the arrows indicate the type of cross-reference. A green arrow represents SYNONYM, a red arrow represents OPPOSITE and a yellow represents COMPARE. Brown arrows represent SEE, a label used both in the section including paradigmatic cross-references and in connection with the idioms. In addition, light blue arrows represent hyperlinked words in the etymological part of the entry and dark blue arrows represent hyperlinked words in definitions (see e.g. the example entry **feminist** in section 3.2).

From the lexicographer's perspective, Figure 2 shows that many entries are treated in a consistent way. For example, there are cross-references in both directions (i.e. reciprocal cross-references) between the adjective headwords **aktiv** and **passiv** and between the noun headwords **aktiv** and **passiv**. As indicated by the colours of the arrows, the cross-references between the word pairs are also of the same type. However, the figure also reveals some shortcomings in the dictionary. For instance, there is a cross-reference (OPPOSITE) from **aktiv** to **inaktiv**, but it is unidirectional.

Finally, through the visualisation it becomes clear that a large proportion of the entries in SO includes no cross-references at all. Most of these entries could easily be enhanced in different ways. For example, the entry **alarmerande** (‘alarming’) could provide information about more or less synonymous adjectives (e.g. **skrämmande** ‘scary’). Furthermore, an entry such as **överviktig** ('overweight') could be enhanced by providing an antonym such as **underviktig** ('underweight'). Likewise, in the entry **animalisk** (‘animal-like’), a cross-reference (COMPARE) to the cohyponym **vegetabilisk** (‘vegetable-like’) could be included. In addition, the content of many entries may become more accessible if individual words in the definitions are hyperlinked. In other words, by using illustrations like the one in Figure 2, lexicographers find different kind of gaps in the lexical database.

4.2 Paradigmatic cross-references in SO

In this section, we will focus on the usage of paradigmatic cross-references placed in certain fields in the microstructure of the dictionary, i.e. the cross-references indicated by the (translated) labels COMPARE, SYNONYM, SEE, and OPPOSITE (also cf. the types mentioned by Svensén 2009:248–251). In total, there are about 38,000 cross-references of this type in the dictionary. The total number is distributed as follows on the four current types: 26,011 COMPARE, 3,746 SYNONYM, 3,552 SEE, and 1,286 OPPOSITE. As already mentioned, the relationship between the headwords connected by cross-reference is indicated by the colour of the arrow. The yellow arrow represents COMPARE, and the green arrow represents SYNONYM. Furthermore, the brown arrow represents SEE and the red arrow represents OPPOSITE.

Before discussing more specific headwords, the choice of labels in SO should be considered. First, the use of abbreviations (i.e., JFR and SYN.; see section 1) is disputable. In order to make things easier for users, especially for L2 learners, these abbreviations should be expanded in future editions of SO. However, even if the labels are expanded, it is not obvious how labels such as COMPARE, SYNONYM, SEE and OPPOSITE should be interpreted. If the dictionary users follow cross-references marked with COMPARE and SEE, they might notice the semantic relationship between the words. Perhaps they also understand the intention of connecting to these closely related words (see section 2 above). However, the SO lexicographers have limited knowledge about this, and a user survey is needed to gain in-depth knowledge of this aspect.

Furthermore, the term *synonymy* usually denotes a meaning relation in which a word can be said to have the same meaning as another word. Principally, they should then have the same set of semantic components. This does not necessarily mean that the words are fully interchangeable in all contexts. Or, as Ullmann stated already in 1962 (p. 141), "In contemporary linguistics it has become almost axiomatic that complete synonymy does not exist". However, the term is common in teaching contexts, especially in second language teaching where it often refers to words that have approximately the same meaning. As already mentioned, there are more than 3,700 SYNONYM cross-references in SO, and a quick look at the words shows that they do not mean exactly the same thing. The two headwords categorised as synonyms often belong to either general or technical language and they often appear in different contexts. In the same way, the synonyms may differ in terms of style. It should also be mentioned that the SO editorial staff, in compiling the second edition, has been very restrictive regarding the use the label SYNONYM, since the headwords are seldom completely interchangeable. Instead, the vaguer COMPARE has been used. In this case, it is possible to discern a change over time concerning the principles for the lexicographical work with SO and its forerunners (and possibly also of the boundaries of the meaning of words like *synonym*). It is not satisfactory that there are differences between the entries in SO depending on the period the entries were compiled.

Thanks to the visualisation software, it is easier than previously to get an overview of the semantic network in SO. Inconsistencies and opportunities for improvement in the dictionary are also more obvious. As an example, the headwords related to the adverb **ganska** ('quite, pretty, rather') are presented in Figure 3.

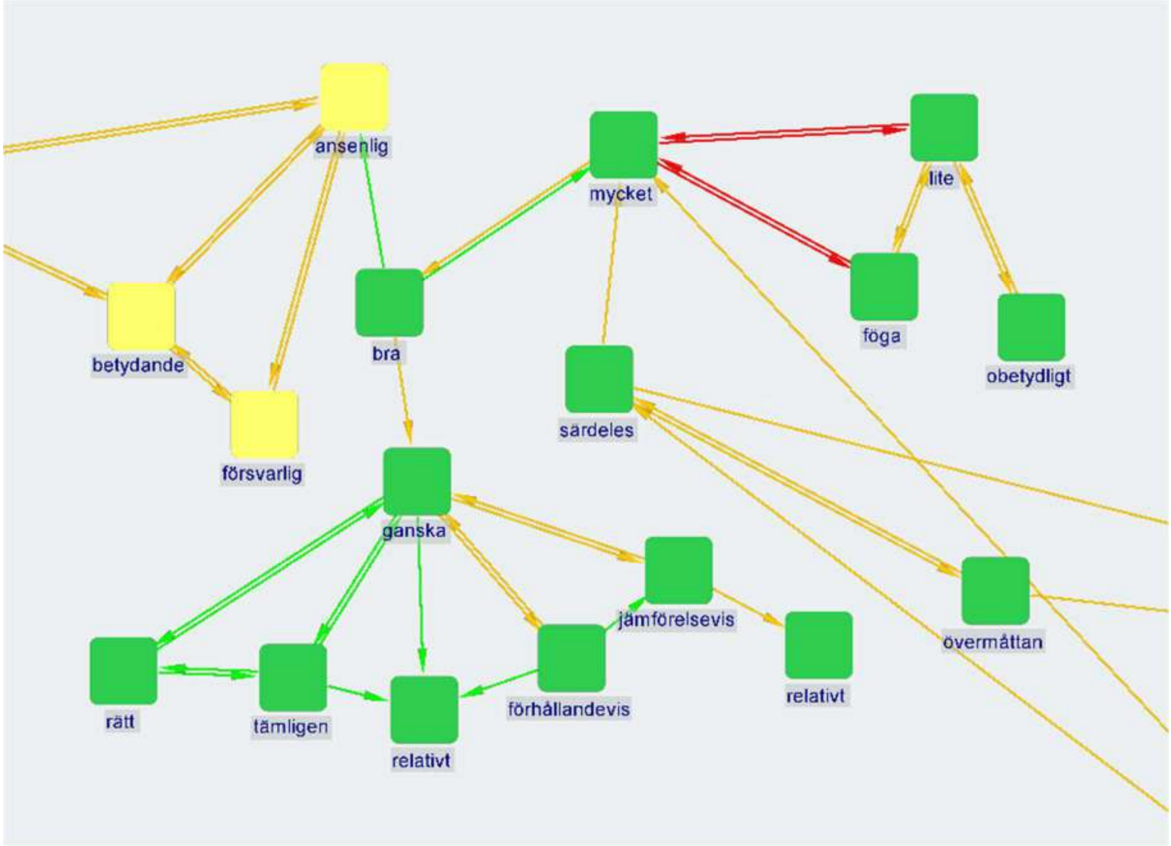


Figure 3: The headword **ganska** and its synonyms etc. in the second edition of SO.

As shown in Figure 3, there are reciprocal cross-references between **ganska** and two other headwords in SO (see the double green synonym arrows to and from the words **rätt** 'fairly' and **tämligen** 'fairly, moderately'). Furthermore, there are reciprocal yellow cross-references of the type COMPARE between **ganska** and **förhållandevis** ('proportionately') and between **ganska** and **jämförelsevis** ('comparatively'). The adverb **relativt** (‘relatively’) is also introduced as a synonym to the actual headword. The alternatives presented to the current adverb are, as already mentioned, important from a production-oriented perspective (see section 2).

Moreover, according to Figure 3, there is no cross-reference from the headword **relativt** to **ganska** – or from **relativt** to any other headword in the dictionary. In other words, the adverb **relativt** could be more clearly related to other adverbs in SO.

In the same figure, it is also possible to overview the cross-references to and from the headword **mycket** (‘very’). According to the dictionary, one of the main senses of the adverb **bra** (‘to a large extent’) corresponds to a sense of **mycket**. The adverb **mycket** is, however, not considered as a synonym to **bra**. The relationship between **bra** and **mycket** is not clarified, but users are encouraged to compare the words. The information provided in the two entries is thus not entirely consistent.

Finally, as shown by Lyons (1977:270–290) among others, it is complicated to discuss the label OPPOSITE because it includes rather disparate categories. The most important types of opposite relations are probably *complementarity* (e.g. the words *dead* - *alive*), *antonymy* (e.g. *hot* - *cold*) and *converseness* (e.g. *husband* - *wife*). As with synonyms, the label OPPOSITE is used in a broad sense in SO, covering at least complementarity and antonymy. And, like synonyms, there are plenty of cases where the information on presumed converse headwords is not consistent.

5. Final remarks

In this paper, we focus on cross-referencing in the second edition of the monolingual contemporary dictionary *Svensk ordbok utgiven av Svenska Akademien* (SO). The cross-references, which create a semantic network, have been investigated by using a software for the visualisation of graph data.

As already mentioned, in preparing the second edition the editorial team has made efforts to use the digital format as much as possible. However, significant work remains before the digital format is fully utilised in SO, especially when it comes to the use of cross-references in the dictionary. A large proportion of entries in SO are completely isolated as they include no cross-references or links to other entries at all.

Overall, there is room for improvement concerning the semantic relations in the dictionary, which are unveiled by the visual network. Hopefully, a more developed version of the visualisation tool could be incorporated into the lexicographers' editing interface.

Furthermore, the relationship between the paradigmatic cross-references and the dictionary's intended target groups and intended areas of use can be discussed. Mother-tongue speakers and advanced learners do not always have the same possibilities for interpreting the information given. In addition, semantically related words that are supposed to support reception and production, respectively, are not necessarily the same words. Moreover, the idea of, for example, 'synonymy' has probably changed over the approx. 40 years that the work with the dictionary database has been going on. The fact that the SO, from now on, is only published electronically can also significantly affect the information category in future editions.

In the future, it would be interesting to compare the semantic network in SO with the networks in other dictionaries. In the same way, it would be of interest to compare the network with other presentations of semantically related word. For Swedish, Åke Viberg and his colleagues have compiled a WordNet, but unfortunately it is not very accessible (see Viberg et al., 2002). A more updated resource is Swesaurus, which has been developed at Språkbanken at the University of Gothenburg (see e.g. Borin & Forsberg, 2014).

Finally, it would also be interesting to investigate whether the user interface could be provided with illustrations to further clarify how the Swedish words are related to each other, from a contemporary as well as from a historical perspective.

6. References

Aitchison, J. (2003). *Words in the Mind. An Introduction to the Mental Lexicon.* 3rd edition. Oxford: Basil Blackwell.

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

Borin, L. & Forsberg, M. (2014). Swesaurus; or, The Frankenstein Approach to Wordnet Construction. In: *Proceedings of the 7th Global WordNet Conference (GWC 2014).* Tartu: Global WordNet Association, pp. 215–223.

Hult, A.-K., Malmgren, S.-G. & Sköldberg, E. (2010). Lexin – a report from a recycling lexicographic project in the North. In A. Dykstra & T. Schoonheim (eds.) *Proceedings of the XIV Euralex International Congress.* Leeuwarden/Ljouwert: Fryske Academy, pp. 800–809.

Järborg, J. (1989). Betydelseanalys och betydelsebeskrivning i Lexikalisk databas. (Unpublished manuscript.) Göteborg.

Lyons, J. (1977). *Semantics.* Vol. 1. Cambridge: Cambridge University Press.

Malmgren, S.-G. (2009). On production-oriented information in Swedish monolingual defining dictionaries. In S. Nielsen & S. Tarp (eds.) *Lexicography in the 21st Century. In honour of Henning Bergenholtz.* (Terminology and Lexicography Research and Practice 12.) Amsterdam/Philadelphia: John Benjamins, pp. 93–102.

Petersson, S. & Sköldberg, E. (2020). To discriminate between discrimination and inclusion: a lexicographer's dilemma. In Gavriilidou, Z. et al. (eds.) *Proceedings of XIX EURALEX Congress: Lexicography for Inclusion, Vol. I.* Alexandroupolis, Greece: Democritus University of Thrace, pp. 381–386.

Rundell, M. (2015). From Print to Digital: Implications for Dictionary Policy and

Lexicographic Conventions. *Lexikos* 25, pp. 301–322.

SO = *Svensk ordbok utgiven av Svenska Akademien* (First ed. 2009, second ed. 2021). Stockholm: Norstedts. Also available as apps and at https://svenska.se/.

Svensén, B. (2009). *A Handbook of Lexicography: The Theory and Practice of Dictionary-Making.* Cambridge: Cambridge University Press.

Ullmann, S. (1962). *Semantics. An Introduction to the Science of Meanin*g. Oxford: Basil Blackwell.

Viberg, Å, Lindmark, K., Lindvall, A. & Mellenius, I. (2002). The Swedish WordNet Project. In Braasch, A. & Povlsen, C. (eds.) *Proceedings of the X EURALEX Congress: Copenhagen-Denmark, August 13-17, 2002.* Proceedings Volume I, pp. 407–412.

Reshaping the Haphazard Folksonomy of the Semantic Domains of the French *Wiktionary*

Noé Gasparini[1], Cédric Tarbouriech[2], Sébastien Gathier[3], Antoine Bouchez[4]

[1,3,4] Institut international pour la Francophonie - Université Jean Moulin Lyon 3, 1C avenue des Frères Lumière CS 78242 - 69372 Lyon Cedex 08 France

[2] Institut de Recherche en Informatique de Toulouse (IRIT), Université de Toulouse & CNRS, France

E-mail: noe.gasparini@univ-lyon3.fr, cedric.tarbouriech@irit.fr, sebastien.gathier@univ-lyon3.fr, antoine.bouchez19@gmail.com

Abstract

Semantic domains are a source of headaches in dictionary projects, and one was built haphazardly in the French edition of the collaborative online project *Wiktionary* called *Wiktionnaire*. *Wiktionnaire* is a lexicographical project that started 17 years ago. It is hosted by the Wikimedia Foundation and edited by a community of volunteers that made it a mature project, but with lacunas, with semantic domains being one of these. Between January 2019 and December 2020, this nomenclature of semantic domains was transformed by a small team with complementary expertise and skills. The team consisted of four people with academic knowledge in linguistics, lexicography and information science, as well as technical skills for coding, proofreading and community management. The strategy was the following: mapping the existing terminology, comparison and extension of the list, documentation, structuring, discussions with the community, deployment, cleaning of remaining irregularities, and monitoring the changes after this process. The result of this two-year operation is a complete reshaping of a messy folksonomy into an innovative lattice nomenclature fully integrated into the *Wiktionnaire* and adopted by the community, but also used in an RDF-based dictionary reusing that data, the *Dictionnaire des francophones*. This paper outlines the context of this work on continually changing content and presents the strategy used by the team, including the major issues and choices encountered during the process.

Keywords: semantic domains; *Wiktionnaire*; *Wiktionary*; folksonomy; collaborative lexicography.

1. Background

Wiktionary is a collaborative multilingual open online collection of lexicographical information (Murano, 2014). The edition in French, *Wiktionnaire*, commenced in March 2004 and has since then seen a constant growth in content and quality (Sajous *et al.* 2014). The population of regular contributors is about 100 people, each contributing on average more than 100 edits monthly. Between 1,500 and 2,000 volunteers edit at least once a month. Most of the regular contributors acquired lexicographic skills as they contributed, without any academic background (Meyer &

Gurevych, 2012a), developing a community-based practice. For an overview of studies about *Wiktionnaire*, see Sajous *et al.* (2020).

One aspect of crafting definitions is to indicate semantic domains for technical terms. In *Wiktionnaire*, such domains are presented at the beginning of definitions, between parentheses. When an editor adds a definition for a lemma, a dedicated code indicating its semantic domain is also added, written between curly brackets. This is a way to transclude a subpage named a *Modèle* (*Template* in English). These templates serve to display a text and categorise pages in *Categories*. For example, the code {{anatomie|fr}} was used to insert the content of the template Modèle:anatomie, resulting in the text "anatomie" being displayed and a link to the page with the lemma being included in Category:Anatomie, forming the French lexicon of anatomical terms.

Before 2020, indicators of semantic domains at the beginning of definitions in *Wiktionary* projects were irregular, and the granularity of subdomains showed a large heterogeneity (Meyer & Gurevych, 2012b). This aspect of *Wiktionnaire* or *Wiktionaries* is rarely studied, and most research on folksonomy focuses on *Wikipedia* and tries to construct an external and independent ontology (Macías-Galindo, 2011).

The evolution, maintenance and reuse of these domain indicators were made complex by the existence of over 400 templates, sometimes with aliases. Most of these templates were poorly documented. There were also more than 10,000 entries with a plaintext indicator rather than a dedicated template. Some pages were added manually to categories, instead of by using a template. The category pages displaying the lists of terms associated with a domain were poorly documented. The structure was not standardised between languages described in *Wiktionnaire.*

Wiktionary was previously used as a corpus to create external tools like the XML-encoded machine-readable version GLAWI (Sajous & Hathout, 2015), or the comparison of new words in dictionaries listed in the DiCo corpus (Martinez, 2013). The results of such studies are rarely shared with contributors and rarely injected back into the *Wiktionnaire* (or *Wiktionary* in another language). This project on semantic domains not only produced an independent taxonomy, but improved the existing structure of *Wiktionnaire* itself. It was crowdsourced applied lexicography.

2. Motivations

In March 2018, the French president Emmanuel Macron presented a plan for promoting the French language and multilingualism. This included a project funded by the Ministère de la Culture [Ministry of Culture] and managed by the Institut International pour la Francophonie at the Université Jean Moulin Lyon 3. The ambition was to create a new dictionary for varieties of French, the *Dictionnaire des francophones* (DDF). It was to be structured as an RDF-based lexicographical database furnished by existing lexicographical resources, including the French entries of *Wiktionnaire*

(Dolar et al., 2020; Steffens et al., 2020). Linked data for lexicography opened a new field, and this project was going to be a first for the French language. A short explanation of this way of organising data is presented by Klimek and Brümmer (2015).

In 2020, a 'Wiktionarian in residence' was recruited to clean up *Wiktionnaire's* content to help the integration of this resource. The purpose was to undertake corrections of general issues but also to clean information structures, including the semantic domains. A dedicated task force emerged to fulfil this mission.

Sébastien Gathier, the resident, is a senior wiki proofreader, mostly for French Wikipedia, Wikidata and Open Food Facts. Noé Gasparini, the DDF project manager, has training in linguistics and language documentation. Antoine Bouchez, an intern for four months, was a lexicographer by training. A skilled Wiktionary contributor, Cédric Tarbouriech, was invited. He is a contributor trained in coding and ontology modelling. The group set regular meetings to work together remotely.

3. Strategy

The strategy developed by the team had several steps: map existing terminology (labels and their aliases); compare and augment this list with domains used in other referential works; build a structure; define each label with short glosses to document and disambiguate domains; discuss with the community to obtain consent to implement this solution; implement the list in the Lua language; prepare scripts to deploy this new code in more than 20,000 pages; correct uncountable irregularities that may remain after deployment; build a 'lexicovigilance' similar to a pharmacovigilance to monitor any adverse effects subsequent to this large transformation of *Wiktionnaire.*

3.1 Mapping existing terminology in *Wiktionnaire*

The *Wiktionnaire* uses templates to transclude content into other pages. These templates may include parameters, such as the language to use to categorise the content. Before 2020, when a contributor wanted to add a new domain for one language they had to create a new template, which was not an easy task. This new template should be documented but in practice they rarely were. Most of these creations were made by a couple of experienced users. Additional isolated templates were created to cater for very specific needs in some languages.

The list of templates used for semantic domain indicators was augmented by some aliases, i.e. shorter names based on traditional abbreviations from printed dictionaries such as 'hist' for 'history'. Some of them were opaque and could be misinterpreted and wrongly used, e.g., 'litt' could be read as indicating the domain of literature or of literary language; 'comm' could be read as the vocabulary of commerce or of communication.

Glossaries are lists of pages gathered around a common hypernym, such as lists of rivers, birds, languages, etc. There are more than 2,000 glossaries in *Wiktionnaire.* Most of them are included under a semantic domain, such as ornithology for the glossaries of bird names. The distinction between glossaries and domains is a grey area. This may lead to confusion when defining new templates. For example, ‘graph theory’ was seen as a glossary but is in fact a lexicon, and vice versa for ‘feelings’.

3.2 Comparison and extension of the list

The original list of domains contained about 400 items. Some very specific subdomains were covered, influenced by contributors’ interests. Other domains stayed barely explored due to the lack of interested contributors. Subdomains of computer science were well described but some sports or scientific domains were missing. Some new domains were added thanks to a comparison with other sources in French, such as the *Larousse illustré* (2014), the *Dictionnaire universel* (2008), and the *Dewey Decimal Classification.* Some of these new domains are banking, bryology, clothing, electromagnetism, geomorphology, leather crafting, immunology, petrology, puppetry, and speleology.

A second phase was initiated with the alignment of items from *Le Grand Dictionnaire Terminologique.* More than 4,000 lexical entries from this resource were given to the *Dictionnaire des francophones* by the *Office québécois à la langue française* [Quebec Board of the French Language], and they were willing to share their own terminology with the team. The *Commission d'enrichissement de la langue française* [Commission to enrich the French language], responsible for the content of the website *FranceTerme*, also shared their classifications to prepare for the integration of this resource in the *Dictionnaire des francophones.* With both of these works more domains were added, such as advertising, archery, brewery, flour production, materials science and engineering, spatial planning, woodworking, and waste management.

New domains were also added by exploring definitions with domain indicators that had no dedicated template in *Wiktionnaire.* The initial list contained 390 domains (April 2019) and the final one contains 615 domains (April 2021). Of these, 578 domains are for the French language and others. Some domains are used for only one or a couple of specific languages, such as ‘cleaning’ (only used for German), or ‘Estonian mythology’ (only for Estonian). Some new domains were suggested but in this new taxonomy are not in use for any languages yet.

3.3 Documenting the lattice structure

Most of the domain templates in *Wiktionnaire* had a short documentation explaining the technical use of the template but not the definition of the concept itself. In order to clarify the terms, short glosses in French were written for each domain.

The initial list was structured with a ramification of categories, in a repeated process of grouping categories together in supercategories or splitting categories into subcategories, with some branches being diversely connected depending on the language described. This structure was regularised and developed as a lattice structure – or, more precisely, a directed acyclic graph – rather than a tree structure, as some domains could have more than one domain above them.

This structure was planned to be explored in DDF through a contributive interface, from the top to the bottom. There are seven top-level domains: technology, arts, alimentation, sports, politics, science and society. They are not supposed to be used as such, but serve as a coarse division of domains to assist in navigating the lattice in the *Dictionnaire des francophones*. In *Wiktionnaire*, this structure has been fully implemented, but is not directly visible to readers.

Direct subdomains include industrial activities, types of arts, types of sports, and academic domains. Those high-level domains are considered as perennial and less inclined to change in the future in comparison with more nested domains. There are five to 26 subdomains directly under the top-level domains. More specific subdomains were not expanded in detail, considering that editors will add new domains when they want to gather the related vocabulary to build them.

3.4 Discussions with the *Wiktionnaire* community

The first step to engage the discussion was to publish a page in *Wiktionnaire* titled Projet:Informations lexicographiques[1] [Project:Lexicographical information] to describe the existing structure. This led to some general observations and offered a way to include more contributors for future discussions. Some parts of the process were presented to the community, mostly when it seemed better to split existing domains to follow the tendency observed in other sources. For example, we suggested a division between psychology and psychoanalysis.

A long-term discussion concerned the lexicons of French law and French history, as we felt they should be separated with a combination of domain indicators and geographical indicators (Law+France and History+France) instead. The community rejected the proposal and both lexicons remained.

Another issue was about the vocabulary of the European Union, considered as a technolect or jargon, depending on the analysis (Gardner, 2016). One of the *Wiktionnaire* contributors was a professional translator, so he had enough knowledge of this vocabulary to reorganise the entries and solve this issue. A dozen contributors

[1] https://fr.wiktionary.org/wiki/Projet:Informations_lexicographiques

shared insights on specific issues such as heraldic subdivision, organ building or social justice. Some comments were about the definition while others were about the structure.

Between January 2020 and May 2021, 18 contributors made at least one modification to the list of domains, in order to correct sentences, add new domains or slightly modify the structure.

3.5 Lua implementation

Lua is a lightweight high-level programming language. It is one of the few programming languages available in a MediaWiki environment, the technical software used by *Wiktionnaire*. It was needed to program advanced behaviours.

The list of domains is written as a Lua table, a structured map linking semantic domains to their information. Each item has a name, a description, an indication of the phrasing to write to make sentences readable by humans, and supercategories in which the domain has to be included. This page is called Module:Lexique/data[2] and it is the unique list of domains for *Wiktionnaire*.

Definition of the domain ‘boulangerie’ (bakery).

```
 ['boulangerie'] = {
      ['description'] = 'La boulangerie désigne la fabrication et la vente de pain et
de viennoiseries.',
      ['determiner'] = 'de la ',
      ['super_categories'] = { 'cuisine' }
   },
```

3.6 Deployment with a dedicated script

A Python script was used to deploy the new system in both articles and categories. Specific issues were documented in maintenance categories. A couple of new domains had been created in the meantime and were added. At the end of the deployment, all old templates and their aliases were deleted to avoid further use, which would result in a confusing coexistence of two incompatible systems. Only the new template *lexique* is now used, and included in 34,000 pages. 96,860 domain tags were added to over 87,692 French definitions.

[2] https://fr.wiktionary.org/wiki/Module:lexique/data

3.7 Dealing with irregularities

After the deployment, a large number of definitions still displayed a free-text domain indicator rather than the new template for the domain. This meant that they were not included in the lexicons. More than 5,000 of those were corrected manually in 2020; this task is still ongoing.

Some cases needed a special investigation. For example, some templates had a parameter that indicated a subdomain, i.e. 'Canadian football' was using the same template as 'Football' with an additional parameter written as "spéc=canadien". It was changed to be two separate parameters in the new template. The same situation occurred with religions, mythologies, and subdomains of law and sports.

Another task was to ferret out missed domains. Some had been added by other contributors during our work and others were used in only a set of languages having very few words. They had to be included or recategorised. An example is the vocabulary of Palaic mythology for a couple of words.

Some *Category* pages had an introduction in plain text that conflicted with the deployment of *Modèle:catégorisation lexique* but included useful information, thus requiring careful revision.

3.8 Monitoring and accompanying the community to change its habits

To avoid a negative response from contributors, careful vigilance was maintained during the six months that followed the deployment to correct any mistakes and explain the changes when necessary.

In addition, a new function was developed to suggest semantic domains with an autocompletion while contributing. The function was developed by a contributor named Darmo117 after the idea was suggested by the community.

After only a few weeks, this lexicovigilance became less necessary as new semantic domains were created and added correctly by contributors outside the team. The new documentation led to the involvement of new contributors into the definition of lexicons.

4. Discussion

This whole process was a success, and the new taxonomy was adopted by both *Wiktionnaire* and *Dictionnaire des francophones*. The lattice allows the inclusion of new branches easily, by any contributor of *Wiktionnaire*. It is dense but still shows some irregularities due to its origins and choices made by the contributors during the process and after. The plan was to structure the semantic domains, considering it might help the readers explore the content and the contributors add new domains. If readers' experience was not monitored, this new organisation seems to have an impact on the addition of new domains during the following months.

This strategy did not come out of the blue, it was based on previous changes of policies encouraged in *Wiktionnaire* such as, for instance, the description of protolanguages or how to describe prototypical pronunciations. The wiki workflow creates a vivid space to discuss the editorial choices made in the past and suggest transformations for any aspect of the project. The refinement of every semantic domain was nonetheless a large-scale change, more ambitious than any previous initiatives, and it had an impact on almost every contributor's habits. As such, it was a first for *Wiktionary* as it was for lexicography.

In 3.6 some metrics were given. Those were not accessible before the inclusion of *Wiktionnaire* in the *DDF* database as the wiki structure is not that easy to query. The adaptation of *Wiktionary* content into an RDF database made possible new exploration of language-centric metrics, impossible with the original multilingual pages.

This new taxonomy is mainly based on the existing one and had to remain close TO IT as the *Dictionnaire des francophones* will import updates of *Wiktionnaire* in the future. Despite a large comparison with existing dictionaries for the French language, our taxonomy remains to be compared and aligned with taxonomy in other languages, such as *SIL Semantic Domains* and *WordNet Domains*. A terminological comparison with more resources could be a way to improve the structure and relations between domains and glossary. It may also help to find any blank spots and suggest new domains to cover.

The structure of the controlled vocabulary in *Wiktionnaire* does not readily allow alignment with other taxonomies. There is no unique identifier for each domain. Nonetheless, each page of category for vocabulary is connected with an entity in the *Wikidata* database, and these could be linked to the related concepts described by the semantic domains. The concepts could then become connected with several ontologies and databases. However, as of April 2021, less than 20% of semantic domains are fully connected with the related concepts. This possibility is still on the roadmap for the team to enhance the semantic domains of *Wiktionnaire.*

This lexicographic process is focused on knowledge engineering and information science, but it aims at producing a semantic structure that is easy to explore rather than a complex graph of relations among domains with a semantic elevation to offer ways to explore the data. This controlled vocabulary of domains is one facet of definition and the semantic structure of the entries. A possible future step may include qualifiers for the relations among the domains, but also with glossaries and significant entries.

5. Conclusion

This experience was successful and could be reproduced on other collaborative crowdsourced projects, such as the *Wiktionaries* written in other languages. In *Wiktionnaire*, a similar process of cleaning, documenting and structuring geographical information started in 2021.

As a conclusion, we want to point out that this strategy would not have been possible without the dedication of a multidisciplinary and multicompetent taskforce during an extensive period of time. This two-year undertaking allowed the cleaning of most of the data, the identification of areas of improvement by comparison with other resources, and the involvement of the *Wiktionnaire* community in the course of the project.

6. Acknowledgements

The authors thank the contributors of *Wiktionnaire,* the French Ministry of Culture for the initiative and funding of *Dictionnaire des francophones* project, Wikimédia France for their funding and the Institut international pour la Francophonie for managing the project and hosting a large part of the team. The authors also want to thank their anonymous reviewers and Robert Scott, Nadia Sefiane, Marya Lambia and Damien Vergnet for their useful comments on the draft.

7. References

Achard-Bayle, G. & Paveau, M.-A. (2008). La linguistique ‘hors du temple’. *Pratiques*, 139/140, pp. 3-16. Available at: https://hal.archives-ouvertes.fr/hal-00516249/document

Dolar, K., Steffens, M. & Gasparini, N. (2020). Dictionnaire des Francophones: A New Paradigm in Francophone Lexicography. In: *Proceedings of XIX EURALEX Congress: Lexicography for Inclusion, Vol. I.* Thrace: Democritus University of Thrace, pp. 23-30. Available at: https://euralex2020.gr/wp-content/uploads/2020/11/EURALEX2020_ProceedingsBook-p023-030.pdf

Gardner, J. (2016). *Misused English words and expressions in EU publications.* Available at: https://www.eca.europa.eu/Other%20publications/EN_TERMINOLOGY_PUBLICATION/EN_TERMINOLOGY_PUBLICATION.pdf

Klimek, B. & Brümmer, M. (2015). Enhancing lexicography with semantic language databases. Kernerman Dictionary News. Available at: https://www.kdictionaries.com/kdn/kdn23_2015.pdf#page=5

Macías-Galindo D., Wong W., Cavedon L., Thangarajah J. (2011). Using a Lexical Dictionary and a Folksonomy to Automatically Construct Domain Ontologies. In: Wang D., Reynolds M. (eds) AI 2011: *Advances in Artificial Intelligence.* AI 2011. Lecture Notes in Computer Science, vol 7106. Springer, Berlin, Heidelberg.

Martinez, C. (2013). La comparaison de dictionnaires comme méthode d'investigation lexicographique. N. Gasiglia (ed.). *Lexique.* 21. Villeneuve-d'Ascq, Presses universitaires du Septentrion, pp. 193-220.

Meyer, C.M. & Gurevych, I. (2012a). Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography. In Sylviane Granger & Magali Paquot (eds), *Electronic Lexicography*, Oxford University Press, pp. 259-292.

Meyer, C.M. & Gurevych, I. (2012b). OntoWiktionary: Constructing an Ontology from the Collaborative Online Dictionary Wiktionary. In: *Semi-Automatic Ontology Development: Processes and Resources.* Maria Teresa Pazienza & Armando Stellato (eds). Information Science Reference, pp. 131-161.

Murano, M. (2014). La lexicographie 2.0 : nous sommes tous lexicographes ?. *Cahiers de recherche de l'École doctorale en linguistique française*, 8, pp. 147-162 Available at: https://www.openstarts.units.it/bitstream/10077/10767/1/9Murano.pdf

Sajous, F., Hathout, N. & Calderone, B. (2014). Ne jetons pas le Wiktionnaire avec l'oripeau du Web ! Études et réalisations fondées sur le dictionnaire collaboratif. In: *4e Congrès Mondial de Linguistique Française.* Les Ulis: EDP Sciences, pp.663-680. Available at: https://halshs.archives-ouvertes.fr/halshs-00969260/document

Sajous F., Navarro E., Gaume B., Prévot L. & Chudy Y. (2010). Semi-automatic Endogenous Enrichment of Collaboratively Constructed Lexical Resources: Piggybacking onto Wiktionary. In: H. Loftsson, E. Rögnvaldsson, S. Helgadóttir (eds). *Advances in Natural Language Processing*, vol. 6233 of Lecture Notes in Computer Science, Springer Berlin/Heidelberg, pp. 332-344. Available at: https://hal.archives-ouvertes.fr/hal-00625326

Sajous F. & Hathout, N. (2015). GLAWI, a free XML-encoded Machine-Readable Dictionary built from the French Wiktionary. *Proceedings of the eLex 2015 conference*, Herstmonceux, England, pp. 405-426.

Sajous F., Calderone, B. & Hathout, N. (2020). Extraire et encoder l'information lexicale de Wiktionary : quel boulot pour étrangler le goulot ! Lexique 27, pp. 121-144. Available at: http://fsajous.free.fr/papers/Lexique27_2020/SajousEtAl2020_Lexique27_ExtraireInformationLexicaleWiktionary.pdf

Steffens, M., Dolar, K., & Gasparini, N. (2020). Structuration de données pour un dictionnaire collaboratif hybride. In: *Terminologie & Ontologie: Théories et Applications. Actes de la conférence TOTh 2019.* Chambéry: Presses Universitaires Savoie Mont Blanc, pp. 413-426.

Automatic Lexicographic Content Creation for Lexicographers

María José Domínguez Vázquez[1], Daniel Bardanca Outeiriño[2], Alberto Simões[3]

[1] Universidade de Santiago de Compostela – ILG, Santiago de Compostela, Spain

[2] Universidade de Santiago de Compostela – Santiago de Compostela, Spain

[3] 2Ai, School of Technology, IPCA, Barcelos, Portugal

E-mail: majo.dominguez@usc.es, daniel.bardanca@rai.usc.es, asimoes@ipca.pt

Abstract

This paper presents *Combinatoria*, a tool for the semi-automatic generation of biargumental valency patterns for nominal phrases, as well as the current development of the tool for describing the passive valency of the noun. First, we describe a set of prototypes developed as exploratory tools for this new approach, together with the lexical and syntactic resources required for the generation of nominal phrases. We will focus especially on lexical resources, their automatic retrieval, and how they assist the lexicographic team in their tasks. This is followed by a description of the tool, the data filtering process, and the presentation of the obtained results. Finally, we include a brief discussion on the usefulness of these generators not only as stand-alone plurilingual dictionaries, but also as integrated resources in other electronic tools.

Keywords: multilingual valency dictionaries; argument patterns; automatic language generation; natural language processing

1. Introduction

The cooperation between lexicography and Natural Language Processing (NLP) has shown that the availability of lexical knowledge is beneficial at different levels (Trap-Jensen, 2018). This interaction, together with new developments in language technologies and the empowerment of the user of lexicographic resources, have significantly influenced the concept of the dictionary itself[1] and the types of tasks that lexicographers must undertake (Maldonado, 2019). According to Villa Vigoni-Theses (2018), the dictionaries of the future "are lexical or linguistic information systems in which existing lexicographic data are conflated, multilingualism and linguistic variety are entrenched [...]" and an essential task for lexicography is "the orderly conflation of data which has been generated automatically by text corpora and specifically processed [...]".

Regarding this context and considering the lack of resources for describing and consulting the valency of a noun, three prototypes for automatic generation of valency

[1] To understand this typological evolution, see, for example, Engelberg & Müller-Spitzer (2013) or Boelhouwer et al. (2017).

patterns were developed to show a new concept of electronic multilingual dictionaries, in this case, automatic and more interactive valency dictionaries (Prinsloo et al., 2011). The three simulators – *Xera*, *XeraWord*, and *Combinatoria* – have been designed as independent lexicographic tools for humans, but may also be integrated into other types of resources and even exported as computational lexicons (see Section 4). The main goal is to create a multilingual platform for describing and consulting the valency of different word classes. These generators also provide an innovative methodological approach: on the one hand, they combine different linguistic theories. – such as Valency Grammar, Prototype Theory, etc. – On the other hand, they implement NLP techniques, WordNet, Wordnet-like lexical databases, and other human-made multilingual resources for automatically generating lexicographic content (see Sections 2 and 3).

This study focuses on the tool *Combinatoria* (2020), a new prototype for automatic generation of biargumental valency patterns for nominal phrases in Spanish, German and French such as "*der Tod der Mutter an Tuberkulose*", "*la muerte del padre de infarto*", or "*la mort du marié par Ébola*". *Combinatoria* is not a stand-alone product; it is closely related to i) the monoargumental simulator *Xera* (2020), whose contents are used as the basis for the generation of nominal phrases with two arguments in *Combinatoria*, and to ii) the monoargumental simulator *XeraWord* (2020) that enables the automatic creation of examples for valency dictionaries in Galician and Portuguese. Although *XeraWord* and *Combinatoria* deal with the description of different languages, the first-mentioned tool allows us to analyse the feasibility of the data access structure – based on onomasiological criteria – and to implement it in the tool *Combinatoria*. The three generators, therefore, feed on each other; not only in terms of description levels and type of linguistic data fed to them, but also share applied analysis procedures and tools (Domínguez et al., 2019). They are free and are updated constantly.

While describing the tool *Combinatoria*, we highlight the role of the applied resources, in particular the set of tools we have developed for the automatic collection and generation of lexicographic content at different stages, as well as the work of the lexicography team (Jakubíček, 2018). Different human tasks are performed to ensure the quality of the automatically gathered data and check their accuracy regarding the dictionary type before being integrated into the generators. This study shows, therefore, how some automation procedures speed up lexicographic work and allow researchers to quickly adapt and design resources.

The paper is organised as follows. Section 2 focuses on the general features of the three language generators – *Xera*, *XeraWord*, and *Combinatoria* – including their description levels as well as the tools and procedures for their development. Section 3 deals with the current state of the project and future work. Section 4 presents the user interface, together with the user's data filtering process and the output of *Combinatoria*. Section 5 suggests possible further applications of this tool in the field of lexicography.

2. The language generators *Xera*, *XeraWord*, and *Combinatoria*

In this section we discuss the three tools that have been developed for the automatic generation of nominal phrases. First, a general description is provided. This is followed up by an explanation of the different procedures implemented during the development of the generators.

2.1. General description

The three generators provide information on the slots opened by a nominal head, that is, the active noun valency. Therefore, a specific slot for a given lexical unit is described considering its syntactic-semantic interface, as well as its combining potential and syntactic-semantic preferences (Engel, 1996; 2004). In opposition to other automatic language generators (Domínguez, 2020), the final goal of the tools is to answer the question of whether a noun *A* contains in its pattern an argument *X*, what their surface realisations are, and how each of them correlates with specific semantic-ontological classes and lexical units. This is the aim of *Xera* and *XeraWord*.

	Xera	***XeraWord***	***Combinatoria***	***CombiContext***
language	es., fr., de.	gl., pt.	es., fr., de.	es., fr., de.
noun valency	active	active	active	passive
nouns	60	10	60	60
patterns	argumental	monoargumental	biargumental with phrasal context	⇨phrasal and sentence context
chronology	first	third	version[1]: second version[2]: fourth	in progress
data access	formal: patterns	conceptual	conceptual	in progress
released				-

Table 1: General description of the generators

The focus is also on the combinatory potential, i.e., describing whether an argument *X* can be combined with another argument *Y,* and what restrictions or preferences determine this combination of arguments. The tool *Combinatoria* can provide this kind of information. It enables the user to obtain examples according to different surface realisations, after selecting the specific semantic role and semantic classes[2]. A new tool is already under development – *CombiContext* – for describing the passive valency of

[2] This relational and ontological approach differentiates *Combinatoria* from databases and annotated corpora such as CPA, Framenet, PropBank or Verbnet.

the nominal phrase, which will display its relationship to other units higher in the dependency hierarchy.

Table 1 summarises the general characteristics of the designed generators. As the starting point for verifying the feasibility of the methodological proposal and, ultimately, the prototypes themselves, 20 nouns in each language have been selected as representatives of different cognitive scenes or semantic fields (Table 2).

MOVEMENT	huida-Flucht-fuite ‖ viaje-Reise-voyage ‖ mudanza-Umzug-déménagement
LOCATION	presencia-Anwesenheit-présence ‖ ausencia-Abwesenheit-absence ‖ estancia-Aufenthalt- séjour
EXPRESSION	conversación-Gespräch-conversation ‖ discusión-Diskussion-discussion ‖ pregunta-Frage-question ‖ respuesta-Antwort-réponse ‖ texto-Text-texte ‖ video-Video- vidéo
AFFECTION	muerte-Tod-mort ‖ aumento-Zunahme-augmentation ‖ dolor-Schmerz-douleur ‖ amor-Liebe-amour
CLASSIFICATION	olor–Geruch-odeur ‖ sabor-Geschmack-saveur ‖ color-Farbe-couleur ‖ (el) ancho-Breite-largeur

Table 2: Nouns selected for generation

The descriptive levels for analysing the combinatory potential and rules of a language unit are common to the three currently available generators (Table 3).

	active valency			**passive valency**
	Xera	***XeraWord***	***Combinatoria***	***CombiContext***
Only specific arguments are included in the argument pattern	✓	✓	+/-	+/-
Semantic description of the arguments: semantic roles	✓	✓	✓	✓
Semantic description of the arguments: ontological features	✓	✓	✓	✓
Syntactic function	✓	✓	✓	✓
Surface realisation	✓	✓	✓	✓
Interaction inside the nominal phrase	✓	✓	✓	✓
Interaction outside the nominal phrase	-	-	-	✓

Table 3: Descriptive levels of the generators

A concrete example of these levels with some quantitative information is shown in Table 4 for the German noun *Diskussion* (*discussion*).

Lemma	• *Diskussion* – Definition and semantic field – Gender – Number	
Quantitative	• monoargumental patterns: 23 • biargumental patterns: 78 • lexical packages: 111	
Syntactic-semantic	• determinant+{adjective}+head+*über*+determinant argument • determinant+{adjective}+head+*zwischen*+determinat+argument$_1$ *über*+determinant+argument$_2$	
Semantic	• Relational	• Semantic role – Role$_1$: someone, who discusses – Role$_2$: what is being discussed
	• Ontological	• Ontological features – Role$_1$: [animate] [human] – Role$_2$: [content] [situation]
Morphosyntactic	• Syntactic function	• subject /object
	• Surface realisation	• *über* / *zwischen*+determinant + noun

Table 4: Example of the information provided in the description

Since the properties of the nominal predicate determine the paradigm of lexical candidates that fit into a valency slot, getting and collecting these paradigmatic lexical units – or the *classe d'objets* according to Gross (2008: 11) – is key for subsequent programming. For the compilation of the lexical packages (see Section 3), it is necessary to consider that this vocabulary list must be filtered in such a way that it corresponds to the lexical units which fit into each of the argument slots of each argument for every surface realisation. Therefore, it is necessary to get and prototype a list of adequate lexical units[3] and encode their combinatorial rules and restrictions. This is dealt with in the next section.

2.2. Tools and procedures for developing the generators

This section provides a general overview of the common procedures applied as well as the tools developed or used to support the generators, relieve the workload of the

[3] To analyse and describe the syntactic-semantic interface we resort therefore to concepts such as semantic roles, ontological features, prototypical lexical units, and semantic classes (Domínguez et al., 2019; Domínguez, 2021).

lexicography team, and speed up the data compilation and revision procedures. Examples of some automation procedures will be presented in Section 3.

The steps and tools applied for developing the generators are summarised below: 1) setting the argument patterns: morphosyntactic and semantic analysis (Table 5), 2) Expansion and translation of lexical data (Table 6), 3) pre-integration into the generators (Table 7), 4) the generators themselves (see Section 4 for an example).

I. Setting the argument patterns: morphosyntactic and semantic analysis					
Goal	**Collected data**	**Tools[1]**			**Human intervention**
		External available	**Own created**	**Open access**	
Collecting the data	Frequency and valency data		PORTLEX	-	Observation and data compilation
		Sketch Engine		-	
Establishment • of the argument patterns: morphosyntactic	Morphosyntactic patterns		PORTLEX	-	Data analysis and compilation according to valency criteria
		Sketch Engine		-	
• of the semantic roles: relational meaning	Patterns with semantic roles	valency dictionaries		+	Lexical prototyping: development of an ontology and annotation of semantic classes
• of the ontological meaning	Patterns with ontological features		bottom- up ontology	+	

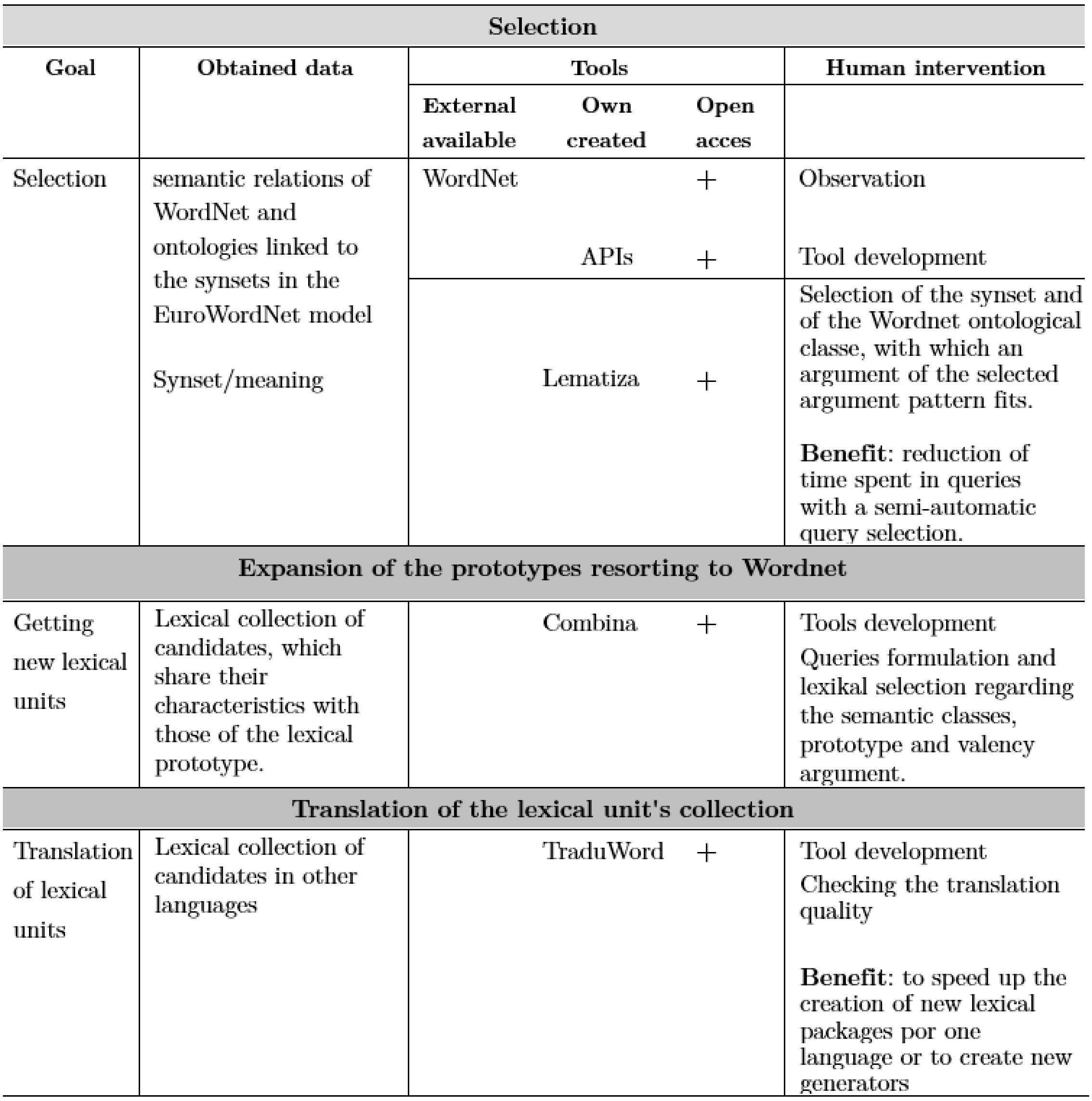
Table 5: Procedures and tools to establish argument patterns

An example of human intervention at this stage is the handling of data provided by Sketch Engine for the German noun *Diskussion* combined with a genitive case[4]. The corpora output cannot be automatically incorporated into the generators because: a) despite its high frequency, some surface realisations do not perform the function of a valency complement – for example, "Diskussion des letzten Jahrs" (*discussion of the last year*) or "Diskussion der letzten Woche" (*discussion of the last week*); b) the genitive of the noun "Diskussion" may express both those who discuss and the topic that is being discussed – for example, "Diskussion der Teilnehmer" (*discussion of the participants*) or "Diskussion der Ergebnisse" (*discussion of the results).*

This simple example illustrates that, in the first instance, frequency is not a crucial factor for selecting the lexical units that fit into a valency slot. In a second stage, frequency does indeed help us to determine lexical prototypes – lexical units that usually fit into a specific slot performing a well-defined semantic role. For example, the Argument$_2$ "what is being discussed" by [die Diskussion+determinant genitive+ Argument$_2$] in the meaning "die Diskussion einer Sache" (*discussion of something*) can

[4] The CQL query was [lemma="Diskussion"][tag="(ART\.(Def|Indef)|PRO.(Dem|Poss).Attr).Gen.*"][tag="ADJ.*"]?[tag="N.*"].

be expressed with *Ergebnis (result)*, *Thema (tema)*, *Frage (question)*, *Begriff (concept)*, *Problem (problem),* etc. We also analyse them according to general ontological features such as {content}, {situation}, etc. (for more information Domínguez, 2021; Domínguez et al., 2019). Once this is done, we are ready to undertake the next phase of the analysis: the expansion and translation of lexical data (Table 6). The aim here is to establish a controlled collection of a considerable number of lexical candidates.

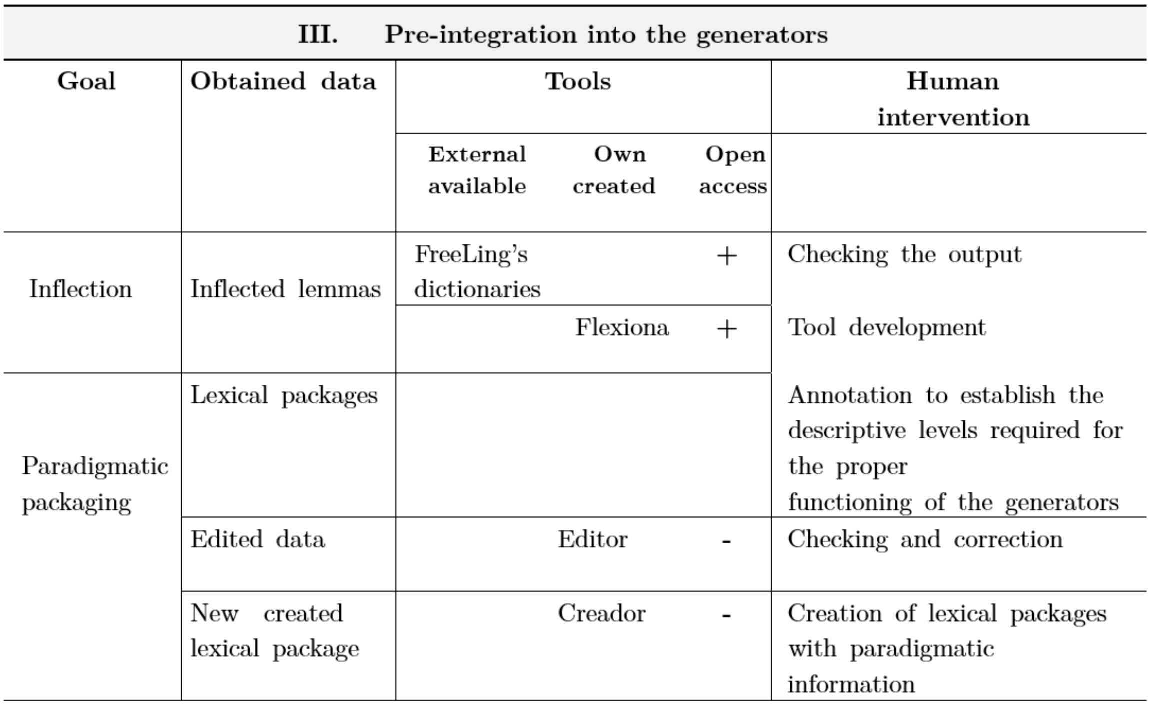


Selection					
Goal	**Obtained data**	**Tools**			**Human intervention**
		External available	**Own created**	**Open acces**	
Selection	semantic relations of WordNet and ontologies linked to the synsets in the EuroWordNet model	WordNet		+	Observation
			APIs	+	Tool development
	Synset/meaning		Lematiza	+	Selection of the synset and of the Wordnet ontological classe, with which an argument of the selected argument pattern fits. **Benefit**: reduction of time spent in queries with a semi-automatic query selection.
Expansion of the prototypes resorting to Wordnet					
Getting new lexical units	Lexical collection of candidates, which share their characteristics with those of the lexical prototype.		Combina	+	Tools development Queries formulation and lexikal selection regarding the semantic classes, prototype and valency argument.
Translation of the lexical unit's collection					
Translation of lexical units	Lexical collection of candidates in other languages		TraduWord	+	Tool development Checking the translation quality **Benefit**: to speed up the creation of new lexical packages por one language or to create new generators

Table 6: Procedures and tools to get and compile new lexical candidates for different languages

In the generators, a valency-based description of the combinatory potential of the noun with a focus on the combinatory meaning (Engel, 2004) is of indispensable value. The question here is not only to find out whether a particular ontological entity fits into a valency slot performing a semantic role, but also which concrete lexical candidates or ontological features fit into it. The expansion procedure should not be underestimated, because diverse automatically generated data is key not only when using the resource, but also for its analysis from a qualitative point of view (Hashimoto et al., 2019; Vicente et al., 2015).

Once we have collected the lexical units that meet the requirements for being integrated into the generators, the steps described in Table 7 below are taken.



III. Pre-integration into the generators

Goal	Obtained data	Tools			Human intervention
		External available	Own created	Open access	
Inflection	Inflected lemmas	FreeLing's dictionaries		+	Checking the output
			Flexiona	+	Tool development
Paradigmatic packaging	Lexical packages				Annotation to establish the descriptive levels required for the proper functioning of the generators
	Edited data		Editor	-	Checking and correction
	New created lexical package		Creador	-	Creation of lexical packages with paradigmatic information

Table 7: Pre-integration procedures and tools

Due to the granularity of the linguistic levels (see Section 2.1; Table 3), the biargumental tool Combinatoria (see Section 4) leads to a total of 9,176 syntactic-semantic argument patterns for the nouns in Spanish, German and French[5], which implies an average of 152 combined structures per noun[6], for example:

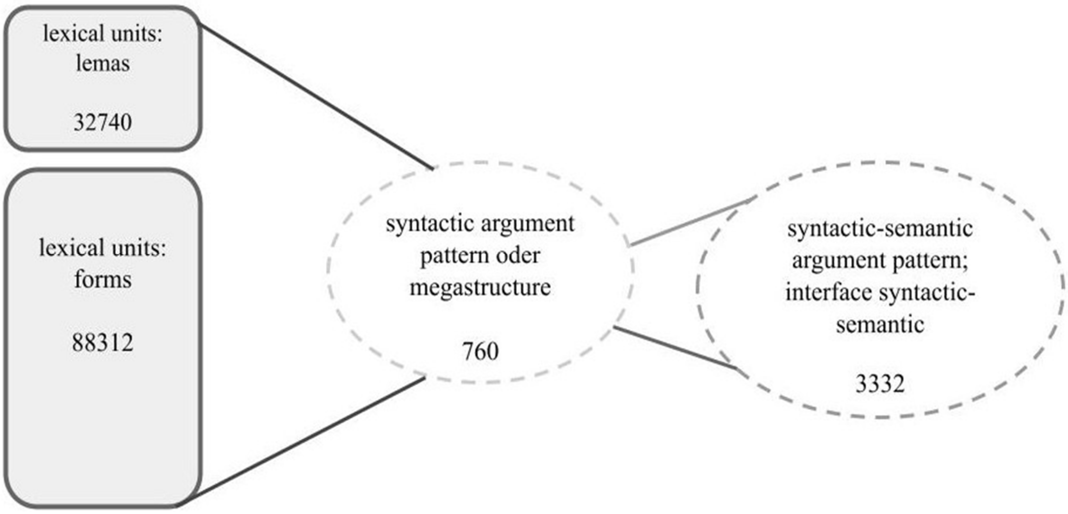
- ['determinant', 'adjective', 'head', 'determinant genitive', 'argument N1G: {human political ideology}', 'über', 'determinant accusative', 'argument N3A: {intellectual meaning}']. Ex: *die alte Diskussion mit dem Faschisten über den Begriff.*
- ['determinant', 'adjective', 'argument N3:{intellectual meaning}', 'head', 'zwischen', 'determinant dative', 'argument N1D: {collective, group}']. Ex: *die rege Definitionsdiskussion zwischen den Delegationen.*

Combinatoria relies on *Xera*, which currently has the following analysed data[7] (Figure 1).

[5] In order to improve the semantic relevance of the combined structures, FastText models (Bojanowski et al., 2017) were also implemented for each language.

[6] Data on April 5, 2021.

[7] Examples for syntactic argument pattern order megastructure are [determinant+ adjective+Diskussion+über+argument N3A], [determinant+adjective+argument N3+ Diskussion], etc. Examples for syntactic-semantic argument pattern or interface syntactic-semantic are [determinant+adjective+Diskussion+über+argument N3A: {intellectual content}], [determinant+adjective+Diskussion+über+argument N3A: {intellectual meaning}], etc. Among the lexical units, the lemmas - for example *decano* (*Dean*) - from forms such as *decano*, *decana*, *decanos*, *decanas* (*Dean*, *Deans*) are differentiated.


lexical units: lemas

32740
lexical units: forms

88312
syntactic argument pattern oder megastructure

760
syntactic-semantic argument pattern; interface syntactic-semantic

3332

Figure 1: Current analysed data as the basis for *Combinatoria*

3. Resources

The tools presented here require a set of linguistic information that describes the different lexical units used in the phrases together with its semantic information, as well as the structure at sentence level for the integration of these units in the possible nominal phrases.

Although we are dividing this section into two parts, one for the description of the lexical resources and a second for the description of the syntactic structures, it is crucial to have in mind that these resources have coupled semantic information, as we will discuss.

3.1. Lexical Resources

The lexical resources used in *Xera*, *XeraWord,* and *Combinatoria* are structured following WordNet senses (in fact, its *synsets*), and based on a custom-tailored ontology derived from WordNet ontologies (see Section 2.2; Domínguez, 2020; 2021). This approach makes it possible to create a variety of phrases with the same or similar concepts, but compiling different words to guarantee the semantic validity of the generated sentences. This aids in the process of bootstrapping data for other languages.

A lexical package (see Table 7) describes a set of related lexical units that, although not interchangeable, have a similar paradigmatic relationship. As an extremely simple example, despite their different meaning, distinct parts of the human body can be used in similar structures – “the pain in my finger” or “the pain in my head” reproduce different meanings but share a common structure.

Each one of these lexical packages includes, for each valency slot of a noun, a unique identifier, a description of the type of object that is being characterised, its classification in the ontology, and a list of lemmas. For each lemma, we link the respective Interlinguistic Index (ILI), used both in WordNet and the Multilingual Central Repository (MCR)[8].

Xera (see Section 2) started with three different languages: Spanish, German, and French. MCR does not include the French and German languages, but their wordnets were imported into the same database, and aligned using their ILI. This process allowed the creation of the original packages.

More recently, the Galician and Portuguese languages have also been included. In order to bootstrap the implementation of new languages, a set of tools were developed that help automate the translation by using WordNet and online translation services. These tools were first implemented in the development of *TraduWord* (see Table 6), which served to validate automatic translations of existing lexical packages and, therefore, to create automatic lexicographic content for lexicographers. The successful implementation of automatic translation of data circumvented the necessity to resource to raw WordNet data and subsequent debugging for every language (Domínguez et al., forthcoming). A concrete example of the implementation of *TraduWord* is the pilot tool *XeraWord,* which supports the Galician and Portuguese languages (see Section 2).

These lexical packages, while being the heart of *Combinatoria*, are useful in other contexts. Therefore, they are being codified using open standards and will be made available, independently of the online tool, in a public GIT repository.

3.2. Syntactic Resources

In the current stage of the project, we are developing a sentence generator, retroactively fed by all the previous work on simple and combined noun phrases. To successfully implement verb generation several previous steps were necessary.

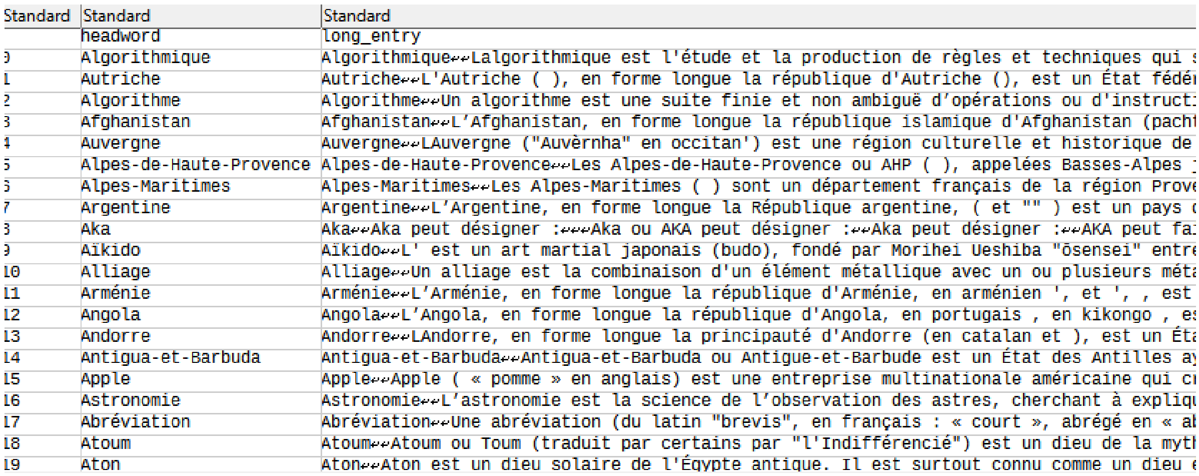
So far, the focus was on semantically filtering appropriate nouns for the combination of noun phrases. At this point, there was no verbal data available in the database of the project. To supply this information, we developed resources based on open-source projects, namely a text chunker and a PoS (Part of Speech) tagger[9] that will allow the extraction of the relevant verbs, adverbs, and adjectives related to the so-called core nouns (Table 3). In this case, all data was extracted from Wikipedia text-only dumps.

Before starting the linguistic analysis of texts from these dumps, the original XML was preprocessed. In this case, the entry per se is the only relevant text we want to feed the NLP tools. Once this has been extracted, the results are stored in a spreadsheet with two columns. This allows us to keep track of the origin of each text (column 2) by

[8] Available at http://adimen.si.ehu.es/web/MCR.

[9] The parser and tagger used are part of the NLP library Spacy: https://spacy.io/

linking it to the headword used by Wikipedia (column 1). An extract from the data is shown in Figure 2.

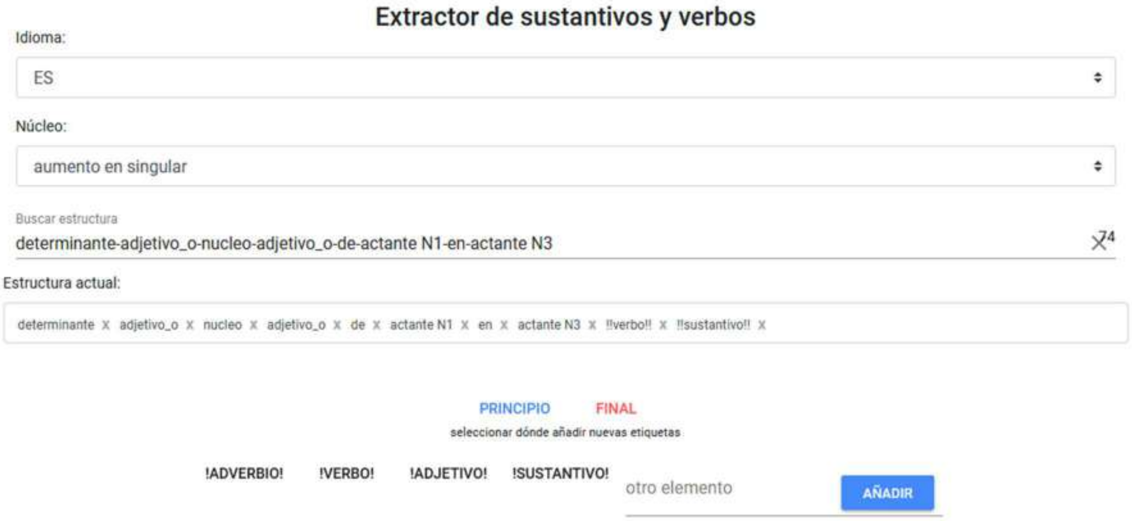


Standard	Standard	Standard
	headword	long_entry
0	Algorithmique	Algorithmique↵↵Lalgorithmique est l'étude et la production de règles et techniques qui
1	Autriche	Autriche↵↵L'Autriche ( ), en forme longue la république d'Autriche (), est un État fédé
2	Algorithme	Algorithme↵↵Un algorithme est une suite finie et non ambiguë d'opérations ou d'instruct
3	Afghanistan	Afghanistan↵↵L'Afghanistan, en forme longue la république islamique d'Afghanistan (pach
4	Auvergne	Auvergne↵↵LAuvergne ("Auvèrnha" en occitan') est une région culturelle et historique de
5	Alpes-de-Haute-Provence	Alpes-de-Haute-Provence↵↵Les Alpes-de-Haute-Provence ou AHP ( ), appelées Basses-Alpes
6	Alpes-Maritimes	Alpes-Maritimes↵↵Les Alpes-Maritimes ( ) sont un département français de la région Prov
7	Argentine	Argentine↵↵L'Argentine, en forme longue la République argentine, (et "") est un pays
8	Aka	Aka↵↵Aka peut désigner :↵↵↵Aka ou AKA peut désigner :↵↵Aka peut désigner :↵↵AKA peut fa
9	Aïkido	Aïkido↵↵L' est un art martial japonais (budo), fondé par Morihei Ueshiba "ōsensei" entre
10	Alliage	Alliage↵↵Un alliage est la combinaison d'un élément métallique avec un ou plusieurs méta
11	Arménie	Arménie↵↵L'Arménie, en forme longue la république d'Arménie, en arménien ', et ', , est
12	Angola	Angola↵↵L'Angola, en forme longue la république d'Angola, en portugais , en kikongo , e
13	Andorre	Andorre↵↵LAndorre, en forme longue la principauté d'Andorre (en catalan et), est un Éta
14	Antigua-et-Barbuda	Antigua-et-Barbuda↵↵Antigua-et-Barbuda ou Antigue-et-Barbude est un État des Antilles a
15	Apple	Apple↵↵Apple (« pomme » en anglais) est une entreprise multinationale américaine qui c
16	Astronomie	Astronomie↵↵L'astronomie est la science de l'observation des astres, cherchant à expliq
17	Abréviation	Abréviation↵↵Une abréviation (du latin "brevis", en français : « court », abrégé en « a
18	Atoum	Atoum↵↵Atoum ou Toum (traduit par certains par "l'Indifférencié") est un dieu de la myt
19	Aton	Aton↵↵Aton est un dieu solaire de l'Égypte antique. Il est surtout connu comme un dieu

Figure 2: Spreadsheet with texts from Wikipedia

The PoS tagging pipeline is then applied to these sentences. The results are reorganised and stored in a tree-like structure that allows retrieval of the data by its frequency with the relevant noun as the central element. This enables the development of a user-oriented tool that lets researchers and language learners visualise the most common PoS tags at each position, together with the most common lemmas, always considering what has already been selected. Therefore, this approach allows autonomous development of new sentences by telling the machine what the desired output structure should have.

This data, together with previously developed work from the *Xera* and *Combinatoria* tools, are currently being used for the development of sentence-capable lexical packages. The new procedure takes up from where the original noun phrase combination phase left off, and the already combined noun phrases are further developed to include verbal constructions. Any modification of the original combined structure is possible with this tool, including the complete overhaul of the elements and data to make a new verbal combination. Four main elements are presented for immediate addition to the combined structures: adverb, verb, adjective, and nouns. Manual addition of other tags already supported by the system is also possible. These tags allow manual construction of combined verbal structures through fixed patterns. The information to fill in these tags can be chosen by following the user interface, as shown in Figure 3. New noun slots may be filled with lexical data from any previous ontological item already classified. A new module is being developed to process verbal combinations that will be called after the original *Combinatoria* module (Figure 3).



Extractor de sustantivos y verbos

Idioma: ES

Núcleo: aumento en singular

Buscar estructura: determinante-adjetivo_o-nucleo-adjetivo_o-de-actante N1-en-actante N3

Estructura actual: determinante x adjetivo_o x nucleo x adjetivo_o x de x actante N1 x en x actante N3 x !!verbo!! x !!sustantivo!! x

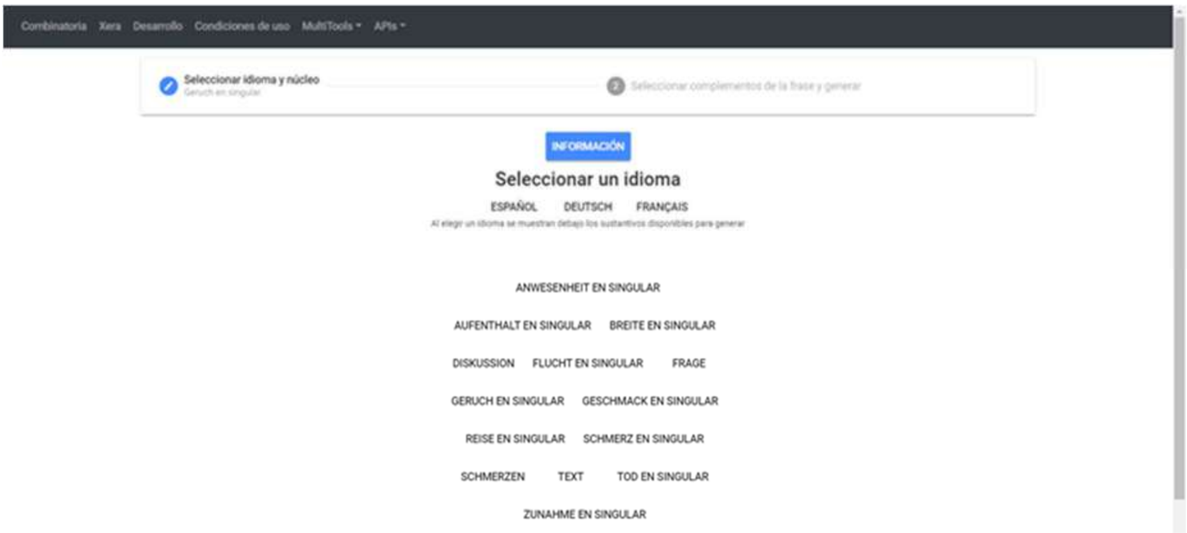
seleccionar dónde añadir nuevas etiquetas

Figure 3: Verbal combinator

Tags marked inside exclamation marks will be processed as the last step, allowing the original *Xera* and *Combinatoria* projects to remain unaltered, while still being called to process already existing tags.

4. Using *Combinatoria*

When using the biargumental tool *Combinatoria*[10] (see Section 2.1), the user must first choose a target language and noun (Figure 4). The information about its meaning and semantic field is displayed as a mouseover effect.



Seleccionar idioma y núcleo
Geruch en singular

Seleccionar complementos de la frase y generar

Seleccionar un idioma

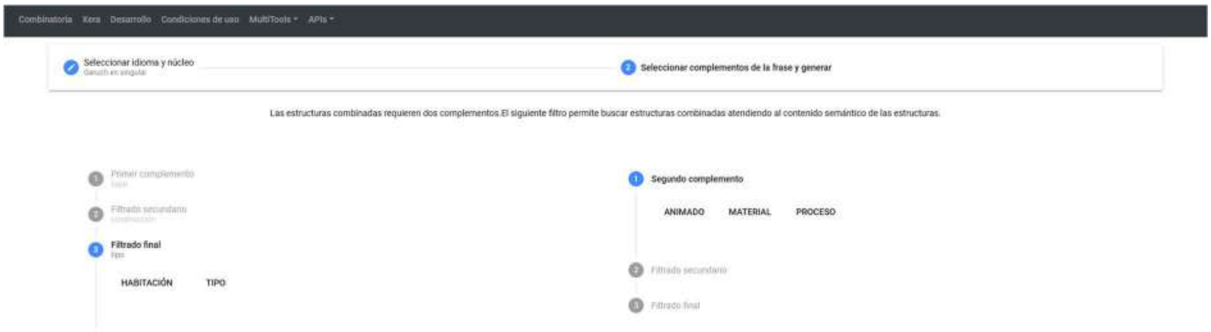
Al elegir un idioma se muestran debajo los sustantivos disponibles para generar

Figure 4: The main interface of *Combinatoira*

[10] Available at http://portlex.usc.gal/combinatoria/

Once the noun has been selected, e.g., *Geruch* (in singular) in German, the user decides which ontological-semantic feature should appear as the first argument. Let us suppose that the user selects *location – building – room* for the first argument, there are now two possible options:

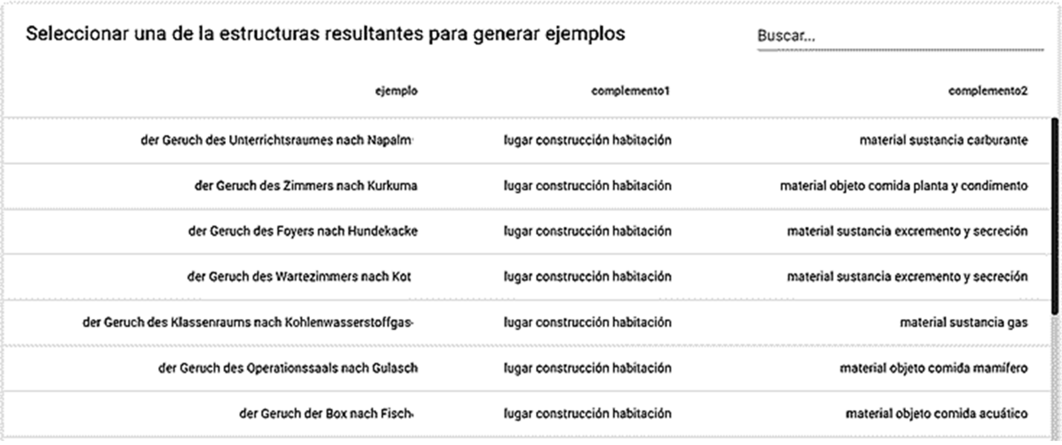
a) As with argument 1, the user tunes the search options for argument 2 in the drop-down menu on the left (Figure 5):



Las estructuras combinadas requieren dos complementos.El siguiente filtro permite buscar estructuras combinadas atendiendo al contenido semántico de las estructuras.

Figure 5: Search options: ontological approach and filtering

b) A list is displayed in the middle of the screen containing all the possible combinations that align with the already selected filter (Figure 6). On the left, the user can see an example of what would be generated when selecting a specific item. The semantic classification for each argument that will be combined is displayed on the right side.

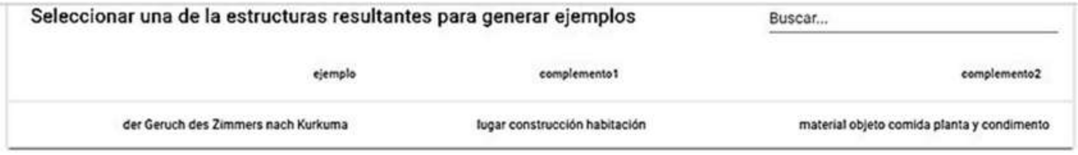


Seleccionar una de la estructuras resultantes para generar ejemplos

ejemplo	complemento1	complemento2
der Geruch des Unterrichtsraumes nach Napalm	lugar construcción habitación	material sustancia carburante
der Geruch des Zimmers nach Kurkuma	lugar construcción habitación	material objeto comida planta y condimento
der Geruch des Foyers nach Hundekacke	lugar construcción habitación	material sustancia excremento y secreción
der Geruch des Wartezimmers nach Kot	lugar construcción habitación	material sustancia excremento y secreción
der Geruch des Klassenraums nach Kohlenwasserstoffgas	lugar construcción habitación	material sustancia gas
der Geruch des Operationssaals nach Gulasch	lugar construcción habitación	material objeto comida mamífero
der Geruch der Box nach Fisch	lugar construcción habitación	material objeto comida acuático

Figure 6: Search options with examples

Continuing with this hypothetical use of the tool, if the second argument is chosen as {*material - object - food - plant - condiment*}, the possible results which are combinations of the selected first and second arguments are shown (Figure 7):



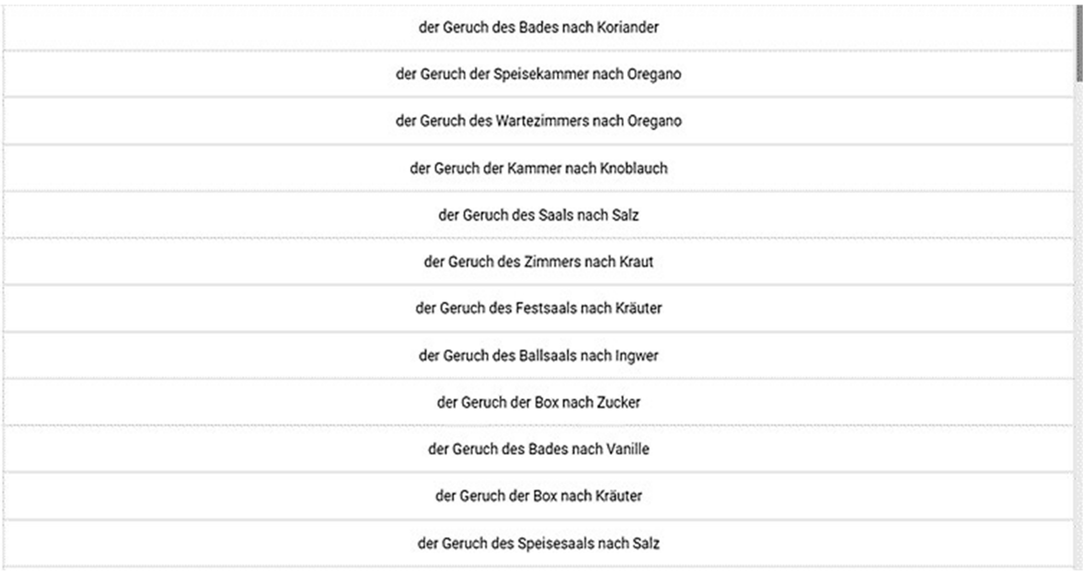
Seleccionar una de la estructuras resultantes para generar ejemplos

Buscar...

ejemplo	complemento1	complemento2
der Geruch des Zimmers nach Kurkuma	lugar construcción habitación	material objeto comida planta y condimento

Figure 7: Structure selected for example generation

Upon clicking on one of the displayed possible combinations, the examples will be generated automatically (Figure 8). It is also worth adding that the generated data follows a principle of predetermined randomness. This randomness affects the lexical representatives of each class, but not the semantic role.



der Geruch des Bades nach Koriander

der Geruch der Speisekammer nach Oregano

der Geruch des Wartezimmers nach Oregano

der Geruch der Kammer nach Knoblauch

der Geruch des Saals nach Salz

der Geruch des Zimmers nach Kraut

der Geruch des Festsaals nach Kräuter

der Geruch des Ballsaals nach Ingwer

der Geruch der Box nach Zucker

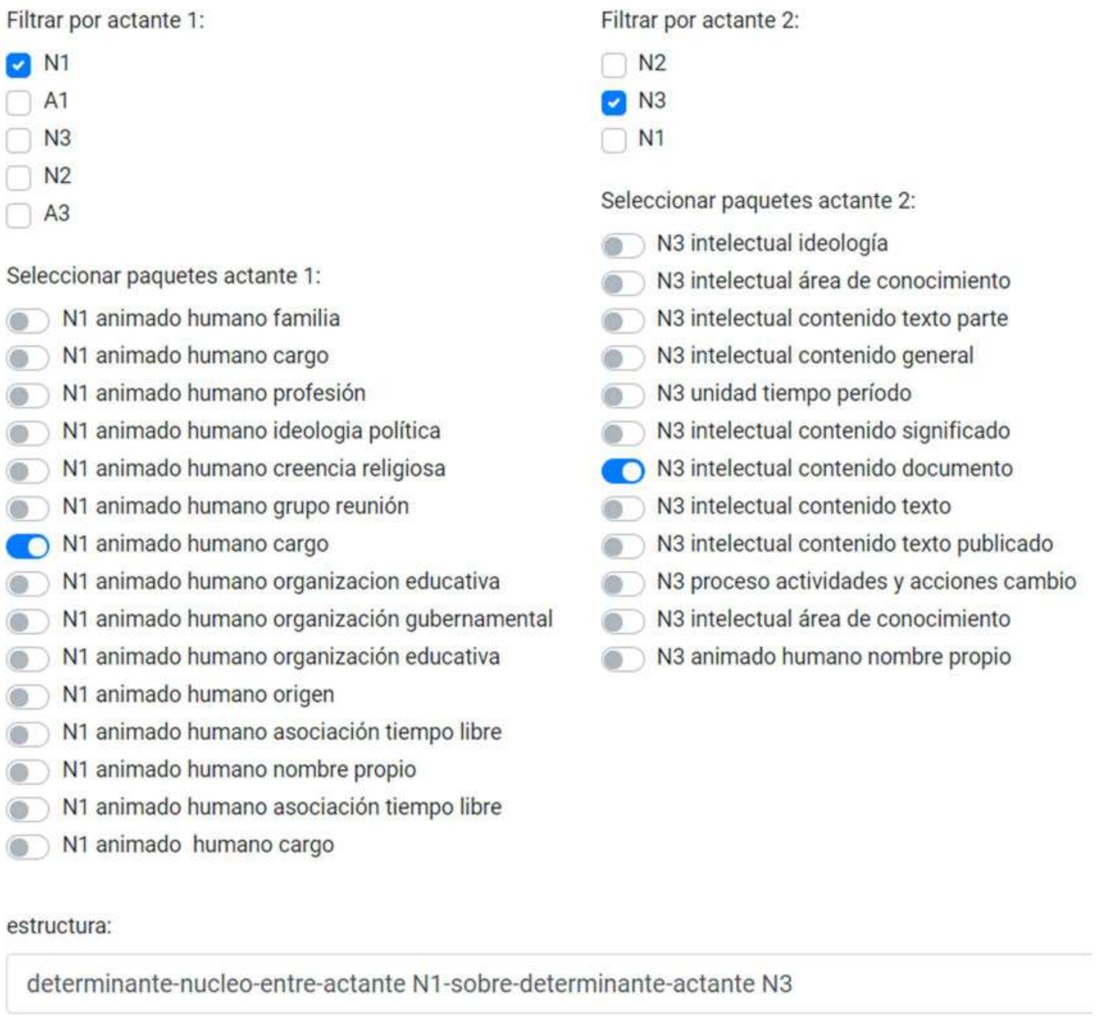
der Geruch des Bades nach Vanille

der Geruch der Box nach Kräuter

der Geruch des Speisesaals nach Salz

Figure 8: Automatic generated examples

The main novelty of the new *Combinatoria*, compared to its first version (Domínguez, 2020; Figure 9), is that it proposes conceptual onomasiological access to the argument pattern of the nominal phrases, as well as standard examples that can guide the user on the type of information that each label refers to. This approach avoids unnecessary valency terminology and formal abbreviations of roles and functions.



Filtrar por actante 1:

☑ N1
□ A1
□ N3
□ N2
□ A3

Seleccionar paquetes actante 1:

○ N1 animado humano familia
○ N1 animado humano cargo
○ N1 animado humano profesión
○ N1 animado humano ideologia política
○ N1 animado humano creencia religiosa
○ N1 animado humano grupo reunión
● N1 animado humano cargo
○ N1 animado humano organizacion educativa
○ N1 animado humano organización gubernamental
○ N1 animado humano organización educativa
○ N1 animado humano origen
○ N1 animado humano asociación tiempo libre
○ N1 animado humano nombre propio
○ N1 animado humano asociación tiempo libre
○ N1 animado humano cargo

Filtrar por actante 2:

□ N2
☑ N3
□ N1

Seleccionar paquetes actante 2:

○ N3 intelectual ideología
○ N3 intelectual área de conocimiento
○ N3 intelectual contenido texto parte
○ N3 intelectual contenido general
○ N3 unidad tiempo período
○ N3 intelectual contenido significado
● N3 intelectual contenido documento
○ N3 intelectual contenido texto
○ N3 intelectual contenido texto publicado
○ N3 proceso actividades y acciones cambio
○ N3 intelectual área de conocimiento
○ N3 animado humano nombre propio

estructura:

determinante-nucleo-entre-actante N1-sobre-determinante-actante N3

Figure 9: The user interface of Combinatoria 1.0

The primary users of our resources are foreign language learners and teachers. It should be highlighted here that the lexeme acquisition is bound up with the learning of its syntactic-semantic frame (Laufer & Nation, 2012) as well as that, in foreign language production, a considerable number of errors lie in the valency domain (Gao & Haitao, 2020; Nied, 2014; Müller-Spitzer et al., 2018).

Although we did not collect a scientifically representative amount of data on the use of these tools, some exploratory experiments with learners of German as a foreign language with A2-B1 level indicate that it takes time to understand the functioning of the tools *Xera* and *Combinatoria*. Taking into account the users' feedback, we are currently exploring the possibility of adding to the general information in the resources a step-by-step guide highlighting each required step. This will avoid unnecessary saturation of the user's interface with explanations and multiple choices. From these preliminary experiments, no preference for formal or conceptual access structure is concluded. Further studies among both learners and teachers are planned to better understand how users want to access the syntactic structures and how to improve the interface. This will also be done for the new *CombiContext* tool, described in Section 3.

5. *Combinatoria* for Lexicographic Work

As key applications of our tools (see Sections 2.2 and 3), and especially for *Combinatoria*, in the field of lexicography, we propose the following:

- As a stand-alone resource: primarily for lexicographic application, *Combinatoria* offers a verified methodological approach and serves as a prototype for further development of plurilingual valency dictionaries in other languages. To improve its usability, the number of units described in the system needs to increase in the future. The automation of analysis procedures, as well as the tools already designed (see Section 2.2) for the compilation and analysis of the lexical units and its semi-automatic translation (see Section 3) facilitates not only the integration of new languages but also the addition of lexical units for each of the prototypes. The first step in that direction has already taken place with the monoargumental tool for Galician and Portuguese *XeraWord* (see Section 3). To use the generators more efficiently in language teaching but also to develop lexicographic resources offering comparative information, it is possible to transform the generators into cross-lingual tools, similarly to the multilingual dictionary *Portlex* (2018).

- As an integrated resource into other dictionaries: It is worth highlighting the usability of the generators themselves as part of the dictionary's microstructure so that instead of static examples there would be dynamic examples, which could be selected by the user according to a specific query. Thus, the dictionary entry and the query itself are individualised.

From the point of view of the lexicographic team and their various tasks, the tools supporting the development of the generators (see Section 2.2) can streamline the human workflow for other projects on the syntactic-semantic interface, and especially in those resorting to WordNet.

6. Conclusions

A valency-based description of the combinatory potential of the noun with a focus on the combinatory meaning (Engel, 2004) is of indispensable value, especially for foreign language teaching and learning.

The question here is not only whether the particular ontological entity can (or cannot) fit into a valency slot in the rendering of a semantic role, but also which concrete lexical candidates or ontological categories can. This is the aim of the *Combinatoria* tool: to present a novel methodological approach for describing the noun valency. As valency resources themselves, the generators described in Section 2 are also innovative in that they enable an individualised selection of examples with specific ontological features as well as their generation *ad libitum.*

The integration of the generators into other lexicographic resources as well as their use as independent multilingual valency dictionaries require further automation of collection and analysis procedures. It is also important to enlarge the scope of the tool by increasing the number of units and performing studies to improve the user interface.

7. Acknowledgments

This contribution has been developed within the framework of research "Generador multilingüe de estructuras argumentales del sustantivo con aplicación en la producción en lenguas extranjeras" – MultiComb (funded by FEDER/Ministry of Economy, Industry and Competitiveness – State Research Agency/Project FFI2017-82454-P, Spain), the research "Ferramentas TraduWord e XeraWord: tradución de caudal léxico e xeración automática da linguaxe natural en galego e portugués" (financed by the University of Santiago de Compostela as part of the programme "Convocatoria de proxectos colaborativos para institutos de investigación da USC. 2020"), and partly funded by Portuguese national funds (PIDDAC), through the FCT – Fundação para a Ciência e Tecnologia and FCT/MCTES under the scope of the project UIDB/05549/2020.

8. References

Boelhouwer, B., Dykstra, A. & Sijens, H. (2017). Dictionary Portals. In P. A. Fuertes-Olivera (ed.) *The Routledge Handbook of Lexicography.* London/New York: Routledge, pp. 754 –766.

Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5, pp. 135–146.

Domínguez Vázquez, M. J., Simões, A., Bardanca Outeiriño, D., Caiña Hurtado, M. & Iglesias Allones, J. (forthcoming). Automatic Generation of Nominal Phrases for Portuguese and Galician Combining Multilingual Resources into XeraWord.

Domínguez Vázquez, M. J. (2021). Zur Darstellung eines mehrstufigen Prototypbegriffs in der multilingualen automatischen Sprachgenerierung: vom Korpus über word embeddings bis hin zum automatischen Wörterbuch. *Lexikos,* 31, pp. 1-31.

Domínguez Vázquez, M. J. (2020). Aplicación de WordNet e de word embeddings no desenvolvemento de prototipos para a xeración automática da lingua. *Linguamática*, 12(2), pp. 71-80.

Domínguez Vázquez, M. J. & Valcárcel Riveiro, C. (2020). PORTLEX as a multilingual and cross-lingual online dictionary. In M. J. Domínguez Vázquez, M. Mirazo Balsa & C. Valcárcel Rivero (eds.) *Studies on multilingual lexicography.* Berlin: de Gruyter, pp. 135-158.

Domínguez Vázquez, M. J., Solla Portela, M. A. & Valcárcel Riveiro, C. (2019). Resources interoperability: Exploiting lexicographic data to automatically generate dictionary examples. In I. Kosem & T. Zingano Kuhn (eds.) *Proceedings of the VI. eLex conference Electronic lexicography in the 21st century*: *Smart Lexicography.* Brno: Lexical Computing CZ s.r.o, pp. 51-71.

Engel, U. (2004). *Deutsche Grammatik – Neubearbeitung.* München: Iudicium.

Engel, U. (1996). Semantische Relatoren. Ein Entwurf für künftige Valenzwörterbücher.

In N. Weber (ed.) *Semantik, Lexikographie und Computeranwendung.* Tübingen: Niemeyer, pp. 223-236.

Engelberg, S. & Müller-Spitzer, C. (2013). Dictionary portals. In R. Gouws et al. (eds.): *Dictionaries. An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography.* Berlin: de Gruyter, pp. 1023-1035.

Fuertes Olivera, P., Niño Amo, M. & Sastre Ruano. A. (2019). Tecnología con fines lexicográficos: su aplicación en los Diccionarios Valladolid-Uva. *RILE. Revista Internacional de Lenguas Extranjeras*, 10, pp. 75-100.

Gao, J. & Haitao. L. (2020). Valency Dictionaries and Chinese Vocabulary Acquisition for Foreign Learners. *Lexikos*, 30, pp. 111-142.

Gross, G. (2008). *Les classes d'objets.* Paris: Presses de l'Ecole normale supérieure.

Hashimoto, T.B., Zhang, H. & Liang, P. (2019). Unifying human and statistical evaluation for natural language generation. In J. Burstein et al. (eds.) *Proceedings of the 2019 Conference of the North American Association for Computational Linguistics: Human Language Technologies.* Minneapolis: Association for Computational Linguistics, pp. 1689-1701.

Jakubíček, M. (2018). Practical Post-Editing Lexicography with Lexonomy and Sketch Engine. *XVIII EURALEX International Congress: Lexicography in Global Contexts.* Ljubljana. http://videolectures.net/euralex2018_jakubicek_sketch_engine/

Laufer, B. & Nation, P. (2012). Vocabulary. In S. M. Gass. Hrsg., Susan M. Gass & A. Mackey (eds.) *The Routledge Handbook of Second Language Acquisition.* London/New York: Routledge, pp. 163-176.

Maldonado, M. C. (2019). Las investigaciones de mercado en lexicografía comercial: un aprendizaje para el mundo académico e investigador. *RILE. Revista Internacional de Lenguas Extranjeras*, 10, pp. 101-118.

Müller-Spitzer, C., Domínguez Vázquez, M.J., Nied Curcio, M., Silva Dias, I. M. & Wolfer, S. (2018). Correct Hypotheses and Careful Reading Are Essential: Results of an Observational Study on Learners Using Online Language Resources. *Lexikos*, 28, pp. 287-315.

Nied, M. (2014). Die Benutzung von Smartphones im Fremdsprachenerwerb und -unterricht. In A. Abel, C. Vettori & N. Ralli (eds.) *Proceedings of the XVI EURALEX International Congress: The User in Focus.* Bolzano/Bozen: Institute for Specialised Communication and Multilingualism, pp. 263-280.

Prinsloo, D. J., Heid, U., Bothma, T. & Faaß, G. (2011). Interactive, dynamic electronic dictionaries for text production. In I. Kosem & K. Kosem (eds.) *Electronic lexicography in the 21st Century: New Applications for New Users, Bled.* Eslovenia: Trojina, Institute for Applied Slovene Studies, pp. 215-220.

Trap-Jensen, L. (2018). Lexicography between NLP and Linguistics: Aspects of Theory and Practice. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (eds.) *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts.* Ljubljana: Ljubljana University Press, pp. 25-37.

Vicente, M., Barros, C., Peregrino, F., Agulló, F. & Lloret, E. (2015). La generación de lenguaje natural: análisis del estado actual. *Computación y Sistemas*, 19, pp. 721–756.

Villa Vigoni-Thesen, V. (2018). Dictionaries for the Future – The Future of Dictionaries. Challenges to Lexicography in a Digital Society. https://www.emlex.phil.fau.eu/files/2019/03/Villa-Vigoni-Theses-2018-English.pdf

Dictionaries and tools

CPA = http://www.pdev.org.uk/

Combina = http://portlex.usc.gal/develop/combina.php

Combinatoria (2020) = *Combinatoria. Prototipo online para la generación biargumental de la frase nominal en alemán, español y francés.* Universidade de Santiago de Compostela. http://portlex.usc.gal/combinatoria

Flexiona = http://portlex.usc.gal/develop/flexiona.php

Framenet = https://framenet.icsi.berkeley.edu/fndrupal/

FreeLing's dictionaries = http://nlp.lsi.upc.edu/freeling/node/1

Lematiza = http://portlex.usc.gal/develop/lematiza/

Portlex (2018) = *Portlex. Dicccionario multilingüe de la valencia del nombre.* Universidade de Santiago de Compostela. http://portlex.usc.gal/portlex/

PropBank = http://verbs.colorado.edu/propbank/framesets-english-aliases/

Sketch engine = https://www.sketchengine.eu

TraduWord = https://ilg.usc.gal/gl/proxectos/interoperabilidade-de-recursos-e-producion-automatica-de-linguaxe-natura

Xera (2020) = *Xera. Prototipo online para la generación automática monoargumental de la frase nominal en alemán, español y francés***.** http://portlex.usc.gal/combinatoria/usuario

XeraWord (2020) = *XeraWord. Prototipo online de xeración automática da argumentación da frase nominal en galego e portugués***.** http://ilg.usc.es/xeraword/en/

Verbnet = https://verbs.colorado.edu/~mpalmer/projects/verbnet.html

WordNet = https://wordnet.princeton.edu

Catching lexemes. The case of Estonian noun-based ambiforms

Geda Paulsen[1,2], Ene Vainik[1], Ahti Lohk[1], Maria Tuulik[1]

[1] Institute of the Estonian Language, Roosikrantsi 6, Tallinn 10119, Estonia

[2] Uppsala University, Thunbergsvägen 3 L, Uppsala 75126, Sweden

E-mail: {geda.paulsen, ene.vainik, ahti.lohk, maria.tuulik} @eki.ee

Abstract

The aim of this study is to test a statistic relying on corpus data, the distributional index (D-index): a statistical benchmark that helps lexicographers judge if a morphological form has been conventionalised to the degree of becoming an independent lexeme. Our focus is on the decategorisation type that originates from a case form of a noun and is directed to an adverb, adposition or adjective. The words or inflected forms corresponding to more than one word class interpretation are in this study termed ambiforms. The analysis compares the D-index levels of ambiforms categorised as nouns and another PoS. The results suggest that for the outcome to be most authentic, the noun-based ambiforms should be analysed without the decategorisation influence, i.e. the D-index analysis should be applied in the pre-PoS-disambiguation stage.

Keywords: form distribution; morphology; lexicography; language technology; Estonian

1. Introduction

An electronic dictionary striving to depict contemporary vocabulary needs to be updated constantly due to the changes that take place in the actual usage of language. Estonian lexicography is developing towards unification of lexical resources (dictionaries and term bases) into a central “super-dictionary”, the EKI Combined Dictionary (CombiDic), with the Ekilex dictionary writing system as its backbone, and lexicographic processes are moving towards a higher degree of automation. (About the recent developments regarding Estonian lexicographic resources, see Tavast et al., 2018; Tavast et al., 2020; Kallas et al., 2020.) Besides monitoring the most recent corpora for neologisms (Langemets et al., 2020), tracking and identifying the degree of grammaticalisation and lexicalisation of existent word forms are essential to attain an adequate overview of language development.

To be able to make a well-grounded decision about a new lexeme candidate, lexicographers need more fine-grained processing of corpus data than simple word frequencies (Paulsen et al., 2019). Blensenius & Martens (2019) argue for the use of word-form relative frequency information derived from existing corpora to improve dictionary content. When it comes to tracking morphological decomposition processes, Hay (2001) states that relative frequency is more elucidative than absolute frequency.

As a solution for capturing decategorising noun forms in Estonian, we suggest a specific statistic predicting a form's degree of salience: the distribution index (D-index). The D-index (DI) calculates the distributional value of nominal case forms as compared to the norm-based relative frequencies of the case forms (Vainik et al., 2021). The aim of this study is to ascertain whether the D-index enables one to detect forms emerging as potentially independent lexemes.

This article is the second report on our ongoing study of the D-index. In our earlier paper, we described the development of the index and tested it on a sample (N = 46) of Estonian noun-based ambiforms (words or inflected forms corresponding to more than one word class interpretation) in 11 (semantic) cases (Vainik et al., 2021). The results were compared to a control group of "ordinary" nouns (N = 26) with an abundant range of case forms displaying a regular distribution of case form frequencies. As a result of this study, we determined the threshold value of the distribution index as an indicator of heightened frequency.

In the present study, we tested the threshold value on a selection of noun-based declined forms that can be expected to be situated at some point in the decategorisation process. Our focus is hence particularly on morphology-based PoS change, i.e. the decategorisation type that originates from a case form of the noun and is directed to an adverb, adposition or adjective (for more about possible PoS combinations in Estonian, see Vainik et al., 2020)[1].

The data for the analysis of noun-based ambiforms were derived from the database of Estonian ambiforms, consisting of approx. 3,500 examples (see Vainik et al., 2020). We will calculate the D-indices of the selected noun-based ambiforms and consider the usability prospects of the distributional identification of case forms. Our main research questions are: Does the threshold of heightened frequency (Vainik, Paulsen & Lohk 2021) capture a form's movement to the status of an independent lexeme? Is it possible to establish other thresholds? What is the impact of corpus preprocessing on the results, i.e. automatic morphological tagging and PoS disambiguation, proceedings that are supported with data from the CombiDic? Can the D-index help to improve corpus tagging systems?

We will begin with a short overview of Estonian nominal morphology and the decategorisation processes related to case endings in Section 2. The methods and data used in the study – the D-index and its calculus, the corpus processing methods, and the data and data processing procedures – are explained in Section 3. Section 4 is devoted to the analysis and discussion of the DI levels of ambiforms with different lexicographic statuses and the effects of the principles of corpus annotation on DI calculations. Section 5 summarises and discusses the results.

[1] The operating of the D-index in practice is described in detail in Vainik, Lohk & Paulsen (2021, this issue).

2. The Estonian case system and inflectional decategorisation processes

In terms of their morphological behaviour, Estonian words can be divided into four main classes: (1) words that can be inflected for mood, time and person (verbs), (2) words that can be inflected for all cases (nominals), (3) words that have no grammatical case forms (some adverb types and some adpositions), and (4) words that have no inflectional forms (some adverb types and adpositions, conjunctions and interjections (Viitso 2003, 32). The Estonian nominals, i.e. nouns, adjectives, numerals and pronouns (and certain participles and infinitives) are inflected for number (singular (SG) and plural (PL)) and case. The semantic cases have functions similar to prefixes or suffixes in many other languages (ibid.). There are three grammatical cases – nominative (NOM), genitive (GEN) and partitive (PART) – and 11 semantic or adverbial cases: illative (ILL), inessive (INE), elative (ELA), allative (ALL), adessive (ADE), ablative (ABL), translative (TRA), terminative (TER), essive (ESS), abessive (ABE) and comitative (COM). (Ibid 32)

In Estonian, the decategorisation processes involving morphological forms are a considerable source of word-class fluidity: common nouns in a (usually semantic) case form may undergo PoS-shift into function words (mainly adverbs and postpositions). The development of nominal case forms into adverbs (or adpositions) is a characteristic feature of Estonian (Grünthal 2003; Karelson 2005; Habicht, Penjam & Prillop 2011). The adverbisation of Estonian nominal case forms can be seen as a type of lexical conversion (Kasik 2015: 40): a (more or less regular) word-formation process. An example of such a process is the adverb *tasuta* 'gratis, without fee', the abessive case form (expressing lack or absence of the noun it is attached to) of the noun *tasu* 'fee' (1):

(1) *tasu* 'reward, pay' > *tasu-ta* [reward-ABE] 'without reward, pay' > *tasuta* 'gratis'

The language internal forces behind morphosyntactic changes are in linguistics approached via two basically opposite notions: grammaticalisation and lexicalisation. While grammaticalisation reflects the development of a lexical item into a marker of a grammatical category (see e.g. Heine & Kuteva 2007, 34), lexicalisation involves a process that adds words with specific content-filled meanings to a language's lexicon (Brinton and Traugott 2005: 18). Both processes influence the natural changes in the lexicon that lexicographers need to observe to give an accurate description in a dictionary.

In our synchronic study of inflectional forms that stand out statistically from the regular frequency patterns, certain grammaticalisation paths of nouns as content words to a function-word usage are observable (> adjective; > adverb; > adposition). There is, however, also the question of a morphological form becoming an independent lexical item, an autonomous dictionary entry[2]. Since the aim of this study is not to give a theoretical explanation of the particular changes behind the (miscellaneous) group of noun-based ambiforms, or to define the stages of grammaticalisation paths, we use the umbrella term *decategorisation* to refer to categorical changes in nominal ambiforms[3].

3. Methods and data

3.1 The distribution index and its formula

The question lexicographers face when analysing a form separating from its lemma is basically: How frequent is frequent enough to establish the form as an independent lexeme? This question is clearly relative: just as the absolute frequencies of lexemes vary, particular forms can also be expected to display different (relative) frequencies. We propose a statistical measure of such relative frequency – the distribution index (DI) – which indicates whether the frequency of a word form fits its normal distribution as a noun form or deviates from it.

The idea behind such an index lies in the assumption that proper nouns tend to have constant distributions along with the case forms (combinations of number and case, e.g. plural elative and singular abessive) in the corpora. If such a constant normal distribution holds, it is possible to predict the frequencies of word forms based on their lemma frequencies. The very idea of the DI is to compare the actual (observed) frequency of a case form in a corpus with its expected frequency. The values of expected and observed frequency should be equal or close as long as the studied form follows the normal distribution. If there is a considerable difference between the values of expected and observed frequencies, one can conclude that the distribution is abnormal.

The hypothesis of constant distribution of word forms was controlled for in a study where the distribution data of case forms from two annotated corpora (the Balanced Corpus of Estonian[4] and the Morphologically Disambiguated Corpus[5]) were compared (Vainik et al., 2021). The distribution of all of the case forms (i.e. 29 combinations of number and case) demonstrated very steady proportions in both corpora (r = 0.999; StDev 0.000). We established these constant proportions of case forms as norms and used them as the basis for calculating the distribution indices (ibid.; Vainik et al., Paulsen 2021).

[2] For a discussion on such forms and their lexicographic status, see Paulsen et al. (2020).

[3] Note that decategorisation of morphological forms is also observable in languages without extensive case morphology, e.g. the plural form of nouns in Swedish (e.g. *blomma* 'flower' > *blommor* 'flowers', see Blensenius & Martens 2019).

[4] https://www.cl.ut.ee/korpused/grammatikakorpus/

[5] https://www.cl.ut.ee/korpused/morfliides/

The DI is calculated according to the following formula:

$$\mathrm{DI} = (\mathrm{Z} - \mathrm{X} \times \mathrm{Y}) \,/\, \mathrm{X}$$

Z = the observed frequency of the word form

Y = the norm of that particular case form (taken from a table of such norms)

X = the frequency of the lemma.

The expected frequency of a word form is calculated as a product of the frequency of the lemma X and the norm of that particular case form (Y). The result of the comparison should be normalised, i.e. the subtraction divided by the frequency of the lemma.

The values of the DI can (theoretically) vary from nearly −1 to 1. Values close to zero indicate normal distribution, and negative values indicate that the word form is underrepresented compared to its expected frequency. Values above zero indicate that the word form occurs more frequently than expected by the norm. On a few occasions, the value can be as high as 0.9, which indicates that the frequency of the lemma and the frequency of case forms are very close: the word occurs mostly in a certain case form. This is a situation far from the normal distribution and such cases can be classified as autonomous or emancipated word forms. These words lack the normal paradigm and can be labelled as uninflected.

In an empirical study that compared the DI of normal case forms and ambiforms, we were able to establish a tentative threshold of DI = 0.130. Values equal to or greater than this clearly show abnormal distributions (Vainik et al., 2021). Values higher than about zero but lower than the threshold show moderate deviation from the normal distribution. Overall, four intervals/ranges can be defined for the stages of DI values (ibid.):

underrepresentation:	−1 ... −0.5
normal distribution:	−0.04 ... 0.04
moderate overrepresentation:	0.05 ... 0.129
critical overrepresentation:	0.13 ... 1

The advantage of the DI is that its values do not depend on the size of the corpus or the position of the lemma or word form in a list of frequencies. The index shows only whether the frequency of the word form follows the normal distribution as a case form in the selected corpus. As a benefit, the behaviour of both rare and frequent word forms can be measured on the same scale of relative frequency. As a result, the DI has the potential to function as a useful heuristic in certain stages of lexicographic work, i.e. when the status of a lexeme as a headword is estimated.

3.2 The corpus and its automatic processing

The study of the distributional index of nominal ambiforms is based on the largest corpus of contemporary Estonian, the Estonian National Corpus 2019, with 1.8 billion tokens[6]. The ENC2019 is lemmatised, tagged and disambiguated with the EstNLTKv.1.6 toolkit (Laur et al., 2020). The EstNLTK[7] is a natural language toolkit targeted explicitly for the Estonian language. The structure of the toolkit is written in the Python programming language and executes basic NLP tasks: tokenisation, morphological analysis (MA), lemmatisation, named entity recognition, etc. (Orasmaa et al., 2016: 2460).

In the case of a morphologically rich language such as Estonian, where different forms may have identical phonological shapes[8], the role of morphological disambiguation (rule-based, probabilistic or neural) is significant for frequency results. The result of the DI analysis hence directly reflects the outcome of the MA analysis, which in the case of Estonian starts with morphological segmentation and proceeds to PoS annotation. The current MA proceedings are based on the Vabamorf analyser, which combines rule-based and statistical models. Its lemmatisation system is mainly a dictionary-based approach, also featuring the Hidden Markov Model for disambiguation of ambiguous output. The problem with this approach is the lack of accuracy and precision with rare words that are not covered by the rules. (see Milintsevich & Sirts, 2020: 158—159.) Particularly problematic is the analysis of grammaticalised and lexicalised words or forms when the morphological tagging of lemmas and PoS is based on an unrenewed dictionary (Koppel, 2020: 59).

The Vabamorf lexicon is incorporated into the EstNLTK toolbox via the Vabamorf morphological analyser. The common ancestor of the Vabamorf lexicon and the morphological database of the Estonian language (MAB) is Ülle Viks's *A Concise Morphological Dictionary of Estonian* (1992). The inflectional patterns of Estonian words are centralised into MAB, which serves all datasets (including the CombiDic) in the dictionary writing system Ekilex[9], the centre to which the databases of the Institute of Estonian language are aggregated (Koppel et al., 2019; Kallas et al., 2020; Tavast et al., 2020).

The primary difference between the Vabamorf lexicon and the MAB is the emphasis

[6] The ENC2019 corpus contains texts collected from various domains. It consists of the Estonian Reference Corpus (texts from the 1990s until 2008 compiled by Tartu University), the Estonian Web (2013, 2017 and 2019), Estonian Wikipedia (2017 and 2019) and Estonian DOAJ (2020). The last data were crawled at the beginning of the year 2020. The ENC2019 is accessible via the Sketch Engine interface (Kilgarriff et al., 2004) at www.sketchengine.eu/ (accessed 24 March 2021).

[7] The EstNLTK toolkit is available at https://github.com/estnltk/estnltk.

[8] An example of form homonymy between nominal and verbal forms in Estonian: *viis* 'five' vs. *viis* 'brought'.

[9] https://ekilex.eki.ee/ (accessed 2 April 2021)

on either formalised morphological rules or lexicographic information. Both systems use both dictionaries and rules, although Vabamorf is focused on rules and the MAB compiles dictionary information that contains morphological paradigms in many different languages, along with frequencies and pronunciations. Moreover, the MAB does not separately perform morphological analysis.

The EstNLTK is used to parse the Estonian National Corpora (the data source in this study) and "Vabamorf is the EstNLTK's brain, heart and liver" (Indrek Hein, personal communication), meaning frequency values of forms are derived from Vabamorf. Updates to Vabamorf's lexicon are made on a daily basis and are immediately available for developers to use in the analyser and for broader use when the creator of Vabamorf, Heiki-Jaan Kaalep, officially updates Vabamorf (this information is based on personal communication with the EKI software developer Indrek Hein).

The Vabamorf analyser gives the rate of correct analyses for at least 97% of the words in texts and produces a list of analyses without the correct analysis for approx. 0.4% of words (Kaalep & Vaino, 2001). Veskis & Liba (2010) report the average accuracy of the morphological disambiguator in the standard 10-fold cross-validation test on the Morphologically Disambiguated Corpus as 96.23%. However, as Jakubíček (2021) points out, PoS tagging (a task depending directly on the morphological analysis) is an NLP task that is poorly evaluated, and its accuracy is conventionally reported on the token level[10], which only gives about 50% sentence accuracy.

In addition to the Vabamorf toolkit, neural models of morphological tagging and disambiguation are currently under development for Estonian. These models, trained on the Universal Dependencies (UD) corpus, have already achieved significant results (see e.g. Tkachenko & Sirts, 2018); however, they are not available for users yet (Kairit Sirts, personal communication). A comparison of DI results based on different morphological analyses would be an interesting task for future research.

3.3 The data and procedures

The data for the analysis of the statistical distribution of ambiforms, i.e. word forms with ambiguous status in respect to their qualification as dictionary headwords and/or their PoS affiliation, derive from a database of approx. 3,500 such ambiguous lexical items[11]. The database is organised into ambitypes according to particular PoS combinations (Vainik et al., 2020). This study focuses on the semantic case forms of nouns (see examples in (2a)); the singular partitive is included because of its participation in a semiproductive construction of "parametric words" (Sahkai, 2008: 173–174; see (2b)):

[10] See https://aclweb.org/aclwiki/POS_Tagging_(State_of_the_art)

[11] At the moment, the database of ambiforms is available at https://drive.google.com/file/d/1ZEchvhupJ_1qS48nFTzSAmkKE_vUsmBJ/view

(2a)	*pilves* [cloud-INE]	'cloudy; stoned'	adverb/adjective
	kõrval [ear-ADE]	'next to'	adverb/adposition
	huvides [interest-PL-INE]	'in the interest of (smb.)'	adposition-like case form
	hetkeks [moment-TRA]	'for a moment'	adverb-like case form
	plussmärgiga [plus.sign-COM]	'positive'	adjective-like case form

(2b) *mõõtu* [size-PART] 'size of (something)' adposition-like case form

The selected test set comprises 965 ambiforms (i.e. roughly one-third of the registered records in our database). The number of possible interpretations of those forms is 2,021 in our initial data table, because each ambiform is associated with at least two PoS affiliations. The data table of ambiforms and their possible interpretations (in terms of PoS and case form) was provided with data on the frequencies of the actual occurrences of their different interpretations in the corpus (ENC2019). The frequency data of a word form and its potential lemma were needed as source data for calculating the DI values.

To generate the summary data table of the DI values for the selected ambiforms, we created an application written in the Python programming language. The input data table (MS Excel) consists of three columns: the first column contains the ambiform, the second the part-of-speech symbol, and the third indicates the morphological form (number + case), if applicable. For each input data triplet (ambiform, part-of-speech and morphological form), an automated HTTP request was made to the text corpus ENC2019 via the Sketch Engine[12] platform. In the DI calculation, we relied on normal distribution rates of the word form and the DI formula. The obtained statistical information, calculated DI and input data were written to a new Excel file.

The results table displays the values of the DI formula components: the absolute frequency of the assumed lemma of the ambiform (X), the frequency of the particular ambiform (Z) and the norm value (Y) for the particular case form of the input ambiform. A label indicating the DI interval was attached to the table, too. The summary table also provides information about the results of automatic morphological analysis in terms of which lemmas in which forms were recognised in each particular case. This additional information provides insights into whether an ambiform has just a single interpretation or if there are possibly several interpretations available: a factor affecting the outcome of DI calculations (see section 4.1 below).

The main data table was further provided with information about the current lexicographic statuses of the ambiforms in the CombiDic (and its underlying database

[12] https://www.sketchengine.eu/

Ekilex[13]), involving three options:

- an ambiform is not included in the dictionary, yet. For this group, we use the label **"Candidates"** in the analysis in Section 4.
- an ambiform is included as a headword but the entry gives no information about its PoS. This group is labelled as **"Underspecified"**.
- an ambiform is included in the CombiDic as a headword and provided with PoS label(s) other than noun, i.e. the decategorisation process has been completed and the form has been approved as an autonomous lexeme. This group is called **"PoS-tagged"**.

4. Applying the D-index to noun-based ambiforms with different tagging statuses in EstNLTK and the CombiDic. The results and influencing factors

The automatic analysis of the ENC2019 corpus reveals that there is not necessarily any correspondence between the lexicographic lexicon (MAB) and the basis for the morphological analysis of EstNLTK, the Vabamorf lexicon (see the description of the interrelations between the different lexicographic and corpus analysing devices in section 3.2). When a case form of a noun has been reinterpreted as an indeclinable word (an adverb, adposition or indeclinable adjective) in the Vabamorf lexicon, the corpus tagging system is forced to "decide" whether to tag a running word in the corpus as a noun or as another part of speech. The result is that if a word form has risen to the status of a dictionary headword (e.g. *kõrval* [ear-ADE] 'next to'), the statistics on its occurrences in a text corpus will be split, too. The discrepancy in PoS-tagging between the CombiDic and the Vabamorf lexicon may be caused by differences in the lists of indeclinable words or the lexicon for the ambiforms with dynamic lexicographic status has not been updated.

In the analysis below, we take advantage of the mismatches in these databases and focus on the noun-based ambiforms from two general angles: (1) cases where the morphological analyser does not tag the ambiforms already decategorised in the MAB with a PoS other than S, and (2) cases where the ambiforms lack a PoS tag but have the status of a headword in lexicographic practice (i.e. in the CombiDic and, accordingly, also in the MAB), and those ambiforms that have no dictionary headword status. Discrepancies in the Vabamorf lexicon and the CombiDic offer the opportunity to study the effect of official decategorisation (interpretations of an ambiform as a noun vs. multiple PoS) on the DI of noun-based ambiforms. In the following, we examine the noun-based ambiforms from two perspectives: corpus processing analysis (4.1) and lexicographic treatment (4.2).

[13] We thank Arvi Tavast for conducting the query on the Ekilex database.

4.1 The impact of morphological analysis and PoS disambiguation on the D-index

In this section, we focus on a set of clearly decategorised ambiforms that are marked as indeclinable headwords in the CombiDic (N = 192), i.e. all of these forms have headword status confirmed with a PoS other than a noun. Some examples of the "PoS-tagged" ambiforms with their DI-values are presented in (3):

(3)	*tasuta*	(DI 0.69)	[fee-ABE]	'free of charge'	(adverb)
	kraesse	(DI 0.24)	[collar-ILL]	'upon smb'	(adverb, adposition)
	süles	(DI 0.37)	[lap-INE]	'in arms'	(adverb, adposition)
	käpas	(DI 0.04)	[paw-INE]	'mastered'	(adverb)

An interesting subset of this group is 51 ambiforms that are still analysed only as case forms of nouns by EstNLTK without alternative interpretations. These ambiforms can be accounted for as the best examples of nouns in the process of decategorisation (a process completed for these forms in the CombiDic and not started yet in the Vabamorf lexicon). The DI analysis of this group allows us to test the previously established threshold value (0.130) of distinctly independent lexemes (see Vainik et al., 2021): the fully decategorised case forms should demonstrate DI values clearly above the threshold. We refer to this small group of ambiforms with discrepant PoS statuses as "Noun".

As a comparison set, we present a group of "PoS-tagged" ambiforms with split PoS analyses (N = 141) to reveal the effects on DI values caused by decategorisation and splitting of the interpretations (nouns and some alternative PoS). These forms may be tagged with several PoS tags by EstNLTK, e.g. *lambist* (noun or adjective) [lamp-ELA] 'randomly', *asjata* (noun, adverb or adjective) [thing-AB] 'pointless'), or the forms may have alternative interpretations as case forms of the same noun due to homonymy (e.g. the forms *mõõtu* [size-ADT] and [size-PART] 'size of' coincide). Therefore, the number of calculated DI is larger (178) than the number of ambiforms in this group (141). We use the label "Noun+" to refer to this group. The DI values in the second group should be lower on average, i.e. fewer items should exceed the threshold of heightened frequency.

The DI results of the two "PoS-tagged" ambiform groups "Noun" and "Noun+" are depicted as a box plot graph in Figure 1. Table 1 presents the descriptive statistics about the compared groups.

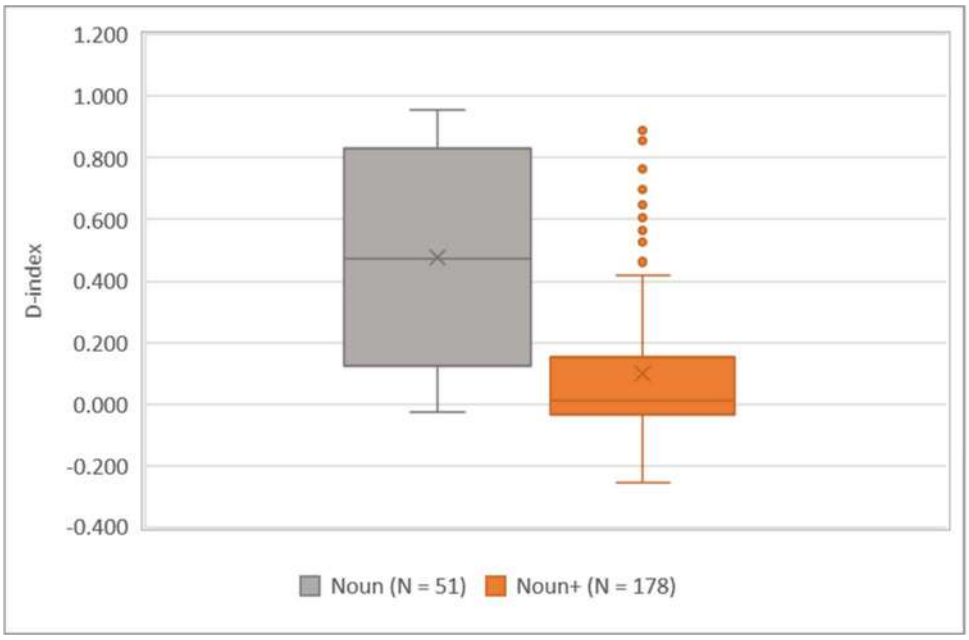


Figure 1: The variation of DI values of the noun-based ambiforms tagged as case forms of nouns ("Noun") and as other PoS in addition to nouns by EstNLTK ("Noun+")

	"Noun"	**"Noun+"**
N	51	178
Max	0.958	0.889
Min	−0.026	−0.256
Median	0.461	0.013
Ave	0.465	0.098
StDev	0.349	0.224

Table 1: Descriptive statistics of "Noun" and "Noun+"

Regarding the threshold level (0.130) established in our previous research (Vainik et al., 2021) distinguishing the forms with critically higher levels of relative salience from those following normal distribution rates or from those overrepresented moderately, the results of the respective samples ("Noun" and "Noun+") show distinct tendencies. The median of "Noun" (0.461) is 35 times higher than the median of the "Noun+" group, and the average value of "Noun" (0.465) exceeds the average of "Noun+" by a factor of 4.7. Outside the boxes, the "Noun+" group shows a noticeably larger variation array, as well as extreme outliers over the upper quartiles as "abnormal" cases in respect to the limitations set by the whiskers. The variability of "Noun" is restrained by the limits of whiskers.

Figures 2 and 3 below present the DI of both groups as dot charts in a descending order. We have highlighted the values closest to the threshold on both diagrams.

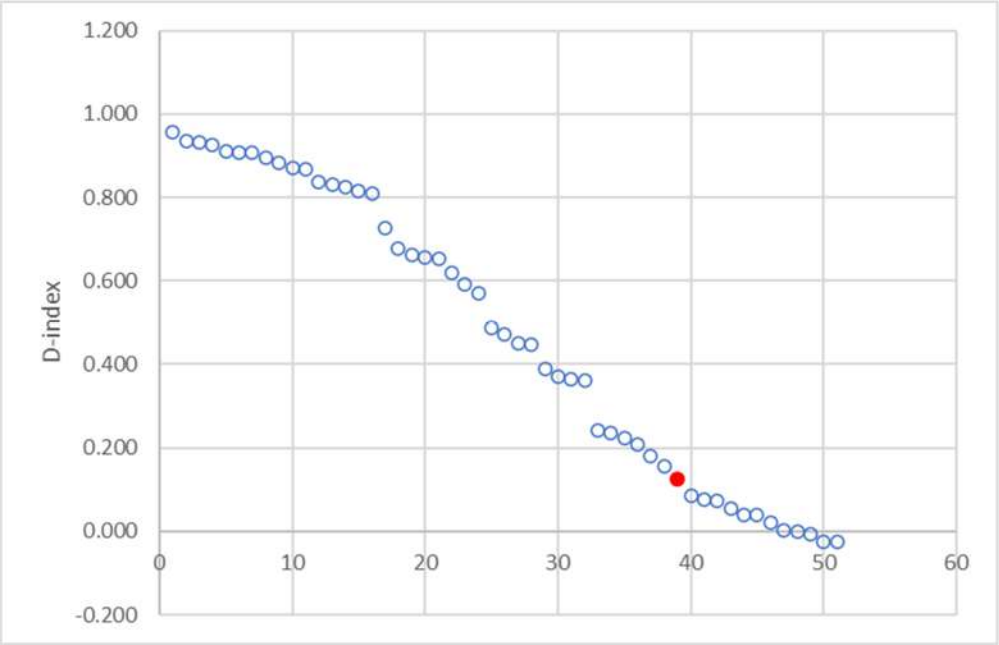


Figure 2: The DI values of the group "Noun" (the 51 "PoS-tagged" ambiforms identified only as case forms of nouns by the EstNLTK morphological analyser)

In Figure 2, we have highlighted the value 0.123 (for word form *lademes* [stratum-INE] 'loads of') as closest to the threshold (0.130). It appears that 75% of the ambiforms in the group "Noun" have indices above the threshold level. This result meets our expectation that the threshold value reveals most of the fully decategorised ambiforms.

The ambiforms with the highest DI values appear to be mostly compounds (see (4)), but there are also forms of some simple words (see (5)).

(4)	*otseloodis*	(DI 0.95)	[straight.level-INE]	'in a straight line'
	eesotsas	(DI 0.93)	[front.end-INE]	'leading'
	erandkorras	(DI 0.93)	[exception.time-INE]	'as an exception'
	üldjuhul	(DI 0.93)	[general.incident-ADE]	'in general'
	eestvõtmisel	(DI 0.91)	[front.taking-ADE]	'on the initiative'
	südametäiega	(DI 0.9)	[heart.whole-COM]	'angrily'
	teosammul	(DI 0.89)	[snail.step-ADE]	'at a snail's pace'
	esirinnas	(DI 0.87)	[forefront-INE]	'in the front lines'
	ahvikiirusel	(DI 0.87)	[monkey.speed-ADE]	'lightning fast'
(5)	*vahendusel*	(DI 0.9)	[medium-ADE]	'via'
	hetkel	(DI 0.62)	[moment-ADE]	'at the moment'
	baasil	(DI 0.45)	[basis-ADE]	'on the basis'

õnneks	(DI 0.57)	[luck-TRA]	'luckily'
süles	(DI 0.37)	[lap-INE]	'on sb.'s lap'
hoolega	(DI 0.36)	[care-COM]	'with care'

However, relative frequency is not a clearly cogent factor leading to the status of a dictionary headword with PoS tags: 25% of the ambiforms with discrepant PoS statuses in the group "Noun" display DI that are below the threshold level:

(6) *esirinnast*	(DI −0.03)	[forefront-ABL]	'from the front lines'
hääles	(DI −0.02)	[sound-INE]	'in tune'
käpas	(DI 0.04)	[paw-INE]	'mastered'
krunnis	(DI 0.05)	[bun-INE]	'in a bun'
mõõdus	(DI 0.07)	[size-INE]	'size'
mängukorras	(DI 0.08)	[play.condition-INE]	'in playing condition'
südamest	(DI 0.09)	[heart-ELA]	'wholeheartedly'

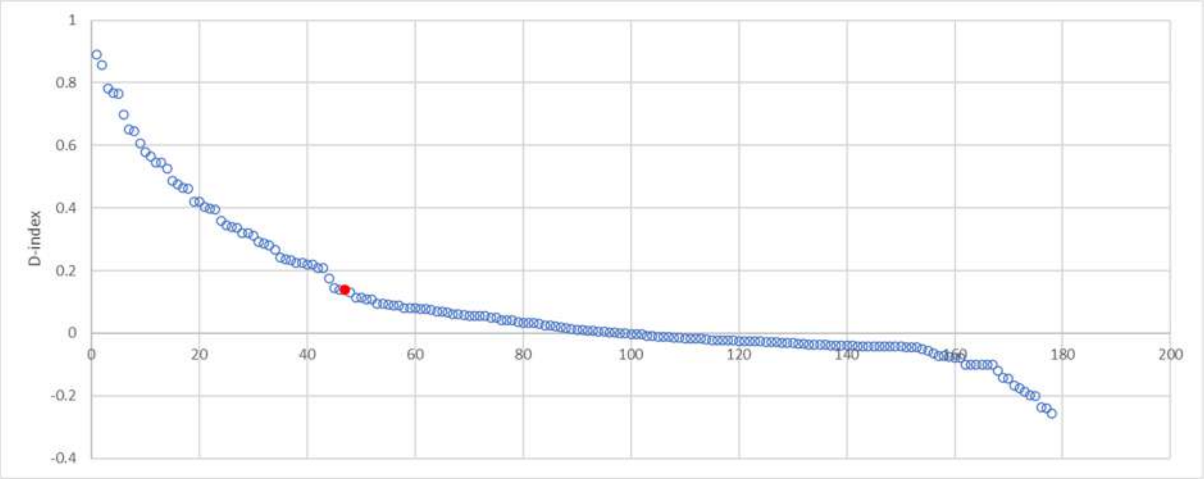


Figure 3: The DI values of the group "Noun+" (the 178 interpretations of the 141 "PoS-tagged" ambiforms labelled with several PoS tags both in the CombiDic and by the EstNLTK morphological analyser)

In Figure 3, we have highlighted the value 0.137, indicating the ambiform *mõõtu* [size-PART] 'size of' as closest to the tentative threshold (0.130). Its position indicates clearly that most of the ambiforms in this group have indices below the threshold; only 27.4% of the ambiforms exceed the level of the threshold. This finding confirms the hypothesis that fewer ambiforms in the "Noun+" group exceed the threshold than in "Noun". Interestingly, the majority of ambiforms in this group (72.6 %) are below the threshold, indicating that split interpretations tend to follow a distribution that is normal or even below normal.

There are, however, some ambiforms with exceptionally high levels of DI in this category (see (7)). There are two explanations for the outstanding DI despite the multiplicity of PoS interpretations: these are either the dominating forms of lemmas with very low corpus frequency (e.g. the descriptive state adverbs *kössis*, *norus*, *kronkus* and *jõllis*: less than 1000), or clearly highly frequent forms from lemmas with high frequency in all forms (e.g. *näiteks* < *näide*, *tasuta* < *tasu* and *täiega* < *täis*).

(7)	*kössis*	(DI 0.89)	[slumped-INE]	'slumped over'
	jommis	(DI 0.86)	[drunk-INE]	'drunk'
	norus	(DI 0.78)	[somberness-INE]	'sombre'
	näiteks	(DI 0.77)	[example-TRA]	'for example'
	tasuta	(DI 0.69)	[charge-ABE]	'free of charge'
	täiega	(DI 0.65)	[full-COM]	'fully'
	kronksus	(DI 0.6)	[curled-INE]	'curled up'
	eos	(DI 0.56)	[seedling-INE]	'at the start'
	jõllis	(DI 0.55)	[bulging-INE]	'bug-eyed'

As a result of the comparison of ambiforms tagged only as case forms of nouns and the ambiforms tagged with more PoS tags than nouns by the EstNLTK morphological analyser, we can conclude that the multiplicity of PoS interpretations (also including homonyms and homographs) generally reduces the DI levels. All in all, the effect of ambiguity followed by the split PoS marking has a considerable effect on the DI of an ambiform and diminishes its reliability as a statistic of relative frequency.

In the following analysis, we will use the set of ambiforms marked as dictionary entries in the CombiDic but interpreted solely as nouns by the EstNLTK (N = 51) as a standard of the DI variation of the good candidates for decategorisation into indeclinable words.

4.2 The impact of the lexicographic status of ambiforms on their D-index

In the following analysis, we will examine the DI variation in two groups of ambiforms based on their lexicographic status. These groups will be set against an external comparison basis, the "Noun" group, representing the ambiforms tagged as case forms of nouns only (see the previous section).

The first group – "Candidates" – consists of 465 ambiforms that are not headwords in the CombiDic at all. These ambiforms originate from different sources, for instance the forms collected during the compilation of the Estonian Collocations Dictionary (2019; see Vainik et al. 2020 for the sources of the database of ambiforms), and can be seen as a possible reserve of new headwords. The question is, do the DI results indicate those ambiforms' critical relative salience and mark them as candidates for entries in the CombiDic? These ambiforms have 516 interpretations by the EstNLTK in our data table, due to form homonymy.

The second group – "Underspecified" – includes the 190 ambiforms in our noun-based ambiform selection that are headwords in the CombiDic but not tagged for PoS. These lexemes are present in the CombiDic in such an underspecified manner as a result of the aggregation processes of the superdictionary (CombiDic) from dictionaries in different formats. Some of these entries were originally subheadwords to main headwords in the Explanatory Dictionary of Estonian (2009); as a way to deal with the decategorising forms of a donor word, the subheadwords had no PoS tags. During the integration process with the CombiDic, all sub-headwords were automatically upgraded to headwords. The PoS-tagging situation of PoS-less headwords constantly changes when the dictionary is updated by lexicographers. These ambiforms have 399 interpretations in the EstNLTK analysis, 206 as case forms of nouns.

The DI variation of the headword candidates and the underspecified headwords in comparison to the set of ambiforms tagged as case forms of nouns by the EstNLTK (see Section 4.1) is presented in Figure 4. The descriptive statistics are given in Table 2.

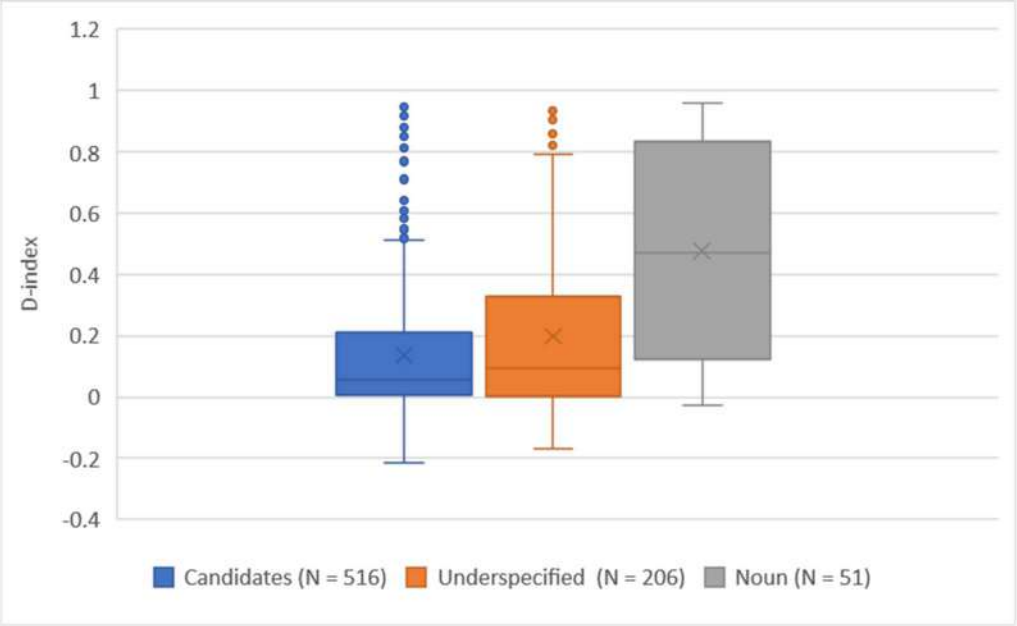


Figure 4: Variance of the DI among two sets of ambiforms: headword candidates and underspecified headwords without PoS tags compared to the ambiforms tagged as case forms of nouns by the EstNLTK

	Candidates	**Underspecified**	**Noun**
N	516	206	51
Max	0.964	0.951	0.958
Min	−0.216	−0.170	−0.026
Median	0.056	0.095	0.471
Ave	0.136	0.198	0.477
StDev	0.208	0.256	0.342

Table 2: Descriptive statistics of headword candidates, underspecified headwords without PoS tags, and the ambiforms tagged as case forms of nouns by EstNLTK morphological analysis

The data in Table 3 reveals that the maximum levels of DI are similar in all three sets, indicating that there are good candidates for decategorisation in each set, regardless of the current lexicographic status of the ambiforms. The average and median are considerably lower in the "Underspecified" group, the ambiforms in headword status without PoS tags, and the lowest in the case of "Candidates". This indicates that the lexicographic status, on average, follows the trend characterised by the relative salience of the word forms.

In relation to the "Noun" sample, the "Candidates" and "Underspecified" groups stand out for showing similar tendencies. These two sets have more tightly grouped DI values: the median results of these sets (0.056 and 0.095) are considerably lower than that of the comparison basis of "Noun" (0.471). Moreover, the average DI of the two analysed groups is 3.5 and 2.4 times lower than that of "Noun". The range of variation outside the box of 50% of the data, however, is much wider in the "Candidates" and "Underspecified" groups than in "Noun"; the extreme outliers over the upper quartiles show "abnormal" cases in these two groups.

The DI values of the headword candidates with no CombiDic headword tags are displayed in a dot chart in Figure 5:

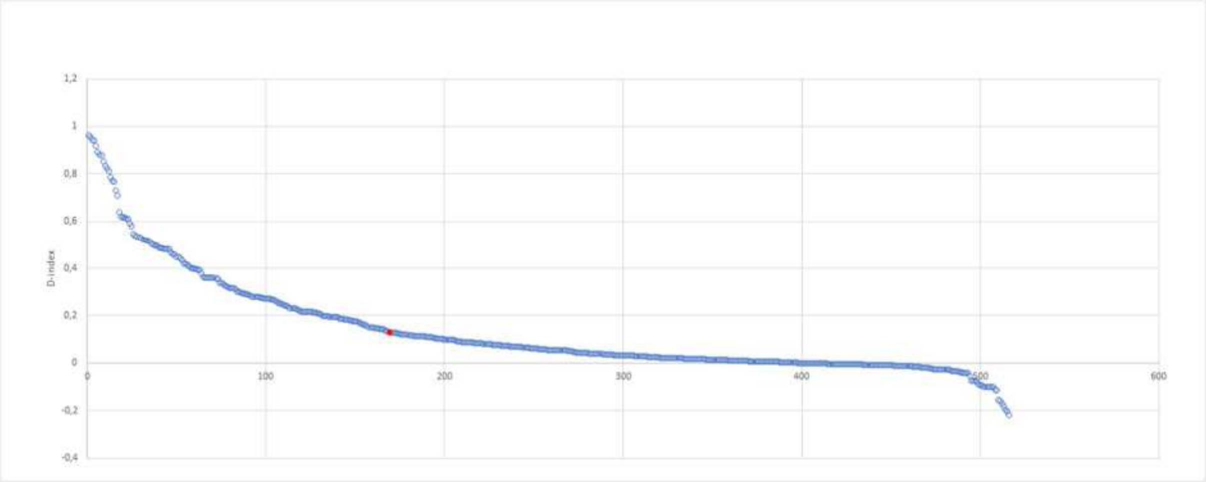


Figure 5: Descending values of the “Candidates” for dictionary headwords

This is a large set of ambiforms (N = 516). The value closest to the threshold (0.129 for the word form *keskmesse* [midpoint-ILL] ‘to the centre’) is highlighted. Only 33% of the ambiforms in this selection exceed the threshold (0.130) and truly qualify as candidates for headwords based on their morphological distribution statistics. Overall, this group shows particularly broad variation, from extremely high DI values (0.964) to negative values down to −2.16, indicating underrepresentation in relation to the expected frequency. At the top of the list are several compound ambiforms (see 8), but there are also non-compound words with exceptionally high DI (9):

(8)	*tikutulega*	(DI 0.96)	[match.light-COM]	‘scrupulously’
	ajajooksul	(DI 0.94)	[time.run-ADE]	‘over time’
	äravahetamiseni	(DI 0.89)	[away.exchange-TER]	‘interchangeable’
	reaalajas	(DI 0.88)	[real.time-INE]	‘in real time’
	vastutasuks	(DI 0.87)	[for.pay-TRA]	‘in return’

(9)	*alustuseks*	(DI 0.95)	[commencement-TRA]	‘for a start’
	nõrkemiseni	(DI 0.92)	[exhaustion-TER]	‘to exhaustion’
	maksvusele	(DI 0.82)	[validity-ALL]	‘validated’

The “Underspecified” ambiforms show a smoother decline in Figure 6. The value closest to the tentative threshold (0.129 for the ambiform *võtmes* [key-INE] ‘à la’) is highlighted. Compared to the “Candidates”, this group has more ambiforms over the threshold: 45% of the calculated DI values. These 93 case forms are good candidates for decategorisation as indeclinable words.

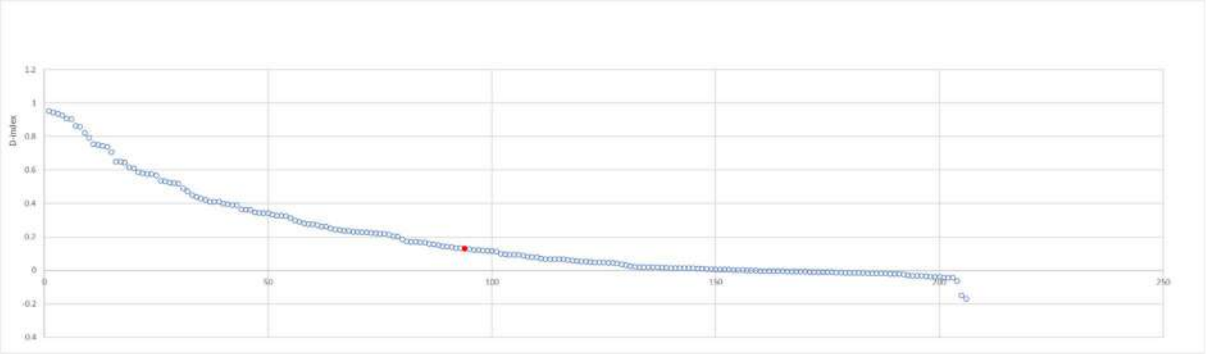


Figure 6: The Descending DI values of the "Underspecified" CombiDic headwords without PoS tags

Similarly to the previous group, "Candidates", the ambiforms with the highest DI are mostly compounds (see (10) and (11)). The ambiforms with DI levels indicating abnormal distributions in the form of underrepresentation (see (12)) are all provided with the comment "used only in negations" in the CombiDic. The reason for that is the emphatic suffix *-gi/-ki* after the case endings, often adding a sense of negation to the stem.

(10)	*üldjoontes*	(DI 0.95)	[common.feature-PL-INE]	'generally'
	esmapilgul	(DI 0.94)	[first.glance-ADE]	'at first glance'
	täismahus	(DI 0.93)	[full.capacity-INE]	'in full'
	lõppkokkuvõttes	(DI 0.9)	[end.conclusion-INE]	'in conclusion'
	eestvedamisel	(DI 0.86)	[front.leading-ADE]	'led by'
	tavamõistes	(DI 0.86)	[ordinary.sense-INE]	'colloquially'
	imeväel	(DI 0.78)	[miracle.power-ADE]	'miraculously'
	noaotsaga	(DI 0.76)	[knife.edge-COM]	'in a pinch'

(11)	*kamaluga*	(DI 0.93)	[cupped hands-COM]	'abundantly'
	mahitusel	(DI 0.9)	[encouragement-ADE]	'with the connivance of sb.'
	kuhjaga	(DI 0.61)	[pile-INE]	'heaped'
	kuubis	(DI 0.57)	[cube-INE]	'cubed'
	moel	(DI 0.57)	[way-ADE]	'in a way'
	sõnul	(DI 0.53)	[word-ADE]	'according to'

(12)	*varjugi*	(DI –0.15)	[shadow-PART-EMPH]	'(not) in the slightest'
	viluvarjugi	(DI –0.17)	[shade.shadow-PART-EMPH]	'(not) in the slightest'
	piiskagi	(DI –0.06)	[drop-PART-EMPH]	'not a drop'

4.3 Implications of morphological and lexicographic PoS tagging status on DI values

An examination of the impact of the morphological analyser on the DI results in Section 4.1 suggests that the most relevant and reliable results of the DI derive from the analysis of ambiforms that are processed as case forms of nouns without splitting the PoS interpretations into noun and additional categories. This suggests that for a realistic outline of the distributional analysis of an ambiform, all of its PoS-readings should be reverted to the noun if possible.

The influence of the headword-labelling situation of ambiforms on their DI levels examined in Section 4.2 raises the question of the relation of lexicographic treatment and ambiforms. We can ask if the DI exposes the lexicographic status of ambiforms, i.e. can the DI predict which word forms are headwords in the combined dictionary? According to our results, the answer is no: the DI variation of ambiforms that are headword candidates (not headwords in the CombiDic) and underspecified ambiforms (headwords without PoS tags) does not show significant differences.

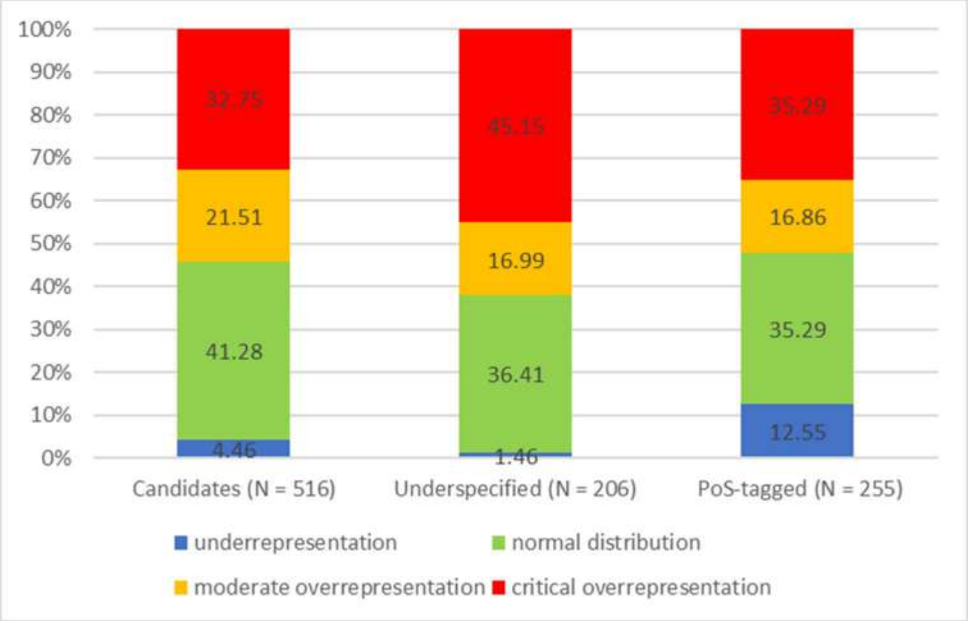


Figure 7: The division of DI results in three data sets: headword candidates, underspecified headwords and PoS-tagged headwords in the CombiDic

The results of the analysis in Sections 4.1–4.2 are summarised in Figure 7. The diagram visualises the division of DI results according to the four degrees of DI values in four data proportions: underrepresentation, normal distribution, moderate overrepresentation, and critical overrepresentation. The three columns represent the examined data from the perspective of their lexicographic status:

- “Candidates” – the ambiforms without headword status in the CombiDic
- “Underspecified” – the ambiforms with headword status but no PoS tags in the

CombiDic

- "PoS-tagged" – the ambiforms with PoS tags other than noun in the CombiDic (this column unites the data analysed in Section 4.1: the case forms of nouns in the EstNLTK morphological analysis ("Noun") and the ambiforms with split PoS analyses ("Noun+")

The proportion of critical and moderate overrepresentation is the highest and the underrepresentation the lowest in the group of underspecified ambiforms, which might indicate why these ambiforms have been given headword status in the CombiDic, although not PoS yet. The headword candidate group has a slightly smaller proportion of critical overrepresentation forms, but the highest proportion of moderate overrepresentation. The group with the expected highest proportion of critical and moderate overrepresentation, the PoS-tagged ambiforms, do not stand out in this respect; surprisingly, this group shows the largest underrepresentation level. It should be noted here that the headword inclusion in the CombiDic has not been related to the statistical distribution of the form so far. For further discussion about the reasons for including word forms with lower-than-normal distribution levels, see Vainik et al. (2021).

After the examination of the ambiform groups with different statuses in morphological analysis and lexicographic practice, we can ask if it is possible to specify any further thresholds in the relatively large area of the critical overrepresentation between the DI values 0.13—1.0. The analysis of the four groups of ambiforms (cf. Figures 3—6) reveals a gap in the line graphs around the value 0.62—0.63. This makes it possible to establish an indicative level of DI of the stage near the indeclinable words. The threshold for ambiforms approaching the characteristics of uninflected words can thus be assigned a provisional value of 0.63.

5. Conclusions

This study aimed to examine the effect of the distributional character of case forms of nouns that have already been or may be decategorised into other parts of speech. We tested the D-index developed a part of this study to detect the deviating frequency of case forms in different settings. PoS-tagging discrepancies between the morphological analyser and the combined dictionary enabled us to study the effect of "inured" and absent decategorisation on the D-index score. The results suggest that for the outcome to be most authentic, the noun-based ambiforms should be analysed without the decategorisation influence, i.e. the D-index analysis should be applied in the pre-PoS-disambiguation stage.

The threshold levels of DI posited in the previous study seemed to function relatively well as indicators of the underrepresentation, normal and moderate and critical overrepresentation of forms. The threshold value of 0.13, the marker of heightened frequency, appears to hold. The analyses of different groups of ambiforms suggest that the upper part of the critical overrepresentation (0.63), as a quite broad stage, could be preserved for the stage of “approaching the characteristics of uninflected words”. A closer study of the ambiforms in this upper area is recommended for future research.

In our opinion, the D-index contributes statistical corpus post-processing information in certain stages of the lexicographic workflow: the specification of a lexeme’s status as a headword and its PoS affiliation. For easy and fast access to a form's D-index, we have developed the Distribution Index Calculator for Estonian. It is a web-based application that retrieves the frequency data of word forms and lemmas from an annotated corpus and retrieves DI statistics on a lexicographer’s workbench (see Vainik et al., 2021).

Since the results of the D-index (and the PoS-tagger) analysis depend on the outcome of morphological dissection, the future development of the natural language processing tasks is also relevant for our purposes. In this article, we have tested one morphological disambiguator available for the Estonian language; the other possibilities are currently the Universal Dependencies PoS Tagger[14] and the TreeTagger[15]. The development of a pre-trained language model, such as Bert, has shown promising results in PoS and morphological tagging of Estonian (see Kittask et al., 2020), which has the potential to also improve the results of the D-index calculus.

In the process of examining the D-index in use, we have determined that “dry” statistical analysis has the potential to give us new knowledge about language. The qualitative study of the groups selected for the analysis in this study and possibly the adjustment of the threshold values of the D-index form an interesting prospect for future research. There are also broader questions arising from this study, for instance: Could the D-index help improve corpus tagging systems? Can it be used in other languages? As an answer to the first question, we suggest that the D-index could help to choose the PoS that is more likely correct in disambiguation processes. The D-index itself is quite readily applicable to other morphologically rich languages, given that the norms of the forms are established.

6. Acknowledgements

We are thankful to the three anonymous referees for their valuable comments on this article. This work was supported by Estonian Research Council grant PSG227.

7. References

Blensenius, K. & von Martens, M. (2019). Improving Dictionaries by Measuring Atypical Relative Word-form Frequencies. In I. Kosem, T. Zingano Kuhn, M. Correia, J. P. Ferreria, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek &

[14] https://cloud.gate.ac.uk/shopfront/displayItem/tagger-pos-et-maxent1

[15] https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

C. Tiberius (eds.). *Proceedings of eLex 2019 conference. 1–3 October 2019. Sintra, Portugal.* Brno: Lexical Computing CZ, s.r.o., pp. 660–675.

Brinton, L. J. & Traugott E. C. (2005). *Lexicalization and language change.* Cambridge: CUP. DOI: 10.1017/CBO9780511615962.

CombiDic = *The EKI Combined Dictionary.* (2020). Hein, I., Kallas, J., Kiisla, O., Koppel, K., Langemets, M., Leemets T., Melts, M., Mäearu, S., Paet, T., Päll, P., Raadik, M., Tiits, M., Tsepelina, K., Tuulik, M., Uibo, U., Valdre, T., Viks, Ü. & Voll, P. Institute of the Estonian Language. Accessed at: Sõnaveeb 2020. https://sonaveeb.ee. (5 March 2021)

The Estonian Collocations Dictionary = *Eesti keele naabersõnad.* (2019). Kallas, J., Koppel, K., Paulsen G. & Tuulik, M., Institute of the Estonian Language. Accessed at: http://www.sonaveeb.ee. (14 February 2020)

Ekilex. Accessed at: https://ekilex.eki.ee/ (20 March 2021)

The Explanatory Dictionary of Estonian = *Eesti keele seletav sõnaraamat* I–VI. (2009). M. Langemets, M. Tiits, T. Valdre, L. Veskis, Ü. Viks, P. Voll (eds.). Institute of the Estonian Language. Tallinn: Eesti Keele Sihtasutus. Accessed at: http://www.eki.ee/dict/ekss/. (5 April 2021)

Grünthal, R. (2003). *Finnic Adpositions and Cases in Change.* Suomalais-Ugrilaisen Seuran toimituksia 244. Helsinki: Finno-Ugrian Society.

Habicht, K., Penjam, P. & Prillop, K. (2011). Sõnaliik kui rakenduslik ja lingvistiline probleem: sõnaliikide märgendamine vana kirjakeele korpuses. *Estonian Papers in Applied Linguistics* 7, pp. 19–41.

Hay, J. (2001). Lexical frequency in morphology: is everything relative? *Linguistics*, 39(6), pp. 1041–1070.

Heine, B. & Kuteva, T. (2007). *The genesis of grammar. A reconstruction.* Oxford: Oxford University Press.

Jakubíček, M. (2021). Morphology is an open problem of NLP. Talk given at the Workshop on Parts of Speech. Tallinn: Institute of the Estonian Language. Available at: https://portaal.eki.ee/component/content/article/101-projektid/3414-workshop-on-the-role-of-parts-of-speech-in-language-technology.html.

Kaalep, H-J. & Vaino, T. 2001. Complete Morphological Analysis in the Linguist's Toolbox. *Congressus Nonus Internationalis Fenno-Ugristarum Pars V*, pp. 9—16, Tartu. Available at: http://www.cl.ut.ee/yllitised/smugri_toolbox_2001.pdf.

Kasik, R. (2015). *Sõnamoodustus* [Word formation]. Tartu: Tartu University Press.

Karelson, R. (2005). Taas probleemidest sõnaliigi määramisel [Once again on the problems of assigning the PoS]. *Estonian Papers in Applied Linguistics* 1, 53—70.

Kilgarriff, A., Rychlý, P., Smrž, P. & Tugwell, D. (2004). The Sketch Engine. *Information Technology*, 105, pp. 116–127.

Kittask, C., Milintsevich, K. & Sirts, K. (2020). Evaluating Multilingual Bert for Estonian. In A. Utka, J. Vaičenonienė, J. Kovalevskaitė & D. Kalinauskaitė (eds.). *Human Language Technologies – The Baltic Perspective.* IOS Press, pp.

19–26. (Frontiers in Artificial Intelligence and Applications). DOI: 10.3233/FAIA200597.

Koppel, K. (2020). *Näitelausete korpuspõhine automaattuvastus eesti keele õppesõnastikele* [Corpus-Based Automatic Detection of Example Sentences for Dictionaries for Estonian Learners]. PhD thesis. Tartu: Tartu University Press.

Koppel, K., Tavast, A., Langemets, M. & Kallas, J. (2019). Aggregating dictionaries into the language portal Sõnaveeb: issues with and without a solution. In I. Kosem, T. Zingano Kuhn, M. Correia, J. P. Ferreria, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek & C. Tiberius (eds.) *Proceedings of the eLex 2019 conference. 1–3 October 2019, Sintra, Portugal.* Brno: Lexical Computing CZ, s.r.o., pp. 434–452.

Langemets, M., Kallas, J., Norak, K. & Hein, I. (2020). New Estonian Words and Senses: Detection and Description. *Journal of the Dictionary Society of North America* 41 (1), pp. 69–82.

Laur, S., Orasmaa, S., Särg, D. & Tammo, P. (2020). EstNLTK 1.6: Remastered Estonian NLP Pipeline. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pp. 7152–7160.

Orasmaa, S., Petmanson, T., Tkatšenko, A., Laur, S. & Kaalep, H-J. (2016). EstNLTK – NLP Toolkit for Estonian. In N. Calzolari, K. Choukri, T. Declerck, M. Grobelnik, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & P. Stelios (eds.) *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016).* Portorož, Slovenia: ELRA, pp. 2460–2466. http://www.lrec-conf.org/proceedings/lrec2016/pdf/332_Paper.pdf

Paulsen, G., Vainik, E., Tuulik, M. & Lohk, A. (2019). The lexicographer's voice: word classes in the digital era. In I. Kosem, T. Zingano Kuhn, M. Correia, J. P. Ferreria, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek & C. Tiberius (eds.) *Proceedings of eLex 2019 conference. 1–3 October 2019, Sintra, Portugal.* Brno: Lexical Computing CZ, s.r.o., pp. 319–337.

Paulsen, G.; Vainik, E.; Tuulik, M. (2020). Sõnaliik leksikograafi töölaual: sõnaliikide roll tänapäeva leksikograafias [On word classes in contemporary lexicography: The lexicographers' view]. *Estonian papers in applied linguistics*, 16, pp. 177–202. DOI: 10.5128/ERYa16.11.

Sahkai, H. (2008). Konstruktsioonipõhine keelemudel ja sõnaraamatumudel [A construction-based model of language and dictionary]. *Estonian Papers in Applied Linguistics*, 4, pp. 177–186.

Tavast A., Koppel K., Langemets M. & Kallas J. (2020). Towards the Superdictionary: Layers, Tools and Unidirectional Meaning Relations. In Z. Gavriilidou, M. Mitsiaki & A. Fliatouras (eds.) *Proceedings of XIX EURALEX Congress: Lexicography for Inclusion*, Vol. 1., Greece: Democritus University of Thrace, pp. 215–223.

Tavast, A., Langemets, M., Kallas, J., & Koppel, K. (2018). Unified Data Modelling for Presenting Lexical Data: The Case of EKILEX. In J. Čibej, V. Gorjanc, I. Kosem & Simon Krek (eds.) *Proceedings of the XVIII EURALEX International*

Congress: EURALEX: Lexicography in Global Contexts. Ljubljana, Slovenia.

Tkachenko, A. & Sirts, K. (2018). Neural Morphological Tagging for Estonian. In Muischnek, K. & Müürisepp K. (eds.). *Human Language Technologies – The Baltic Perspective.* IOS Press. (Frontiers in Artificial Intelligence and Applications), pp. 166–174. DOI: 10.3233/978-1-61499-912-6-166.

Vainik, E., Paulsen, G. & Lohk, A. (2020). A typology of lexical ambiforms in Estonian. In Z. Gavriilidou, M. Mitsiaki & A. Fliatouras (eds.) *Proceedings of XIX EURALEX Congress: Lexicography for Inclusion*, Vol. 1. Alexandroupolis, Greece: Democritus University of Thrace, pp. 119–130.

Vainik, E.; Paulsen, G. & Lohk, A. (2021). Käändevormist sõnaks: mida näitab sagedus? [From inflected form to a word: the role of frequency]. Accepted by *Estonian Papers in Applied Linguistics*, 17.

Vainik, E.; Lohk, A. & Paulsen, G. (2021). The Distribution Index Calculator for Estonian. *Proceedings of eLex 2021 conference.* 5–7 July 2021, Brno, Czechia. Brno: Lexical Computing CZ, s.r.o.

Veskis, K.; Liba, E. (2010). Automatic Tagger Evaluation. Syntax assignment report. NGSLT (Nordic graduate school on language technology) NLP course 2008. Available at: http://teataja.ee/veskis-liba-syntax-assignment-modified.pdf

Viitso, T-R. (2003). Structure of the Estonian language: Phonology, morphology, and word formation. In M. Erelt (ed.) *Estonian language.* Tallinn: Estonian Academy Publishers, pp. 9–92.

Viks, Ü. (1992). *Väike vormisõnastik. I: Sissejuhatus & grammatika; II: Sõnastik & lisad* [A Concise Morphological Dictionary of Estonian. I: Introduction & Grammar; II Dictionary and Appendices]. Tallinn.

MOR*Digital*:

The Advent of a New Lexicographic Portuguese Project

Rute Costa[1], Ana Salgado[2], Anas Fahad Khan[3], Sara Carvalho[1,4],

Laurent Romary[5], Bruno Almeida[1,6], Margarida Ramos[1],

Mohamed Khemakhem[7], Raquel Silva[1], Toma Tasovac[8]

[1] NOVA CLUNL, Centro de Linguística da Universidade NOVA de Lisboa, Portugal
[2] Academia das Ciências de Lisboa, Portugal
[3] Istituto Di Linguistica Computazionale 'A. Zampolli', Italy
[4] CLLC, Centro de Línguas, Literaturas e Culturas da Universidade de Aveiro, Portugal
[5] Inria, team ALMAnaCH, France
[6] ROSSIO Infrastructure, Portugal
[7] Arcascience, France
[8] BCDH – Belgrade Center for Digital Humanities
E-mail: rute.costa@fcsh.unl.pt, anasalgado@campus.fcsh.unl.pt, fahad.khan@ilc.cnr.it, sara.carvalho@ua.pt, laurent.romary@inria.fr, brunoalmeida@fcsh.unl.pt, mvramos@fcsh.unl.pt, medkhemakhemfsegs@gmail.com, raq.asilva@gmail.com, ttasovac@humanistika.org

Abstract

MOR*Digital* is a newly funded Portuguese lexicographic project that aims to produce high-quality and searchable digital versions of the first three editions (1789; 1813; 1823) of the *Diccionario da Lingua Portugueza* by António de Morais Silva, preserving and making accessible this important work of European heritage. This paper will describe the current state of the art, the project, its objectives and the methodology proposed, the latter of which is based on a rigorous linguistic analysis and will also include steps necessary for the ontologisation of knowledge contained in and relating to the text. A section will be dedicated to the various investigation domains of the project description. The output of the project will be made available via a dedicated platform.

Keywords: digital humanities; GROBID-Dictionaries; legacy dictionary; lexicography; ontologies; standards

1. Introduction

The *Diccionario da Lingua Portugueza* by António de Morais Silva, hereafter referred to as Morais, constitutes a considerable piece of cultural heritage since it marks the beginning of modern Portuguese lexicography, serving also as a model for all subsequent lexicographic production throughout the 19th and 20th centuries. In this paper, we present MOR*Digital*, a newly funded Portuguese lexicographic project, which was successfully submitted to the IC&DT 2020 Projects Call under the scientific area of 'information sciences computing', which falls under 'languages and literatures –

linguistics, subarea computer sciences and information sciences'. The project will be funded over the next three years (2021–2024).

The MOR*Digital* project aims to produce high-quality and searchable digital versions of the first three editions (1789; 1813; 1823) of Morais in order to preserve this important European heritage work while also making it accessible. These digital versions will be converted into structured data and made publicly available with the purpose of guaranteeing the preservation of this legacy resource. After an introduction to the dictionary itself, we provide a general outline of the project and detail its main objectives, focusing on the importance of using standards and formats for interoperability purposes. We then explore the research methodology adopted. This methodology for the creation of an open-access Portuguese language dictionary is based on a comprehensive understanding of lexical units and the privileging of a strictly linguistic analysis to create future ontologies that adequately represent the lexical data in the study, in addition to making them accessible and reusable.

This project aims to make a substantial contribution to the scientific community and aspires to apply innovative computational methodologies to digitise lexicographic texts and coding based on a comprehensive analysis of lexicographic articles and their components.

This paper is organised as follows: the first (and current) section introduces and outlines the article. Section 2 reviews the theoretical framework and existing standards. In Section 3, we historically frame our object of study. Section 4 introduces the Morais dictionary. Section 5 describes the MOR*Digital* project, the methodology, as well as tools and formats. Finally, in Section 6, we highlight our future work and present concluding remarks.

2. Theoretical Framework

European lexicography can boast a long tradition of theoretical and descriptive work on dictionaries and especially in the case of historical dictionaries, as is discussed in several works, amongst which Zgusta (1971), Wiegand (1984), Quemada (1987), Atkins and Rundell (2008), Tarp (2008), Durkin (2019) and Considine (2019). These authors have approached lexicography from either a theoretical or methodological perspective, helping to bring to light the paradigm shift we witness in the convergence between lexicography, computational linguistics, digital humanities, and ontologies.

In Portugal, this scientific activity around lexicography work is present in Villalva & Williams (2019), Salgado et al. (2019), Salgado & Costa (2019), Lino (2018), Silvestre (2016), Gonçalves & Banza (2013), Correia (2009) and Verdelho (2003), among others. The *European Dictionary Portal*[1] points to the existence of four online Portuguese dictionaries and a portal. Despite being electronic, most of these resources are

[1] http://www.dictionaryportal.eu/en/

structured and formalised according to a paper-based methodology, and therefore do not fully explore their digital potential. In turn, the *Dicionário Aberto*, one of the dictionaries available on the portal, differs from our objectives, even though it is based on a historical dictionary. This is because the researchers' primary focus (Simões & Farinha, 2009) was not so much preserving the original source but mainly modernizing the dictionary. Thus, and according to the available data, there are no dynamic, open-access resources based on Portuguese heritage dictionaries, so efforts must be made to provide this accessibility to recognised heritage value sources in the form of searchable, dynamic resources.

Lexicography has undergone a radical change in the past two decades, especially with technological advances, the fall of many publishers, as well as the changes introduced into their business models (Rundell, 2010: 170). This paradigm shift is also directly related to the advancement of digital humanities, which quickly became an aggregator of several scientific disciplines. Although the first definitions of the term 'digital humanities' were limited to humanities computing (Terras & Vahouette, 2013), today, these definitions are far from being universally accepted (Gold & Klein, 2016). Instead, the term now covers a variety of lines of research belonging to a number of different disciplines, and is characterised by the use of tools, computational methods and standards, implying, above all, a new general perspective of the humanities in response to the epistemological challenges that these changes impose.

The perspective underpinning the construction of lexical resources that we propose in this project presupposes rethinking the methodologies of the Portuguese lexicographic tradition, perceiving lexicography, terminology, ontologies and computational linguistics as an integral part of the digital humanities, which will imply a paradigm shift in the construction of dictionary resources. In this new paradigm, ontologies will play a key role in organising and representing linguistic and metalinguistic knowledge, bringing added value by providing greater logical consistency in the representation of data (Carvalho et al., 2018; Almeida et al., 2019), as well as supporting its operationalisation and, therefore, its preservation in the long term.

The European lexicographic scenario is currently quite heterogeneous, both in what concerns the types of existing lexicographic resources and their particular structural component, which relates to how the data are represented, the adopted models, as well as the respective applied formats. Each format has its own syntax and vocabulary, defined according to certain parameters to enable the reusability of the lexicographic content. The diversity of incompatible formats creates severe problems in the digital landscape, making it impossible to interconnect resources and their respective metadata and lexical data. Herein lies the importance of following compatible standards and formats such as LMF (ISO 24613: 2008), TEI Lex-0[2] (Tasovac and Romary et al., 2018) and Ontolex-Lemon (McCrae et al., 2017).

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

3. Historical Background

Diccionario da Lingua Portugueza by António de Morais Silva was elaborated during the Age of Enlightenment. This century brought a renewal in several fields of knowledge, namely those concerning the description of living languages, at a time when Latin was still the language of instruction. Dictionaries were perceived as metalinguistic instruments. The 17th century marked a very prolific period in terms of lexicographic production, especially with regard to the French dictionary production (for example, *Dictionnaire françois, contenant les mots et les choses, plusieurs nouvelles remarques sur la langue françoise* (1680) by Father Richelet or *Dictionnaire universel* (1690) by Antoine Furetière), which served as a model for all subsequent lexicographic works.

Portuguese lexicography benefited from this moment, especially with the Morais dictionary's publication in 1789, which inaugurated modern Portuguese lexicography. This dictionary followed the publication of the third edition of the *Vocabolario degli Accademici della Crusca* (1691), the *Dictionnaire de l'Académie Française* (1694) and the *Vocabulario Portuguez and Latino* (1712–1728) by Father Rafael Bluteau. The latter marked the transition between the Latin-Portuguese dictionary and the first Portuguese monolingual dictionary [Morais] (Silvestre, 2008, p. 7), thus paving the way for the emergence of a new way of working in lexicography that would influence subsequent publications, such as the *Diccionario da lingoa portugueza* (1793), published by the Lisbon Science Academy and the *Elucidário das Palavras, Termos e Frases* by Joaquim de Santa Rosa de Viterbo (1798). As Verdelho (2003: 473) mentions, Morais 'laid the foundation to all the lexicographic genealogy developed over the last 200 years' and, according to Biderman (1984: 5), referring to the second edition, 'constitutes a milestone in Portuguese-language lexicography'.

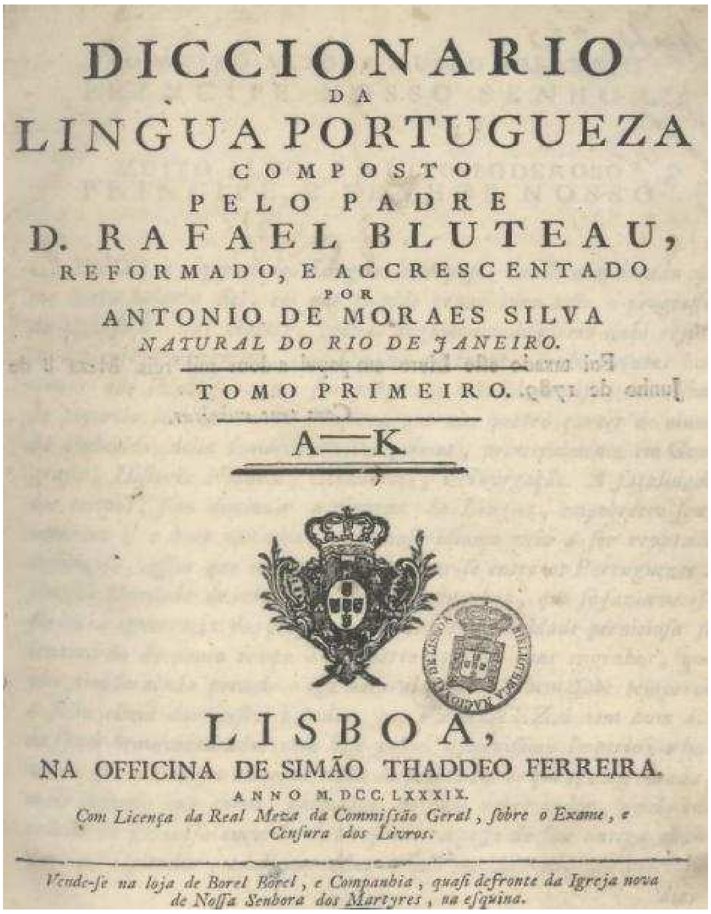
Despite all this, lexicographic production arises late in Portugal when compared with that of other countries. The publication of dictionaries in vernacular languages was already proliferating throughout Europe, as can be seen from the publishing timelines of other monolingual dictionaries.[3]

4. Morais Dictionary

The first edition of the known Morais dictionary is entitled in its main edition (1789) *Diccionario da Lingua Portugueza composto pelo Padre D. Rafael Bluteau Diccionario da Lingua Portugueza composto pelo Padre D. Rafael Bluteau, reformado, e accrescentado por Antonio de Moraes Silva, natural do Rio de Janeiro* [Diccionario da Lingua Portugueza composed by Father D. Rafael Bluteau, retired, and accredited by

[3] Such as the *Tesoro de la lengua castellana*, española by Sebastián de Covarrubias in 1611, which, in addition to being the first Spanish monolingual dictionary, is the first European one. Other examples include the first edition of the *Vocabolario degli Accademici della Crusca*, which was compiled in Florence and printed in Venice in 1612, as well as the French dictionaries mentioned before.

Antonio de Morais Silva, born in Rio de Janeiro], as seen in Figure 1.



DICCIONARIO
DA
LINGUA PORTUGUEZA
COMPOSTO
PELO PADRE
D. RAFAEL BLUTEAU,
REFORMADO, E ACCRESCENTADO
POR
ANTONIO DE MORAES SILVA
NATURAL DO RIO DE JANEIRO.
TOMO PRIMEIRO.
A—K

LISBOA,
NA OFFICINA DE SIMÃO THADDEO FERREIRA.
ANNO M. DCC. LXXXIX.
Com Licença da Real Meza da Commiſsão Geral, ſobre o Exame, e Cenſura dos Livros.

Vende-ſe na loja de Borel Borel, e Companhia, quaſi defronte da Igreja nova de Noſſa Senhora dos Martyres, na eſquina.

Figure 1: Frontispiece of Morais (1789), first volume

The information that immediately stands out concerns the authorship attribution, since Morais does not claim to be the author, assigning this condition to Bluteau, author of the *Vocabulario Portuguez and Latino*. However, Morais recognises in the '*Prólogo ao Leitor*' [Prologue to the Reader] that the additions he brought to the dictionary are quite relevant. Morais further developed Bluteau's work and systematically took into account most of the entries and definitions. Verdelho (2003) considers this attitude inevitable, which, in reality, reflects, '*o que todos os dicionaristas não podem deixar de fazer ao retomar e renovar a nomenclatura dos seus predecessores, uma espécie inevitável de 'plágio por ordem alfabética*' [what all dictionary-makers cannot fail to do when resuming and renewing the nomenclature of their predecessors, an inevitable kind of 'plagiarism in alphabetical order'].

As mentioned above, Morais represents the first modern work to systematise the lexicon of the Portuguese language, a model and example for all the ones that followed. It was also, for almost two centuries, a work of mandatory consultation for Portuguese language, both in Portugal and in Brazil. As Correia (2009) observes, the Morais dictionary '*tornou-se uma referência incontornável para o estudo da evolução do léxico* *do Português, tendo constituído, simultaneamente, um elemento de normalização e mesmo de padronização da língua*' [has become an essential reference for the study of the evolution of the Portuguese lexicon, having simultaneously constituted an element of normalisation and even of language standardisation].

The first edition was first published in two volumes: first, from the letters A to K, in a total of 752 pages, and then, from the letters L to Z, with 541 pages. The work was printed at Simão Thaddeo Ferreira's publishing house, in Lisbon.

The following two editions (1813; 1823) are considered new dictionaries, due to both their enrichment and the updating. The second edition, corrected and enlarged in two volumes (A–E; F–Z), was also published in Lisbon, in Typographia Lacerdina. Morais claims the authorship of the dictionary on the title page, where the work is presented as the *Diccionario da Lingua Portugueza, recopilado dos vocabularios impressos ate agora, e nesta segunda edição novamente emendado, e muito accrescentado, por Antonio de Moraes Silva natural do Rio de Janeiro* [*Diccionario da Lingua Portugueza*, compiled from the vocabularies printed so far, and in this second edition, again amended and incredibly enriched by Antonio de Moraes Silva]. The same happened to the third edition, coordinated by Pedro José de Figueiredo, who expanded it from five to six thousand articles, as stated in the title.

The author died the following year, in 1824. The work continued to be published and enhanced over the years until 1949. From then to 1959, in 12 volumes, the tenth edition was prepared, under the coordination of Augusto Moreno, Cardoso Júnior and José Pedro Machado, but maintaining Morais as the author.

Even though the Morais dictionary is available on some web pages (e.g. CEPESE[4]), it is provided as a PDF document, resulting from the digitisation of the work on paper. This format does not take great advantage of the digital environment and its potential, since it does not allow advanced searches. It is this issue that we intend to explore in our project.

5. MOR*Digital*

5.1 The Project

As stated in the introduction, the main goal of MOR*Digital*[5] is to encode the selected editions of *Diccionario de Lingua Portugueza* by António de Morais Silva. MOR*Digital* aims to promote accessibility to cultural heritage while fostering reusability and contributing towards a greater presence of lexicographic digital content in Portuguese through open tools and standards. MOR*Digital* follows a new paradigm in lexicography,

[4] https://www.cepese.pt/portal/pt/bases-de-dados/dicionario/apresentacao

[5] MORDigital – Digitalização do *Diccionario da Lingua Portugueza* de António de Morais Silva [PTDC/LLT-LIN/6841/2020]

which results from the convergence of lexicography, terminology, computational linguistics, and ontologies as an integral part of digital humanities and Linked (Open) Data.

In this project, we connect data and metadata within the same lexicographic resource and between different resources, through the Web of Data, which is based on principles structured around the use of RDF, URIs and SPARQL, a language for querying and retrieving information. Underlying the formalisation and application of the standards is the linguistic and lexicographic knowledge that permeates the entire project and contributes to the necessary systematisation of data and metadata. Being a project dedicated to Portuguese, it has the added value of bringing a historical resource into the LLOD cloud in a language that is still underrepresented.

Retrodigitising historical dictionaries into machine-readable dictionaries poses several challenges that the scientific community has tried to resolve by creating tools, different formats, and establishing standards, following the FAIR[6] principles for modelling lexical resources and making them available.

Our starting point will be the Morais digitisations available as PDF at the Portuguese National Library and the Brasiliana Library[7]. However, the lack of quality of the available PDF may lead us to undertake a new digitisation process of Morais. High-quality digitisation is required to use GROBID-Dictionaries (Khemakhem, Foppiano, Romary, 2017, Khemakhem et al., 2019), a machine learning system for converting PDF into the TEI/XML format and structuring the content of the digitised versions of the dictionaries.

Following current open data best practices, the main goal is to put forward a methodology that can be replicated in other legacy paper dictionaries, using tools that allow the automatic extraction of lexicographic content, as well as the modernisation of the spelling in an automated way.

5.2 Methodology

MOR*Digital* proposes to: (i) analyse all components that comprise the dictionary's macro- and microstructure; ii) identify, organise and describe the different levels of linguistic knowledge to apply the aforementioned standards systematically; (iii) develop methodologies that can be replicated for other applications and test the alignment of the different encodings of Morais; (iv) participate in reviewing the corresponding standards as members of the standard bodies and scientific forums; (v) propose best practices for harmonising the encoding of lexicographic resources; (vi) make Morais available via an open-access platform.

[6] Findable, Accessible, Interoperable, Reusable; cf. Wilkinson et al. (2016).

[7] http://dicionarios.bbm.usp.br/pt-br/dicionario/edicao/2

Our methodology is based on 5 central axes:

(1) high-quality retrodigitisation of Morais and automatic structuring of the lexical content for the creation of a computer-readable resource;

(2) lexicographically-oriented language description;

(3) Morais encoding, using the TEI Lex-0 specifications mapped to the LMF standard and their respective serialisations, as well as to OntoLex-Lemon;

(4) creation of an ontology for alignment purposes;

(5) and conception of a platform for Morais, enriched with both lexicographic and ontological modules.

All defined tasks will be accomplished successively and managed through subtask assignments, which will be carried out either simultaneously or sequentially, depending on their nature.

We will initiate by surveying the dictionary sources and by a prior evaluation of the quality of digitised versions of these sources (paper to text), for the extraction of lexical information (text to structure). Firstly, this involves transforming the native encoding format into a TEI/XML compliant one (the encoding will be based on TEI standards according to the TEI Lex-0 specification) and LMF metamodels into advanced techniques for semi-structured text acquisition.

The result will be a model of a historical dictionary whose entries are structured in a standard format, namely TEI Lex-0. We plan to adapt the system's cascading architecture to allow the extraction of the different TEI constructs corresponding to the lexicographic structures and conventions. The outcome is a chain of cascading machine learning models, trained and evaluated against manually annotated data. Once the source is digitised, further corrected and marked-up, it will be compared to precedent and subsequent versions, and a series of queries will be conducted to extract all available information about labels. We will then convert TEI Lex-0 datasets into RDF by means of the W3C recommendation for publishing lexicons as Linked Data, namely OntoLex-Lemon. More specifically, we intend to test the implementation of the lexicography module of the Lexicon Model for Ontologies (lexicog)[8], which was recently specified by the Ontology Lexicon community group of the W3C. This will allow for the publication of the Morais datasets as LOD graphs, enabling further NLP applications.

A further step will be the creation of an ontology of all the previously identified and systematised labels (e.g. domain, register, grammar, among others). This will be

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

implemented by resorting to Protégé[9], a free, open-source ontology editor. The ontology will be represented in OWL.

The next step is the alignment of the dictionary versions, which will be carried out in stages: i) alignment of the entries; ii) alignment of the senses; iii) alignment of other lexicographic content.

During the testing phase, formally controlled tests will be carried out to discover errors and bugs that need to be resolved. Finally, we will build a platform that integrates all Morais versions while also mapping the different heterogeneous annotation models, in order to provide access to high-quality digital lexicographic content enhanced by ontologies.

Thus, the search functionalities will include basic and advanced queries, namely searches by lexical relations. A specialised team will be hired to build and develop the interface. Its robustness will be tested according to the types of functionalities defined on validation tests. The alignment between the various editions will be searchable, and the scanned pages made available. In another module, where there will be considerable investment by the team, it is intended that the lexicographic content can be deconstructed and organised in the form of an ontology. We will develop advanced search engines (search for entries by different labels or lexical relations). As part of the aforementioned platform, we will include a section to promote training for the sustainable development of lexicographic resources. This will foster both the qualification of Portuguese lexicographers as well as the users' linguistic knowledge. Moreover, this will provide quality data for researchers.

We aim for our lexical resources to maintain the original spelling. However, making a resource available to the public today, and considering the prevalence of search engines, requires the modernisation of the spellings, especially at the lemma level. The original spelling of the lemma will have to be aligned with more current spellings. To this end, the original forms will be noted as a lemma, but we will first match them with the most current spellings and simultaneously work on their encoding in the XML annotation file. This topic represents the added value of enabling reduplication in other related works, since the correspondences between the lexical units and their respective coding can be reused. We will subsequently create a correspondence between the MOR spellings and the spellings in accordance with the 1945 Luso-Brazilian Convention[10] and the 1990 Portuguese Spelling Agreement[11], taking advantage of work previously developed by one of the team members on the *Vocabulário Ortográfico da Língua Portuguesa* (VOLP-ACL) [Portuguese Language Spelling Vocabulary] of the Lisbon Science Academy[12]. The result will allow the end-user to search the current spellings,

[9] https://protege.stanford.edu

[10] http://www.portaldalinguaportuguesa.org/?action=acordo&version=1945

[11] https://dre.pt/application/file/a/403254

[12] Available at https://www.volp-acl.pt/

with which he/she is familiar, and find the entry corresponding to the old spelling, which will thus remain faithful to the original.

The way we look at Morais transcends the traditional concept of dictionary and is in line with the evolution of e-lexicography itself. We will take advantage of standard formats and linked data technologies for encoding dictionaries, which will allow us to abandon, once and for all, the editorial perspective that is still present in most digital resources. To achieve our goal, we also believe it is necessary to put forward methodologies for improving the quality of lexicographic descriptions.

At the end of the project, we expect to have encoded a vital heritage dictionary, compliant with the most advanced standards for scholarly digital editions and made available via an open licence. The versions will be accessible and searchable through an advanced interface, which will enable the selective querying of text by lemma and type of lexicographic content. The source data will be made available separately from the querying interface, both for research and long-term preservation. Thus, the project will have significantly contributed towards the analysis and annotation of dictionaries through computer-assisted processes.

6. Concluding remarks

This project will represent a substantial contribution to the scientific community, aiming to create innovative and data-driven computational methods for text digitisation and encoding, based on a comprehensive analysis of lexicographic articles and their respective components. Tests on automatic text capture will refine processes and techniques, advancing the state of the art regarding semantic annotation of semi-structured documents. A rigorous linguistic treatment will make it possible to organise and structure the lexicographic components, and to elicit lexical relationships between various elements. The linking mechanisms of the resulting structured dictionary to other resources will constitute a prototype that can be replicated in other works, namely in the Portuguese-speaking world.

MOR*Digital* will be a user-friendly, open-access web interface, equipped with a robust research system that will not only facilitate the search on a more traditional lexicographic perspective but will also allow undertaking research on various types of structured lexicographic and terminological information (Costa et al., 2020). Combining semasiological and onomasiological approaches applied to the three editions of Morais will be possible via the inclusion of ontologies (e.g. diasystematic marking, namely domain labels, registers and part of speech categories). This method will make a new type of dictionary emerge which will contribute to creating a digital linguistic resource that is central to digital humanities. End-users will be predominantly scholars dealing with language and historical issues.

7. Acknowledgements

This paper is supported by (1) the MOR*Digital – Digitalização do Diccionario da Lingua Portugueza de António de Morais Silva* [PTDC/LLT-LIN/6841/2020] project financed by the Portuguese National Funding through the FCT – Fundação para a Ciência e Tecnologia (2) Portuguese National Funding through the FCT – Fundação para a Ciência e Tecnologia as part of the project Centro de Linguística da Universidade NOVA de Lisboa – UID/LIN/03213/2020 and (3) the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015 (ELEXIS) (European Lexicographic Infrastructure).

8. References

Almeida, B., Costa, R. & Roche, C. (2019). The names of lighting artefacts: extraction and representation of Portuguese and Spanish terms in the archaeology of al-Andalus. *Revue TAL, 60*(3), pp. 113–137.

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

Biderman, M. T. C. (1984). A Ciência da Lexicografia. *Alfa*, n. 28, Brasil: São Paulo, pp. 1-26.

Carvalho, S., Costa, R. & Roche, C. (2018). The Role of Conceptual Relations in the Drafting of Natural Language Definitions: an Example from the Biomedical Domain. In I. Kernerman & S. Krek (eds.), *Proceedings of the LREC 2018 Workshop 'Globalex 2018 – Lexicography & WordNets'*. Miyazaki: European Language Resources Association (ELRA), pp. 10–16. ISBN 979-10-95546-28-3.

Considine, J. (2019). *The Cambridge World History of Lexicography.* Cambridge: Cambridge University Press.

Correia, M. (2009). *Os Dicionários Portugueses.* Coleção: O Essencial Sobre Língua Portuguesa. Lisboa: Editorial Caminho.

Costa, R., Carvalho, S., Salgado, A., Simões, A. & Tasovac, T. (2020). Ontologie des marques de domaines appliquée aux dictionnaires de langue générale. In Xavier Blanco (ed.), La lexicographie en tant que méthodologie de recherche en linguistique *Revue de Philologie Française et Romane - Langue(s) & Parole*, n. 55. Mons: Edition du CIPA, pp. 201-230.

Durkin, P. (ed.) (2019). *The Oxford Handbook of Lexicography.* ISBN: 9780199691630. DOI: 10.1093/oxfordhb/9780199691630.001.0001.

Gold, K. M. & Klein L. F. (eds.) (2016). *Debates in the Digital Humanities.* Mineápolis: University of Minnesota Press.

Gonçalves, M. F. & Banza, A. P. (2013). Fontes de metalinguísticas para a história do português clássico – O caso das Reflexões sobre a Lingua Portugueza. In M. F. Gonçalves e A. P. Banza (coord.), *Património Textual e Humanidades Digitais: da antiga à Nova Filologia*, pp. 73–111. Col. Biblioteca – Estudos & Colóquios, Série ebook, n. 1. Évora: CIDEHUS.

ISO 24613. 2008. *Language resource management – Lexical markup framework (LMF).* Geneva: ISO.

Khemakhem, M., Galleron, I., Williams, G. Romary, L. & Suárez, P. J. O. (2019). How OCR Performance Can Impact on the Automatic Extraction of Dictionary Content Structures. In *19th Annual Conference and Members' Meeting of the Text Encoding Initiative Consortium.* Austria: Graz. https://hal.archives-ouvertes.fr/hal-02263276.

Khemakhem, M., Foppiano, L. & Romary, L. (2017). Automatic Extraction of TEI Structures in Digitized Lexical Resources Using Conditional Random Fields. In *Proceedings of eLex 2017 Conference: Electronic lexicography in the 21st century: Lexicography from Scratch.* Netherlands: Leiden, pp. 598–613.

Lino, T. (2018). Portuguese lexicography in the internet era. In P. Fuertes-Oliveira (ed.), *The Routledge Handbook of Lexicography.* Abingdon: Routledge, [n.a.]. ISBN 9781138941601.

McCrae, J. P., Bosque-Gil, J., Gracia, J., Buitelaar, P. & Cimiano, P. (2017). The Ontolex-Lemon model: development and applications. In *Proceedings of eLex 2017 Conference: Electronic lexicography in the 21st century: Lexicography from Scratch.* Netherlands: Leiden, pp. 587–597.

Morais: Silva, António de Morais (1789). *Diccionario da lingua portugueza composto pelo padre D. Rafael Bluteau, reformado, e accrescentado por Antonio de Moraes Silva, natural do Rio de Janeiro*, 2 vols. Lisboa: Officina de Simão Thaddeo Ferreira. [For the purpose of this project, other editions will be consulted.]

Quemada, B. (1987). Notes sur lexicographie et dictionnairique. *Cahiers de Lexicologie*, v. 51, n. 2, pp. 229–242. Paris.

Rundell, M. (2010). What future for the learner's dictionary? I. J. Kernerman & P. Bogaards (eds.), *English Learners' Dictionaries at the DSNA 2009.* Jerusalem: Kdictionaries, pp. 169–175.

Salgado, A. Costa, R. & Tasovac, T. (2019). Improving the consistency of usage labelling in dictionaries with TEI Lex-0. In *Lexicography: Journal of ASIALEX 6* (2), pp. 133–156. DOI: https://doi.org/10.1007/s40607-019-00061-x.

Salgado, A. & Costa, R. (2019). Marcas temáticas en los diccionarios académicos ibéricos: estudio comparativo. *RILEX. Revista sobre investigaciones léxicas 2 (2)*, pp. 37–63. DOI: http://dx.doi.org/10.17561/rilex.v2.n2.2.

Silvestre, J. P. (2008). *Bluteau e as Origens da Lexicografia Moderna.* Lisboa: INCM.

Silvestre, J. P. (2016). Lexicografia. In A. M. Martins & E. Carrilho (eds.). *Manual de Linguística Portuguesa.* Berlin: De Gruyter Mouton, pp. 200–223.

Simões, A. & Farinha, R. (2009). Dicionário Aberto: um recurso para processamento de linguagem natural. In *Viceversa: Revista Galega de Traducción*, v. 16. Spain: Vigo, pp. 159–171.

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

Tasovac, T. & Romary, L., et al. (2018). *TEI Lex-0: A baseline encoding for lexicographic data.* Version 0.8.6. DARIAH Working Group on Lexical Resources. https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html.

Terras, M., Nyhan, J. & Vahouette, E. (eds.) (2013). *Defining Digital Humanities: A*

Reader. London: Ashgate.

Verdelho, T. (2003). O Dicionário de Morais Silva e o Início da Lexicografia Moderna. *História Da Língua e História Da Gramática – Actas do Encontro*: 473–490. Braga: ILCH, Universidade do Minho.

Villalva, A. & Williams, G. (2019). *The Landscape of Lexicography.* Lisboa–Aveiro: Centro de Linguística da Universidade de Lisboa–Universidade de Aveiro.

Wiegand, H. E. (1984). On the Structure and Contents of a General Theory of Lexicography. In R. R. K. Hartmann (ed.), *LEXeter'83 Proceedings*, pp. 13–30.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Sci. Data3*:160018. DOI: 10.1038/sdata.2016.18.

Zgusta, L. (1971). *Manual of Lexicography.* Prague: Academia/The Hague: Mouton.

Mudra's Upper Sorbian-Czech dictionary – what can be done about this lexicographic "posthumous child"?

Michal Škrabal[1], Katja Brankačkec[2]

[1] Charles University, Institute of the Czech National Corpus,
Panská 7, 110 00 Praha 1, Czech Republic
[2] Institute of Slavonic Studies of the Czech Academy of Sciences,
Valentinská 1, 110 00 Praha 1, Czech Republic
E-mail: michal.skrabal@ff.cuni.cz, brankatschk@slu.cas.cz

Abstract

Jiří Mudra, among his numerous selfless activities, was a Czech *doyen* of Sorbian studies. He had been working for decades on an Upper Sorbian-Czech dictionary but, unfortunately, had not finished his work on it at the time of his death. Presently, we are considering completing Mudra's project. The material collected by Mudra is undoubtedly valuable for us, providing us with a launchpad for further work; still, it is necessary to challenge it with the current data and a modern lexicographic approach. The paper presents the proposed individual methods aimed at finishing the main body of the dictionary.

Every lexicographer works with the data and tools available in his or her time – and Mudra was certainly no exception. There is, therefore, no reason to maintain exaggerated reverence towards his dataset where it is in apparent conflict with the current language reality. The aim is not to foster Mudra's cult, but to acknowledge his admirable initiative and enthusiasm. The best way to do so is to complete his dictionary with all the possibilities currently offered to us and make it available – as the first academic dictionary in this language combination – to Czech users.

Keywords: Upper Sorbian-Czech dictionary; completion; Jiří Mudra

1. Introduction

In recent years, there has only been a small amount of literature on Upper Sorbian[1] (US) lexicography from the methodological perspective (Itoya, 2013; Šěrakowa, 2009; Pohončowa, 2008; Pohončowa & Šołćina, 2007), let alone on *digital* lexicography (Bartels et al., 2021). The following two sections summarise the leading publications on US lexicographic works from the past 30 years, including Jiří Mudra's unfinished US-Czech dictionary. The plan to complete this dictionary is then revealed in Chapter 2. The final chapter answers the question asked in the title of our paper.

[1] As our interest lies in US, we omit Lower Sorbian in our paper, referring only to these sources: Leszcyński (2013), Szpila (2014) and Bartels et al. (forthcoming).

1.1 US lexicography and dictionaries nowadays

Sorbian lexicography only stands on the threshold of the digital-born era. In the most recent paper on it (Bartels et al., 2021), we even find the categoric statement that "there is not even one Sorbian dictionary based on a systematic and extensive analysis of written texts". The authors describe their plans and their initial experience with the first project of digital lexicography targeting Sorbian. The project aims to identify neologisms and changes in the usage of the Upper and Lower Sorbian lexicon. For this purpose, a new corpus of Sorbian texts published by the Domowina publishing house (which covers an overwhelming majority of all officially published Sorbian texts) starting from 2019 is being created.

However, there are some dictionaries available digitally – both in Upper and Lower Sorbian. The most important of them is "Soblex",[2] which aggregates material from the following printed dictionaries and other language tools:

- the 5th edition of the US-German spelling dictionary (Völkel, 2005), initially prepared by P. Völkel (1931–1997). The first edition comes from 1970; further, revised editions appeared in 1976, 1979, and a larger version with spelling rules in 1981. The 5th edition, a revised and enlarged version with the new spelling rules, was prepared by T. Meškank (2005), also with a CD-ROM version (2008). It was edited again in 2014, and is now being processed in a modern way for a wholly revised 8th edition based on modern lexicographical methods that is envisaged to be published during the 2030s. Meanwhile, there also seems to be an internal review of the 7th edition, planned for 2022 (Bartels et al., 2021);
- the German-US dictionary of neologisms (Jenč et al., 2006). This latest printed dictionary (not considering reissues) is designed to extend the two-volume German-US dictionary published earlier (Jentsch et al., 1989, 1991). These older volumes should soon expand the Soblex infrastructure;
- the dictionary for native speakers in US schools (Hajduk-Veljkovićowa, 2017);
- the dictionary of Sorbian names (Meškank, 2017);
- various terminological dictionaries for US schools (1995–2017);
- a tool integrating US into the Linux OS (including US interface, spellchecker, dictionary of synonyms, among others) developed by E. Werner at the Leipzig University;[3]
- the US text corpus "Hotko", prepared by the Sorbian institute in Bautzen (2.3);
- the application "SorbOrto", developed by G. Nagora and G. Müller;
- the series produced for the US broadcasting programme "Rěčne kućiki" ('Language columns');[4]

[2] https://soblex.de

[3] https://hsb.l10n.kde.org/

[4] https://hornjoserbsce.de/kuciki/ and https://www.mdr.de/serbski-program/rozhlos/recny-

- the database of geographic exonyms prepared by the Sorbian Institute and the Witaj Language Centre;[5]
- the most recent add-on is a translating system "sotra"[6] ('sister') in both US-German-US directions. It works on the basis of the statistical translation software "Moses" that gradually learns from a parallel corpus prepared by the Witaj Language Centre.

There is also a set of older printed dictionaries available in digital form, e.g., the online version of the two-volume German-US dictionary (Jentsch et al., 1989, 1991) on the portal hornjoserbsce.de,[7] covering more than 36,000 entries. This website is closely linked to the Soblex dictionary and will soon become a part of it (Bejmak et al., forthcoming).

The older dictionaries by Kral (1927) and Pful (1866) were prepared in a digital version and published by the Sorbian Institute in 2006.[8] Unfortunately, this project has not been developed further technically, and some users may experience difficulties while trying to reach the server.

The small digital US-Czech dictionary (Martínek & Brankačkec, 2005) could also be helpful for our purposes. It was created by manually choosing about 10,000 entries from the US spelling dictionary by Völkel (1981) and translating the German part into Czech. A modest US-Czech dictionary (but also a LS-Czech one and even a Polabian-Czech glossary) is also available at D. Krčmařík's personal website,[9] unfortunately without any more detailed information.

A crowdsourced multilingual dictionary Glosbe[10] also includes US. It is not surprising that the largest amount of data can be found in the US-German (6,674 phrases; 1,133 examples) and US-English (6,454 phrases; 231 examples) sections. The US-Czech part, considering its limited list of entries (4,863 phrases; 252 examples – cf. US-Polish: 5,414/233), may be inspirational principally in terms of web design and engaging the lay public into the project; as regards the data itself, it is deficient in too many ways.

Another crowdsourced project – Wiktionary – also has its US version[11] (the LS version doesn't exist yet), with 4,176 entries.

kucik/index.html

[5] https://www.serbski-institut.de/os/Geografiske-mjena-hornjoserbsce/

[6] https://soblex.de/sotra/

[7] https://hornjoserbsce.de/dow/

[8] http://www.serbski-institut.de:8180/dict/online

[9] http://slovnik.vancl.eu/abc/index.php

[10] https://hsb.wiktionary.org/wiki/H%C5%82owna_strona

[11] https://app.glosbe.com (All statistics from the portal were valid as of 8 April 2021.)

A few remarks about the metalexicographic aspect: New dictionaries elicit only a little attention among language professionals, and there is a rather languorous discussion about these developments. Most of the dictionaries are reviewed only once, e.g., the English-US (Wornar, 2007) and the US-English (Stone, 2005) dictionaries were only examined by Szpila (2008), albeit in "unusual" detail. The German-US dictionary was reviewed twice (Lewaszkiewicz, 2008; Šěrakowa, 2009). A more vivid discussion can be observed on the so-called "new Völkel" (2005; see Pohončowa & Šołćina, 2006 for its review) that reflects changes in the US orthography, especially of more recent loanwords from German and English and with a change even in the alphabetic order: the grapheme *ć* was formerly arranged after *t*; newly, it follows *č*.

1.2 Jiří Mudra's US-Czech dictionary

Jiří Mudra (1921–2009) was a doyen of Sorbian studies in former Czechoslovakia. Besides undertaking numerous activities in organising, propagating, and interpreting Lusatian literature and other cultural artifacts (Kaleta, 2011), he was active in Sorbian linguistics. He is co-author of a four-volume US textbook (Mudra & Petr, 1982–1989), and for many decades he also worked on a US-Czech dictionary, which unfortunately remained unfinished. He left behind a manuscript of approximately 22,000 entries ranging from *a* to *smyknyć*, plus its digitised version. In his paper, regrettably too general and brief, Mudra himself characterised his planned *chef d'ouvre* (Mudra, 1999). He intended to make a dictionary that would be helpful for all Czech speakers interested in US. It should have been medium-sized, with approximately 30–40 thousand entries covering all grammar words and the most frequent content words in US; the frequency criterion could not be taken into account reliably due to the lack of empirical data. Although there were some US-Czech dictionaries, they were outdated, and their lists of entries were very limited along with having a primitive microstructure (Páta, 1920; Mohelský, 1948), an thus they would suffice to only cover users' elementary needs. Presumably, this was the primary impetus for Mudra to start work on his dictionary. Besides its own excerptions, it was based on then-recent US dictionaries (Völkel, 1981; Budarjowa, 1990; Korjeńk, 1995; Mehrowa & Pawlikowa, 1995; Trofymovyč, 1974; Jentsch et al., 1989, 1991) as well as some older ones (Jakubaš, 1954; Kral, 1927; even Pful, 1866), along with US grammar books (Faßke & Michalk, 1981; Šewc-Schuster, 1976, 1984). However, no corpora data were available to Mudra, unlike the current situation, which gives us a chance to finish the project.

2. Completing Mudra's dictionary

In this chapter, we present our plan to complete the main body of Mudra's dictionary. Naturally, the suggestions described below are mere theses – partly due to the limited space of the paper, and partly because we do not present a ready-made style guide for the lexicographic team (this should be created later, but we are still in the initial phase of the project). We will successively deal with various aspects: the current state of the project (2.1) and its possible pitfalls (2.6), technical issues (2.2), other potential data sources (2.3), specific proposals for changes in both the macrostructure and microstructure of the dictionary (2.4) as well as purely practical issues, such as the workflow of the whole project (2.5). We constantly consider the target users (especially 2.1); no matter how we define them – until the resulting dictionary will have been made available to the users, Mudra's work is not complete, thus *de facto* worthless.

2.1 Mudra's dictionary from our perspective

Currently, Mudra's lifelong lexicographic work looks like this:

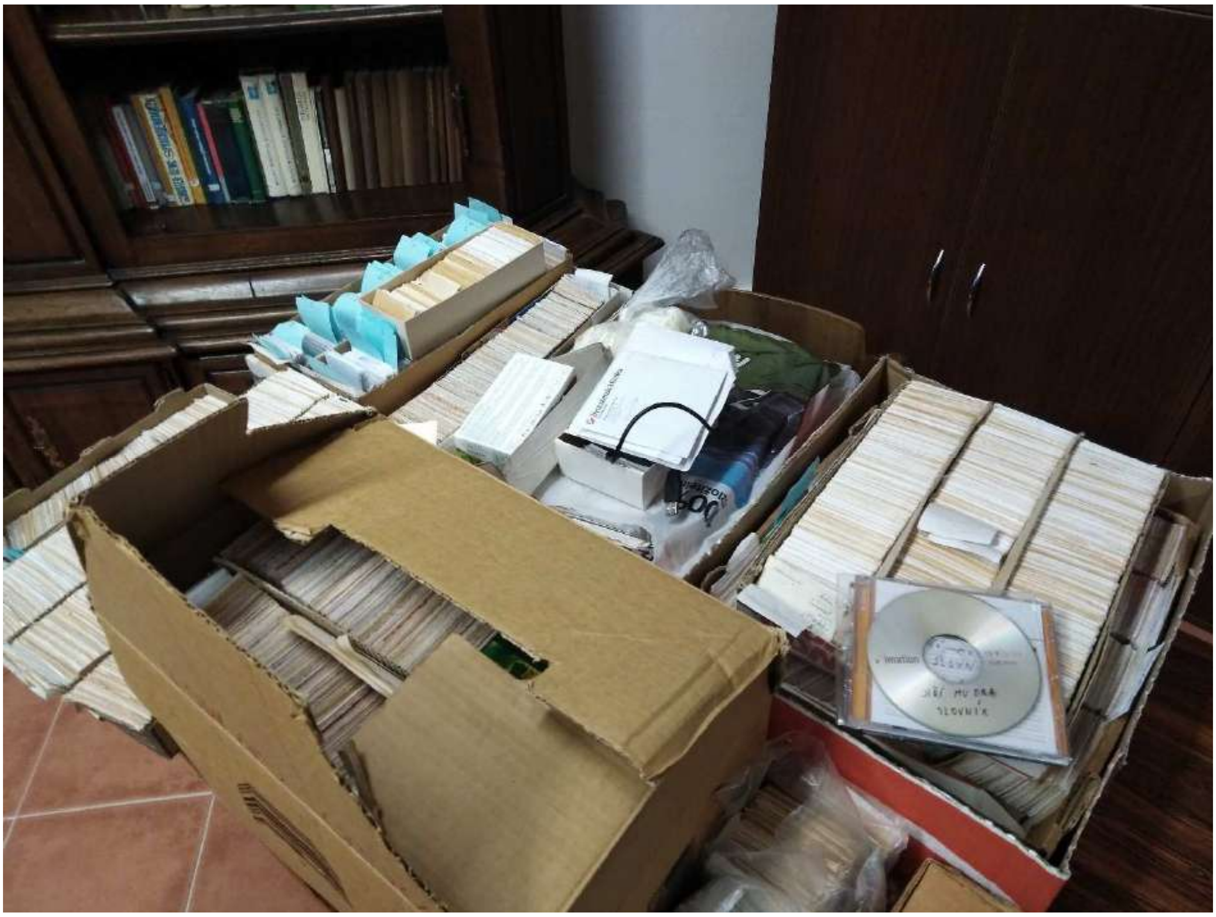


Figure 1: Jiří Mudra's lexicographical legacy

Such a view may arouse nostalgia for old-school lexicography, albeit along with regret about the incompleteness of the project. It does not matter how much data Mudra has collected and how good this data is. The work done to this point is practically useless until the dictionary starts to serve its users: specifically, if it is the only dictionary of its kind or if no decent dictionary exists in this language combination. Of course, the question arises whether it might not be better to start again from scratch, without the "burden" of pre-corpus lexicography. However, we do not want to do this – for several reasons. Firstly – and above all – we do not have sufficient human resources for a completely new project. Also, the US is still a low-resource language, with no representative sample of the current language of adequate quality available (see the Hotko corpus in 2.3). Furthermore, Mudra's material is too valuable to ignore, although not corpus-based. In the vast majority of it, it adequately represents the desired *equivalence* between the US and Czech lexicons. Any inaccuracies and errors can simply be revised; this task is definitely easier than pointless – not harmless! – drudgery. Last but not least, we consider Mudra an indisputable personality of Sorbian studies in Czechia, who deserves respect and credit. These are the main reasons why we prefer completing Mudra's dictionary, considering ourselves as editors of his preprocessed material, successors to his unfinished work. It does not mean that the work will proceed without a professional lexicographic and critical approach. The assembled data needs to be revised and supplemented with new knowledge from new sources (at least those from the last twenty years) which were not available to Mudra or deliberately excluded by him. In addition, it is necessary to process the rest of the alphabet (the rest of the letter S, followed by the letters Š–Ž).[12] Eventually, the outcome needs to be passed on in an acceptable form to users as soon as possible. We already consider the paper dictionary to be an obsolete form, preferring an electronic one, which, among others, allows the publishing of new entries progressively as they emerge or to update already published entries easily.

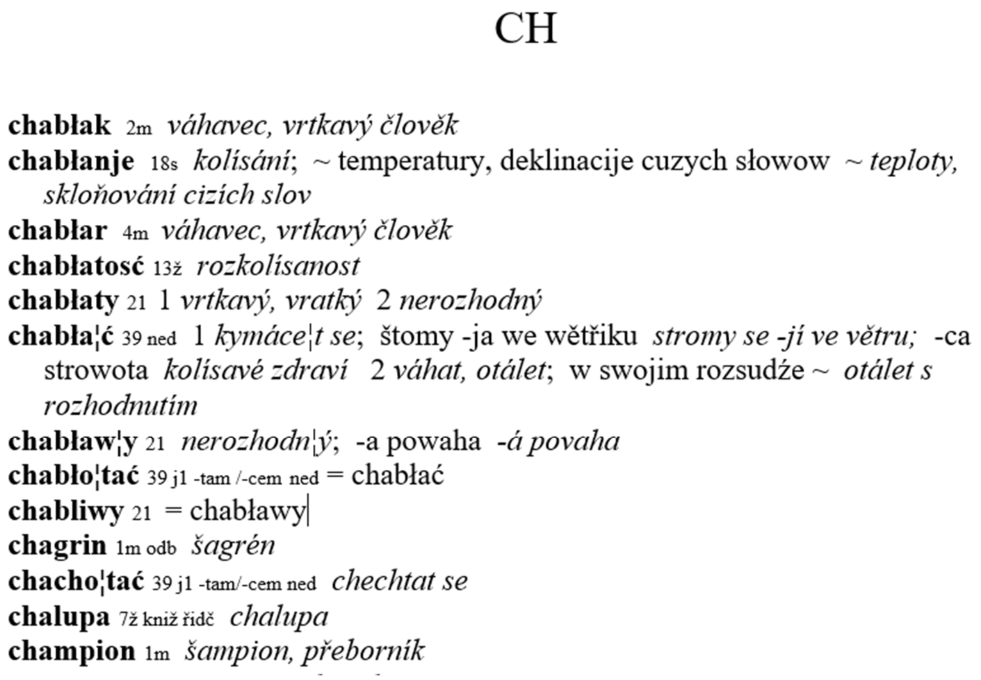
We define the target users broadly enough due to the specificity of the language combination. In such a situation, a universally designed dictionary accumulates the functions of different types of dictionaries (terminological, phraseological, vernacular, etc., but often conversation books too; cf. Mudra, 1999: 260). Therefore, we primarily aim at any Czech-speaking person interested in US (enthusiasts, linguists and other humanities scholars, translators, etc.) as well as Sorbian people interested in Czech. The dictionary concept should correspond to this: it should be universal enough to serve the practical needs of US students, along with those of people traveling to Lusatia, translators of US fiction, etc. Besides, the project's openness to the general public (2.5) explicitly counts on its active participation and can respond flexibly to its needs.

2.2 Technical issues

The CD in the lower right part of the photograph above suggests that the data has been digitised. Indeed, we have an MS Word file with the semi-finished dictionary,

[12] As Völkel suggests (approximately 52,000 entries on 661 pages), the remaining part covers almost *one-third* of the whole dictionary. Hopefully, this is sufficient proof we intend not just to publish the main body of Mudra's dictionary without making any contribution, but we understand our engagement as an equal partnership with a deceased colleague.

which looks like this:



CH

chabłak 2m *váhavec, vrtkavý člověk*
chabłanje 18s *kolísání*; ~ temperatury, deklinacije cuzych słowow ~ *teploty, skloňování cizích slov*
chabłar 4m *váhavec, vrtkavý člověk*
chabłatosć 13ž *rozkolísanost*
chabłaty 21 1 *vrtkavý, vratký* 2 *nerozhodný*
chabła¦ć 39 ned 1 *kymáce¦t se*; štomy -ja we wětřiku *stromy se -jí ve větru;* -ca strowota *kolísavé zdraví* 2 *váhat, otálet*; w swojim rozsudźe ~ *otálet s rozhodnutím*
chabław¦y 21 *nerozhodn¦ý*; -a powaha *-á povaha*
chabło¦tać 39 j1 -tam /-cem ned = chabłać
chabliwy 21 = chabławy
chagrin 1m odb *šagrén*
chacho¦tać 39 j1 -tam/-cem ned *chechtat se*
chalupa 7ž kniž řidč *chalupa*
champion 1m *šampion, přeborník*

Figure 2: An example from Mudra's unfinished dictionary in digitised form (MS Word), cf. Mudra, 1999: 261–262

As can be seen in the figure, the microstructure of the entry is quite simple.[13] The lemma is followed by grammatical information (numeric or alphameric code assigns the lexeme to the appropriate class of words, irregular and problematic forms appear) and a stylistic marker. The semantic part follows, i.e., an overview of meanings – or their Czech equivalents– supplemented sometimes by a few significant collocations or idioms. Figure 2 is an example of a typical lexicographic production in the Czech environment in the second part of the 20th century: an apparent effort to save space as much as possible is made due to the limited space of the printed dictionary. Nevertheless, we no longer have these limits today. Therefore, we can enrich entries with missing features (see 2.4 below), add more collocations, replace the opaque system of ciphers and codes with explicit metalanguage, etc.

We feel that continuing to work in MS Word or another text editor is an anachronism now, especially if numerous dictionary writing systems (DWS) are available, which have become the current lexicographic standard. The final choice has not yet been made, although some team members have good experience with the DWS TshwaneLex while working on another dictionary (Škrabal, 2016).

[13] Even simpler is the microstructure of entries in Mudra & Petr (1989), which is understandable, for it is not a full-fledged dictionary, just a practical tool for Czech students of US (it covers approximately 13,000 entries used in Mudra and Petr's textbook of US). Mudra intended it as a predecessor to a regular dictionary (Mudra, 1999: 260).

We primarily propose an online dictionary, and its website[14] should be responsive so that it can be used on various devices, not just a desktop computer. In the future, a stand-alone application for smartphones should be created too. A printed version is not foreseen, although processing the data does not necessarily exclude it. However, it will be a minor mode of using the data in a print-on-demand scenario if a potential applicant makes an explicit request. Instead, we imagine using printed materials for didactic purposes: the teacher prints out only a pertinent fragment from the dictionary for students, such as topic-related vocabulary for a relevant lesson, or lexemes belonging to the same word family, etc. This can be handled simply via tick-box features in most DWS and the subsequent filtering out of the selected entries.

2.3 Other data sources

Mudra's materials are in themselves partly a compilation (in the best sense of the word) of various dictionaries (1.1–1.2), supplemented by long-term excerpts, but these are inherently selective. Therefore, they should be challenged by other sources available today, mainly corpus data, and ideally, that of parallel corpora. Currently, the Hotko v2 corpus of US is available (via the KonText interface, see 2.5 below). It contains journalistic (57%), fiction (23%), religious, and scientific texts from the middle of the 19th century to the present that were scanned and OCR-ed, but not corrected. Besides, the corpus is neither lemmatised nor morphologically annotated, which can complicate the search. Some of the data, e.g., a number of dictionaries (12%), old texts reflecting historical spelling, or numerous German-language fragments, are unusable for our purpose. The total corpus size is currently 43.9 million tokens, including punctuation; more than half of the texts (54%) date from the 1990s onwards. Post-war texts seem relevant to us (with exceptions mentioned above), and such a subcorpus contains no more than 31 million tokens.

Later, it will also be possible to start using data from the parallel US-Czech corpus, which is planned as an extension of the InterCorp (IC) corpus (Čermák & Rosen, 2012). We currently have the first three texts ready for alignment (one of which is a translation from US, and two into US). We believe that IC is a good investment, as not only the US-Czech subcorpus is being created, but – with a suitable selection of texts[15] – also the subcorpora of other US-X language combinations. These can then serve as a dataset for other translation dictionaries.

[14] We are counting here on our own website although, in theory and after mutual agreement, it would be possible to connect to the already existing infrastructure, such as the above-mentioned Soblex dictionary, that nowadays is the US-German-US dictionary only. The portal could be extended by several language modules (Czech, Polish, …), with US as a pivot language.

[15] The ideal dataset would primarily consist of as many US originals as possible, but these are only seldom translated into foreign languages. For example, Jurij Brězan's novel *Stary nan* can appear in IC in the Czech, Russian, or Slovak translations, while the US translation of Douglas Adams *The Hitchhiker's Guide to the Galaxy* could be aligned to another 25 languages along with the English original.

2.4 Amendments in the macrostructure and microstructure of Mudra's dictionary

The basic data structure (Fig. 2) will be converted to DWS, and other desirable elements and attributes will extend the DTD structure. It is necessary, among other tasks, to predefine various grammatical and semantic classes of words, etc. Frequencies should play an important role, as they were not taken into account in pre-corpus times, at least in the Czech environment.[16] After all, Mudra himself was aware of the importance of frequency data; in his paper (1999: 261) he expressed regret about the non-existence of both a US frequency dictionary and a dictionary of spoken US. He also could not rely on corpus data in his work, as the first US publicly available corpora did not appear until 2013 (Hotko v1). Nevertheless, we believe that he would appreciate the corpus as an amazing tool for lexicographers.

We do not want to overestimate the frequency data, being well aware of the problems of corpus data (2.3) that should thus be treated with caution. However, it is possible to cover frequency information in several ways:

- by dividing the lexicon into a reasonable number of frequency bands (e.g., *very common – common – unusual – sparse/idiosyncratic*);
- by extracting the top frequency lexemes (e.g., the top 1,000 words) reliably even on a relatively small corpus, such as Hotko, and then marking them directly in the dictionary with a proper graphic means (suitable mainly for didactic purposes);
- by stating the frequency data (i.p.m. preferably) explicitly for selected words.

We want to extend the number of collocations, especially idioms; the exemplification should be the central part of the revisited entry. Appropriately selected examples from IC will be quoted along with their Czech translations; examples from non-parallel corpora will be provided with ad-hoc translations. Besides, the user can find more exemplification directly in a given corpus via a hypertext link.

The pronunciation is entirely neglected by Mudra as he considers it to be regular, perhaps with some rare exceptions and in the case of loanwords. We see a wise solution in audio recordings directly from native speakers, whether professionals or amateurs.[17] This is the area where we believe crowdsourcing can be applied most effectively (see also 2.5 below).

[16] With an exception of a few markers denoting either the uncommon or obsolete usage of words.

[17] Cf. the US section of the pronunciation guide Forvo.com: https://forvo.com/languages/hsb/.

The pragmatic aspect of word usage is also essential: our dictionary should inform the users about the word's specific place within the lexicon and even warn them of possible negative reactions (e.g., to offensive and vulgar words). Usage notes are not a common phenomenon in the Czech lexicographic tradition (Šemelík & Škrabal, 2019), and we understand their inclusion in our dictionary as partial repayment of this debt. In a specific US-Czech combination, notes can also be used to alert false friends explicitly or, in general, any potentially problematic places for a Czech user.

As far as the macrostructure of the dictionary is concerned, it will also undergo some revisions and add-ons. Mudra avoided some (from his point of view) problematic groups of words in the list of entries, such as Germanisms or vulgarisms, tending towards a literary language that may sound somewhat artificial today. We cannot identify with this protective approach, as it does not correspond to our descriptive basis. The lexicographic description should be in accordance with the language reality. Any language taboo is also inadmissible in our eyes: vulgar words have a valid place in the lexicon of each language and should be therefore described, although with special treatment and means (be it markers, usage notes, or other features).

Above all, the current colloquial form of US should be bolstered, not withdrawing from new loanwords from both German and English. Naturally, data from the US spoken corpora would be precious, yet, this is still a long way off, and we must be content even with data from written corpora and the most recent US dictionaries such as Völkel (2005) and Jenč et al. (2006).

2.5 Workflow

Our project's workflow depends mainly on the amount of funding available within the appropriate grant. We expect the involvement of two academic workplaces (the Institute of Slavonic Studies of the Czech Academy of Sciences and Institute of the Czech National Corpus) and one civic body (Society of Friends of Lusatia). The former will provide the relevant know-how and infrastructure[18] while engaging the lay public is an opportunity to use the "wisdom of crowds" (Surowiecki, 2005). Amateur volunteers can contribute to us in various ways: be it by notifying us of errors in the dictionary, suggesting changes or additions in the list of entries or individual entries, recording the pronunciation, etc. We also want to involve Czech students of Sorbian studies[19] (or Sorbian students of Czech); they could learn the basics of practical

[18] ISS has numerous and long-term experience in the field of Slavic lexicography. The Russian-Czech electronic dictionary database with the Large Czech-Russian Dictionary (http://slovnik.slu.cas.cz), as well as the digital portal in Old Slavonic *Gorazd* (http://gorazd.org/gulliver/), are digitally accessible to the public. Besides its major project (Czech National Corpus, aimed at studying the Czech language), ICNC provides infrastructure for parallel corpora, including the emerging US component (see 2.3 above), and it also hosts the Hotko corpus via the KonText interface (Machálek, 2014).

[19] Currently implemented at the Faculty of Arts, Charles University in the form of an optional

lexicography and related linguistic disciplines (lexicology, corpus linguistics, etc.) during a suitably designed course or a hands-on workshop.

2.6 Problematic issues

In addition to the already mentioned need for grant support, the most severe difficulty concerns the copyright to Mudra's work. Negotiations with the heirs of these rights are not without problems. The idea of the dictionary as a tangible book artifact is too entrenched in laypeople, and its virtual form is difficult to accept. (Although more and more people do not even use paper dictionaries, which are too cumbersome for them, and prefer online resources.) The fear of insufficient acknowledgment of Mudra's life-long effort, of the appropriation of his work and merits probably also plays a role. However, these fears seem odd: all of Mudra's data used by us will be appropriately marked in the dictionary (by an icon, cipher, etc.), and his name will, of course, be listed *first* in the list of used sources. Besides, we want to place Mudra's biographical profile on the dictionary's website, which would summarise his life and Sorabist career (1.2).[20]

Despite certain complications, we believe in the successful outcome of the negotiations. Without this, Mudra's work – undoubtedly remarkable and worthy of respect – will remain a mere fragment (Fig. 1 and 2) instead of fulfilling its purpose and receiving the attention and recognition it deserves.

3. Conclusion

To answer the question in the title of this paper: what can be done about Mudra's lexicographic "posthumous child"? A lot – under two basic assumptions: a) that we obtain the permission of the copyright heirs for the free (yet not arbitrary) handling of data, and b) that we obtain sufficient grant support. In such a case, we can complete the unfinished work, which otherwise would end up irretrievable and useless, and thus give Mudra's project appropriate credit in the Czech Sorabist milieu, even among the youngest generation which no longer remembers or knew the doyen. It is evident that the material assembled by Mudra (as by anyone else) needs to be approached critically, with hands not bound by exaggerated piety. The aim is to reconcile this material with both the additional data sources and the modern lexicographic approach, as we have tried to describe in this paper. Only in this way can the resulting dictionary begin to fulfil its primary purpose, i.e., to serve its users as effectively as possible. There may not be a vast number of them (we estimate it as in the hundreds to low thousands), but this does not reduce their need for quality lexicographic tools.

practical course in US. In the first year, it was attended by approximately 10 students; in distance learning during the pandemic, the number of students decreased to 3–5.

[20] Regrettably, Jiří Mudra does not have his own article on either the Czech or US Wikipedia yet; the most complete biographical source is an anthology, published posthumously in homage to Mudra (Kaleta, 2011).

4. Acknowledgements

We are grateful to anonymous reviewers for their constructive feedback.

5. References

Bartels, H. & Kaulfürst, F. & Szczepański, M. & Wölke, S. (2021). Das Monitoring des sorbischen Schrifttums. Grundlagen und erster Jahresbericht eines neuen Forschungsvorhabens. *Lětopis* 68(1), pp. 77–136.

Bayer, M. (2006). *Sprachkontakt deutsch-slavisch. Eine kontrastive Interferenzstudie am Beispiel des Ober- und Niedersorbischen, Kärntnerslovenischen und Burgenlandkroatischen.* Frankfurt am Main: Peter Lang.

Bejmak W. & Schmiedelowa, A. & Szczepański, M. & Wölkowa, S. (forthcoming). *Němsko-hornjoserbski słownik na Hornjoserbsce.de.* Bautzen: Serbski institute, Accessed at: https://obersorbisch.de/dow/info (22 March 2021).

Budarjowa, L. (1990). *Wörterbuch Obersorbisch-deutsch / Słownik hornjoserbsko-němski.* Bautzen: Domowina-Verlag.

Čermák, F. & Rosen, A. (2012). The case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics*, 17(3), pp. 411–427.

Faßke, H. & Michalk, S. (1981). *Grammatik der obersorbischen Schriftsprache der Gegenwart. Morphologie.* Bautzen: Domowina-Verlag.

Hajduk-Veljkovićowa, L. (2017). *Kak to jenož praju? Přiručka za bohatši słowoskład za wyšu šulu a gymnazij.* Budyšin: Domowina.

Hornjoserbsce.de/obersorbisch.de. Bautzen: Serbski institut. Accessed at: https://obersorbisch.de/dow/info (22 March 2021).

Itoya, B. (2013). Badania metod kodyfikacji leksyki górnołużyckiej w zestawieniu z dorobkiem metaleksykografii i leksykografii czeskiej. In M. Milewska-Stawiany & S. Wölkowa (eds.) *Leksikologiske přinoški II. IV. Seminar serbskeje słowotwórby a leksiki. IV Seminarium Słowotwórstwa i Słownictwa Łużyckiego. Uniwersytet Gdański / Serbski institut 31. 5. – 1. 6. 2012.* Bautzen, pp. 102–115.

Jakubaš, F. (1954). *Hornjoserbsko-němski słownik. Obersorbisch-deutsches Wörterbuch.* Bautzen

Jenč, H. & Pohončowa, A. & Šołćina, J. (2006). *Němsko-hornjoserbski słownik noweje leksiki. Deutsch-obersorbisches Wörterbuch neuer Lexik.* Bautzen: Domowina-Verlag.

Jentsch, H. & Michalk, S. & Schierack, I. (1989, 1991). *Deutsch-Obersorbisches Wörterbuch A–K/ L–Z.* Bautzen: Domowina-Verlag.

Kaleta, P. (2011). *Jiří Mudra. K 90. výročí narození českého vlastence a přítele Lužických Srbů (s albem fotografii).* Praha / Budyšin: Společnost přátel Lužice / Maćica serbska.

Korjeńk, J. (1995). *Terminologija za předmjet fyzika: němsko-hornjoserbsce, hornjosebsko-němsce.* Budyšin: Domowina.

Kral, J. (1927). *Serbsko-němski słownik hornjołužiskeje serbskeje rěče. Sorbisch-deutsches Wörterbuch der Oberlausitzer sorbischen Sprache.* Budyšin: Maćica serbska.

Krčmařík, D. *HLS slovník.* Accessed at: http://slovnik.vancl.eu/abc/index.php (22 March 2021).

Leszcyński, R. (2013). *Podręczny słownik polsko-dolnołużycki/dolnołużycko-polski. Pśirucny słownik pólsko-dolnoserbski/dolnoserbsko-pólski.* Budyšin: Domowina.

Lewaszkiewicz, T. (2008). Recenzje. Helmut Jentsch / H. Jenč, Anja Pohontsch / A. Pohončowa, Jana Schulz / J. Šołćina, Deutsch-obersorbisches Wörterbuch neuer Lexik/ Němsko-hornjoserbski słownik noweje leksiki. Domowina-Verlag / Ludowe nakładnistwo Domowina, Bautzen / Budyšin 2006. *Slavia Occidentalis* 65, pp. 163–164.

Machálek, T. (2014). *Kontext – rozhraní pro vyhledávání v korpusech.* Praha: Filozofická fakulta Univerzity Karlovy. https://kontext.korpus.cz (8 April 2021).

Martínek, F. & Brankačkec, K. (2005). *Hornjoserbsko-čěski słownik.* Accessed at: http://www.serbski-institut.de:8180/dict/ and: https://www.serbski-institut.de/de/Online-Publikationen/ (22 March 2021).

Mehrowa, H. & Pawlikowa, B. (1996). *Terminologija za předmjet geografia: němsko-hornjoserbsce, hornjosebsko-němsce.* Budyšin: Domowina.

Meškank, T. (2017). *Serbske předmjena. Serbske pśedmjenja. Sorbische/wendische Vornamen.* Bautzen: Witaj-Sprachzentrum.

Mohelský, V. (1948). *Mluvnice hornolužické srbštiny a slovník hornosrbsko - český = Rěčnica hornjołužiskeje serbšćiny a słownik hornjoserbsko-čěski.* Olomouc: V. Tomek.

Mudra, J. & Petr, J. (1982, 1987, 1989). *Učebnice lužické srbštiny.* Praha: Filozofická fakulta Univerzity Karlovy.

Mudra, J. (1999). K problematice hornjoserbskeho-čěskeho słownika. *Lětopis* 46. *Wosebity zešiwk*, pp. 260–263.

Páta, J. (1920). *Kapesní slovník lužicko-česko-jihoslovanský a česko-lužický.* Praha: Českolužický spolek „Adolf Černý“.

Pful, K. B. (1866). *Łužiski serbski słownik.* Budyšin: Maćica serbska.

Pohončowa, A. (2008). Neologizmy w hornjoserbskej spisownej rěči a jich leksikografiske fiksěrowanje. In: M. Milewska-Stawiany & S. Wölkowa, S. (eds.) *Leksikologiske přinoški: III. Seminar serbskeje słowotwórby.* Bautzen: Domowina, pp. 43–49.

Pohončowa, A. & Šołćina, J. (2007). Strategije modernizowanja hornjoserbskeho słowoskłada. *Rozhlad* 57, 1, pp. 5–9.

Pohončowa, A. & Šołćina, J. (2006). Nowy prawopisny słownik – wšitko jasne? (= Das neue Orthographiewörterbuch – alles klar?). *Serbska šula* 59(1), pp. 11–12.

Schuster-Šewc, H. (1976, 1984). *Gramatika hornjoserbskeje rěče I, II.* Budyšin: Domowina.

Serbski Institut Budyšin (2021). *HOTKO: hornolužický textový korpus, verze 2 z 6. 3. 2021.* Praha: Ústav Českého národního korpusu. Accessed at: https://www.korpus.cz (22 March 2021).

Serbski Institut Budyšin (2021). *SobLex. Hornjoserbsko-němski słownik. V3.03.06 14.01.2021.* Available online: https://www.soblex.de (22 March 2021).

Siatkowska, E. & Leszczyński, R. (2002). *Słownik polsko-górnołużycki i górnołużycko-polski. Pólsko-hornjoserbski a hornjoserbsko-pólski słownik.* Warszawa: Uniwersytet Warszawski, Instytut Filologii Słowiańskie.

Surowiecki, J. (2005). *The Wisdom of Crowds: Why the Many are Smarter than the Few and How Collective Wisdom Shapes Business, Economies, Societies, and Nations.* New York: Doubleday.

Szpila, G. (2008). Edward Wornar: Jendźelsko-hornjoserbski šulski słownik. English-Upper Sorbian Learner's Dictionary. Recensija. *Lětopis* 55(1), pp. 158–163.

Szpila, G. (2014). Rafał Leszczyński: Podręczny słownik polsko-dolnołużycki/dolnołużycko-polski. Pśirucny słownik pólsko-dolnoserbski/dolnoserbsko-pólski. *Lětopis* 61(2), pp. 157–162.

Šemelík, M. & Škrabal, M. (2019). Poznámky k poznámkám. Usage notes v českém lexikografickém prostředí. *Naše řeč* 102(1), pp. 25–35.

Šěrakowa, I. (2009). Moderne hornjoserbske słownistwo we wužitnym słowniku. *Rozhlad* 59(6), pp. 222–224.

Škrabal, M. (2016). 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.

Šołćina, J. (1999). Recenzija: Rafał Leszczyński: Hornjoserbsko-pólski a pólsko-hornjoserbski słownik pozdatnych ekwiwalentow / Górnołuzycko-polski i polsko-górnołuzycki słownik ekwiwalentów pozornych, Warszawa 1996. *Lětopis* 46(1), pp. 157–161.

Trofymovyč, K. K. (1974). *Hornjoserbsko-ruski słownik = Verchnelužicko-russkij slovar'.* Budyšin / Moskva: Domowina / Russkij jazyk.

Völkel, P. (1970–1981). *Prawopisny słownik. Hornjoserbsko-němski słownik. Obersorbisch-deutsches Wörterbuch.* Bautzen: Domowina.

Völkel, P. (2005). *Prawopisny słownik. Hornjoserbsko-němski słownik. Obersorbisch-deutsches Wörterbuch.* Bautzen: Domowina.

Wölkowa, S. & Pohončowa, A. & Šołćina, J. Accessed at: https://hornjoserbsce.de/kuciki/ (22 March 2021).

Wornar, E. (2007). *Jendźelsko-hornjoserbski šulski słownik. English-Upper Sorbian Learner's Dictionary.* Budyšin: Domowina.

Living Dictionaries: An Electronic Lexicography Tool for Community Activists

Gregory D. S. Anderson[1], Anna Luisa Daigneault[1]

[1]Living Tongues Institute for Endangered Languages, 4676 Commercial St SE, #454.
Salem, OR 97302, USA

E-mail: gdsa@livingtongues.org, annaluisa@livingtongues.org

Abstract

Living Dictionaries are comprehensive, free online technological tools integrating audio, images and other multimedia that can assist endangered and other language communities, providing a simple way to create high-quality multilingual documentation records. The platform is a progressive web application functioning within any Internet browser on any computer or mobile device, Android or iOS. If needed, Living Dictionaries can be created, managed and edited using only smartphones or tablets, which can function as complete workstations for recording and entering linguistic data and other multimedia. Living Dictionaries may be public or private and may include written entries with translations and example sentences in multiple languages and scripts, audiovisual files, parts of speech and semantic domains, morphosyntactic linguistic analysis and be tagged with other metadata. The platform is free because for almost all minority language communities the costs related to producing high-quality linguistic materials can be insurmountable. A moral imperative of the 21st century is the decolonisation and democratisation of linguistic resources. Online dictionaries should reflect the user communities, tailored to suit their needs as well as curated by citizen-linguists. Community resources have greater uptake and engagement by communities if they take a primary role in developing them.

Keywords: dictionary; language technology; endangered languages; lexicography; web application

1. Introduction

Technology can be “disruptive” because it can forever change the way people operate in their daily lives. But what if technology could also “disrupt” language bias and privilege? What if access to certain language technologies could help challenge language hierarchies and give endangered languages a fighting chance of survival? With over 3,000 languages in danger of being lost before the end of the century, we know there is a need to act quickly. Living Dictionaries[1] address the urgent need to provide comprehensive, free online technological tools that can assist endangered language communities simultaneously in conservation efforts and revitalisation programs by providing a simple way to create high-quality language documentation records. The Living Dictionaries platform can accommodate everyone from seasoned field linguists to emerging language activists in developing countries. The platform is free to use, and

[1] Available worldwide online at https://livingdictionaries.app/

the intended target audience of this web app is inclusive, diverse and multilingual.

2. The Advantages of Creating Digital Dictionaries in the 21st Century

The advantages of online dictionaries have been well-known for some time. Dmitrova et al. (2009: 77) discussed such features as their wide accessibility, the possibility for them to be continuously updated as well as corrected and edited, or the creation of an online community of real-time users in multiple different locations, which can lead to real-time editing and updating of the dictionary. Lew and de Schryver (2014: 345) concur that "[o]nline dictionaries as well as dictionary apps can be updated as often as needed, and all users can instantly benefit from the improved content or features right from the moment these become available." Dmitrova et al. (2009: 77) also commented on a key feature of online digital dictionaries: no restrictions on the size. Indeed, the old dictionary-making paradigm was dependent on printing restrictions, content limits, page layouts, alphabetisation and other 'corporate' concerns, where when updating dictionaries "the editors usually had to grapple with the dilemma of what to sacrifice in order to make space for the new items," (Lew and de Schryver, 2014: 345). Today, these types of bottom-line concerns are largely irrelevant, and powerful search functionality and the relatively low cost of database storage has obviated the challenges of the past. As Lew and de Schryver (ibid.) put it: "[t]he digital revolution has changed that, and now items are in fact very rarely removed when digital dictionaries are updated." Other innovative advantages of electronic dictionaries include: "the option to hear new words being pronounced, being able to copy over foreign scripts one would be hard pressed to type in, the interconnectivity with other resources (such as corpora), and the fact that one stays within the same (digital) medium, rather than having to move back and forth between the screen and a book on one's desk" (Lew & de Schryver, 2014: 347). Furthermore, we now benefit from the possibility of integrating large numbers of photos and other audiovisual multimedia, the ability to accommodate sign as well as oral languages, and perhaps most importantly, the capacity to address the vast gap in digital resource availability that disproportionately impacts minority communities worldwide. A multimedia online dictionary platform such as Living Dictionaries accommodates the needs of twenty-first century users of such tools by using the latest technologies to produce tools that in the long run can become encyclopaedic in nature.

3. The Impacts of Colonisation on Under-Represented Languages

Colonialism has had a deep impact on most countries of the world. The legal and social status of minority and under-represented languages, as well as the resources that support them, are characterised by unequal distribution and injustice in almost every polity across the globe. The linguistic consequences of colonialism entail in some cases the nearly complete elimination of most of the original languages spoken on a conquered territory, and the nearly complete domination of the colonial language, e.g., Russian in Siberia (Anderson, 2017), English in the US and Australia, and Spanish in most countries in Central and South America. In other cases, this means the enfranchisement of a group who acquired power within the colonial structure and have held it in the postcolonial period, and who in a similar neocolonial hegemonic manner promote their language as a national one over others also spoken in the country, (e.g., Setswana in Botswana, Burmese in Myanmar or Hindi in India) or regionally within a section of the country, e.g., Hausa in northern Nigeria. In some countries, constructed national languages have been vigorously promoted at the cost of others in the country, e.g., in Indonesia or Philippines, which have rebranded de-ethnicised versions of languages of the just mentioned type as national ones, whether a neocolonialist hegemonic language (Filipino) or a former urban/trade lingua franca (colloquial Malay > Bahasa Indonesia). In Melanesia, colonial-era contact languages were adopted as national ones and are promoted at the expense of others, leading to a decline in linguistic diversity over time. With very few exceptions, most nation-states favour a single language of one of these types over all others spoken in their territory. This institutionalised disenfranchisement has resulted in half of the world's languages presently undergoing an active shift towards dominant languages, and another 40% or so being threatened in such a way that this process will likely begin soon.

The main reason dominant language groups use to justify continued disenfranchisement of the minority languages of their countries is that it is too costly to support all languages. They also believe a subtractive language policy is the best means for ensuring a kind of national sense of self and to maintain territorial integrity. Both reasons are false. The latter belief is rooted in a continuation and naturalization of European Romantic/Herderian notions creating an ideal of one nation, one people, one language. With regards to the financial impacts of multilingualism, the actual costs of maintaining language diversity have been shown to be not nearly as high as imagined (Grin, 2003). The mindset regarding linguistic diversity thus needs to evolve: diverse languages need to be seen as resources that empower nations and not weaken them.

While for nations the financial cost of supporting multilingualism is not preventative in the way typically imagined, for almost all minority language communities the costs related to producing high-quality linguistic materials can be insurmountable. As activists in the field of endangered language documentation globally, we know this is to be true. Thus, we have created a state-of-the-art dictionary-builder that we have made available free of charge to all users. Through the Living Dictionaries platform, the Living Tongues Institute has approached solutions to the massive global language extinction crisis by attempting to obviate institutionalised barriers that prevent equal status and equitable treatment of all forms of linguistic communication. Training local people to conduct language documentation and revitalisation work and build dictionaries for their own communities is a core, long-term aspect of our approach.

A moral imperative of the 21st century is the decolonisation and democratisation of linguistic resources, as colonised peoples have often been forcibly resettled, assimilated and disenfranchised from their own heritage. Indeed, it can be almost impossible for marginalised people in some parts of the world to even access documented knowledge about their languages. Prinsloo (2019: 218), citing CCURL 2014, succinctly summarises one of the realities facing many minority language communities as follows: "[u]nder-resourced languages suffer from a chronic lack of available resources (human-, financial-, time- and data-wise), and of the fragmentation of efforts in resource development. This often leads to small resources only usable for limited purposes [...] without much connection with other resources and initiatives."

Now, through the accessibility of online digital media collections, scholars and activists have a great opportunity (and indeed a duty) to connect communities with the data they are entitled to. Under-represented languages need online resources to thrive in the digital era because people need to be able to easily store, reference and share content in their languages. To be sure, the Internet is a place where linguistic hierarchies in theory could be potentially upended, subverted and reinvented according to the needs of individuals and communities. Technologists and digital lexicographers must thus be publicly inclusive when it comes to minority languages and take a positive stance towards multilingualism. We advocate for an inclusive, citizen science approach to digital lexicography. Living Dictionaries address the obvious need to provide comprehensive, free access to robust technological resources. This platform provides an easy-to-use framework for systematically storing and sharing dictionary data in thousands of endangered languages, thus increasing their viability for survival in the long-term. This comes with significant implications: studies in North America and Australia show that language revitalisation leads to better mental health, better performance in schools, and expanded economic opportunity (Whalen et al., 2016).

4. Citizen Science: The Future of Lexicography

The very concept of a dictionary has changed in this new era. Lew and de Schryver observe (2014: 342),

> "[a]s dictionaries moved from the bookshelves gradually onto [...] internet servers, and now mobile devices, they found themselves as it were in the same league as utility and productivity software, which in turn encouraged a more pragmatic and less ideological or dogmatic view of dictionaries. This trend was only strengthened as users themselves started getting involved in bottom-up dictionary-making."

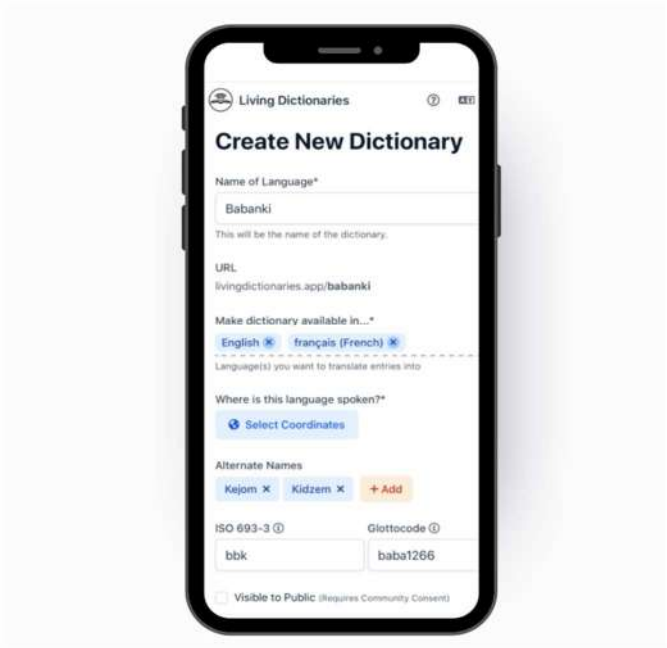
Online dictionaries can now reflect the user communities in a meaningful way, they can be tailored to suit their needs as well as curated by citizen-linguists who wish to build resources for their languages. No longer the exclusive domain of academic expert authorities and state-sanctioned language academies, digital dictionaries of the electronic era indeed belong to the realm of the collective intellectual property of language communities themselves. We strongly feel that for endangered and threatened minority languages, the future of lexicography is crowd-sourced citizen science.[2] Community resources developed by community members are almost certain to have greater uptake and engagement by communities if they take a primary role in developing these resources themselves. Speaking about (South) Africa, Prinsloo (2019: 220) reminds us that "[w]hat is emphasised and encouraged today is the urge to compile dictionaries for African languages in Africa, by Africans, for Africans", see also Prinsloo et al. (2017). This includes taking into consideration, among other things, that the complex grammatical structures of many African languages differ rather significantly from those of other major European and Asian languages (Van Wyk, 1995). During our online and in-person training workshops at the Living Tongues Institute for Endangered Languages, language activists who are facing rapid language loss have enthusiastically voiced their desire to create and maintain their own digital resources. We have created the Living Dictionaries platform with them in mind, optimising it for global remote collaboration, ease of use and accessibility on mobile devices, and we integrate community user feedback into the design and programming of the tool.

The Living Tongues Institute stands at the intersection of linguistics and activism, with the capacity to launch technological solutions that help aspiring language activists and scholars alike. Our team has adopted a vertically integrated approach to language documentation, in which local language consultants learn transferable digital and scientific research skills to eventually become research assistants, colleagues, and ambassadors for their languages. By facilitating in-person and online workshops during which we train local indigenous language activists to record and edit words and phrases in their native languages, we have developed a strong strategy that prioritises documentation as well as professional empowerment. Documenting languages is not only important to the scientific field of linguistics, but also to speech communities who are urgently looking for tools to combat language loss, and it is also crucial to conserving humanity's intangible heritage. It is up to our generation to use the tools of globalisation to empower those who have been disenfranchised. We consider this project a humanitarian mission that requires collaboration between scientists and local activists to make a difference. By pairing technology with our passion to document endangered languages, our platform is positioned to make a big impact on this field. The work we do is essential to help bolster the contemporary linguistic identity of the communities we serve and ensure a future for them. The materials and resources we create in collaboration with citizen-linguists will become the driving force that helps our descendants revitalise their languages in the future.

[2] Note that this does not mean that we advocate for the use of search engines to replace dictionaries, an alarm sounded among others by participants at Australex 2019, who fear it is becoming widely believed that dictionaries are no longer needed. Rather, we advocate for providing an easy to use, multimedia online digital dictionary tool that can create quality, multilingual (or monolingual) lexicographic resources for the widest possible range of languages worldwide.

5. Living Dictionaries: Set-up and Design Considerations

While much of our user community grapples with limited Internet connectivity and digital literacy, they regularly have access to smartphones and other mobile devices that can function as complete workstations for recording and entering linguistic data and other multimedia. Living Dictionaries are fully creatable, manageable and editable using mobile technology alone. The platform is a web-based application that functions within any Internet browser on any device, whether it is Android or iOS. The software works seamlessly across all mobile devices and tablets as well as desktop computers, and a service worker allows some features to be used offline in locations with limited Internet connectivity (more details on this below).



Create New Dictionary

Name of Language*

Babanki

This will be the name of the dictionary.

URL

livingdictionaries.app/babanki

Make dictionary available in...*

English ✕ français (French) ✕

Language(s) you want to translate entries into

Where is this language spoken?*

Select Coordinates

Alternate Names

Kejom ✕ Kidzem ✕ + Add

ISO 693-3 ⓘ

bbk

Glottocode ⓘ

baba1266

□ Visible to Public (Requires Community Consent)

Figure 1: A mobile mock-up view of the creation of the Babanki Living Dictionary

Once a user registers for an account on the platform, they may create a new Living Dictionary right away, and become a manager of that dictionary. All of this, as well as the entry functions described below, can be done on mobile or desktop. Figure 1 shows the mobile view of the digital information required to create a new Living Dictionary for Babanki, a Grassfields Bantu language spoken by under 40,000 people in Cameroon (the depiction is based on how the process looks in a Chrome browser on an iPhone 6+). The dictionary creation process can take as little as a couple of minutes, or a bit longer if the dictionary manager needs to search online for the metadata relevant to their language project. We made the set-up process very user-friendly and fast so that activists can easily start their dictionary projects with as few bottlenecks as possible, and no institutional red tape. They do not have to go through the website administrators or through any type of approval process to get started.

Among the information requested to create a dictionary is the name of the language, a string of data which in turn automatically populates the ending of the URL of the new dictionary. The name attributed to the dictionary itself can be modified by the manager at any time in the left sidebar “Settings” tab of their Living Dictionary. For example, communities may wish to modify the spelling or add an additional name in parentheses to the dictionary, to reflect contemporary ways of referring to the language. The URL, however, cannot be changed after it is established because it becomes hardcoded into the website.

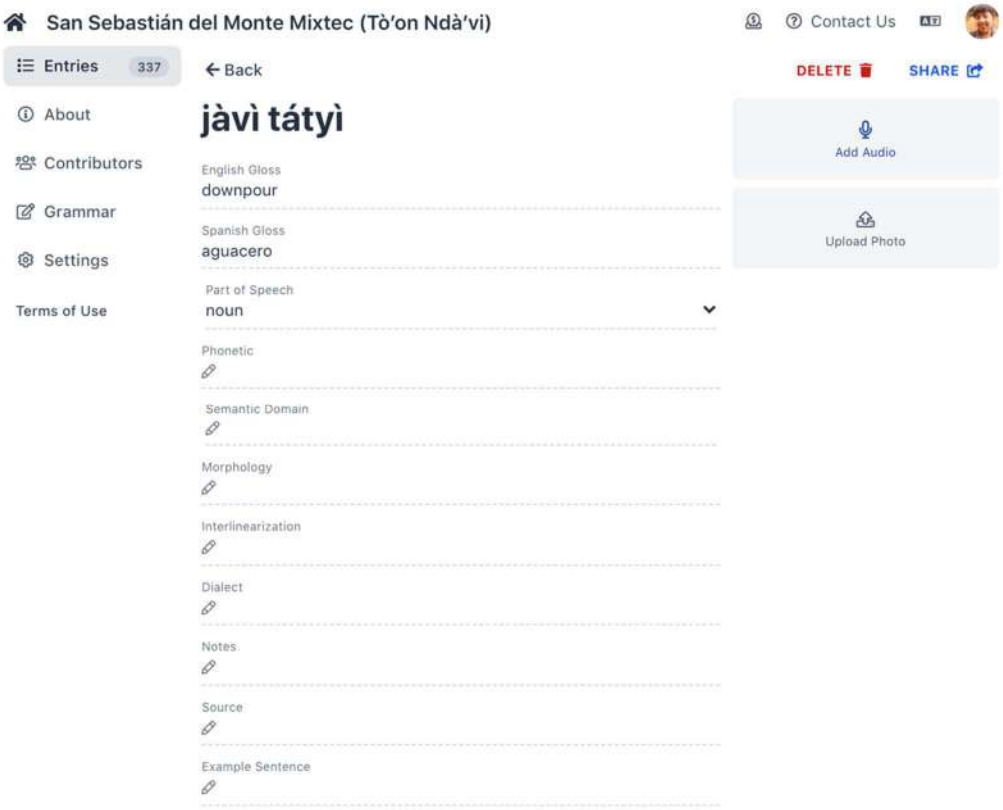
Next, the dictionary manager is prompted to add glossing languages to the project. In the above example, since the Babanki language is spoken in Cameroon, English and French glossing languages are included here. This is done by choosing from a list of over 300 useful glossing languages that are worldwide in scope. We curated the list based on the dominant regional languages that users might need for their glosses. Then, geo-coordinates are requested under the prompt “Where is this language spoken?” to display the language on the Living Dictionaries homepage map. The manager may manually enter latitude and longitude coordinates or search our digital map (using an integrated MapBox plug-in) to drop a pin in the general area, or perhaps the exact village, where the language is spoken. This geo-location step is optional, and this data may be amended later by the platform administrators. We are currently working on the ability to drop multiple geo-pins as well as create polygons to better represent regions where languages are spoken, since many users have requested such options. User feedback and suggestions help drive our design process, and we value the input from dictionary managers on the platform.

After that, the dictionary manager may fill out “alternate names” for the language by typing them in one by one and hitting *enter* to lock them in. Many languages are known by multiple names in the linguistic literature and may also have various endonyms. We designed this naming aspect to be inclusive to all the possible naming conventions of the language, so there is no limit to how many alternate names one can list here in this step. They may also be typed in any script that is Unicode-compliant. All the “alternate names” will be used to tag the dictionary, which helps improve the search engine optimisation (SEO) of each Living Dictionary on the Internet, as well as assist people in searching for dictionaries on our homepage using any of the possible alternate names. The final steps in the Living Dictionary creation process include typing in the ISO 693-3 code and the Glottocode associated with the language. This also helps SEO, in case people are searching for online linguistic resources by one of those codes. Adding these codes is optional because 1) people may not know these codes or be aware that they even exist for their languages, and 2) some under-represented languages do not yet have these codes.

Lastly, the dictionary manager must decide whether the Living Dictionary will be “visible” to the public or not, by checkmarking a box indicating that they have community consent to put representations of this language online. The default setting for new dictionaries is “not visible to the public” which we consider to be a “private” mode. We designed it this way for various reasons: we want to be sure that the language community has given their consent for the language being represented online, and we also want to give people the option of building their resources privately at their own pace before letting the rest of the world know that the Living Dictionary exists. It is important to note that a private Living Dictionary is not password-protected, but merely unlisted and not accessible to anyone who does not have the link. If made "visible" the Living Dictionary will be available for browsing on our public list of dictionaries on the platform's homepage and will also be displayed on our map (if geo-coordinates are provided in the set-up process). The "visible to the public" option may be activated at any time using the "Settings" tab on the left sidebar. Many Living Dictionary managers populate their dictionaries privately with data, recordings and images and then switch the setting to "visible" when they are ready. At any time, whether the dictionary is set to private or public, a user may copy-paste the URL of the dictionary itself and share it with their friends, colleagues and relatives. Anyone who has been given the link can then view and browse it without having to type in a password or register for an account. Viewers cannot modify the Living Dictionary unless they are registered as a collaborator or manager of the project. Language communities own their own linguistic content on the platform. It is important to us that the intellectual property rights related to linguistic and cultural content remain in the hands of the native speakers and dictionary creators who work together to build the dictionaries on the platform. In terms of adding entries and multimedia to a Living Dictionary, this can be done on the platform by adding individual text entries and recording audio directly onto the platform. If the dictionary manager already has a large amount of text data in a .CSV, .PDF or .DOC file, they may request a batch import spreadsheet template from our team by using our "Batch Import Request" form found on the platform. It is also possible to merge two existing dictionaries once the data structure and any issues pertaining to orthography and duplicate content have been assessed by stakeholders and platform administrators.

Below is an individual lexeme entry page view from the San Sebastián del Monte Mixtec (Tò'on Ndà'vi) Living Dictionary, an indigenous language of Mexico. The possible fields to fill out in the data structure[3] of the Living Dictionary are as follows: lexeme, English gloss, Spanish gloss, part of speech, phonetic transcription, semantic domain, morphology, interlinearisation, and an example sentence using the lexeme.

[3] The data structure of Living Dictionary entries can be found here: https://gist.github.com/jwrunner/b8e658e3551f204225305d482f6743b2



jàvì tátyì

English Gloss
downpour

Spanish Gloss
aguacero

Part of Speech
noun

Phonetic

Semantic Domain

Morphology

Interlinearization

Dialect

Notes

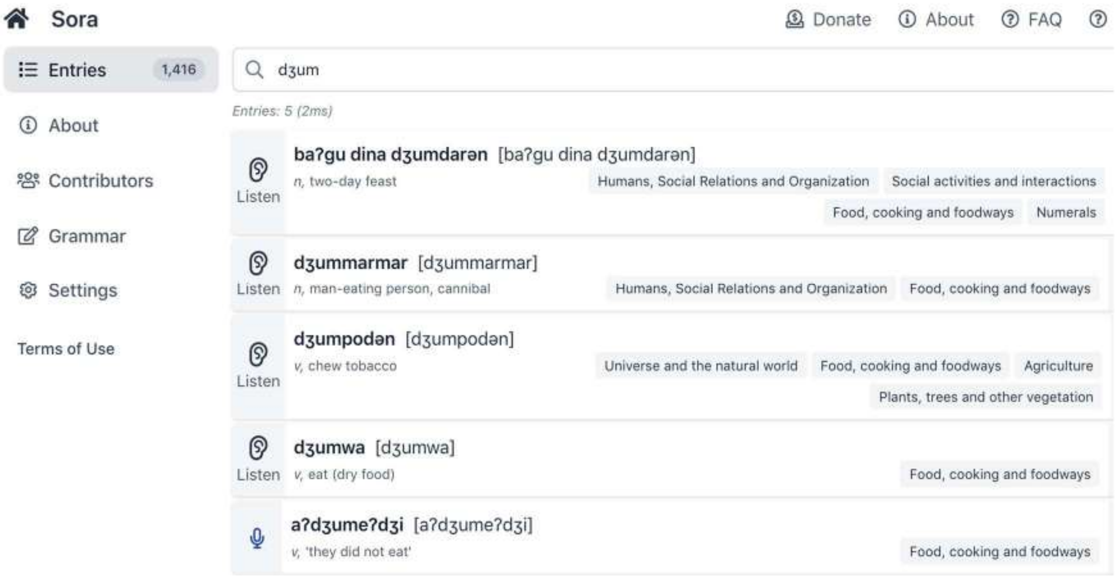
Source

Example Sentence

Figure 2: Lexeme entry page for the word for “downpour” in San Sebastián del Monte Mixtec Source: https://livingdictionaries.app/san-sebastian-del-monte-m/entries/list

The Living Dictionaries platform is a “progressive web application” (PWA) that functions as a website and behaves like a mobile app on smartphones. PWAs do not require the user to download and install any software from the Internet. A Living Dictionary instead lives and caches data on the user’s device, and it also updates automatically from the Web. When launched from the user’s home screen, service workers enable a PWA to synchronise with the server and load text data instantly, regardless of the network state, so a user can be online or not. PWAs must be served from a secure origin and therefore live on HTTPS (and not http:). They are known to be secure, reliable and fast. Once a new digital dictionary has been created online, it can later be accessed and used offline, as well as modified. Text entries may be edited offline, and changes will automatically be uploaded to the cloud when the user is online. While one must currently be online to access and edit images and audio, plans are underway to make multimedia editing accessible offline in the future.

As Lew and de Schryver (2014: 342) aptly commented, “[m]odern dictionaries in the form of apps or online services are probably better seen as collections of structured data and code, rather than hardware.” This observation certainly applies to the Living Dictionary platform, which is programmed using HTML, CSS, Javascript and ReactJS with Svelte integration, and uses Google Firebase on the backend as a cloud-hosted database. The language data, audio recordings and images are stored in the cloud. The code is currently stored on a private GitHub repository, with plans to make it open source in the coming years. The administrators have access to the backend from anywhere in the world. We partnered with the tech company Algolia to improve the platform's search engine capability on mobile and on desktop. The Algolia search integration allows users to search a Living Dictionary very efficiently, as well as use new filters that can search by categories such as part of speech, semantic domain, speaker name, or the presence of other kinds of tags. One can also use the powerful search bar (located in the centre right above the language data) to locate entries by lexeme, morpheme, part of speech, or semantic domain and other parameters. Search results are displayed alphabetically. It is important to note that users can easily search for any morphemes that are embedded inside lexemes. This is a very important search feature in polysynthetic languages such as Sora, where users may want to yield search results related to morphemes inside words, and alphabetical considerations are therefore inconvenient. As Figure 3 illustrates, searching for the morpheme 'dʒum' (eat) in the Sora Living Dictionary yields a list of results that contains the 'dʒum' inside of words and phrases, and not just at the beginning of an entry.



dʒum

Entries: 5 (2ms)

baʔgu dina dʒumdarən [baʔgu dina dʒumdarən]
n, two-day feast
Humans, Social Relations and Organization | Social activities and interactions | Food, cooking and foodways | Numerals

dʒummarmar [dʒummarmar]
n, man-eating person, cannibal
Humans, Social Relations and Organization | Food, cooking and foodways

dʒumpodən [dʒumpodən]
v, chew tobacco
Universe and the natural world | Food, cooking and foodways | Agriculture | Plants, trees and other vegetation

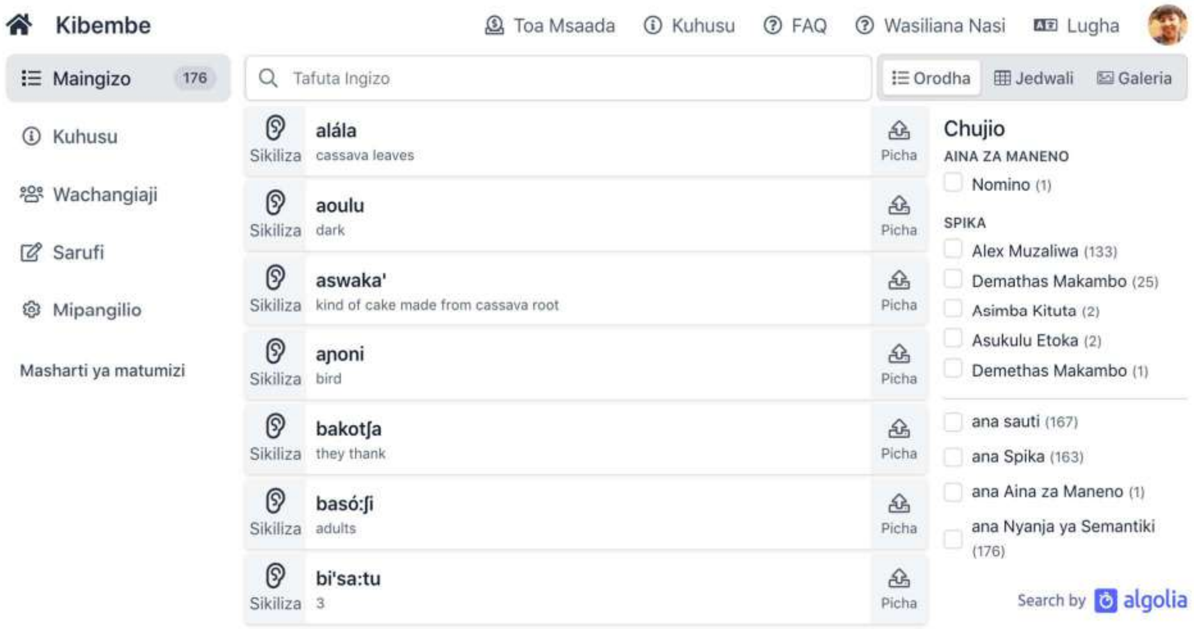
dʒumwa [dʒumwa]
v, eat (dry food)
Food, cooking and foodways

aʔdʒumeʔdʒi [aʔdʒumeʔdʒi]
v, 'they did not eat'
Food, cooking and foodways

Figure 3: Search results for Sora morpheme 'dʒum' (eat) in the Sora Living Dictionary

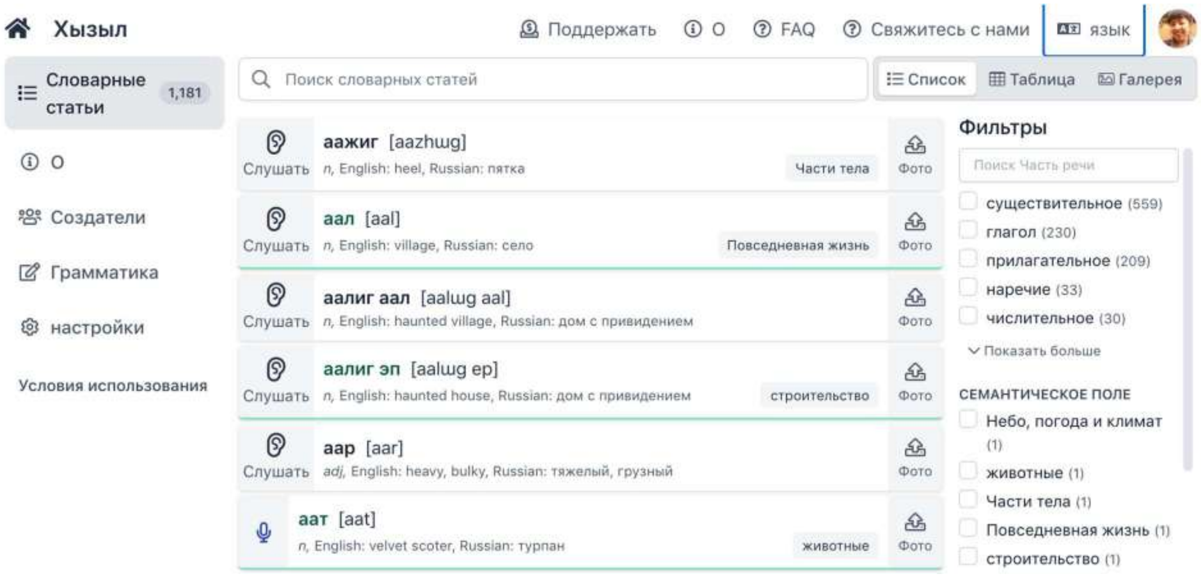
Source: https://livingdictionaries.app/sora/entries/list

The Living Dictionary website interface is currently available for use in English, Spanish, French, Portuguese, Hebrew, Russian, Bahasa Indonesia, Malay and KiSwahili, with Modern Standard Arabic, Tagalog, Zulu, Shona, Amharic, Hausa, Hindi, Assamese, Oḍia and Bengali interfaces coming online in 2021. A dictionary user can click on the top-right "Language" button to toggle between interface languages to display the website in the available languages (see Figures 4, 5 and 6). All functionality and features, including extensive dropdown menus for semantic domains and parts of speech are represented in the various interface languages. The website remembers the user's choice of language interface preference and automatically displays the website in this language upon the user's return. At any point in navigating the web platform, the user may toggle between interface languages without having to leave the website at all. The platform also allows for nearly three hundred built-in glossing languages, covering most languages that function as a local, national or regional language of wider communication. The Living Dictionaries not only elevate threatened languages but allow for them to be explored in multilingual online environments, tailored for the usage needs of specific communities.



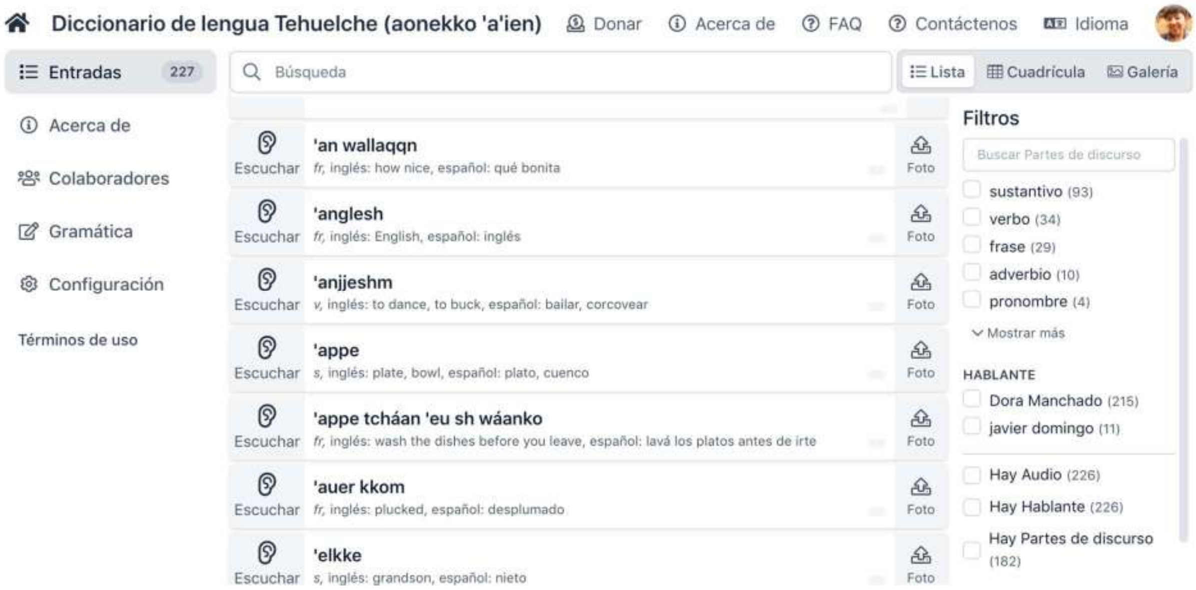
alála	cassava leaves
aoulu	dark
aswaka'	kind of cake made from cassava root
aɲoni	bird
bakotʃa	they thank
basó:ʃi	adults
bi'sa:tu	3

Figure 4: The Kibembe Living Dictionary displayed in the KiSwahili interface. Source: https://livingdictionaries.app/kibembe/entries/list



аажиг [aazhɯg]	*n*, English: heel, Russian: пятка	Части тела
аал [aal]	*n*, English: village, Russian: село	Повседневная жизнь
аалиг аал [aalɯg aal]	*n*, English: haunted village, Russian: дом с привидением	
аалиг эп [aalɯg ep]	*n*, English: haunted house, Russian: дом с привидением	строительство
аар [aar]	*adj*, English: heavy, bulky, Russian: тяжелый, грузный	
аат [aat]	*n*, English: velvet scoter, Russian: турпан	животные

Figure 5: The Xyzyl Living Dictionary, displayed in the Russian (Cyrillic) interface. Source: https://livingdictionaries.app/xyzyl/entries/list



Escuchar 'an wallaqqn fr, inglés: how nice, español: qué bonita Foto

Escuchar 'anglesh fr, inglés: English, español: inglés Foto

Escuchar 'anjjeshm v, inglés: to dance, to buck, español: bailar, corcovear Foto

Escuchar 'appe s, inglés: plate, bowl, español: plato, cuenco Foto

Escuchar 'appe tcháan 'eu sh wáanko fr, inglés: wash the dishes before you leave, español: lavá los platos antes de irte Foto

Escuchar 'auer kkom fr, inglés: plucked, español: desplumado Foto

Escuchar 'elkke s, inglés: grandson, español: nieto Foto

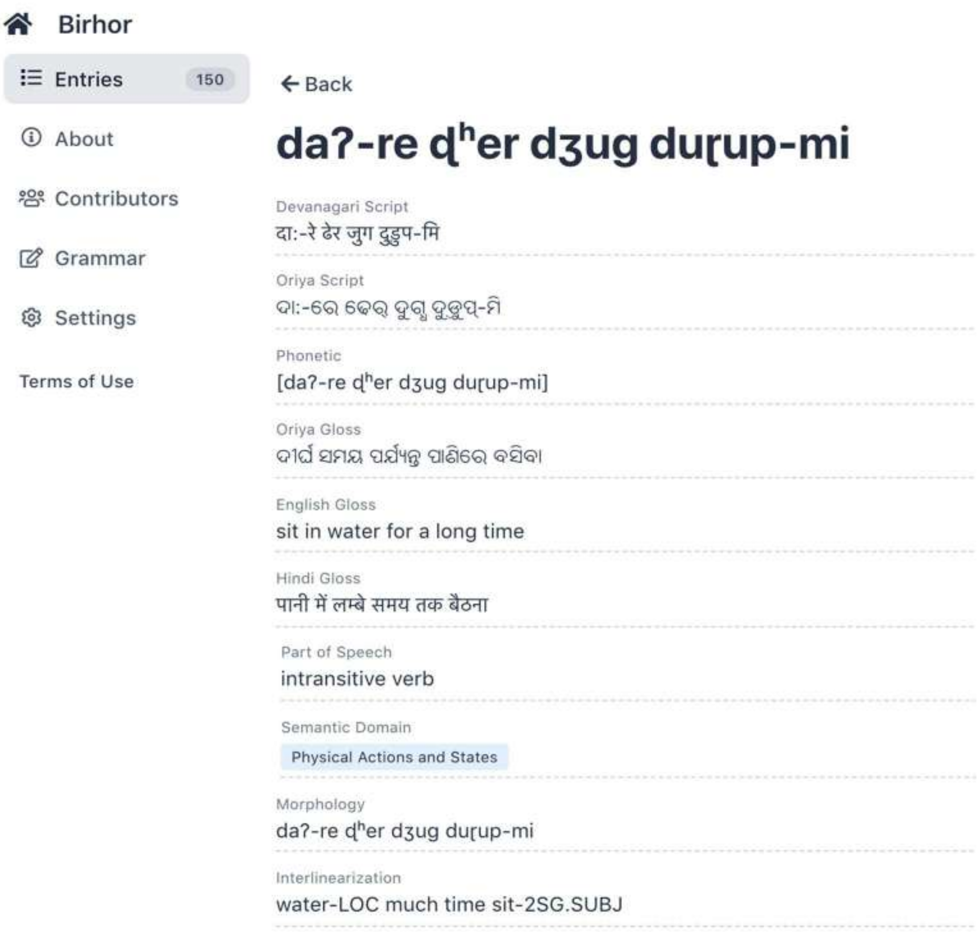
☐ sustantivo (93)
☐ verbo (34)
☐ frase (29)
☐ adverbio (10)
☐ pronombre (4)

☐ Dora Manchado (215)
☐ javier domingo (11)

☐ Hay Audio (226)
☐ Hay Hablante (226)
☐ Hay Partes de discurso (182)

Figure 6: The Tehuelche (aonekko'aien) Living Dictionary displayed in the Spanish interface. Source: https://livingdictionaries.app/80CcDQ4DRyiYSPIWZ9Hy/entries/list

Each dictionary can also include up to five glossing languages so that users may search for terms across regionally dominant and other relevant languages. For example, for Living Dictionaries for the tribal languages of the Munda family of India, glossing languages include English, Hindi and Oḍia (and sometimes other languages like Assamese or Bengali) so that users may search for terms in various languages. A Living Dictionary can also display up to five writing systems for an entry, which is useful for dictionaries where multiple competing scripts are used to represent a language. An example of one such project is the Birhor Living Dictionary (see Figure 7 for a sample entry), which is a multilingual resource that contains multiple glossing languages (English, Oḍia and Hindi) and multiple scripts (Devanagari and Oḍia). Another project, the Sora Living Dictionary, also includes an array of scripts and glossing languages (see Figure 8). In short, Living Dictionaries are designed explicitly with maximal inclusivity and unrestricted multilingualism in mind.



daʔ-re ɖʰer dʒug duɽup-mi

Devanagari Script
दा:-रे ढेर जुग दुड़ुप-मि

Oriya Script
ଦା:-ରେ ଢେର୍ ଦୁଗ୍ ଦୁଡୁପ୍-ମି

Phonetic
[daʔ-re ɖʰer dʒug duɽup-mi]

Oriya Gloss
ଦୀର୍ଘ ସମୟ ପର୍ଯ୍ୟନ୍ତ ପାଣିରେ ବସିବା

English Gloss
sit in water for a long time

Hindi Gloss
पानी में लम्बे समय तक बैठना

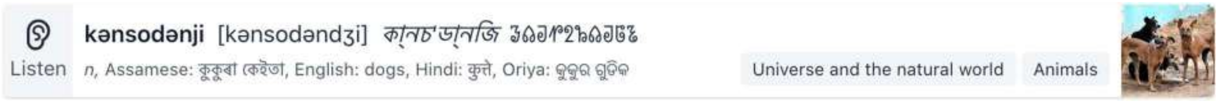
Part of Speech
intransitive verb

Semantic Domain
Physical Actions and States

Morphology
daʔ-re ɖʰer dʒug duɽup-mi

Interlinearization
water-LOC much time sit-2SG.SUBJ

Figure 7: *Entry view* for the phrase ‘sit in water for a long time’ (Birhor Living Dictionary).
Source: https://livingdictionaries.app/birhor/entries/2SIhhZQdAr8ZfLaXI8f9



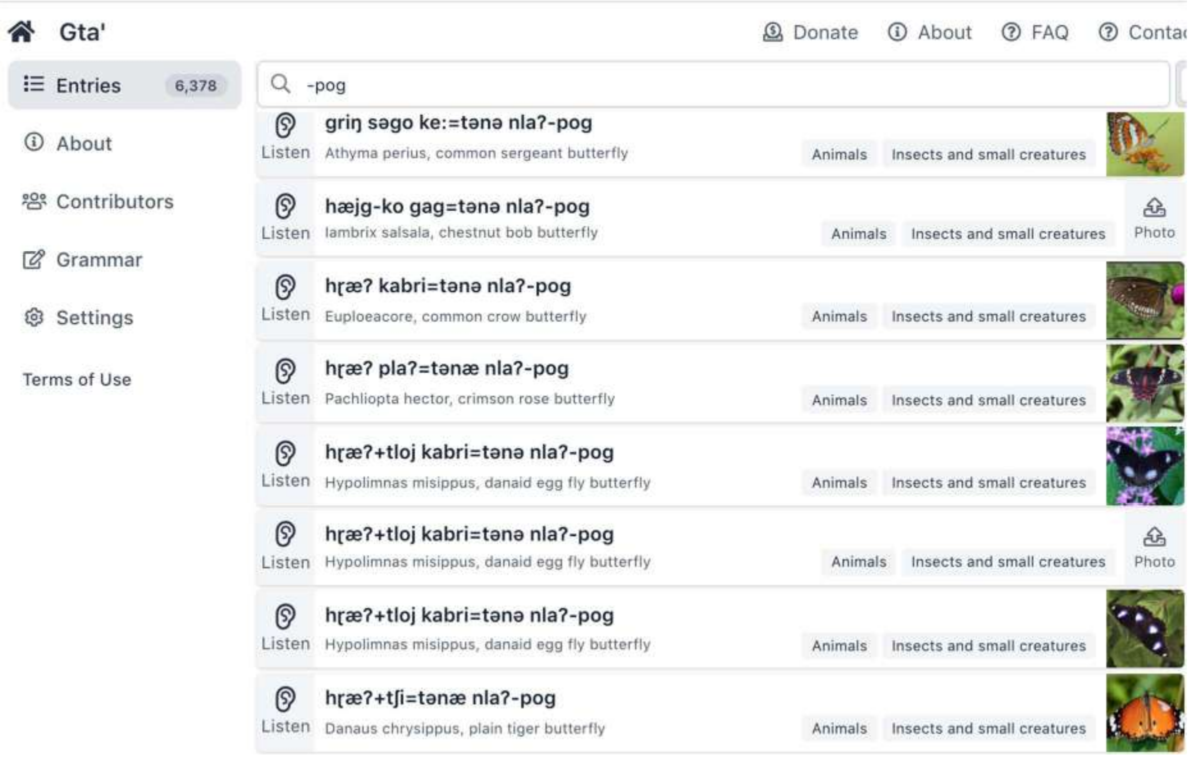
kənsodənji [kənsodəndʒi] কা্নচ'ডা্নজি 𑃟𑃧𑃙𑃐𑃦𑃔𑃧𑃙𑃠𑃤
n, Assamese: কুকুৰা কেইতা, English: dogs, Hindi: कुत्ते, Oriya: କୁକୁର ଗୁଡିକ

Universe and the natural world | Animals

Figure 8: The Sora word “dogs” in the Sora Living Dictionary (displayed in *list view*).
Source: https://livingdictionaries.app/sora/entries/wPKHVIbyQgJhEVI1mCcI

Figures 7 and 8 also show the types of information that can be provided for each entry in a Living Dictionary: headword, phonetic transcription, representation in different scripts, glosses into different languages, part of speech, semantic domain, morphology, interlinearisation, dialect name, audio recording and image file. All are optional metadata depending on the needs of the user, except for the headword. Not displayed in these entries are other optional fields, such as a sample sentence that contains the entry headword alongside a gloss of the sample sentence.

Living Dictionaries may be adjusted depending on what data the user wants to see. They may be viewed through three different types of visualisation: *list view, table view* and *gallery view* (settings that are available near the top right-hand corner of the "Entries" page). Each different setting provides the user with different ways of visualising and navigating the data inside the dictionary. *List view* (Figure 9) displays the data in a traditional dictionary list, *table view* (Figure 10) shows a spreadsheet of data, and *gallery view* (Figure 11) only pulls in entries with accompanying images.



griŋ səgo ke:=tənə nlaʔ-pog
Athyma perius, common sergeant butterfly
Animals | Insects and small creatures

hæjg-ko gag=tənə nlaʔ-pog
Iambrix salsala, chestnut bob butterfly
Animals | Insects and small creatures

hɽæʔ kabri=tənə nlaʔ-pog
Euploeacore, common crow butterfly
Animals | Insects and small creatures

hɽæʔ plaʔ=tənæ nlaʔ-pog
Pachliopta hector, crimson rose butterfly
Animals | Insects and small creatures

hɽæʔ+tloj kabri=tənə nlaʔ-pog
Hypolimnas misippus, danaid egg fly butterfly
Animals | Insects and small creatures

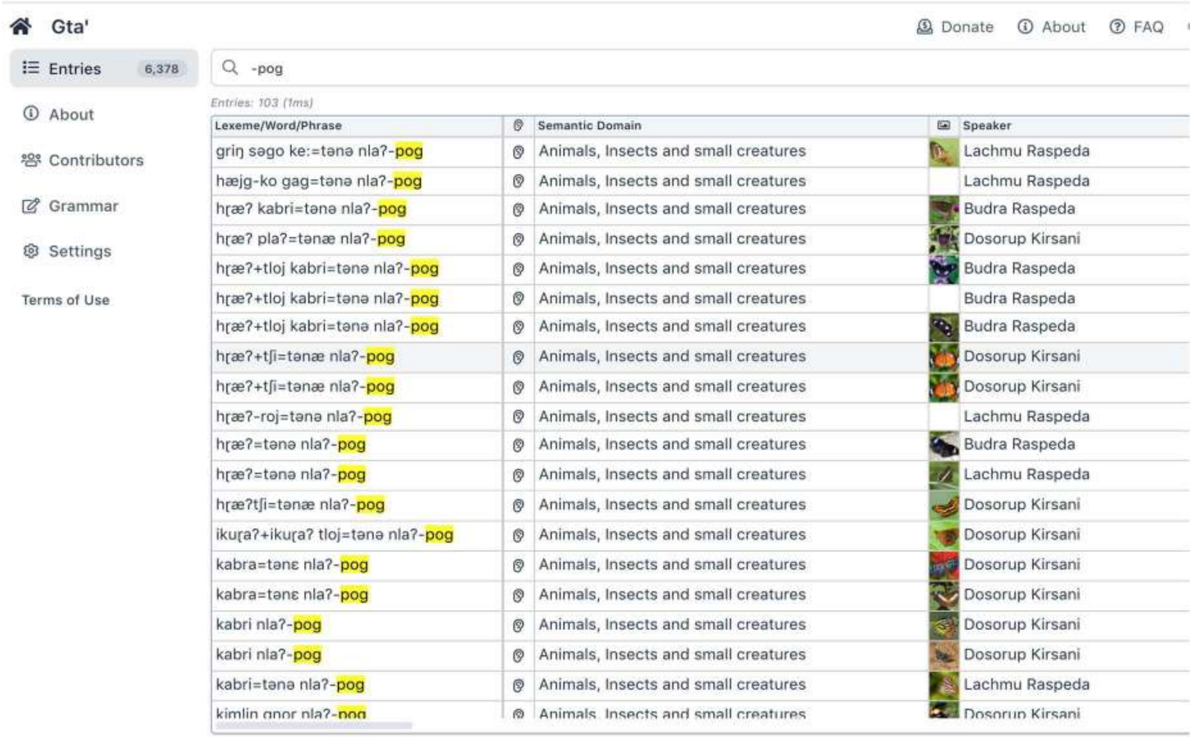
hɽæʔ+tloj kabri=tənə nlaʔ-pog
Hypolimnas misippus, danaid egg fly butterfly
Animals | Insects and small creatures

hɽæʔ+tloj kabri=tənə nlaʔ-pog
Hypolimnas misippus, danaid egg fly butterfly
Animals | Insects and small creatures

hɽæʔ+tʃi=tənæ nlaʔ-pog
Danaus chrysippus, plain tiger butterfly
Animals | Insects and small creatures

Figure 9: *List view* display of Gtaʔ morpheme -pog 'bug' in the Gtaʔ Living Dictionary

Source: https://livingdictionaries.app/gta/entries/list

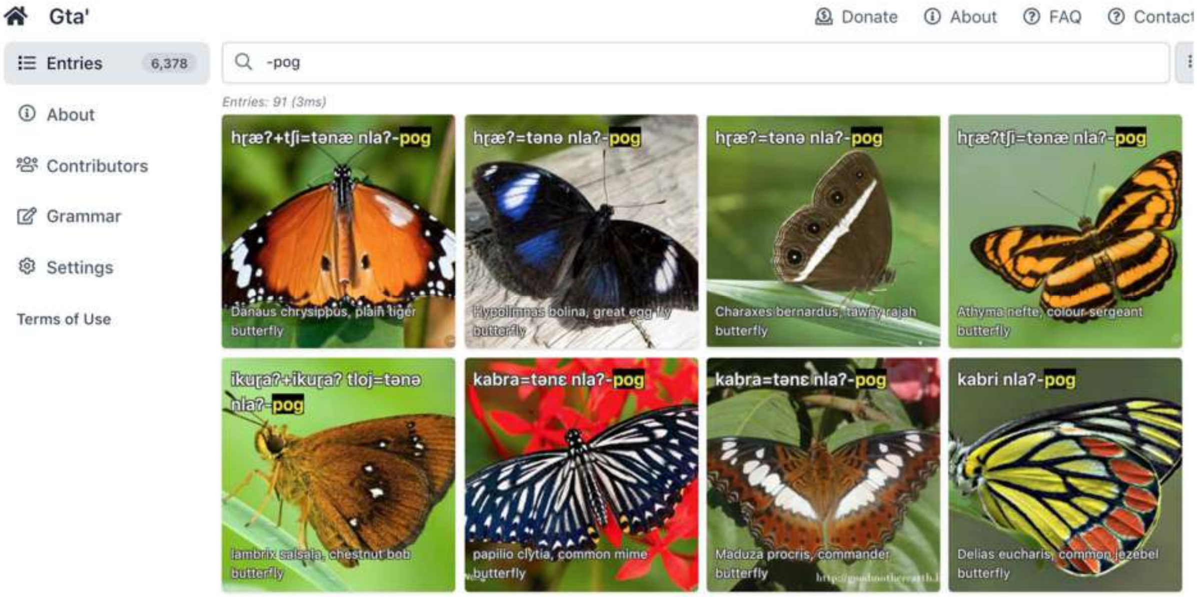


Entries: 103 (1ms)

Lexeme/Word/Phrase	Semantic Domain	Speaker
griŋ səgo ke:=tənə nlaʔ-pog	Animals, Insects and small creatures	Lachmu Raspeda
hæjg-ko gag=tənə nlaʔ-pog	Animals, Insects and small creatures	Lachmu Raspeda
hɽæʔ kabri=tənə nlaʔ-pog	Animals, Insects and small creatures	Budra Raspeda
hɽæʔ plaʔ=tənæ nlaʔ-pog	Animals, Insects and small creatures	Dosorup Kirsani
hɽæʔ+tloj kabri=tənə nlaʔ-pog	Animals, Insects and small creatures	Budra Raspeda
hɽæʔ+tloj kabri=tənə nlaʔ-pog	Animals, Insects and small creatures	Budra Raspeda
hɽæʔ+tloj kabri=tənə nlaʔ-pog	Animals, Insects and small creatures	Budra Raspeda
hɽæʔ+tʃi=tənæ nlaʔ-pog	Animals, Insects and small creatures	Dosorup Kirsani
hɽæʔ+tʃi=tənæ nlaʔ-pog	Animals, Insects and small creatures	Dosorup Kirsani
hɽæʔ-roj=tənə nlaʔ-pog	Animals, Insects and small creatures	Lachmu Raspeda
hɽæʔ=tənə nlaʔ-pog	Animals, Insects and small creatures	Budra Raspeda
hɽæʔ=tənə nlaʔ-pog	Animals, Insects and small creatures	Lachmu Raspeda
hɽæʔtʃi=tənæ nlaʔ-pog	Animals, Insects and small creatures	Dosorup Kirsani
ikuɽaʔ+ikuɽaʔ tloj=tənə nlaʔ-pog	Animals, Insects and small creatures	Dosorup Kirsani
kabra=tənɛ nlaʔ-pog	Animals, Insects and small creatures	Dosorup Kirsani
kabra=tənɛ nlaʔ-pog	Animals, Insects and small creatures	Dosorup Kirsani
kabri nlaʔ-pog	Animals, Insects and small creatures	Dosorup Kirsani
kabri nlaʔ-pog	Animals, Insects and small creatures	Dosorup Kirsani
kabri=tənə nlaʔ-pog	Animals, Insects and small creatures	Lachmu Raspeda
kimlin gnor nlaʔ-pog	Animals, Insects and small creatures	Dosorup Kirsani

Figure 10: *Table View* display of Gtaʔ morpheme -pog ‘bug’ in the Gtaʔ Living Dictionary

Source: https://livingdictionaries.app/gta/entries/list



Entries: 91 (3ms)

hɽæʔ+tʃi=tənæ nlaʔ-pog
Danaus chrysippus, plain tiger butterfly

hɽæʔ=tənə nlaʔ-pog
Hypolimnas bolina, great egg fly butterfly

hɽæʔ=tənə nlaʔ-pog
Charaxes bernardus, tawny rajah butterfly

hɽæʔtʃi=tənæ nlaʔ-pog
Athyma nefte, colour sergeant butterfly

ikuɽaʔ+ikuɽaʔ tloj=tənə nlaʔ-pog
Iambrix salsala, chestnut bob butterfly

kabra=tənɛ nlaʔ-pog
papilio clytia, common mime butterfly

kabra=tənɛ nlaʔ-pog
Maduza procris, commander butterfly

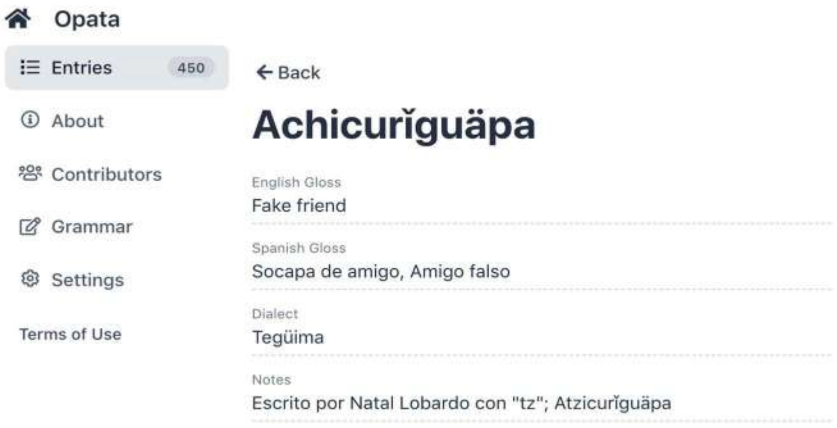
kabri nlaʔ-pog
Delias eucharis, common jezebel butterfly

Figure 11: *Gallery view* display of Gtaʔ morpheme -pog ‘bug’ in the Gtaʔ Living Dictionary

Source: https://livingdictionaries.app/gta/entries/list

Search and use of entries in a platform such as Living Dictionaries are freed of the linear constraints of traditional dictionaries. As Lew and de Schryver (2014: 350) put it “[t]he user of a digital dictionary is no longer constrained by either the formal (spelling or phonology) or semantic criteria as the organizing principle. It is now perfectly possible to combine formal and semantic relations and utilise both types in navigating the lexical material.” One key feature we have included in this are the tagging of entries according to semantic domains. The use of semantic domains as an organisational search principle is grounded in insights of cognitive linguistics (Langacker, 1987; Clausner & Croft, 1999; see also Bowers & Romary, 2018: 97) and allows for the generation of specific subsets of lexical entries to facilitate instruction in formal or informal educational settings in language revitalisation programs. Semantic domains are a sensitive issue because they often overlap and may be difficult to delineate. Our system allows for flexibility, and thus there is no limit to the number of semantic domains that can be used to tag entries. Users can also search by one or various semantic domain “filters” to yield tailored sets of results related to their domains of inquiry.

Ideologies of what is a ‘proper’ linguistic variety to be used are not relevant to the Living Dictionaries. Decisions guiding what dialects are represented (or not) within a Living Dictionary are community-driven. A digital dictionary may be created for any variety, whether it is oral or signed, recognised as a separate distinct language or ‘just’ a dialect, patois, Creole, pidgin, or any other lectal designation. Living Dictionaries can accommodate as many dialects or variants as desired by the community members creating the tool. For example, Zapotec and Mixtexc communities in Mexico may wish to have a separate dictionary for each dialect, and therefore each dictionary will contain data from a specific dialect rather than showcasing multiple dialects. In the example below, the Mexican/American research team that created the first-ever Living Dictionary for the inactive indigenous language Opata (the name given to two closely related Uto-Aztecan tongues, Tegüima and Eudeve) decided it was best to group resources for both Opata varieties into one dictionary. They accomplished this by tagging the entries with the dialect names Tegüima and Eudeve (Figure 12).



Achicurǐguäpa

English Gloss
Fake friend

Spanish Gloss
Socapa de amigo, Amigo falso

Dialect
Tegüima

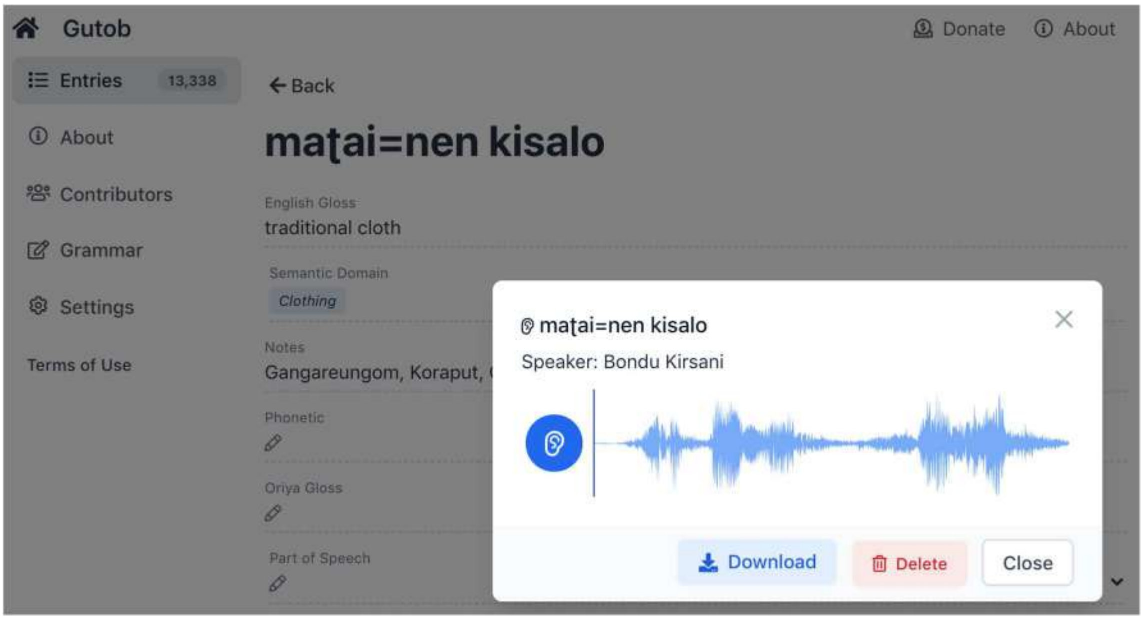
Notes
Escrito por Natal Lobardo con "tz"; Atzicurǐguäpa

Figure 12: An entry from the Opata Living Dictionary tagged as the “Tegüima” dialect.
Source: https://livingdictionaries.app/opata/entries/yK1Yi17Fivn37BWDMima

6. Usage and Remote Collaboration

In terms of usage, there are currently close to 300 activists working on over 200 different Living Dictionaries on the platform, and more joining every week. In terms of dictionary size, recently created Living Dictionaries contain anywhere from a handful to several hundred entries, while many other Living Dictionaries that have been developed over the course of many years contain over 10,000 entries. Altogether, the platform contains over to 250,000 entries and is growing each week.

One of the strengths of the Living Dictionaries is that they allow people to hear pronunciations of the words and phrases (Figure 13). We strongly encourage dictionary managers to upload audio files, or record audio content directly into the platform when possible, by using the microphone on their desktop or mobile device. If a dictionary manager does not speak the language fluently, we encourage them to locate a fluent speaker who can record audio entries later. Each dictionary, and each entry within a dictionary, is shareable with a unique URL that can be easily shared on social media or hyperlinked on other websites.



maţai=nen kisalo

English Gloss
traditional cloth

Semantic Domain
Clothing

Notes
Gangareungom, Koraput, (

Phonetic

Oriya Gloss

Part of Speech

maţai=nen kisalo

Speaker: Bondu Kirsani

Figure 13: The audio waveform entry for "maţai=nen kisalo" in the Gutob Living Dictionary

Source: https://livingdictionaries.app/gutob/entries/KTJzdxbcYxtZsjRI2fTt

Remote collaboration is possible and encouraged on the platform. Many existing Living Dictionaries have collaborators who work on different aspects of the work: some work on the text entries while others undertake the recording of the words and phrases based on the written data has been added to the system. There is no limit to the number of collaborators in a Living Dictionary. A dictionary manager may invite other collaborators to join the dictionary directly through the platform itself by using the "Invite Manager" or "Invite Collaborator" feature. Dictionary managers may add, edit or delete content. Contributors are project collaborators who can also add and edit but cannot delete any content. The latter feature is designed for students and interns who may be working on the project as digital assistants, and they need to be able to safely work on content without deleting any of it by accident. The Living Dictionaries platform is engineered to have multiple collaborators logged into the system and editing a dictionary project at the same time, in real-time. The collaborators can be working remotely in different places in the world and see the exact same changes that are being made without even having to refresh their browsers, within seconds. There is no limit to the number of people who can be logged into a project at once, but we suggest that a team coordinates its strategy so that multiple people are not trying to edit the exact same entries at the same time.

7. The Future of Living Dictionaries

The platform is built to make ongoing relevant contributions to an increasingly dynamic world. As such, we continue to innovate and roll out new features on a regular basis. In 2021, we are releasing an updated International Phonetic Alphabet (IPA) Chart Picker on the platform, so that users may easily locate and select phonetic characters when they are creating (or editing) entries. It will be a great help for activists who need to be able to type effectively in IPA without leaving the platform. This year, we will also be launching our video integration feature, in which dictionary managers can directly record videos within entries, or link to existing YouTube videos, without ever leaving the platform. We are also working on displaying links to ecological databases within entries about species, which will help create a global network linking linguistic knowledge to other relevant databases. In 2020, we collaborated with the Ethno-Ornithology World Atlas to discuss and enact ways in which traditional ecological knowledge about birds can better interface with existing scientific online resources. Our intention is to keep these kinds of interdisciplinary discussions flowing so that our platform may become increasingly encyclopaedic over time. We also regularly meet with indigenous leaders, experts and scholars to discuss new opportunities for collaboration and avenues for language revitalisation that include Living Dictionaries.

Our long-term development roadmap includes expanding and improving features on the platform like speed optimisation, offline mode functionality, audio analysis and rolling out important new features such as export functionality (so that dictionary managers can retrieve their data in CSV, XML, JSON, PDF and other formats) and further multimedia integration. Based on user feedback, we intend to explore ways to integrate lists of culturally specific prompts by allowing users to draw from existing elicitation lists to start their dictionary projects from scratch. Users have also requested the implementation of an image API (Application Programming Interface) that would allow them to use relevant copyright-free images from sources such as Creative Commons directly in the platform. We intend to expand storage capacity exponentially over time and implement language localisation of the dictionary interface into two dozen additional dominant languages to serve the widest audience of endangered language activists possible. We plan to implement notifications to increase real-time contributions and collaboration between users and begin regional campaigns to attract hundreds of new users and contributors worldwide. This will be done by demonstrating the software at regional and international gatherings of linguists and language activists to maximise the potential user groups as well as rolling out comprehensive training videos and webinars in various languages to assist contributors on the Living Dictionary platform.

In summary, the Living Tongues Institute has developed practical, web-based software (found at the URL LivingDictionaries.app) that can help people build a dictionary from the ground up. Moving forward, our team will continue to build and refine this framework for global application and deploy the platform at scale to serve all the world's endangered languages. This project can help mitigate the global language extinction crisis by opening the door to linguistic documentation for all, expanding access to cultural equity and self-determination. As an online platform that presently houses dictionaries for over 200 languages, it utilises the safety and flexibility of remote collaboration between dictionary managers. We are committed to maintaining this platform for decades to come so that the work of language activists may live on and benefit our descendants, community stakeholders, educators and scholars.

Acknowledgments

We gratefully acknowledge the support of the following grants that helped support the creation of Living Dictionaries for Sora, Remo, Birhor, Gtaʔ, Gutob and other Munda languages spoken in India: "Sora Typological Characteristics: Towards a Re-Evaluation of South Asian Human History" (National Science Foundation Linguistics Grant Award NSF/BCS #1844532, 2019-2022), "Documenting the Fragile Knowledge Domains of the Birhor People" (The Zegar Family Foundation, 2018-2020), "Documentation of Gutob, an endangered Munda language" (National Science Foundation, Documentation of Endangered Languages Grant Award #1500092, 2015-2017), "Documentation of Hill Gta'–a seriously endangered Munda language" (National Endowment for the Humanities, Documentation of Endangered Languages Grant Award #PD-50025-13, 2013-2015) and "Documentation of Remo" (National Science Foundation, Documentation of Endangered Languages Grant Award #0853877, 2010-2011). We acknowledge web developer Jacob Bowdoin for his extensive contributions in programming the Living Dictionaries platform, and we thank our Living Tongues colleagues Dr. Luke Horo, Dr. Bikram Jora and Opino Gomango for their many contributions to the Living Dictionaries for Munda languages mentioned above. We acknowledge the work of Dr. K. David Harrison and Jeremy Fahringer who collaborated with us previously on a different desktop version of this platform, Talking Dictionaries (spearheaded by Living Tongues Institute for Endangered Languages and housed at Swarthmore College). We recognise the efforts of many volunteers and interns who have contributed to making the Living Dictionary platform multilingual and who have edited and updated data on the platform. For the interface languages featured in this article, we thank the linguists and translators Dr. Michael Karani (KiSwahili interface), Dr. Denis Tokmashev (Russian interface), Amanda Chao Benbassat and Mónica Bonilla Parra (Spanish interface), Crisofia Langa da Camara (Portuguese interface) Yustinus (Yanche) Ghanggo Ate (Bahasa Indonesia interface), Nur Hidayah Binte Sunaryo (Bahasa Melayu interface) and Dana Melaver and Daniel Bögre-Udell (Hebrew interface) for their assistance. Lastly and very importantly, we acknowledge the tireless and ongoing efforts of the hundreds of language activists, speakers and scholars who are building their Living Dictionaries on our platform.

References

Langacker, R. W. (1987). *Foundations of Cognitive Grammar: Theoretical Prerequisites*. Vol. 1. Stanford, CA: Stanford University Press.

Anderson, G. D. S. (2017). Consequences of Russian Linguistic Hegemony in (Post-)Soviet Colonial Space. In Ramazan Korkmaz and Gürkan Dogan (eds.) *Endangered Languages of the Caucasus and Beyond*, pp. 1-16. Leiden: Brill.

Prinsloo, D. J., Prinsloo, J. V. & Prinsloo, D. (2017). African Lexicography in the Era of the Internet. *The Routledge Handbook of Lexicography.* Pedro A. Fuertes Olivera (Ed.). London: Routledge, pp. 487-502.

Dimitrova, L., Koseska(-Toszewa), V., Dutsova, R. & Panova, R. (2009). Bulgarian-Polish Online Dictionary — Design and Development. In V. Koseska-Toszewa, L. Dimitrova & R. Roszko (eds.) *Representing Semantics in Digital Lexicography: Innovative Solutions for Lexical Entry Content in Slavic Lexicography.* Warsaw, Institute of Slavic Studies, Polish Academy of Sciences, pp. 78-88.

Prinsloo, D J. (2019). A perspective on the past, present and future of lexicography with specific reference to Africa. In M. Gürlek, A. N. Çiçekler and Y. Taşdemir (eds.) *Asialex 2019.* Istanbul. 217-230.

Rundell, M. (2012). It works in practice but will it work in theory? The uneasy relationship between lexicography and matters theoretical. In R. Vatvedt Fjeld & J. M. Torjusen (eds.). *Proceedings of the 15th Euralex International Congress.* 7-11 August 2012. Oslo.

Clausner, T. C. & Croft, W. (1999). Domains and image schemas. *Cognitive Linguistics* 10, pp. 1–32.

Lew, R. & de Schryver, G.-M. (2014). Dictionary users in the digital revolution. *International Journal of Lexicography* 27(4), pp. 341-59.

Grin, F. (2003). The economics of language planning. *Current Issues in Language Planning* 4, pp. 1-66.

Van Wyk, E.B. (1995). Linguistic Assumptions and Lexicographical Traditions in the African Languages. *Lexikos* 5, pp. 82-96.

Whalen, D. H., Moss M. & Baldwin D. (2016). Healing through language: Positive physical health effects of indigenous language use. *F1000Research* 2016, **5**, pp. 852.

Websites:

ilc.cnr.it/ccurl2014 Accessed at: http://www.ilc.cnr.it/ccurl2014/ (8 April 2021)

livingdictionaries.app. Accessed at: https://livingdictionaries.app/ (8 April 2021)

slll.cass.anu.edu.au. Accessed at: http://slll.cass.anu.edu.au/centres/andc/australex-2019 (8 April 2021)

Dictionaries:

Akumbu, Pius. (2021). *Babanki Living Dictionary.* Living Tongues Institute for Endangered Languages. https://livingdictionaries.app/babanki

Anderson, G.D.S., Jora, B. & Gomango, O. with Raspeda, B. Raspeda, P., Raspeda, L., Raspeda, K., Raspeda, K. Kirsani, D., Majhi, B., Majhi, S., Golpeda, B., & Mondal, I. (2021). *Gta' Living Dictionary.* Living Tongues Institute for Endangered Languages. https://livingdictionaries.app/gta

Anderson, G.D.S., Jora, B. & Gomango, O. with Sisa, T and Sisa, K., Kirsani, B., Kirsani, R. & Mondal, I. (2021). *Gutob Living Dictionary.* Living Tongues Institute for Endangered Languages. https://livingdictionaries.app/gutob

Anderson, G.D.S. & Harrison, K.D. with Muzaliwa, A., Makambo, D., Kituta A. & Etoka, A. (2021). *KiBembe Living Dictionary.* Living Tongues Institute for Endangered Languages. https://livingdictionaries.app/kibembe

Anderson, G.D.S., Gomango, O., Harrison, K.D. & Horo, L. with Gomango, S., Gomango, M., Gomango, O., Roita, Z., Roita, M., Sabar, T., Mondal, I. (2021). *Sora Living Dictionary.* Living Tongues Institute for Endangered Languages. https://livingdictionaries.app/sora

Anderson, G.D.S., Tokmashev, D., Fahringer, J., Harrison, K.D. & Tabatkin, M.M. (2021). *Xyzyl Living Dictionary.* Living Tongues Institute for Endangered Languages. https://livingdictionaries.app/xyzyl

Cortés Torres, F.D, Mantenuto, I., Espinoza, A.D, Pimentel Rojas, I., Anderson, G.D.S & Daigneault, A.L. (2021). *San Sebastián del Monte Mixtec (Tò'on Ndà'vi) Living Dictionary.* Living Tongues Institute for Endangered Languages. https://livingdictionaries.app/san-sebastian-del-monte-m

Jora, B., Anderson, G.D.S., Lakra, S. K., & Markki, A. with Birhor, A., Birhor, B., Birhor, B. Birhor, K., Birhor, K. Birhor, M., Birhor, M. Birhor, S. (2021). *Birhor Living Dictionary.* Living Tongues Institute for Endangered Languages. https://livingdictionaries.app/birhor

Fundación TEGUIMA-OPATA & Daigneault, A.L. (2021). *Opata Living Dictionary.* Living Tongues Institute for Endangered Languages. https://livingdictionaries.app/opata

Manchado, D., Domingo, J., Duval, N., Arias García, N. Berger, T., Daigneault, A.L. (2021). *Diccionario de lengua Tehuelche (aonekko 'a'ien).* Living Tongues Institute for Endangered Languages. https://livingdictionaries.app/80CcDQ4DRyiYSPIWZ9Hy

Visionary perspectives on the lexicographic treatment of easily confusable words: *Paronyme – Dynamisch im Kontrast* as the basis for bi- and multilingual reference guides

Petra Storjohann[1]

[1] Leibniz-Institut für Deutsche Sprache, R5-13, 68161 Mannheim, Germany
E-mail: storjohann@ids-mannheim.de

Abstract

The German e-dictionary documenting confusables *Paronyme – Dynamisch im Kontrast* contains lexemes which are similar in sound, spelling and/or meaning, e.g. *autoritär/autoritativ*, *innovativ/innovatorisch*. These can cause uncertainty as to their appropriate use. The monolingual guide could be easily expanded to become a multilingual platform for commonly confused items by incorporating language modules. The value of this visionary resource is manifold. Firstly, e-dictionaries of confusables have not yet been compiled for most European languages; consequently, the German resource could serve as a model of practice. Secondly, it would be able to explain the usage of false friends. Thirdly, cognates and loan word equivalents would be offered for simultaneous consultation. Fourthly, users could find out whether, for example, a German pair is semantically equivalent to a pair in another language. Finally, it would inform users about cases where a pair of semantically similar words in one language has only one lexical counterpart in another language. This paper is an appeal for visionary projects and collaborative enterprises. I will outline the dictionary's layout and contents as shown by its contrastive entries. I will demonstrate potential additions, which would make it possible to build up a large platform for easily misused words in different languages.

Keywords: contrastive lexicography; bilingual paronyms; easily confused words; false friends; multilingual platform

1. Introduction

Electronic lexicographic resources are often shaped by modularity and their potential to be extended. To some extent, allowance is already made at the draft stage for linear and/or vertical expansion, which is then realised at different times and in various stages of development. In elexiko, for example, the online dictionary of modern German (www.elexiko.de), and also in the paronym dictionary *Paronyme – Dynamisch im*

Kontrast (freely accessible in OWID or in OWID$_{plus}$[1]), which documents easily confused expressions in contemporary German usage, work packages were defined from the very beginning which anticipated the successive addition of new dictionary rubrics or were intended to augment existing content with supplementary linguistic or subject-related information, in addition to the continuous development of word entries. If we disregard e-dictionaries that are supplemented with new information after they have been retrospectively digitised, then little detailed attention has been paid to how more recent electronic language resources might be productively extended after completion. At the end of 2021, *Paronyme – Dynamisch im Kontrast* will be complete, with around 360 contrastive entries (about 800 individual lemmas) comprising expressions that can create linguistic uncertainty as a result of their formal and/or semantic similarity to one another. The new dictionary will fill a gap in the lexicographic landscape. For the first time, we will have at our disposal a rigorously corpus-based reference work on the phenomenon of paronymy, which will provide help in situations of linguistic uncertainty on multiple descriptive levels through its contrastive entries with dynamic display options. As such, it is aimed primarily at native speakers. However, we know from email enquiries that there is also interest among those learning German as a foreign language[2].

This paper is intended to demonstrate how it is possible to extend a reference work in a valuable way and to show, using hypothetical examples, how an existing monolingual resource could be transformed into a bilingual or multilingual platform for native speakers, for learners of German as a foreign language, and for other second-language learners, thereby appealing to additional groups of users. The development options outlined in what follows would constitute a considerable step forward for comparative, bilingual, and language-learning lexicography. The corpus-based principles underpinning the dictionary and the dynamic display of information on two descriptive levels provide users with the potential to undertake comprehensive comparisons of headwords across languages according to their own needs. It is worth emphasising that the types of extension considered in this paper are relatively easy to implement since the underlying structures have already been established, experience exists in using them, and extensive corpora are available for numerous other European languages.

2. On the Treatment of German Confusables and Paronyms

In every language there are terms that are easily mistaken or confused. Often these are words that are separated by just one or two letters, sometimes also differing with respect to their prefixes or suffixes. As Room (1979: 1) has pointed out, "we say one

[1] OWID is a lexicographic platform combining 11 different German reference guides with a unified search. OWID$_{plus}$ is an experimental platform for diverse multilingual lexical-lexicographic data.

[2] Some websites and university language centres are already linking to the paronym dictionary. See, for example, https://www.sprachenzentrum.fu-berlin.de/slz/sprachen-links/deutsch/wortschatz/index.html.

word when we mean another, half-comprehend or misunderstand words, and encounter unfamiliar and 'hard' words daily. In short, we confuse words". Speakers of German confuse words for different reasons, such as close semantic meaning between near-synonyms (e.g. *kalt/kühl* (*cold/crisp*)). Confusion also occurs due to similarity or identity of spelling among homographs, e.g. der *Band*/das *Band*/die *Band* (*volume*/*ribbon*/*band*) or because the words are identical in sound, as is the case with homophones such as *Leib*/*Laib* (*body/loaf*). However, these rather prototypical cases do not account for a full classification of commonly confused terms, and the reasons for them and the effects of the confusion are rather complex. As well as lexical confusion, confusables can also be the result of grammatical confusion, such as difficulties arising from varying inflection, the usage of neologisms and loan words, and uncertainties surrounding word formation patterns, congruence and variable genders of nouns, to name but a few causes. Klein (2018) provides a detailed account of different cases of lexical confusion for German.

Paronyms are a specific group of confusables. They are usually pairs of lexical items that, in different ways, exhibit similarities in their meaning and/or form of expression. A large proportion of these are adjectives (*sportlich-sportiv*, *autoritär-autoritativ*), but paronyms also include verbs (*kodieren-kodifizieren-coden*, *referieren-referenzieren*) and nouns (*Methode-Methodologie-Methodik*). These kinds of words lead in some cases to uncertainty and confusion in usage among native speakers as well as learners of German as a foreign language, which may in turn cause misunderstandings, and numerous discussions on internet forums testify to this[3].

2.1 Confusables in German Linguistics and Lexicography

The most comprehensive theoretical approach to paronymy so far is offered by Lăzărescu (1999). His model treats paronyms from a structuralist point of view, accounting for language as a formal and logical system, and is not based on empirical evidence in real communicative situations. Looking at this relation from a language learner's perspective and with approaches used in translation studies, Lăzărescu developed an elaborate model based on strict formal criteria, primarily word formation and syntax. He aimed to establish clear-cut boundaries between paronymy and other phenomena of lexical confusion, such as homographs, homophones, lexical alternatives, false friends, etc. Still, fundamentally his model was not based on large amounts of empirical evidence of language use and consisted of the following main categories: phonetic-orthographic aspects (*Föhn*/*Fön*), morphological aspects (*Kinderliebe*/*Kindesliebe*), syntactic aspects (*schuld*/*schuldig*) and stylistic aspects (*essen*/*fressen*).

[3] See the largest question and answer platform in Germany, gutefrage.net, or the forum Deutsch als Fremdsprache.

Since then, the phenomenon of paronymy has not attracted much attention, either from a corpus linguistic or from a cognitive linguistic perspective. Until today, this rather complex phenomenon is still widely under-researched and we still lack a definition of the phenomenon from a usage-based perspective incorporating cognitive aspects. Currently, investigations also focus on research into paronymy as a complex lexical-conceptual phenomenon, aiming to develop an empirically driven classification of paronyms using diverse genres of language evidence and including written and spoken texts (cf. Mell et al., 2019). Today, we are concerned with an empirically sound, usage-guided investigation of commonly confused words based on large corpus data. So far, we have gained valuable insights into functions in specific contextual instances, communicative functions, thematic domains, discourse and style, text types and degrees of semantic similarity or contrast between easily confused words. Furthermore, speakers' attitudes can be expressed through their choice of paronyms, while encyclopaedic knowledge and cultural experience also play a key role in the use and interpretation of specific discourse-bound word pairs. These influential elements can be detected through collocations and grammatical constructions in context. They are more or less conveyed meta-linguistically and are therefore explicit in written communication. Overall, defining and classifying paronyms is a complex matter. Paronymy is not a lexical relation but a dynamic conceptual relation with cognitive implications which are visible on a linguistic level. In order to develop a full model, the identification of communicative functions and influences on lexical confusions is necessary. The effects of lexical misuse open up a number of questions concerning misunderstanding or semantic change.

With regard to German lexicography, this phenomenon has already received some attention (cf. Klein, 2018), but not since corpus data have been shaping the lexicographic landscape. Confusing words are, in fact, not systematically documented in standard dictionaries although they have been of interest to a small number of lexicographers for over a hundred years. The first dictionary by Wustmann (1891), rather random and limited in scope, followed a prescriptive tone, pointing to the correct usage of alternative plural forms. Specialised dictionaries were subsequently written by larger publishing companies such as Duden (Müller, 1973) and PONS (Pollmann, Wolk, 2010), appearing as new lexicographic authorities trying to disentangle all types of frequently confused words. Traditionally, their entries often instruct users or inform them about the "correct" use, the "correct" choice of lexemes or recommend avoiding certain terms. Some of the cases described in prescriptive reference guides behave differently in authentic language due to semantic change. Here, Hanks's (2013) observation with respect to English also holds true for German:

> These standards were based on the ill-defended assumptions that earlier forms of a language are somehow more 'correct' than contemporary forms and that etymology guarantees meaning. (Hanks, 2013: 514)

A dictionary solely dedicated to genuine German paronyms and accounting for them from a contemporary descriptive perspective, however, was never systematically compiled until 2018.

2.2 Paronyms, Cognates and Equivalents in other Languages

On closer inspection of paronym cases and the headwords identified in the paronym dictionary project, it becomes apparent that these pairs – and also groups of paronyms (such as *provokant-provozierend-provokativ-provokatorisch* or *patriarchalisch-patriarchal-patriarchisch*) – constitute a very heterogeneous category. The linguistic uncertainty they provoke can be traced back to different causes, and the contextual confusion between words takes different forms among native speakers and learners of German. The latter group encounters very different problems depending on their own first language. As well as native expressions such as *knöchern-knochig-knöch(e)rig*, *lebenslang-lebenslänglich*, *fachkundig-fachkundlich*, paronyms include technical expressions (*kardiologisch-kardial*, *linguistisch-lingual, Parodontose-Parodontitis*) and loanwords (*fiktiv-fiktional*, *Anarchie-Anarchismus*), that is, expressions borrowed from other languages. It is precisely these expressions that frequently have cognates – words derived from the same etymon – in other European languages. If native speakers lack the relevant linguistic or encyclopaedic knowledge, then more significant communication problems can arise with these technical and borrowed expressions. If these terms differ in meaning, for example, in English and French, but less so in form, then they will cause difficulties for native speakers[4]. Particularly tricky are words where differences cause well-known mistranslations, especially so-called ‘false friends’. If we compare German and English, for example, we quickly realise that the German adjectives *sensibel-sensitiv* do not correspond at all in meaning with the formally equivalent English pair *sensible-sensitive.* The question may also arise whether the English word *muscular* can be translated as both *muskulär* and *muskulös.* In exactly the same way, German learners of English will wonder about the contexts in which *versichern* is the most appropriate translation for *assure, ensure,* or *insure.*

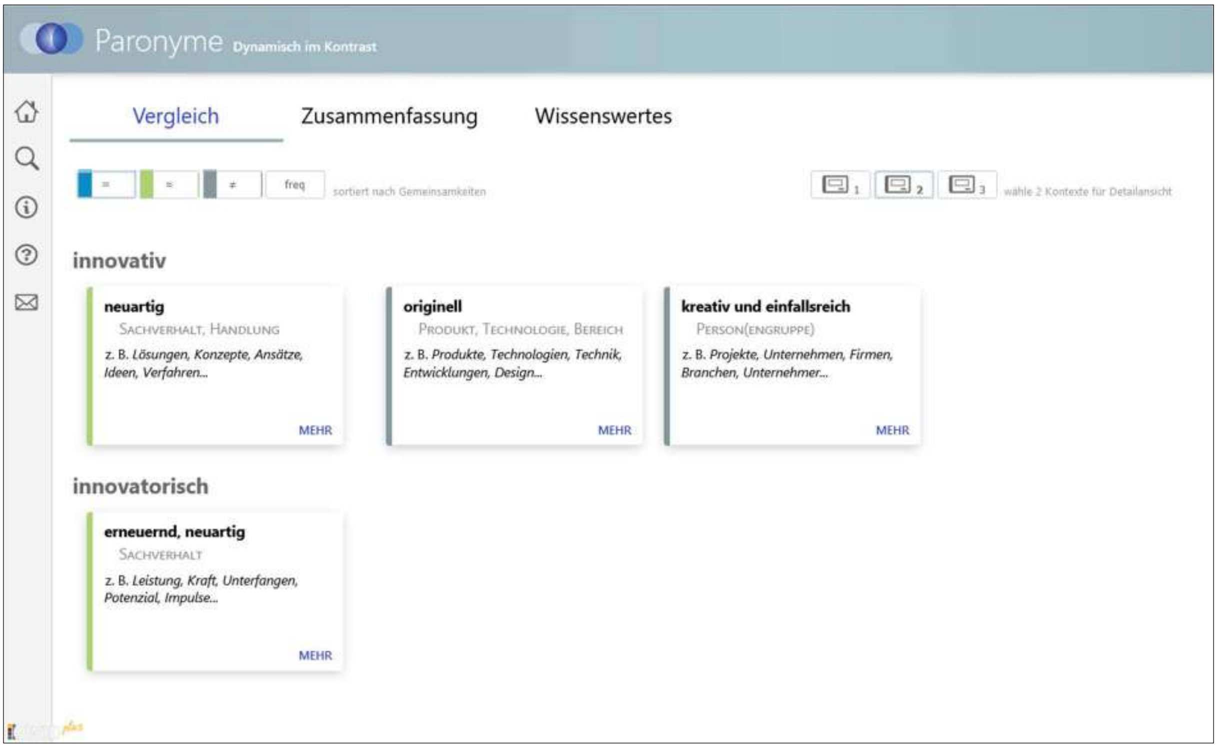
3. The Monolingual Dictionary *Paronyme – Dynamisch im Kontrast*

Approximately 2,000 more or less common German paronyms were identified using corpus-based methods and then analysed and edited (Schnörch, 2015). The most frequent of them were included in the new paronym dictionary. The discrepancy between the number of potential headwords identified and the actual number of dictionary entries (around 350) is explained by the fact that a large proportion of the words were compounds or negations of other headwords. For example, the entry *Technik-Technologie* alone accounts for 65 compounds attested in the corpus e.g.,

[4] This applies in this case both to German-speaking learners of English or French, and to English or French speakers who are learning German.

Antriebstechnik-Antriebstechnologie, *Atomtechnik-Atomtechnologie*, *Computertechnik-Computertechnologie*.

The lexicographic practice involved in the dictionary *Paronyme - Dynamisch im Kontrast* (Storjohann, 2018; 2019) is notable, among other things, for the work on a relatively balanced corpus and the combination of complementary corpus-based methods with editorial analysis and interpretation. In the description of the lexicographic data, the empirical and descriptive approach takes into account conceptual referential aspects of the word's meaning as well as the connected documentation of linguistic and extra-linguistic information. As is appropriate to the object of enquiry, all the information is presented in contrastive entries of up to four headwords in a dynamic descriptive model, as the following example demonstrates. For entries such as *innovativ-innovatorisch* there is a relationship of similar meaning when the two expressions are applied to subject matter that can be characterised as '*neuartig'*, e.g. *Ansätze* (*approaches*), *Ideen* (*ideas*), *Denken* (*thinking*). They are not completely conterminous, because only *innovativ* can be used to refer to actions (*denken*, *arbeiten*, *gestalten*). Furthermore, in contrast to *innovatorisch*, the word *innovativ* can be used to describe products, technologies, and fields as '*originell'* (*original*) and people as '*kreativ'* and '*einfallsreich'* (*creative* and *inventive*). Hence, there are both similarities and differences in this case (see Figure 1).


Paronyme Dynamisch im Kontrast
Vergleich
Zusammenfassung
Wissenswertes
=
≈
≠
freq
sortiert nach Gemeinsamkeiten
1
2
3
wähle 2 Kontexte für Detailansicht
MEHR
MEHR
MEHR
MEHR

innovativ

neuartig
SACHVERHALT, HANDLUNG
z. B. *Lösungen, Konzepte, Ansätze, Ideen, Verfahren...*

originell
PRODUKT, TECHNOLOGIE, BEREICH
z. B. *Produkte, Technologien, Technik, Entwicklungen, Design...*

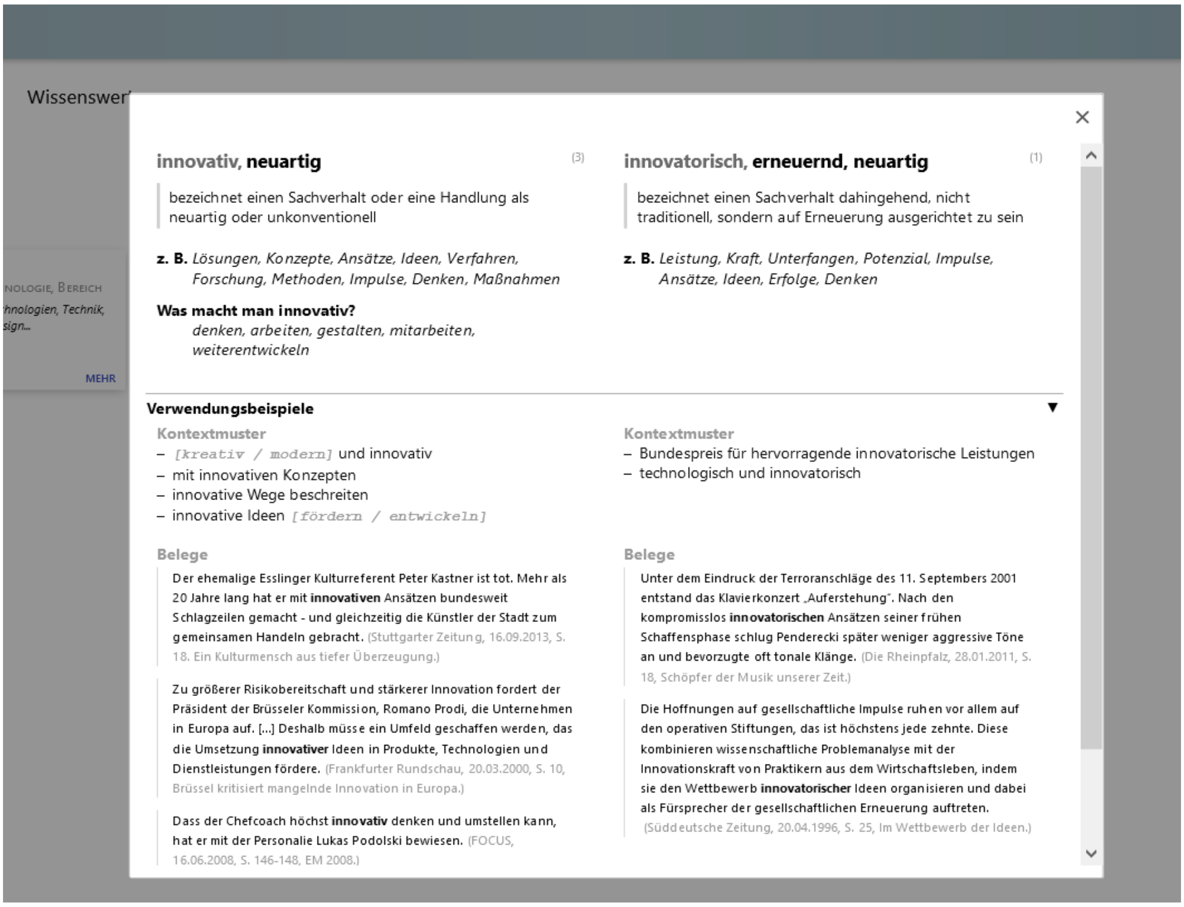
kreativ und einfallsreich
PERSON(ENGRUPPE)
z. B. *Projekte, Unternehmen, Firmen, Branchen, Unternehmer...*

innovatorisch

erneuernd, neuartig
SACHVERHALT
z. B. *Leistung, Kraft, Unterfangen, Potenzial, Impulse...*

Figure 1: Entry for *innovativ* and *innovatorisch* in the comparative outline view

The individual contextual uses (shown in boxes/tiles) are listed horizontally for each headword and connected to one another vertically in cases of meaning overlaps in usage with the partner term. Colour is used to clearly mark semantically similar usage or usage that occurs only for one term or that is divergent; these are either positioned directly beneath one another or are offset[5]. In this way, it is possible to see at a glance the number of available senses, as well as the semantic overlaps and differences between the relevant paronyms. This form of presentation provides both a compact comparative overview, arranged according to different linguistic parameters[6], and a means of navigation to the detailed view. In the detailed view, which represents an additional level of description, up to three senses can be selected per mouse click in order to study additional information – such as collocations, attested examples, and synonyms – directly alongside one another (see Figure 2).



innovativ, neuartig (3)

bezeichnet einen Sachverhalt oder eine Handlung als neuartig oder unkonventionell

z. B. *Lösungen, Konzepte, Ansätze, Ideen, Verfahren, Forschung, Methoden, Impulse, Denken, Maßnahmen*

Was macht man innovativ?
denken, arbeiten, gestalten, mitarbeiten, weiterentwickeln

Verwendungsbeispiele ▼

Kontextmuster
- *[kreativ / modern]* und innovativ
- mit innovativen Konzepten
- innovative Wege beschreiten
- innovative Ideen *[fördern / entwickeln]*

Belege

Der ehemalige Esslinger Kulturreferent Peter Kastner ist tot. Mehr als 20 Jahre lang hat er mit **innovativen** Ansätzen bundesweit Schlagzeilen gemacht - und gleichzeitig die Künstler der Stadt zum gemeinsamen Handeln gebracht. (Stuttgarter Zeitung, 16.09.2013, S. 18. Ein Kulturmensch aus tiefer Überzeugung.)

Zu größerer Risikobereitschaft und stärkerer Innovation fordert der Präsident der Brüsseler Kommission, Romano Prodi, die Unternehmen in Europa auf. [...] Deshalb müsse ein Umfeld geschaffen werden, das die Umsetzung **innovativer** Ideen in Produkte, Technologien und Dienstleistungen fördere. (Frankfurter Rundschau, 20.03.2000, S. 10, Brüssel kritisiert mangelnde Innovation in Europa.)

Dass der Chefcoach höchst **innovativ** denken und umstellen kann, hat er mit der Personalie Lukas Podolski bewiesen. (FOCUS, 16.06.2008, S. 146-148, EM 2008.)

innovatorisch, erneuernd, neuartig (1)

bezeichnet einen Sachverhalt dahingehend, nicht traditionell, sondern auf Erneuerung ausgerichtet zu sein

z. B. *Leistung, Kraft, Unterfangen, Potenzial, Impulse, Ansätze, Ideen, Erfolge, Denken*

Kontextmuster
- Bundespreis für hervorragende innovatorische Leistungen
- technologisch und innovatorisch

Belege

Unter dem Eindruck der Terroranschläge des 11. Septembers 2001 entstand das Klavierkonzert „Auferstehung". Nach den kompromisslos **innovatorischen** Ansätzen seiner frühen Schaffensphase schlug Penderecki später weniger aggressive Töne an und bevorzugte oft tonale Klänge. (Die Rheinpfalz, 28.01.2011, S. 18, Schöpfer der Musik unserer Zeit.)

Die Hoffnungen auf gesellschaftliche Impulse ruhen vor allem auf den operativen Stiftungen, das ist höchstens jede zehnte. Diese kombinieren wissenschaftliche Problemanalyse mit der Innovationskraft von Praktikern aus dem Wirtschaftsleben, indem sie den Wettbewerb **innovatorischer** Ideen organisieren und dabei als Fürsprecher der gesellschaftlichen Erneuerung auftreten. (Süddeutsche Zeitung, 20.04.1996, S. 25, Im Wettbewerb der Ideen.)

Figure 2: Detailed view of the semantically similar uses of *innovativ* und *innovatorisch* to mean '*neuartig*' (*new*).

The attested examples of usage are selected and edited in such a way that, in cases where the context is similar, each headword is documented together with the same collocation. The aim of this is to illustrate contextual interchangeability in cases of synonymous contexts (here, for example, *innovative/innovatorische*

[5] Individual contextual senses that are offset from one another become particularly important when the two expressions are used in different contexts, not just one of the headwords as in the case shown in Figure 1 (cf., for example, the entry *autoritativ-autoritär* in the paronym dictionary).

[6] For example, a menu makes it possible to sort the uses in tiles also by frequency of occurrence.

Ansätze/approaches, *Ideen/ideas*). In this way, overlaps are illustrated concretely in real language. In the same way, different reference objects are manifested linguistically through collocations, thereby highlighting the differences between the paronyms and demonstrating those differences with examples. For instance, there is an attested example which illustrates the use of *innovativ* with the verb *denken* (*to think*) which does not appear for *innovatorisch*, because the latter term does not occur directly together with verbs. This is the main point of difference in this specific usage.

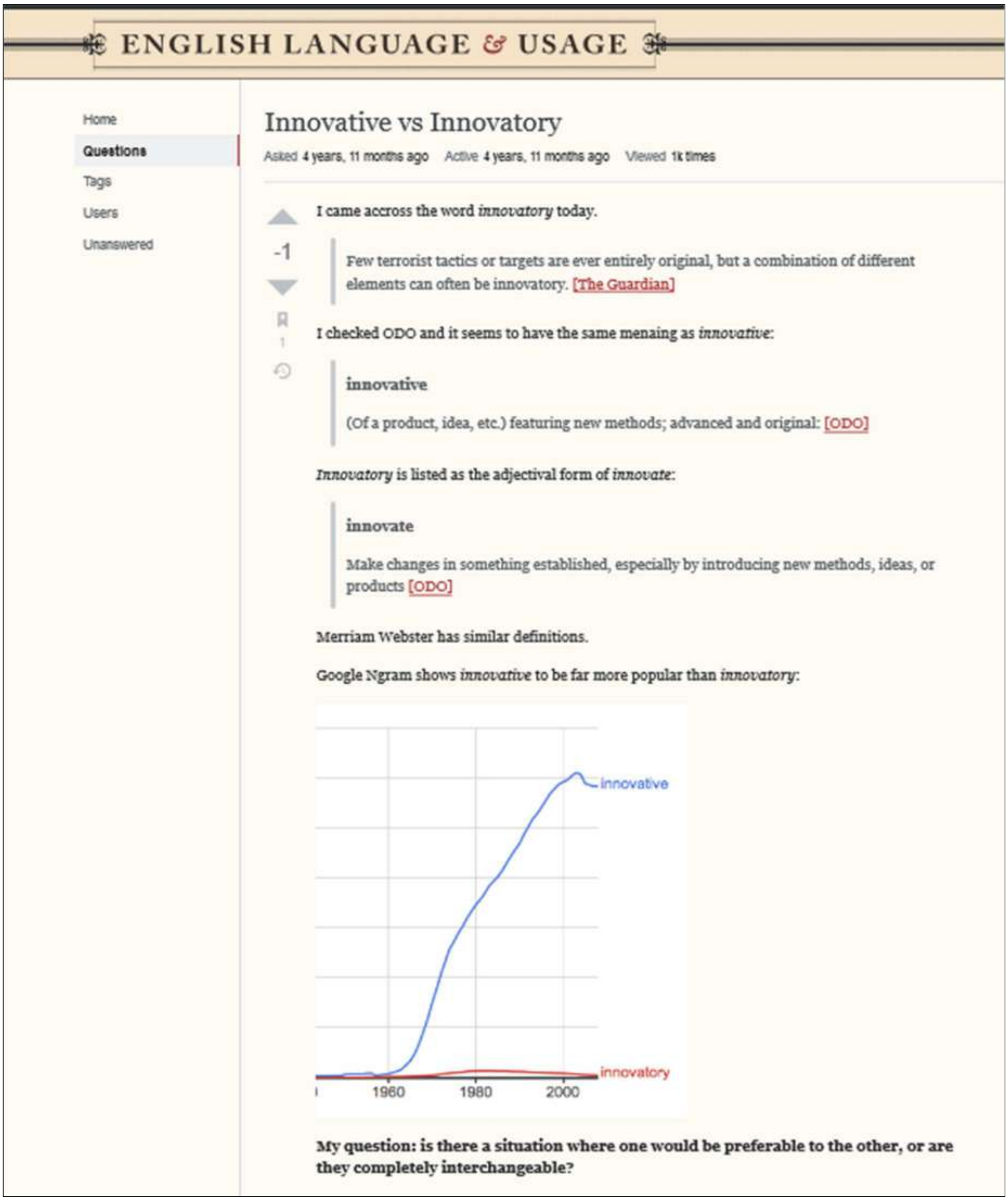
4. A Possible Bilingual Dictionary

While the new paronym dictionary enables us to check the fine differences in use between German expressions that are similar in form, there are scarcely any comparable reference works for other languages. However, there is certainly awareness in other languages of the phenomenon of paronymy. For English, Room‘s (2000) *Dictionary of Confusable Words* is a work that is similar in content. However, it is not comparable to the German paronym dictionary because of its very small scope, its depth of description, presentation, methodology, and the fact that it is not up to date. A brief glance at the headword list in the German paronym dictionary reveals formal equivalents in English, with some examples given in Table 1.

German	**English**
Akzeptanz-Akzeptatibilität	*acceptance-acceptability*
anarchisch-anarchistisch	*anarchical-anarchistic*
autoritär-autoritativ	*authoritarian-authoritative*
elektrisch-elektronisch	*electric-electronic*
fiktiv-fiktional	*fictional-fictitious*
human-humanitär	*human-humanitarian*
innovativ-innovatorisch	*innovative-innovatory*
konzeptuell-konzeptionell	*conceptual-conceptional-conceptive*
legislativ-legislatorisch	*legislative-legislatorial*
minimal-minimalistisch	*minimal-minimalistic*
mysteriös-mystisch-mythisch-mythologisch	*mysterious-mystic-mythical-mythological*
originell-original-originär	*original-originative-originary*
sensibel-sensitiv	*sensible-sensitive*
unsozial-asozial-antisozial	*unsocial-asocial-antisocial*

Table 1: Examples of German paronym pairs and their formal equivalents in English

An examination of internet forums also shows that English native speakers experience similar uncertainty to German native speakers with these kinds of pairs (cf. *innovative-innovatory* in Figure 3).



ENGLISH LANGUAGE & USAGE

Innovative vs Innovatory

Asked 4 years, 11 months ago Active 4 years, 11 months ago Viewed 1k times

I came accross the word *innovatory* today.

> Few terrorist tactics or targets are ever entirely original, but a combination of different elements can often be innovatory. [The Guardian]

I checked ODO and it seems to have the same menaing as *innovative*:

> innovative
>
> (Of a product, idea, etc.) featuring new methods; advanced and original: [ODO]

Innovatory is listed as the adjectival form of *innovate*:

> innovate
>
> Make changes in something established, especially by introducing new methods, ideas, or products [ODO]

Merriam Webster has similar definitions.

Google Ngram shows *innovative* to be far more popular than *innovatory*:

My question: is there a situation where one would be preferable to the other, or are they completely interchangeable?

Figure 3: Native speaker’s question about *innovative*/*innovatory* in the forum english.stackexchange

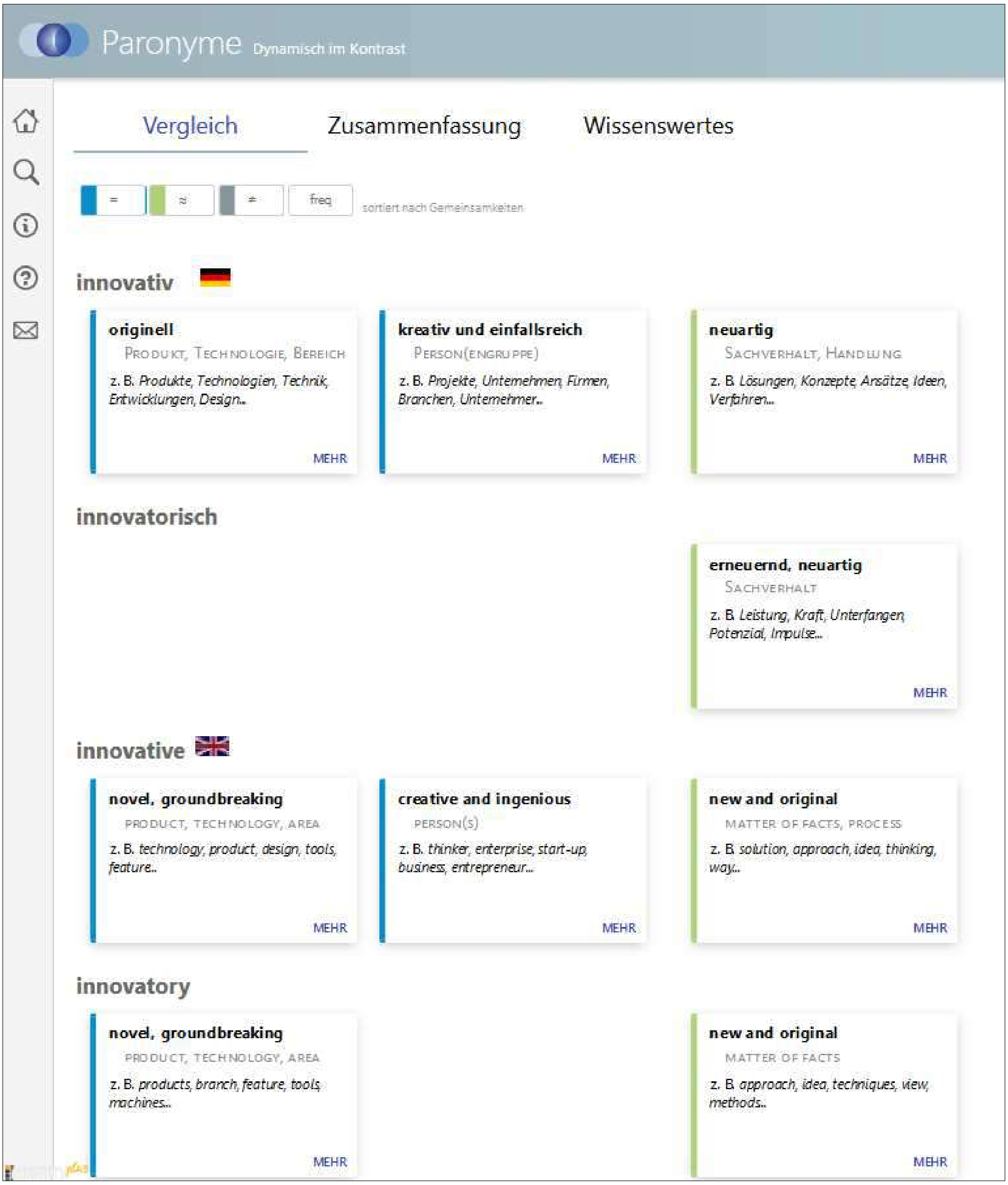
In addition to native speakers’ uncertainty concerning the appropriate use of these expressions, German-speaking learners of English may also wonder about the most suitable translation for the two German adjectives *innovativ-innovatorisch*. In the standard bilingual resources, they will come across *innovative* and *innovatory.* However, they will learn nothing about exactly which contexts they are used in, whether they behave analogously to *innovativ-innovatorisch*, or whether the same referential differences apply to the description of products, fields, and people, or the like. And by the same token, English-speaking learners of German as a foreign language may be uncertain as to whether the German adjectives *innovativ* and *innovatorisch* can be used as translations for the English terms. Despite the different perspectives of the users, very similar questions arise. If the German paronym dictionary were to be supplemented with a bilingual component, all the preceding questions could be answered. We would require only an extension of the content of the existing monolingual dictionary following the same principles of lexical analysis and documentation in order to achieve a reliable bilingual comparison.

If we synchronise the entries for (Gm.) *innovativ-innovatorisch* and (Eng.) *innovative-innovatory* and connect them to one another, then they can be looked up both as individual lemmas and as a pair, and both monolingually and bilingually. As such, we can cater for all the groups of users mentioned previously and their different requirements. Figure 4 shows what such an entry might look like. The terms are arranged as pairs in individual languages, one beneath the other. The optimum here would undoubtedly be to have an additional option to arrange them by lemma, so that *innovative*, for example, could be positioned directly under *innovativ*, and *innovatory* under *innovatorisch.* Flexible modes of presentation already exist, which take into account different organising principles, sorting individual usages, for example, according to their frequency.

The hypothetical dictionary entry constructed for illustrative purposes was created using Sketch Engine and an English-language web corpus. This is not a completely comparable word analysis, since the underlying data and methods of analysis for the German adjective pair differ (cf. Storjohann, 2021). The purpose of Figure 4 is simply an illustrative representation of a possible bilingual resource, rather than the complete accuracy of the English lexicographic content [7]. The English corpus is very comprehensive and reflects the language of everyday public communication. Thus, it can be assumed that the meanings identified do not diverge too greatly from linguistic reality. As with *innovativ-innovatorisch*, both similarities and differences were discovered for *innovative-innovatory.* However, they are positioned slightly differently for the two pairs. The uses marked in blue demonstrate strong overlaps and can be considered essentially identical. The green contexts have similarities, but with small semantic nuances. For example, there is a key difference between *innovative-innovatory*

[7] It is possible that the content of the English articles might be different if more representative data was used. Since Figure 4 presents a hypothetical word-entry which follows the structure of an existing German entry from the paronym dictionary, it includes information exclusively in German, such as the terminology '*z. B.*', '*Belege'* or '*Kontextmuster'*. These would have to be in English.

in the context of '*new and original*' insofar as *innovative* occurs with verbs that denote processes.



innovativ

originell
PRODUKT, TECHNOLOGIE, BEREICH
z. B. *Produkte, Technologien, Technik, Entwicklungen, Design..*
MEHR

kreativ und einfallsreich
PERSON(ENGRUPPE)
z. B. *Projekte, Unternehmen, Firmen, Branchen, Unternehmer..*
MEHR

neuartig
SACHVERHALT, HANDLUNG
z. B. *Lösungen, Konzepte, Ansätze, Ideen, Verfahren...*
MEHR

innovatorisch

erneuernd, neuartig
SACHVERHALT
z. B. *Leistung, Kraft, Unterfangen, Potenzial, Impulse...*
MEHR

innovative

novel, groundbreaking
PRODUCT, TECHNOLOGY, AREA
z. B. *technology, product, design, tools, feature..*
MEHR

creative and ingenious
PERSON(S)
z. B. *thinker, enterprise, start-up, business, entrepreneur...*
MEHR

new and original
MATTER OF FACTS, PROCESS
z. B. *solution, approach, idea, thinking, way...*
MEHR

innovatory

novel, groundbreaking
PRODUCT, TECHNOLOGY, AREA
z. B. *products, branch, feature, tools, machines...*
MEHR

new and original
MATTER OF FACTS
z. B. *approach, idea, techniques, view, methods..*
MEHR

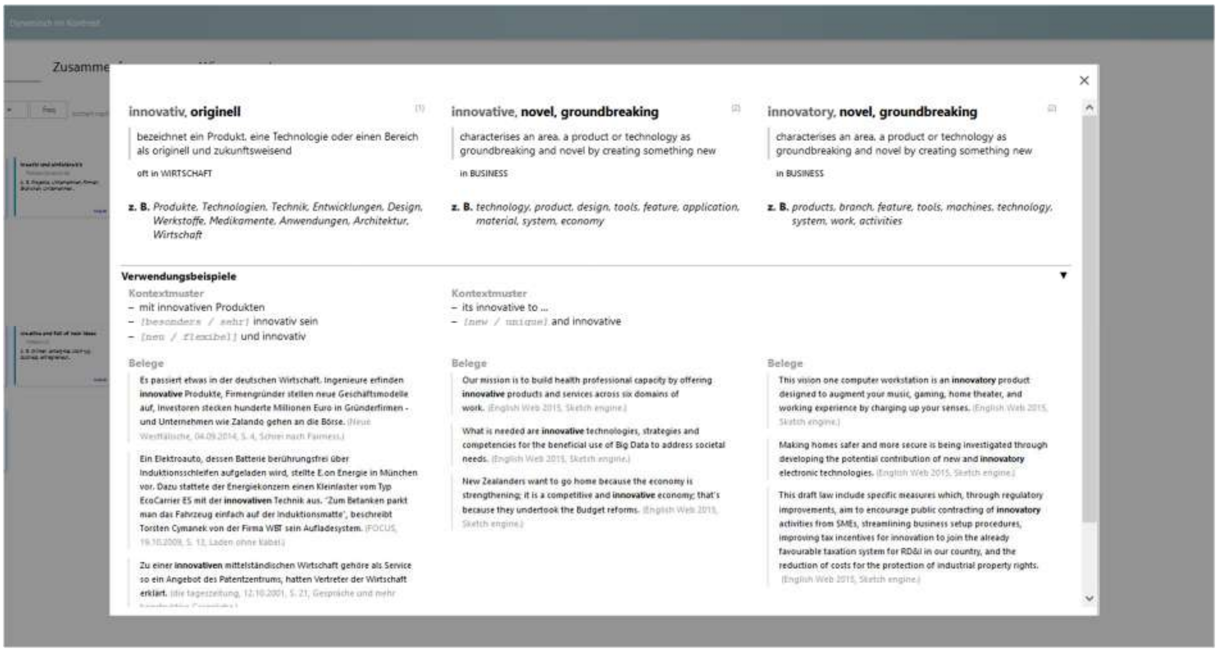
Figure 4: A hypothetical cross-language dictionary entry in the comparative outline view

By contrast, *innovatory* does not modify any verbs, so that the reference PROCESS does not appear in the short paraphrase. *Innovativ* in the sense of '*neuartig*' and *innovatorisch* in the sense of '*erneuernd, neuartig*' differ from one another in their short paraphrase and in the fact that, analogous to the English pair, verbs that denote actions and processes can only be further characterised as *innovativ*. In this case, the information HANDLUNG (PROCESS) is also missing from the reference underneath. Leaving aside possible divergences as a result of insufficient alignment between the two corpora, this outline view nonetheless reveals the following:

- the uses exhibited by each adjective
- the relationship between the individual headwords in each pair
- the relationship between the two pairs.

Detailed views of multiple contexts and usages can also be compared between the two languages. As an example, contexts have been chosen here which refer to products, technology, or fields and which were attested for three of the four adjectives (*innovativ* '*originell'*, *innovative* and *innovatory* '*novel, groundbreaking'*), but not for *innovatorisch* (see Figure 5).

This form of parallel view preserves the direct comparison of paraphrases, domains/referential frames, collocations, and illustrative examples which again show selected and, as far as possible, analogous contextual partner terms being used with the corresponding adjectives (here *Produkte/products*, *Technik/technology*, *Wirtschaft/economy*, *activities*). The parallel placement of information highlights the strong commonalities. Similarly, the less strongly overlapping contexts can also be considered in more detail in this way.



innovativ, **originell**

bezeichnet ein Produkt, eine Technologie oder einen Bereich als originell und zukunftsweisend

oft in WIRTSCHAFT

z. B. *Produkte, Technologien, Technik, Entwicklungen, Design, Werkstoffe, Medikamente, Anwendungen, Architektur, Wirtschaft*

innovative, **novel, groundbreaking**

characterises an area, a product or technology as groundbreaking and novel by creating something new

in BUSINESS

z. B. *technology, product, design, tools, feature, application, material, system, economy*

innovatory, **novel, groundbreaking**

characterises an area, a product or technology as groundbreaking and novel by creating something new

in BUSINESS

z. B. *products, branch, feature, tools, machines, technology, system, work, activities*

Verwendungsbeispiele

Kontextmuster
- mit innovativen Produkten
- [besonders / sehr] innovativ sein
- [neu / flexibel] und innovativ

Belege

Es passiert etwas in der deutschen Wirtschaft. Ingenieure erfinden **innovative** Produkte, Firmengründer stellen neue Geschäftsmodelle auf, Investoren stecken hunderte Millionen Euro in Gründerfirmen - und Unternehmen wie Zalando gehen an die Börse. (Neue Westfälische, 04.09.2014, S. 4, Schrei nach Fairness.)

Ein Elektroauto, dessen Batterie berührungsfrei über Induktionsschleifen aufgeladen wird, stellte E.on Energie in München vor. Dazu stattete der Energiekonzern einen Kleinlaster vom Typ EcoCarrier ES mit der **innovativen** Technik aus. "Zum Betanken parkt man das Fahrzeug einfach auf der Induktionsmatte", beschreibt Torsten Cymanek von der Firma WBT sein Aufladesystem. (FOCUS, 19.10.2009, S. 12, Laden ohne Kabel.)

Zu einer **innovativen** mittelständischen Wirtschaft gehöre als Service so ein Angebot des Patentzentrums, hatten Vertreter der Wirtschaft erklärt. (die tageszeitung, 12.10.2001, S. 21, Gespräche und mehr

Kontextmuster
- its innovative to ...
- [new / unique] and innovative

Belege

Our mission is to build health professional capacity by offering **innovative** products and services across six domains of work. (English Web 2015, Sketch engine.)

What is needed are **innovative** technologies, strategies and competencies for the beneficial use of Big Data to address societal needs. (English Web 2015, Sketch engine.)

New Zealanders want to go home because the economy is strengthening; it is a competitive and **innovative** economy; that's because they undertook the Budget reforms. (English Web 2015, Sketch engine.)

Belege

This vision one computer workstation is an **innovatory** product designed to augment your music, gaming, home theater, and working experience by charging up your senses. (English Web 2015, Sketch engine.)

Making homes safer and more secure is being investigated through developing the potential contribution of new and **innovatory** electronic technologies. (English Web 2015, Sketch engine.)

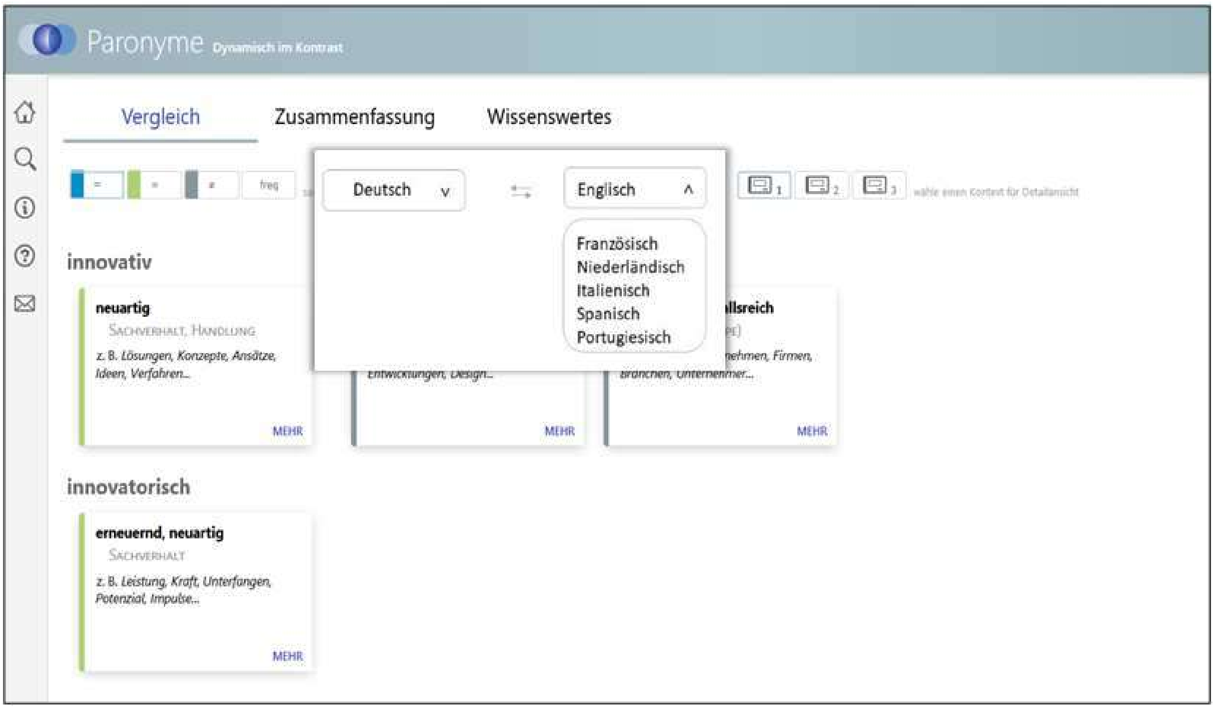
This draft law include specific measures which, through regulatory improvements, aim to encourage public contracting of **innovatory** activities from SMEs, streamlining business setup procedures, improving tax incentives for innovation to join the already favourable taxation system for RD&I in our country, and the reduction of costs for the protection of industrial property rights. (English Web 2015, Sketch engine.)

Figure 5: Hypothetical cross-language dictionary entry in detailed view

5. A Multilingual Reference Work and Portal

It is not a big step from a bilingual reference work to the construction, with additional modules, of a comprehensive multilingual resource. Here, we would restrict ourselves to

European languages since, as has already been emphasised, it is apparent that there are a series of word pairs among loanwords, including internationalisms, which may also create uncertainty or confusion in other European languages; possible examples would include: Gm. *effektiv-effizient*, Eng. *effective-efficient*, Fr. *efficace-effectif-efficient*, It. *efficace-efficiente*. It is difficult to assess the extent to which these kinds of word pairs occur in the individual languages; at the same time, there is a clear interest in comparability. There is not always an equivalent pair in each European language for each paronym pair with a Latin or French root. However, even this information is valuable from a learner's perspective. Even translation equivalents that do not exhibit any paronyms in the chosen language (e.g. Fr. *innovant–novateur* for Gm. *innovativ-innovatorisch*) can be indicated, and important information provided about translation problems[8]. The user chooses any two languages for comparison from a menu (see Figure 6).



innovativ

neuartig
SACHVERHALT, HANDLUNG
z. B. *Lösungen, Konzepte, Ansätze, Ideen, Verfahren...*

Entwicklungen, Design...

llsreich

nehmen, Firmen, Branchen, Unternehmer...

innovatorisch

erneuernd, neuartig
SACHVERHALT
z. B. *Leistung, Kraft, Unterfangen, Potenzial, Impulse...*

Figure 6: Possible language selection via a menu

Corresponding content is overlaid dynamically, such as the choice of language-specific lists of headwords in the search options (on the left in the menu in Figure 6). The number of potential users is increased considerably with the addition of multiple languages. In order to build multilingual data in a consistent way and create a complex multilingual reference system, it is not enough to simply put individual dictionaries together. The exact lexicographic content to be shown, the level of description at which it can be presented, and how the interfaces are best realised would all have to be considered with a freshly conceived, common entry architecture.

[8] This would mean that, ultimately, this was, in terms of content, much more than a pure paronym dictionary.

Of course, the foreign language perspective (for example, translation) comes more strongly into focus each time languages are compared in a reference work. In this context, it is possible to imagine extensions involving foreign language learning exercises, in which these dictionary resources are not limited to their reference function, but are instead extended to become a work portal. Having multiple bilingual dictionaries available, which could be selected in a targeted way, would open up a wide-ranging, convenient, and flexible reference space for this specific phenomenon, creating a comprehensive multilingual language resource which documents easily confused, specifically paronymic expressions from a comparative European perspective.

To construct this kind of resource would require comparable corpora or completely parallel corpora, as well as comparable methods of analysis. Sketch Engine would provide a tool to investigate more accurately relationships across languages by means of multilingual collocation profiles. Naturally, there are limits to what can be presented with the current form of the paronym dictionary. Already, if the language menu were to be supplemented with a third or fourth language it would scarcely be possible to maintain a compact overview of the content. In the detailed parallel view, it is also not possible to select any number of boxes and arrange them next to one another. For that, new and creative solutions would be necessary.

6. Conclusion

Studying linguistic data has been shaping our understanding of how meanings are constructed through context. German user studies (e.g. Müller-Spitzer, 2014) of online dictionaries have increased our focus on users' interests, and hypertext structures have transformed the way we present lexicographic information. It should be noted that the ideas presented above are mere lexicographic fiction, and so far, no user study has been conducted, and the scope of potential users has not been discussed thoroughly. Moreover, options as to linking such a specific resource with more general dictionaries to complement those or integrating it into even larger lexicographic enterprises have not been considered yet. Certainly, one should pursue such endeavours first. Instead, this paper reflects on possibilities. It is a genuine call for tools which extend across languages and across existing resources in order to deal with cases of linguistic uncertainty.

Paronyme – Dynamisch im Kontrast is a project which is structured monolingually and which has chosen new forms of presentation in order to compare, in different ways, the linguistic patterns and structures of easily confused words. In the process, it has been able to take into account flexible reference queries determined by the user. The Design Thinking approach that has been used offers the potential as a platform, both conceptually and in terms of language technology, to document in lexicographic form the different results of contrastive lexical analyses. As such, it does not have to remain limited to a single language and is able to make allowances for a range of different applications. In this way, the spectrum of lexicographic description can be expanded from a monolingual reference work to one that is bilingual or multilingual. At the moment, it is possible to undertake dynamic comparisons between words and their uses within the German language; it is conceivable that, by analogy, corresponding comparisons could be offered beyond that, between two or more languages chosen at will. In addition, it could be a model for similar digital dictionaries with a principal focus on linguistic comparison, for example, synonym or antonym dictionaries. What has been outlined here is a visionary digital paronym network, but it is worth emphasising that this is an eminently realisable vision of a resource that would be of inestimable value.

7. References

Forum Deutsch als Fremdsprache. Accessed at https://www.deutsch-als-fremdsprache.de/austausch/forum/read.php?4,45474. (20 May 2021)

English-stackexchange. Accessed at: https://english.stackexchange.com/questions/287083/innovative-vs-innovatory. (20 May 2021)

elexiko. *Online-Wörterbuch zur deutschen Gegenwartssprache.* Accessed at: https://www.owid.de/docs/elex/start.js. (20 May 2021)

gutefrage.net – *Deutschlands größte Frage-Antwort-Plattform.* Accessed at https://www.gutefrage.net/. (20 May 2021)

Klein, W. P. (2018). *Sprachliche Zweifelsfälle im Deutschen. Theorie, Praxis, Geschichte.* Berlin & Boston: de Gruyter.

Lăzărescu, I. (1999). *Die Paronymie als lexikalisches Phänomen und die Paronomasie als Stilfigur im Deutschen.* Bukarest: Anima.

Mell, R., Schnörch, U. & Storjohann, P. (2019). Korpussemantische Einflussfaktoren auf Eigenschaften und Funktionen von Paronymen. *Deutsche Sprache* 1/2019, pp. 53–67.

Müller, W. (1973). *Leicht verwechselbare Wörter.* Duden Taschenwörterbücher Vol. 17. Mannheim: Bibliographisches Institut.

Müller-Spitzer, C. ed. (2014). *Using Online Dictionaries.* Berlin & Boston: de Gruyter.

OWID. *Online-Wortschatz-Informationssystem Deutsch* (2008ff.) Accessed at: https://www.owid.de/. (20 May 2021)

OWID*plus*. Accessed at: https://www.owid.de/plus/index.html. (20 May 2021)

Paronyme – Dynamisch im Kontrast. Accessed at: https://www.owid.de/parowb/. (20 May 2021)

Pollmann, Ch. & Wolk, U. (2010). *Wörterbuch der verwechselten Wörter. 1000 Zweifelsfälle verständlich erklärt.* Stuttgart: Pons.

Rooms, A. (2000). *Dictionary of Confusable Words.* London: Routledge.

Schnörch, U. (2015). Wie viele Paronympaare gibt es eigentlich? Das Zusammenspiel aus korpuslinguistischen und redaktionellen Verfahren zur Ermittlung einer Paronymstichwortliste. *Sprachreport* 4, pp. 16–26.

Sketch Engine. (Korpus English Web 2015 enTenTen15). Accessed at https://www.sketchengine.eu/. (20 May 2021)

Storjohann, P. (2021). Korpusmethoden zur Erarbeitung eines Wörterbuches leicht verwechselbarer Ausdrücke. In S. Steinmetz, D. Strömsdörfer, M. Willmann & N. Wulff (eds.) *Deutsch weltweit - Grenzüberschreitende Perspektive auf die Schnittstellen von Forschung und Vermittlung* (=MatDaF 104). Göttingen: Universitätsverlag, pp. 207–288.

Storjohann, P. (2019). Paronyme – Dynamisch im Kontrast. Ein kognitiv ausgerichtetes, multifunktionales, dynamisches Nachschlagewerk. *Deutsche Sprache* 1, pp. 82–94.

Storjohann, P. (2018). Commonly Confused Words in Contrastive and Dynamic Dictionary Entries. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (eds.) *Proceedings of the XVIII EURALEX International Congress, EURALEX 2018.* Ljubljana: Znanstvena založba, pp. 187–197.

Wustmann, G. (1891). *Allerhand Sprachdummheiten: kleine deutsche Grammatik des Zweifelhaften, des Falschen und des Häßlichen.* Grunow: Leipzig.

Designing the ELEXIS Parallel Sense-Annotated Dataset in 10 European Languages

Federico Martelli[1], Roberto Navigli[1], Simon Krek[2], Jelena Kallas[3], Polona Gantar[4], Svetla Koeva[5], Sanni Nimb[10], Bolette Sandford Pedersen[8], Sussi Olsen[8], Margit Langemets[3], Kristina Koppel[3], Tiiu Üksik[3], Kaja Dobrovoljc[2], Rafael-J. Ureña-Ruiz[9], José-Luis Sancho-Sánchez[9], Veronika Lipp[11], Tamás Váradi[12], András Győrffy[11], Simon László[11], Valeria Quochi[14], Monica Monachini[14], Francesca Frontini[14], Carole Tiberius[13], Rob Tempelaars[13], Rute Costa[6], Ana Salgado[6] [7], Jaka Čibej[2] and Tina Munda[2]

[1]Sapienza NLP Group, Department of Computer Science, Sapienza University of Rome, Italy
[2]Artificial Intelligence Laboratory, Jožef Stefan Institute, Slovenia
[3]Institute of the Estonian Language, Estonia
[4]Faculty of Arts, University of Ljubljana, Slovenia
[5]Institute for Bulgarian Language, Bulgarian Academy of Sciences, Bulgaria
[6]NOVA CLUNL, Centro de Linguística da Universidade NOVA de Lisboa, Portugal
[7]Academia das Ciências de Lisboa, Portugal
[8]University of Copenhagen, Denmark
[9]Centro de Estudios de la Real Academia Española, Spain
[10]Society for Danish Language and Literature, Copenhagen, Denmark
[11]Hungarian Research Centre for Linguistics, Institute for Lexicology, Hungary
[12]Hungarian Research Centre for Linguistics, Institute for Language Technologies and Applied Linguistics, Hungary [13]Instituut voor de Nederlandse Taal, The Netherlands [14]Istituto di Linguistica Computazionale "A. Zampolli", Centro Nazionale delle Ricerche, Italy
Email: federico.martelli@uniroma1.it, roberto.navigli@uniroma1.it, simon.krek@ijs.si, jelena.kallas@eki.ee, apolonija.gantar@ff.uni-lj.si, svetla@dcl.bas.bg, kaja.dobrovoljc@ijs.si, lipp.veronika@nytud.hu, varadi.tamas@nytud.hu, simon.laszlo@nytud.hu, gyorffy.andras@nytud.hu, valeria.quochi@ilc.cnr.it, monica.monachini@ilc.cnr.it, francesca.frontini@ilc.cnr.it, jaka.cibej@ijs.si, tina.munda@ijs.si, bspedersen@hum.ku.dk, saolsen@hum.ku.dk, margit.langemets@eki.ee, kristina.koppel@eki.ee, tiiu.yksik@eki.ee, rute.costa@fcsh.unl.pt, anasalgado@campus.fcsh.unl.pt, carole.tiberius@ivdnt.org, rob.tempelaars@ivdnt.org

Abstract

Over the course of the last few years, lexicography has witnessed the burgeoning of increasingly reliable automatic approaches supporting the creation of lexicographic resources such as dictionaries, lexical knowledge bases and annotated datasets. In fact, recent achievements in the field of Natural Language Processing and particularly in Word Sense Disambiguation have widely demonstrated their effectiveness not only for the creation of lexicographic resources, but also for enabling a deeper analysis of lexical-semantic data both within and across languages. Nevertheless, we argue that the potential derived from the connections between the two fields is far from exhausted. In this work, we address a serious limitation affecting both lexicography and Word Sense Disambiguation, i.e. the lack of high-quality sense-annotated data and describe our efforts aimed at constructing a novel entirely manually annotated parallel dataset in 10 European languages. For the purposes of the present paper, we concentrate on the annotation of morpho-syntactic features. Finally, unlike many of the currently available sense-annotated datasets, we will annotate semantically by using senses derived from high-quality lexicographic repositories.

Keywords: Digital lexicography; Natural Language Processing, Computational Linguistics, Corpus Linguistics; Word Sense Disambiguation.

1. Introduction

The fields of lexicography and Word Sense Disambiguation (WSD), i.e. the area of Natural Language Processing (NLP) concerned with identifying the meaning of a word in a given context (Bevilacqua et al., 2021), are increasingly interconnected. The reasons for this are manifold. On the one hand, since the so-called statistical revolution of the late 1980s, lexicography has benefited greatly from the development and constant refinement of automatic approaches for the lexical semantic analysis of vast amounts of textual data (Johnson, 2009). On the other hand, by its very nature WSD relies heavily on the availability of wide-coverage sense repositories such as monolingual and multilingual dictionaries or lexical knowledge bases (LKBs), e.g. WordNet[1] (Miller et al., 1990) and BabelNet[2] (Navigli & Ponzetto, 2012). Furthermore, modern lexicography and WSD are inextricably tied to corpora, i.e. large collections of written text in machine-readable form. Indeed, while lexicographers analyse corpora to identify and record relevant linguistic phenomena for the purpose of creating and updating dictionaries, WSD exploits corpora in multiple ways, such as learning effective unsupervised dense representations (Devlin et al., 2019; Conneau et al., 2020), or producing training and test data to be used in supervised approaches (Vial et al., 2019; Huang et al., 2019; Bevilacqua & Navigli, 2020; Blevins & Zettlemoyer, 2020; Conia & Navigli, 2021) by annotating them in a manual, semi-automatic or fully-automatic fashion.

Interestingly, both fields suffer from the paucity of standardised manual sense-annotated data in different languages, especially low-resource ones. In fact, the majority of high-quality sense-annotated datasets focus primarily on English. This is the case, for example, with SemCor (Miller et al., 1993) and those datasets introduced in the context of the Senseval and SemEval competitions (Edmonds & Cotton, 2001; Snyder & Palmer, 2004; Pradhan et al., 2007; Navigli et al., 2007). The few exceptions (Agirre et al., 2010; Navigli et al., 2013; Moro & Navigli, 2015) included just a limited number of instances in languages other than English. To cope with this shortcoming, several attempts have been made to bootstrap multilingual sense-annotated datasets. Pasini & Navigli (2017); Scarlini et al. (2019); Barba et al. (2020); Procopio et al. (2021) all addressed the lack of sense-annotated data in languages other than English via cross-lingual label propagation. Recently, Pasini et al. (2021) proposed XL-WSD, a large-scale multilingual evaluation framework for WSD, extending the Senseval and SemEval datasets using an automatic approach. However, despite the efforts undertaken, existing datasets are either not entirely manually curated, or they lack coverage in terms of languages, domains and parts of speech, or they rely on outdated sense inventories, which severely hampers their effectiveness.

In order to successfully address the aforementioned limitations, we have initiated the creation of a novel, manually-curated dataset, which will feature five annotation layers, i.e. tokenisation, sub-tokenisation, lemmatisation, POS tagging and Word Sense Disambiguation. The dataset will be available in 10 European languages, namely Bulgarian, Danish, Dutch, English, Estonian, Hungarian, Italian, Portuguese, Slovene and Spanish. Importantly, in contrast to existing manually annotated datasets, we will annotate our dataset using high-quality sense inventories. This will enable the highest possible number of sense instances to be covered. Moreover, we will also link the annotated instances to a multilingual sense repository, namely, BabelNet, so as to allow WSD systems to use our dataset as a challenging new evaluation benchmark. In what follows, we first describe how we constructed the dataset; next, we illustrate the annotation process focusing on the first four annotation layers, and finally we describe the sense inventories which we will use to semantically annotate our dataset.

[1] https://wordnet.princeton.edu/

[2] https://babelnet.org/

Language	Tokens	Unique Lemmas	Nouns	Verbs	Adjs	Advs
Bulgarian	33,994	6,683	7,892	3,970	3,313	1,157
Danish	32,524	6,832	7,322	3,099	2,626	1,677
Dutch	34,923	6,488	7,142	3,004	2,833	1,020
English	34,228	6,297	6,716	2,946	2,818	1,079
Estonian	37,693	6,112	8,189	3,327	2,310	1,487
Hungarian	29,657	7,457	6,930	2,485	3,561	1,173
Italian	39,067	6,371	7,864	3,022	2,961	1,368
Portuguese	38,723	6,260	7,372	3,181	2,757	1,302
Slovene	31,237	6,688	7,550	2,579	3,820	1,077
Spanish	37,693	6,112	8,189	2,806	3,141	1,140

Table 1: Number of tokens, unique lemmas and open-class parts of speech.

2. Construction of the dataset

In this section, we illustrate the construction of our dataset. This process was divided into two steps: i) the automatic extraction of candidate sentences from a wide-coverage parallel corpus, and ii) the manual validation of sentences to be included in our dataset, according to specific linguistic criteria. In what follows, we detail the two phases.

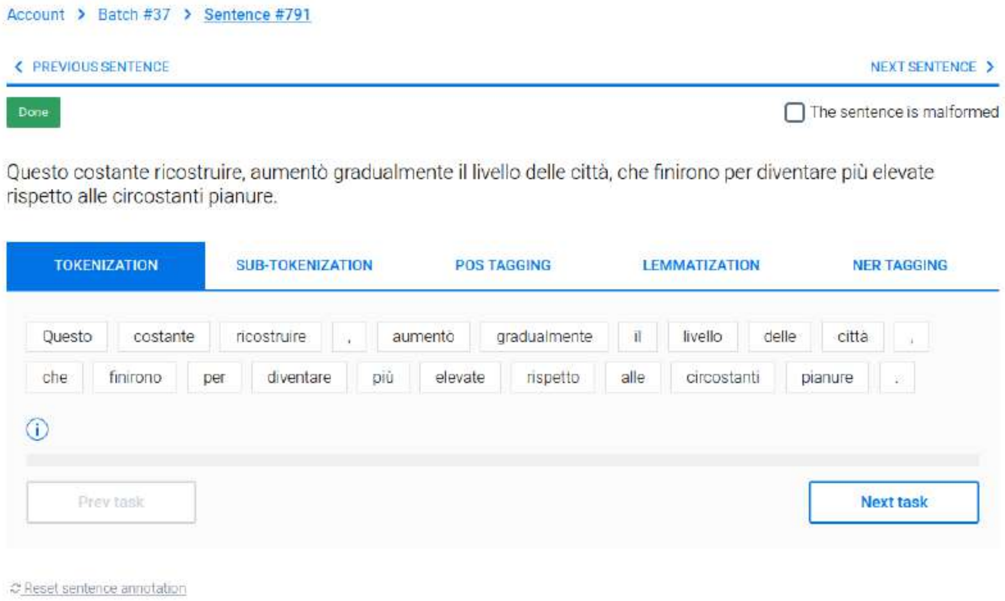
2.1 Automatic extraction of candidate sentences

First, we automatically extracted a set of sentences from WikiMatrix[3] (Schwenk et al., 2019), a large open-access collection of parallel sentences derived from Wikipedia using an automatic approach based on multilingual sentence embeddings. WikiMatrix covers 85 languages and includes 135 million parallel sentences for 1,620 language pairs, out of which 34 million are aligned with English. The corpus is divided into different files, each containing sentence pairs in a specific language combination. We performed our extraction in the following steps: i) we considered only language combinations such that the first language was English and the second was one of the other target languages; ii) we computed an overlap matrix which, given an English sentence s_e, showed the number of the selected WikiMatrix datasets which contained s_e and its corresponding translation into a target language; ii) we extracted the first 2,500 English sentences with the highest overlap across all language combinations.

2.2 Manual validation of parallel sentences

After completion of the first step, we manually validated the automatically extracted sentences according to specific formal and lexical-semantic criteria. In particular, we removed incorrect punctuation, morphological errors, notes in square brackets and etymological information typically provided in Wikipedia pages. Furthermore, in an effort to obtain a satisfying semantic coverage, we filtered out sentences which did not contain

[3] https://ai.facebook.com/blog/wikimatrix/



Questo costante ricostruire, aumentò gradualmente il livello delle città, che finirono per diventare più elevate rispetto alle circostanti pianure.

Questo | costante | ricostruire | , | aumentò | gradualmente | il | livello | delle | città | , | che | finirono | per | diventare | più | elevate | rispetto | alle | circostanti | pianure | .

Figure 1: Annotation interface used for the morpho-syntactic layers (the NER-tagging annotation was not performed at this stage).

at least 5 words, out of which at least two were polysemous. Subsequently, in order to obtain datasets in the other nine target languages, for each selected sentence in English we retrieved the corresponding WikiMatrix translation into each of the other languages. If no translation was available, we translated the English sentence manually. After the translation process, we reviewed the final dataset automatically and manually. As a result, we obtained a dataset composed of 2024 sentences for each target language.

3. Annotation

In this section, we describe the annotation process and highlight some significant linguistic peculiarities impacting on the annotation. We divided the annotation process into two phases. In the first, we focused on tokenisation, sub-tokenisation, lemmatisation and POS tagging. In the second phase we will annotate our dataset with senses derived from the specified inventories. In this paper, we will concentrate on the first phase only. In order to carry out our annotation, we used the ad hoc interface illustrated in Figure 1, developed at Babelscape[4], a Sapienza University spinoff company. In order to minimise the impact of subjectivity and ensure data consistency, we outlined specific criteria which we now detail. First, as a general guideline, we decided to follow the Universal Dependencies[5] (UD) standard for each language, so as to allow for a consistent annotation of lexical-semantic instances across languages. Importantly, we annotated both concepts and named entities. Furthermore, we normally included sub-tokenisation in almost all cases in which the token was composed of two or more distinct lemmas. As we shall see, sub-tokenisation was particularly challenging, especially when dealing with Germanic languages such as Dutch and Danish. Another challenge was posed by adjectival participles, which are derived from verbs but used as adjectives. In these cases, each annotator was required not only to consider the UD annotation standard, but also to thoroughly analyse the context in

[4] https://babelscape.com/
[5] https://universaldependencies.org/

which such instances occurred and their grammatical function in order to provide the correct tag. Table 1 reports the number of tokens, unique lemmas and the open-class part-of-speech distribution for each of the target languages.

In the following subsections, we focus on significant linguistic issues encountered during the annotation process, and provide reasons for our tagging choices.

3.1 Bulgarian

The simple and derived words, including the proper names, contractions, abbreviations and numerical expressions, were automatically annotated with the Bulgarian language processing chain (Koeva et al., 2020). This ensured the correct tokenisation, part-of-speech tagging and lemmatisation of homonymous verb particles, personal and possessive pronouns, derived numerals and proper names which were not present in the morphological dictionary. The main effort during the manual evaluation and correction was directed towards the re-annotation of multiword named entities as proper names. There are fixed multiword named entities which do not change either in terms of word order or grammar (*Yuzhna Amerika* 'South America') and semi-fixed multiword named entities which also have fixed word order but their constituents in Bulgarian can undergo certain paradigmatic changes within certain grammatical categories (for example, *Britanski muzey* 'British museum' – singular, indefinite, and *Britanskiya muzey* 'the British museum' – singular, definite). Some parts of the multiword names which can be used separately as common nouns had to be marked as proper nouns (for example, all constituents at the organization name *Evropeyski socialen fond* 'European Social Fund', etc.). The lemmas of semi-fixed multiword names in many cases were re-annotated because they differed from the lemmas of the corresponding simple words (for example, the lemmas of the words *ruskata* 'Russian' and *pravoslavna* 'orthodox' from the named entity *Ruskata pravoslavna carkva* 'Russian Orthodox Church' were changed from singular masculine to singular feminine).

3.2 Danish

The most prominent challenge in the Danish dataset was how to deal with compounds, which, as for most Germanic languages, are quite common and relatively dynamically generated, and more importantly: they are written as a single word. Our decision across all 10 languages was that conventionalized compounds found in the dictionary of the language should be kept as such, while compounds not found in the dictionary should be split into lemmas included in the dictionary, so as to enable them to be semantically tagged. For Danish we used the Danish Dictionary (DDO). When splitting compounds with a binding element, e.g. 's' in *helbredsanliggender* (health matters), we decided to keep the binding element during the subtokenisation and POS-tagging phase and to remove it in the lemmatisation phase. A further problem pertaining to compounds concerned the quite frequent phenomenon where two compounds that share a head are split and one head is left out, as in *certificeringsog revisionsmyndighed* (certification and audit authority). One possibility was to insert the head for both parts *certificeringsmyndighed og revisionsmyndighed* in the subtokenisation phase, but we decided that the head in the second part suffices for the disambiguation task and consequently we only annotated *certificerings-*'. The DDO was also used in the cases of participles used as adjectives.

Participles with adjective entries in the dictionary were annotated as such, e.g. *udstrakt* (Eng. outstretched, fig: extensive), while those that had only verb entries in the dictionary were annotated as verbs, e.g. *samlede* (Eng. lit: assembled, fig: total).

3.3 Dutch

Similarly to Danish, compounds also represented a specific challenge for Dutch. In this case, a compound was initially subtokenised if it did not occur in the Van Dale dictionary[6]. Later, this criterion was slightly relaxed and some other transparent compounds were also subtokenised, as we observed that a substantial number of compounds would not otherwise be found in the sense inventory. As in Danish, the binding element of compounds was kept in the subtokenisation phase, but removed in the lemmatisation one. Overall, 620 compounds were subtokenised in the Dutch dataset, mostly in two parts, but sometimes even in three or four parts (e.g. *laryngotracheobronchopneumonitis laryngo*; *tracheo*; *broncho* and *pneumonitis*).

An important subclass of compounds in Dutch is formed by the separable complex verbs. These are combinations of a verb and some other word. Examples are *aanvallen* ‘to attack’, *plaatsvinden* ‘to take place’. They sometimes behave as one word (*het kan plaatsvinden* ’it can take place’) and sometimes as two (*wanneer vindt het plaats?* ’when does it take place?’). Separable complex verbs are a known problem in corpus linguistics in Dutch and they presented another challenge for the annotation task. According to the UD guidelines, which are based on a lexicalist view of syntax, separable verbs should be annotated as separate words if they are written as separate words and the dependency relation should be used to identify them. After long discussions, it was decided to deviate from the UD guidelines and to consistently lemmatise separable complex verbs with the ‘complex’ lemma, regardless of whether the parts were separated or not. The latest version of the Alpino parser[7] also does this and lemmatises separable complex verbs with the ‘complex’ form, including an underscore to mark that it can occur as one word or as two, e.g. *aan_vallen.* The advantage of lemmatising with the complex verb is that the whole verb will be automatically looked up in the semantic annotation phase. This is important, as the meaning of separable complex verbs is not always compositional. Moreover, in some instances the parts of a separable complex verb are not even existing Dutch lemmas, as in the case of *aanmoedigen* ‘encourage’, which can be split into *aan* and *moedigen*, but where *moedigen* cannot occur on its own.

3.4 English

In the annotation of the English dataset, the scarce English-specific UD guidelines were complemented with querying the two largest manually annotated English UD treebanks – EWT (Silveira et al., 2014) and GUM (Zeldes, 2017), especially for resolving lexicon-based linguistic issues. Among others, these included the tokenisation of compounds (e.g. *cease-fire*), lemmatization of group names (e.g. *Muslims*), classification of determiner-like words (e.g. *its*), and the classification of various types of verb particle (e.g. *speed up*). Where there were discrepancies between the two treebanks, which was often the case with

[6] https://zoeken.vandale.nl/

[7] The Dutch UD corpora consist of data annotation with the Alpino annotation tools and guidelines, but do not yet include this. https://github.com/rug-compling/alpino

the under-specified lemmatisation layer, specific guidelines were drafted to consolidate the annotation of various phenomena, such as demonyms (e.g. lemma *American* of the form *American*), inflected adjectives (e.g. lemma *low* of the form *lower*) and personal pronouns (e.g. lemma *they* of the form *them*). In accordance with the general ELEXIS guidelines and the reference English treebanks, the constituents of multi-word named entities were annotated as PROPN regardless of their original POS class, with function words as an exception (e.g. *United.*PROPN *States.*PROPN *of.*ADP *America.*PROPN).

3.5 Estonian

The manual validation of the tokenisation, lemmatisation and POS tagging of the Estonian dataset generally followed the Estonian-specific UD annotation guidelines. Estonian uses 16 universal POS categories (all UD categories except PART). Regarding lemmatisation and POS tagging we relied also on the EKI Combined Dictionary[8], the biggest lexicographic database for modern Estonian compiled in the Institute of the Estonian Language. In the tokenisation phase manual correction was necessary in the case of nonconventionalised compounds (e.g. *puuja juurviljad* (fruits and vegetables)), conventionalised compounds were left as one token. For words with splitting element *Shakespeare'i* (Shakespeare's) we kept splitting elements during the subcategorisation, but removed it in the lemmatisation phase.

The most problematic was POS tagging, since Estonian UD POS tags are very different from other morphological annotators developed for Estonian (e.g. estNLTK)[9], and also from POS nomenclature used in the EKI Combined Dictionary. UD-specific parts of speech are AUX and DET. Conjunctions are also split into CCONJ and SCONJ. On the other hand, the degrees of comparison of adjectives are analysed as ADJ, while it is common for Estonian to analyse positive, comparative and superlative degrees as separate parts of speech.

According to UD annotation lemmas *olema* (to be), *ei*, *ära* (not), and modal verbs were annotated as AUX. Participles used predicatively were annotated as verbs; participles used attributively were annotated as adjectives. Abbreviations for single words were assigned the part of speech of the full form. Acronyms for proper names such as NATO were tagged as proper nouns. Foreign words were annotated as X.

3.6 Hungarian

With regard to lemmatisation and POS tagging in general we relied on the Hungarian UD guidelines[10,11] and the Magyar értelmező kéziszótár (ÉKsz. 2002) *Concise Explanatory Dictionary of Hungarian.* Regarding tokenisation, we followed the *Rules of Hungarian Orthography*, 12th edition (2015). We had to deal with the following problems in the manual correction of the result of the UD-based automatic annotation process (tokenisation, lemmatisation, POS tagging) in the Hungarian texts. First of all, the Hungarian UD POS-system is very different from the standard Hungarian POS-system

[8] http://sonaveeb.ee
[9] https://github.com/estnltk/estnltk/tree/version_1.6
[10] https://universaldependencies.org/treebanks/hu_szeged/index.html
[11] https://github.com/dlt-rilmta/panmorph

that is represented in the main explanatory dictionaries. This made the correction of the automatic POS-tagging difficult. Specific problems arose because of the lack of such categories in the UD POS-system as *igekötő* (particles or prefixes linked to verbs) and *igenév* (participles, adverbial participles and infinitives). In our explanatory dictionaries, words in the *igenév* POS-category are processed under the VERB lemmas, from which they are formed. For example, in this sentence: *A bolygót meglátogató két űreszköz...* ('The first of two spacecraft to visit the planet...'), *meglátogató* is a particle formed from the verb *meglátogat* ('to visit'). In the dictionary, there is no such lemma as *meglátogató* (because we can form participles from almost every verb). However, in the sentence this word behaves like an ADJ (attributive role), thus we have to tag it as an ADJ according to the general guidelines, even if there is no *meglátogató* ADJ in Hungarian. On the other hand, the UD POS category determiner is missing from the standard Hungarian POS-categories. (Instead, we use other categories like PRON (*egyik* ('one (of the)'), NUM (*sok* 'many')). In the annotation, we kept the DET category only for articles (*a*, *az* 'the', *egy* 'a, an'). Besides this, under the UD POS tag ADP (adposition), in Hungarian we only have postpositions. In addition, the POS tag PART is applied for only two words: *meg* and *utol.* Regarding tokenisation and lemmatisation, we had to deal with general, non-language specific problems: the multi-word proper names had to be analysed as a whole, despite the fact that they might have contained common noun elements, too. Another problematic case was presented by the ellipsis in complex compounds, in which the lemmatisation depends on whether we take the missing words into consideration, or not. In Hungarian, an agglutinative language, we also had such problems as suffixes attached to symbols (e.g. %-*át*), and suffixes after quotation marks (e.g. "xyz"-*t*). Also, the orthographic rules influence tokenisation: *Schmidt-távcső* ('Schmidt telescope') is a compound lemma (one token), a NOUN, thus, the PROPN-element is "lost" in it. (Analogy: *Kossuth-szakáll* 'a type of beard which was made famous by Lajos Kossuth' – not a PROPN).

3.7 Italian

The manual correction/validation of the morphosyntactic layers of the Italian dataset generally followed the Italian-specific UD annotation guidelines[12] and the praxis established in the Italian treebanks[13]. When clashes with project-level indications arose, it was decided to adhere to the UD praxis, with a few exceptions as follows: a) abbreviations are treated as single words that may contain punctuation (e.g. *U.S.A.*, *UE*) except when they indicate units of measure, in this case they are annotated as SYM as in the rest of the datasets; and b) foreign words are annotated as X in titles and long expressions (i.e. when they are incidentals). As for POS annotation, base infinitives used as nouns and participles used predicatively are annotated as verbs, even when the subject is implied; participles used attributively are annotated as adjectives instead. Possessive adjectives are always tagged as determiners, while predeterminers and quantifiers are tagged as such if no other determiner is present, adjective otherwise. As for POS-tags, it is worth noting that AUX is also used for copulas, so that the verb *essere* "to be" is almost always an AUX. Subtokenisation in Italian UD is required in the following cases: 1) complex prepositions (i.e. combined/fused with the definite article, e.g. *nella* "in the.fem", *del* "of the.masc"); and 2) verbal forms with enclitic pronouns (e.g. *dammelo* "give-to me-it", *mangiandolo*

[12] https://universaldependencies.org/it/index.html

[13] i.e. https://universaldependencies.org/treebanks/it_isdt/index.html, https://universaldependencies.org/treebanks/it_partut/index.html

"eating-it"). Given that they are quite frequent in training data, manual correction was not often required for these aspects. As for lemmatisation, articles and pronouns were lemmatised with their base form (i.e. singular masculine); adjectives with the positive, singular, masculine forms, except for irregular comparative and superlative forms.

Finally, regarding the incidence of manual corrections needed, lemmatisation required a considerable effort, as the automatic lemmas assigned were often wrong, especially for homographs and irregular and infrequent words.

3.8 Portuguese

One of the major challenges in annotating the Portuguese dataset was presented by lemmatisation in a dictionary that did not always abide by the same annotation criteria applied to corpora. We decided to always annotate the lemma as being the canonical form during the lemmatisation process, ignoring some of the lexical items identified that occur as a headword in a dictionary. For instance, the personal pronoun *ela* (she), the Portuguese feminine form of *ele* (he), is registered as an entry in the Portuguese dictionary, and the recorded lemma is the canonical form *ele*. The option we decided on guarantees better data consistency and coherency. In dictionaries, cases of this type often turn out to be cross-referenced to the canonical form, e.g., the definite article *a* [the Portuguese feminine form of 'the'] is a cross-reference to *o* [the Portuguese masculine form of 'the'], which strengthens our decision.

Another decision we took concerned the forms corresponding to degrees of adjectives and adverbs. Although in the Lisbon Academy of Sciences dictionary we find comparative and superlative forms as headwords, e.g., *pior* (worse; worst), we considered the positive form as a lemma according to Universal Dependencies recommendations. Generally, for the part-of-speech tagging, we used the Universal Dependencies (UD) in its current version 2.7 (Zeman et al., 2020). Nevertheless, we did not adhere to the UD criterion for abbreviations. Lexical items such as *km* (kilometre) and *m* (meter) were tagged as abbreviations as previously agreed by all ELEXIS team members, rather than as nouns, as UD suggests. It is important to note that we labelled some past participles as adjectives rather than as verbs when they served an adjectival function in the analysed sentences.

As for the subtokenisation, contractions were broken into smaller units, for example, *da* (*de* + *a*) [preposition *de* (of) + the feminine article form *a* (the)]. However, in the case of *desde* (since), which is a contracted form (< prep. Latin *de* + *ex*), we preferred instead to keep it as a preposition, as recognised by Portuguese grammar and dictionaries.

3.9 Slovene

The Slovene dataset was automatically tokenised, lemmatised and tagged with the CLASS LA tagger (Ljubešić & Dobrovoljc, 2019), which was developed for processing South Slavic languages. The tagger proved to be a highly accurate tool, although some corrections were needed.

For Slovene, two POS tagsets are generally used, the default JOS (Erjavec et al., 2010)/ Multext-East system (Erjavec, 2017), and UD (Dobrovoljc et al., 2017). Taggers usually struggle with two major differences between the systems. One difference lies in the distinction between the categories AUX and VERB in case of the omnipresent verb *biti* ('to be'). In the UD system, the AUX category is assigned when 'to be' is used as an auxiliary or a linking verb (e.g. *Večina prebivalcev je* AUX *katolikov.* The majority of the population are AUX Catholic.), and the category VERB when it is used as a lexical verb (*Njihov glavni štab je* VERB *v Tel Avivu.* Their headquarters are VERB in Tel Aviv.). In real life, the distinction between these is not always clear-cut; however, to solve the dilemmas, we consulted the detailed UD-POS tagging guidelines for Slovene (*ibid.*).

The other major difference is the use of categories DET vs PRON. In UD, the DET category is assigned to pronouns when used as modifiers in nominal (or other) phrases, and PRON when they are used as heads. Other notable issues include the use of CCONJ vs ADV (*Ali* ADV *so te razlike neposredni vzrok za debelost ali* CCONJ *pa njena posledica, je še odprto vprašanje.* Whether CCONJ these differences are the direct cause or CCONJ the result of obesity has yet to be determined unequivocally.); ADP vs ADV (*Sklepali so, da je okoli* ADP *Urana sistem obročev.* They concluded that there must be a ring system around ADP Uranus. vs *V naravi povprečno živi okoli* ADV *20 let.* The life expectancy in the wild is approximately ADV 20 years.).

In order to obtain as many content words as possible, such words being the only ones considered in the lexical-annotation phase, components of named entities that were not proper nouns were assigned the part of speech they belong to in their simple, common sense (e.g. *Evropska* ADJ *unija* NOUN; European ADJ Union NOUN). This decision is in line with the Slovene UD guidelines, but contrary to the practice of most of the project participants. As for lemmatisation, the preposition *s* ('with') was oftentimes automatically lemmatised as *biti* ('to be'), and prepositions, when occupying the first place in a sentence, were lemmatised with the capital letter, all of which was manually corrected. There were no errors in tokenisation.

3.10 Spanish

The main revision points on the annotated Spanish dataset were the lemmatisation of infrequent or rare words, verb infinitive lemmas with adjectival tags and non-toponym non-anthroponym PROPN sequences readjusting into (chiefly postmodified) common noun phrases, e.g. *Tribunal Supremo* "Supreme Court". Lemmatisation followed the standard practice of Spanish linguistics: infinitive for verbs and masculine singular form for other inflectional elements (N, ADJ, PRON, DET, but not for PART), even when some of their forms were dictionary entries, usually for alphabetical reasons or retrievability purposes.

Some functional tags also needed correction in traditional fuzzy zones of Spanish syntax such as NOUN-ADJ triplets and DET-PRON subsystems, correlative structures (e.g. *tan... como "as... as"*) and, very occasionally, complement-relative clausal misanalyses of *que* "who, that".

Tokenisation followed UD guidelines and the only subtokenised elements were verbal forms with oblique pronouns like *matarlo* "kill him". As a rule, general complex elements such as multiple verbs (compound tenses, aspectual or catenative structures and the like), comitatives and the only two amalgamated ADP-DET remnants in Spanish (*al, del*) are kept exactly as tokenised – the former split and the latter two groups not subtokenised.

Language	**Resource**
Bulgarian	Dictionary of Modern Bulgarian
Danish	DanNet (The Danish WordNet)
Dutch	Open Dutch WordNet
English	English WordNet
Estonian	EKI Combined Dictionary
Hungarian	The Explanatory Dictionary of the Hungarian Language
Italian	PAROLE-SIMPLE-CLIPS + ItalWordNet
Portuguese	Dictionary of the Lisbon Academy of Sciences
Slovene	sloWNet
Spanish	Spanish Wiktionary

Table 2: Sense inventories

4. Sense inventories

We now describe the sense inventories which we will use to annotate our dataset semantically, as shown in Table 2. Importantly, during the semantic annotation validators will be able to improve the coverage and quality of the specified sense inventories, for instance, by adding new entries or improving already existing ones.

4.1 Bulgarian

The Dictionary of Modern Bulgarian (DMB, *Rechnik na savremenniya balgarski knizhoven ezik*) was published in three volumes between 1955 and 1959 by the Bulgarian Academy of Sciences. In addition to the general vocabulary the dictionary includes some obsolete words, words gradually moving into the passive vocabulary, and foreign words which are widely used in modern Bulgarian. Each entry is structured in a specific way according to the part of speech of the headword and it represents the major senses accompanied with quotations. The headword is followed by a forms section, a grammar section, a stylistic section and an etymology (where relevant). An entry may also include compounds, phrases, and derivatives (secondary lemmas) based on the headword. Today, the dictionary is in the process of its first major revision. The update is revising and extending the DMB, adjusting the vocabulary to cover the missing senses from the ELEXIS *multilingual parallel sense-annotated dataset*, to label some senses as obsolete, to include some new borrowings in the language, and to replace the obsolete quotations. As of March 2021 the dictionary covers 60,744 headwords, 68,387 lemmas and secondary lemmas, 78,569 sense definitions and 80,520 quotations coming mainly from classic literature and periodicals.

4.2 Danish

The Danish sense inventory applied for the annotation task consists of by the Danish wordnet, DanNet (Pedersen et al., 2009). DanNet currently contains 70,000 synsets corresponding roughly to the same number of word senses, covering nouns, verbs and adjectives. The wordnet follows the Princeton WordNet standard, but is compiled semi-automatically from a Danish source, namely The Danish Dictionary (DDO), and linked to the senses in the dictionary (Pedersen et al., 2009). Approx 10,000 of the synsets are also linked to Princeton WordNet (Pedersen et al., 2019). DanNet is currently being extended to cover a broader number of word senses (Nimb et al., 2021), still in essence relying on the sense inventory of DDO as a basis, but aiming towards partly clustering very subtle meaning distinctions inherited from the source (Pedersen et al., 2018). DanNet was chosen for the annotation task first of all because it allows us to publish the annotation sense inventory as open source, but also because we want to test the lexical coverage as well as the operability of the wordnet for such a task. Based on the feedback and results, missing lemmas and senses will subsequently be added to the wordnet and further integrated into a future Danish language resource for AI purposes to be developed in COR (The Central Word Register for Danish)[14], a collaborative project between the Society for Danish Language and Literature, the Danish Language Council, Centre for Language Technology at UCPH and the Danish Agency for Digitisation.

4.3 Dutch

Open Dutch WordNet is a Dutch lexical semantic database. It was created by removing the proprietary content from Cornetto[15]. A large portion of the Cornetto database originated from the commercial publisher Van Dale[16] preventing it from being distributed as open source. In order to create Open Dutch WordNet, all the synsets and relations from WordNet 3.0 were used as a basis and existing equivalence relations between Cornetto synsets and WordNet synsets were exploited in order to replace WordNet synonyms by Dutch synonyms. Concepts that were not matched through hyperonym relations to the WordNet hierarchy were added, as well as manually created semantic relations from Cornetto. The synonyms, concepts and relations were limited to those on which there were no copyright claims. In addition, the inter-language links in various external resources were used to add synonyms to the resource (Postma et al., 2016).

4.4 English

The English WordNet[17] is an open-source derivation from Princeton WordNet (Miller, 1995), a widely used lexical network of the English language grouping words into synsets and linking them according to different semantic relations between them. In its second release, the English WordNet 2020 (McCrae et al., 2020) introduced a substantial number of changes compared to the original database, including the integration of contributions from other projects, such as Colloquial WordNet (McCrae et al., 2017), enWordNet (Rudnicka et al., 2015) and Open Multilingual WordNet (Bond & Paik, 2012). This resulted in the introduction of several new manually-validated synsets (120,054 in total), lemmas (163,079), senses (211,864) and definitions (120,059), as well as the development of clear guidelines for future community-driven additions to the database, which is planned to be released annually.

[14] https://cst.ku.dk/english/projects/the-central-word-register-for-danish-cor/
[15] http://www2.let.vu.nl/oz/cltl/cornetto
[16] https://www.vandale.nl/
[17] https://en-word.net/

4.5 Estonian

EKI Combined Dictionary[18] is the biggest lexicographic database of modern Estonian compiled in the Institute of the Estonian Language. The current description of Estonian headwords in Ekilex includes definitions, semantic types, parts of speech, inflected forms, collocations, government patterns, semantic relations, related words, etymology, usage examples, and translations. As of April 2021, Ekilex contains about 160,000 words and phrases in Estonian. For this task's development, a total of 7,044 Estonian lemmas and 14,870 senses were extracted from Ekilex. Ekilex allows the annotation sense inventory to be published as open source.

4.6 Hungarian

The Explanatory Dictionary of the Hungarian Language (*A magyar nyelv értelmező szótára*, abbr. ÉrtSz.) was compiled in the Research Institute for Linguistics, Hungarian Academy of Sciences in seven volumes between 1959 and 1962. ÉrtSz. covers Hungarian literary language of the 19th century, as well as the written and spoken standard Hungarian of the first half of the 20th century, with a total of 60,000 entries. The main source of input was a corpus of about six million examples collected since the end of the 19th century. Entries included pronunciation (where it differed from what could be expected on the basis of spelling) and an aid to the hyphenation of compound words. Each sense unit is illustrated by a few examples: citations from the classical Hungarian literature and example sentences created by the lexicographers. In terms of the fine sense discrimination and sophisticated sense definitions, it stands out from the genre of a desk dictionary and is closer in its ambitions to unabridged dictionaries, particularly as regards the treatment of function words and detailed treatment of verb senses. This is one of the best used dictionaries from a professional point of view, but its vocabulary and the examples are old-fashioned.

4.7 Italian

The Italian Sense Inventory was produced by combining two existing openly available lexical resources, namely PAROLE SIMPLE CLIPS (PSC)[19] and ItalWordNet (IWN)[20]. PSC, developed within two subsequent European projects PAROLE and SIMPLE, is a large lexical database for the Italian language. In the semantic layer, the main basic blocks are semantic units, Usems, which are provided with definitions and examples, and linked to the SIMPLE Ontology and also to other Usems through a rich set of semantic relations (Bel et al., 2000). ItalWordNet (IWN) is a lexical semantic database for the Italian language started within the context of the EuroWordNet project and then subsequently enlarged and refined within national projects until 2012. It is mapped and linked to the Princeton WordNet – thus also indirectly, to BabelNet – and is also available in the Open Multilingual Wordnet format (Quochi et al., to appear). The two resources have been partially aligned, so that a subset of IWN synsets are linked to PSC corresponding Usems. In order to produce the current sense inventory, the two resources were queried for all the target lemmas present in the Italian dataset and a list of corresponding Usems

[18] http://sonaveeb.ee
[19] http://hdl.handle.net/20.500.11752/ILC-88
[20] http://hdl.handle.net/20.500.11752/ILC-62

from PSC and IWN synsets were retrieved together with their definitions, examples and original IDs. Where a mapping between the two resources was available, a unique sense was produced, merging the two definitions into a single one. The resulting sense inventory contains 4,424 lemmas for a total number of 11,298 senses.

4.8 Portuguese

The *Dicionário da Língua Portuguesa* (DLP) is a scholarly dictionary of the Portuguese language being developed by the Lisbon Academy of Sciences. DLP is a retro-digitised dictionary created by converting the *Dicionário da Lingua Portuguesa Contemporânea*, last published in 2001. Currently, the DLP is being prepared under the supervision of the Instituto de Lexicologia e Lexicografia da Língua Portuguesa (ILLLP) in collaboration with researchers and invited collaborators. Between 2015 and 2016, some preparatory work for the Portuguese Academy digital dictionary was performed through the ILLLP, and a database was developed by a team working in NLP at the University of Minho (Simões et al., 2016), which now includes IPCA and NOVA CLUNL (Salgado et al., 2019). This project is supported by a Community Support Fund – Fundo de Apoio à Comunidade (FAC) – by the Portuguese National Funding through the FCT – Fundação para a Ciência e Tecnologia. For the development of this task, a total of 4,031 lemmas (lemma, part of speech, and definitions) were extracted from DLP.

4.9 Slovene

As a Slovene sense inventory, we used the current version of the Slovene wordnet – sloWNet 3.1 (Fišer & Sagot, 2015). This is an open-source lexical database containing the complete Princeton WordNet 3.0 and 71,803 Slovene literals, 33,546 of which were manually validated. The literals were inserted automatically from several existing language resources, comprising two bilingual dictionaries, a few domain-specific resources, parallel corpora, as well as Wikipedia. The 4,919 content word lemmas appearing in the dataset were validated and corrected, if necessary, during the WSD annotation process.

4.10 Spanish

To come up with a freely distributable dataset, the Spanish lexical fragment of Wiktionary[21] was chosen to tag Spanish texts. Wiktionary is a multilingual free dictionary, being written collaboratively on the web. A dump as of late 2020 was filtered to sort out non-semantic information (etymology, morphology, pronunciation, etc.) and about 92,000 lemmas with more than 140,000 senses were kept. Wiktionary has been shown (Ahmadi et al., 2020) to exhibit a great deal of overlap with the reference Spanish dictionary (Real Academia Española & Asociación de Academias de la Lengua Española, 2014), so standard coverage is envisaged.

5. Conclusions

In this work, we addressed a major shortcoming affecting both lexicography and Word Sense Disambiguation, namely the paucity of manual sense-annotated data. We

[21] https://es.wiktionary.org/wiki/Wikcionario:Portada

described how we plan to design a novel manually curated dataset available in 10 European languages, i.e. Bulgarian, Danish, Dutch, English, Estonian, Hungarian, Italian, Portuguese, Slovene and Spanish, focusing on the morpho-syntactic annotation layers. We have now finalised the annotation of the morpho-syntactic layers and, as next step, we will annotate our dataset with senses derived from the aforementioned high-quality sense inventories. We argue that, thanks to our dataset, both scientific communities will be provided with a very effective resource which, on the one hand, will enable lexicographic phenomena to be investigated both within and across languages, and on the other hand, can be used as a new evaluation benchmark for WSD systems.

Acknowledgments

The authors gratefully acknowledge the support of the ELEXIS project No. 731015 under the European Union's Horizon 2020 research and innovation programme.

6. References

Agirre, E., De Lacalle, O.L., Fellbaum, C., Hsieh, S.K., Tesconi, M., Monachini, M., Vossen, P. & Segers, R. (2010). SemEval-2010 task 17: All-words word sense disambiguation on a specific domain. In *Proceedings of the 5th international workshop on semantic evaluation*. pp. 75–80.

Ahmadi, S., McCrae, J.P., Nimb, S., Khan, F., Monachini, M., Pedersen, B.S., Declerck, T., Wissik, T., Bellandi, A., Pisani, I., Troelsgárd, T., Olsen, S., Krek, S., Lipp, V., Váradi, T., Simon, L., Gyorffy, A., Tiberius, C., Schoonheim, T., Moshe, Y.B., Rudich, M., Ahmad, R.A., Lonke, D., Kovalenko, K., Langemets, M., Kallas, J., Dereza, O., Fransen, T., Cillessen, D., Lindemann, D., Alonso, M., Salgado, A., Sancho, J., Ureña-Ruiz, R., Zamorano, J.P., Simov, K., Osenova, P., Kancheva, Z., Radev, I., Stankovic, R., Perdih, A. & Gabrovsek, D. (2020). A Multilingual Evaluation Dataset for Monolingual Word Sense Alignment. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk & S. Piperidis (eds.) *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*. European Language Resources Association, pp. 3232–3242. URL https://www.aclweb.org/anthology/2020.lrec-1.395/.

Barba, E., Procopio, L., Campolungo, N., Pasini, T. & Navigli, R. (2020). MuLaN: Multilingual Label propagatioN for word sense disambiguation. In *Proc. of IJCAI*. pp. 3837–3844.

Bel, N., Busa, F., Calzolari, N., Gola, E., Lenci, A., Monachini, M., Ogonowski, A., Peters, I., Peters, W., Ruimy, N., Villegas, M. & Zampolli, A. (2000). SIMPLE: A General Framework for the Development of Multilingual Lexicons. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*. Athens, Greece: European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2000/pdf/61.pdf.

Bevilacqua, M. & Navigli, R. (2020). Breaking through the 80% glass ceiling: Raising the state of the art in Word Sense Disambiguation by incorporating knowledge graph information. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. pp. 2854–2864.

Bevilacqua, M., Pasini, T., Raganato, A. & Navigli, R. (2021). Recent Trends in Word Sense Disambiguation: A Survey. In *Proc. of IJCAI*.

Blevins, T. & Zettlemoyer, L. (2020). Moving Down the Long Tail of Word Sense Disambiguation with Gloss Informed Bi-encoders. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. pp. 1006–1017.

Bond, F. & Paik, K. (2012). A survey of wordnets and their licenses. *Small*, 8(4), p. 5.

Conia, S. & Navigli, R. (2021). Framing Word Sense Disambiguation as a Multi-Label Problem for Model-Agnostic Knowledge Integration. In *Proceedings of the EACL*.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, É., Ott, M., Zettlemoyer, L. & Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. pp. 8440–8451.

Devlin, J., Chang, M., Lee, K. & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In J. Burstein, C. Doran & T. Solorio (eds.) *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, pp. 4171–4186. URL https://doi.org/10.18653/v1/n19-1423.

Dobrovoljc, K., Erjavec, T. & Krek, S. (2017). The universal dependencies treebank for Slovenian. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*. pp. 33–38.

Edmonds, P. & Cotton, S. (2001). Senseval-2: overview. In *Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*. pp. 1–5.

Erjavec, T. (2017). MULTEXT-East. In *Handbook of Linguistic Annotation*. Springer, pp. 441–462.

Erjavec, T., Fiser, D., Krek, S. & Ledinek, N. (2010). The JOS Linguistically Tagged Corpus of Slovene. In *LREC*. Citeseer.

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.

Huang, L., Sun, C., Qiu, X. & Huang, X.J. (2019). GlossBERT: BERT for Word Sense Disambiguation with Gloss Knowledge. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. pp. 3500–3505.

Johnson, M. (2009). How the statistical revolution changes (computational) linguistics. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?* pp. 3–11.

Koeva, S., Obreshkov, N. & Yalamov, M. (2020). Natural Language Processing Pipeline to Annotate Bulgarian Legislative Documents. In *Proceedings of the 12th Language Resources and Evaluation Conference, European Language Resources Association*. pp. 6988–6994.

Ljubešić, N. & Dobrovoljc, K. (2019). What does neural bring? analysing improvements in morphosyntactic annotation and lemmatisation of Slovenian, Croatian and Serbian. In *Proceedings of the 7th workshop on balto-slavic natural language processing*. pp. 29–34.

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

McCrae, J.P., Wood, I.D. & Hicks, A. (2017). The Colloquial WordNet: Extending Princeton WordNet with Neologisms. In *LDK*.

Miller, G.A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, 38(11), pp. 39–41.

Miller, G.A., Beckwith, R., Fellbaum, C.D., Gross, D. & Miller, K. (1990). Introduction to WordNet: an Online Lexical Database. *International Journal of Lexicography*, 3(4).

Miller, G.A., Leacock, C., Tengi, R. & Bunker, R.T. (1993). A semantic concordance. In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.

Moro, A. & Navigli, R. (2015). SemEval-2015 task 13: Multilingual all-words sense disambiguation and entity linking. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*. pp. 288–297.

Navigli, R., Jurgens, D. & Vannella, D. (2013). SemEval-2013 task 12: Multilingual word sense disambiguation. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. pp. 222–231.

Navigli, R., Litkowski, K.C. & Hargraves, O. (2007). SemEval-2007 task 07: Coarse-Grained English all-words task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*. pp. 30–35.

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.

Nimb, S., Pedersen, B.S. & Olsen, S. (2021). DanNet2: Extending the coverage of adjectives in DanNet based on thesaurus data. In *Proceedings of the 11th Global Wordnet Conference*. pp. 267–272.

Ogilvie, S. (2020). *The Cambridge Companion to English Dictionaries*. Cambridge University Press.

Pasini, T. & Navigli, R. (2017). Train-O-Matic: Large-scale supervised word sense disambiguation in multiple languages without manual training data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. pp. 78–88.

Pasini, T., Raganato, A. & Navigli, R. (2021). XL-WSD: An Extra-Large and Cross-Lingual Evaluation Framework for Word Sense Disambiguation. In *Proc. of AAAI*.

Pedersen, B.S., Aguirrezabal Zabaleta, M., Nimb, S., Olsen, S. & Rørmann, I. (2018). Towards a principled approach to sense clustering – a case study of wordnet and dictionary senses in Danish. In *Proceedings of Global WordNet Conference 2018 Singapore: Global WordNet Association*.

Pedersen, B.S., Nimb, S., Asmussen, J., Sørensen, N.H., Trap-Jensen, L. & Lorentzen, H. (2009). DanNet: the challenge of compiling a wordnet for Danish by reusing a monolingual dictionary. *Language Resources and Evaluation*, 43(3), pp. 269–299.

Pedersen, B.S., Nimb, S., Olsen, I.R. & Olsen, S. (2019). Linking DanNet with Princeton WordNet. In *Global WordNet 2019 Proceedings, Wroclaw, Poland*.

Postma, M., van Miltenburg, E., Segers, R., Schoen, A. & Vossen, P. (2016). Open Dutch WordNet. In *Proceedings of the Eight Global Wordnet Conference*. Bucharest, Romania.

Pradhan, S., Loper, E., Dligach, D. & Palmer, M. (2007). SemEval-2007 task-17: English lexical sample, srl and all words. In *Proceedings of the fourth international workshop on semantic evaluations (SemEval-2007)*. pp. 87–92.

Procopio, L., Barba, E., Martelli, F. & Navigli, R. (2021). MultiMirror: Neural Cross-lingual Word Alignment for Multilingual Word Sense Disambiguation. In *Proc. of IJCAI*. Online.

Quochi, V., Bartolini, R. & Monachini, M. (to appear). ItalWordNet goes open. *Special Issue on Linking, Integrating and Extending Wordnets. Linguistic Issues in LanguageTechnology. Linguistic Issues in Language Technology. LiLT*, 10(4).

Real Academia Española & Asociación de Academias de la Lengua Española (2014). *Diccionario de la lengua española*. Espasa Calpe, vigesimotercera edición edition.

Rudnicka, E.K., Witkowski, W. & Kaliński, M. (2015). Towards the methodology for extending Princeton WordNet. *Cognitive Studies| Études cognitives*, (15), pp. 335–351.

Salgado, A., Costa, R., Tasovac, T. & Simões, A. (2019). TEI Lex-0 In Action: Improving the Encoding of the Dictionary of the Academia das Ciências de Lisboa. In I. Kosem, T. Zingano Kuhn, M. Correia, J.P. Ferreira, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek & C. Tiberius (eds.) *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference*. pp. 417–433.

Scarlini, B., Pasini, T. & Navigli, R. (2019). Just "OneSeC" for producing multilingual sense-annotated data. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. pp. 699–709.

Schwenk, H., Chaudhary, V., Sun, S., Gong, H. & Guzmán, F. (2019). Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. *arXiv preprint arXiv:1907.05791*.

Silveira, N., Dozat, T., de Marneffe, M.C., Bowman, S., Connor, M., Bauer, J. & Manning, C.D. (2014). A Gold Standard Dependency Corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*.

Simões, A., Almeida, J.J. & Salgado, A. (2016). Building a Dictionary using XML Technology. In M. Mernik, J.P. Leal & H.G. Oliveira (eds.) *5th Symposium on Languages, Applications and Technologies (SLATE)*, volume 51 of *OASIcs*. Germany: Schloss Dagstuhl, pp. 14:1–14:8.

Snyder, B. & Palmer, M. (2004). The English all-words task. In *Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*. pp. 41–43.

Taghipour, K. & Ng, H.T. (2015). One million sense-tagged instances for word sense disambiguation and induction. In *Proceedings of the nineteenth conference on computational natural language learning*. pp. 338–344.

Vial, L., Lecouteux, B. & Schwab, D. (2019). Sense Vocabulary Compression through the Semantic Knowledge of WordNet for Neural Word Sense Disambiguation. In *Global Wordnet Conference*.

Yuan, D., Richardson, J., Doherty, R., Evans, C. & Altendorf, E. (2016). Semi-supervised Word Sense Disambiguation with Neural Models. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. pp. 1374–1385.

Zeldes, A. (2017). The GUM Corpus: Creating Multilayer Resources in the Classroom. *Language Resources and Evaluation*, 51(3), pp. 581–612.

Zeman et al. (2020). *Universal Dependencies 2.7*. URL http://hdl.handle.net/11234/1-3424. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Semi-automatic building of large-scale digital dictionaries

[‡]Marek Blahuš, [‡]Michal Cukr, [†‡]Ondřej Herman,
[†‡]Miloš Jakubíček, [†‡]Vojtěch Kovář, [†‡]Marek Medveď

[†]Natural Language Processing Centre
Faculty of Informatics, Masaryk University

[‡]Lexical Computing
{firstname.lastname}@sketchengine.eu

Abstract

This paper presents a novel way of creating dictionaries by using a particular post-editing workflow, all of which is carried out in the context of building a set of three bilingual dictionaries – Tagalog, Urdu and Lao dictionaries with translations into English and Korean. The dictionaries were created completely from scratch without reusing any existing content and in a completely automatic manner, amounting to 50,000 headwords, out of which 15,000 headwords were subject to subsequent manual post-editing.

In the paper we discuss the post-editing methodology that we used and its impact on the overall lexicographic workflow. We describe the web corpora that were built specifically for the purpose of building these three dictionaries as well as their annotations (such as PoS tagging and lemmatisation) and tools that were used for the corpus annotation and for automating individual entry parts and the post-editing thereof. Most of the automatic drafting and post-editing relied on a backbone consisting of the Sketch Engine corpus management system and Lexonomy dictionary editor

We also detail the overall amount of work involved in each post-editing step, the technical and managerial difficulties faced alongside in the project, and the major technological issues that still need improvement in the post-editing scenario.

Keywords: post-editing lexicography; dictionary drafting; Sketch Engine

1. Introduction

Contemporary lexicography is based on using large text corpora to reflect the real use of a language as much as possible. Sometimes the corpora are only used as an additional tool helping lexicographers compile the entries, while other projects use corpora very extensively, generating large parts of the entries automatically and then post-editing, or correcting them. One of the most advanced procedures in the latter direction is the "Million-click dictionary" (MCD) method described in (Baisa et al., 2019) and (Jakubíček et al., 2020).

This paper reports on three related dictionary projects compiled using the MCD method, which are currently completed and signed-off. Looking back at the projects, we discuss the strengths and weaknesses of the approach, the errors made and lessons learned, and the overall resources needed to finish the projects.

2. About the Dictionaries

The three dictionaries are bilingual dictionaries from Tagalog, Urdu and Lao to English and further to Korean. Each dictionary consists of 15,000 manually post-edited entries and an additional 35,000 entries produced only automatically. Each post-edited entry contains:

- Pronunciation;
- possible word forms;

- sense disambiguators;
- translations into English and Korean;
- examples and their translation into English and Korean;
- an image (if appropriate);
- collocations;
- synonyms.

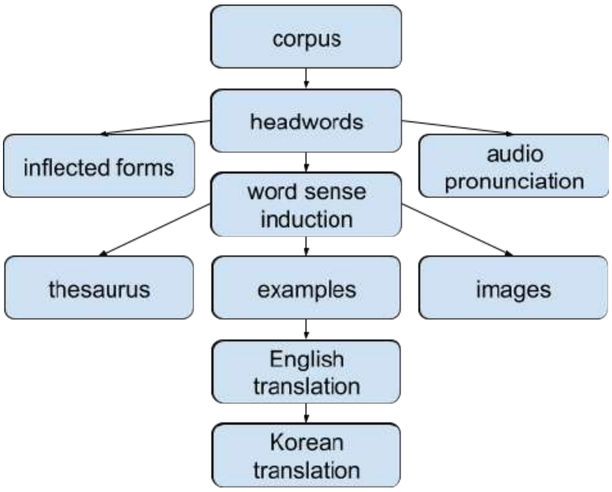


Figure 1: Workflow of the dictionary post-editing process

Following the MCD method, we divided entry creation into phases according to the entry parts above, and for each phase (except pronunciation) we generated data automatically from a large web corpus. Then the parts of the entries were manually cleaned and corrected by native speakers of the respective source languages; translations were proofread by translators. Each entry was in one phase at a time – the automatic data for the next phase were generated only after manual correction of the data in the previous phase. The overall workflow is demonstrated in Figure 1 and each of the steps is described in detail in the next section.

All post-editing steps have been implemented within the Lexonomy dictionary editor (Měchura, 2017) – typically as a custom editing widget, a small piece of JavaScript code a dictionary user can upload to set an editing form for an entry. This mechanism has proven to be sufficiently flexible and versatile to allow us to easily prepare a dedicated editing interface for a particular entry part.

3. Post-editing Workflow

3.1 Corpus processing

Three web corpora were created for the purposes of automatic dictionary drafting for each of the source languages using the methodology described in (Jakubíček et al., 2013). The sizes of the corpora were 230 (Tagalog), 265 (Urdu) and 120 (Lao) million words. Clearly, the sizes of the corpora are not overwhelming and represented a serious issue for automation, but we were simply unable to crawl more quality data from public websites. Our hypothesis is that for these languages most online content is published through non-open social networks instead of publicly available websites. Additionally, for all three languages, the internet contains a substantial amount of machine-translated content that has to be avoided as far as possible. For example, the Tagalog corpus contained 650 million words after initial boilerplate removal and partial deduplication; however, subsequent semi-automatic analysis of the data identified almost two-thirds of the data as machine-translated.

Each of the corpora was automatically part-of-speech tagged and lemmatised by different tools:

- for **Tagalog**, we used the Stanford tagger (Toutanova et al., 2003)[1] and lemmatised using an in-house improved version of a Tagalog stemmer[2],
- for **Urdu**, we used RFTagger (Schmid & Laws, 2008) to improve the tagging and lemmatisation output of the IIIT Hyderabad Urdu Parser[3],
- for **Lao**, we used RFTagger and a custom segmenter. Lao is not a flective language, thus lemmatisation was not relevant.

Additionally, we developed a word sketch grammar for each of the languages so that we could use the functions of the Sketch Engine (Kilgarriff et al., 2014) corpus management system.

3.2 Headwords

Headwords were automatically drafted by taking top words (lemmas) sorted by document frequency and having editors go over the list during post-editing. The classification manual used by the editors for Tagalog is provided in Figure 3. Editors labelled the headwords using the flag functionality in Lexonomy, as illustrated in Figure 2.

Additionally, the top 1,000 n-grams were also post-edited in order to cover the most salient multi-word expressions.

3.3 Inflected Forms

Inflected forms were automatically generated based on the automatic lemmatisation of the corpus. The editors reassigned inflected form where the lemmatiser incorrectly identified the base form.

3.4 Pronunciation

Pronunciation is the only part of the entry that was not automated. This is so for two reasons:

1. it would not be possible to “post-edit” the automatic recordings since there is no efficient way for a human to improve an automatically produced pronunciation in the form of an audio stream; and,
2. a manual recording can be carried out very quickly, so the potential gains of automation are rather limited.

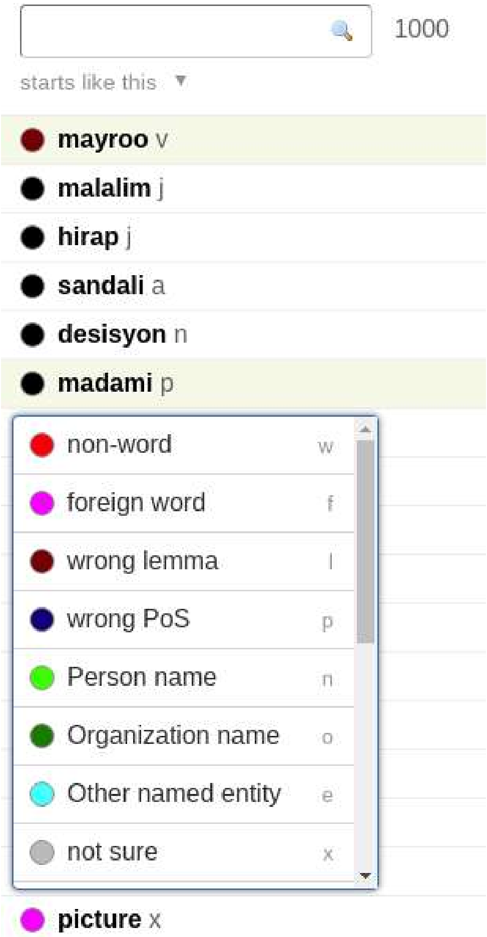


Figure 2: Using flags for headword classification in Lexonomy

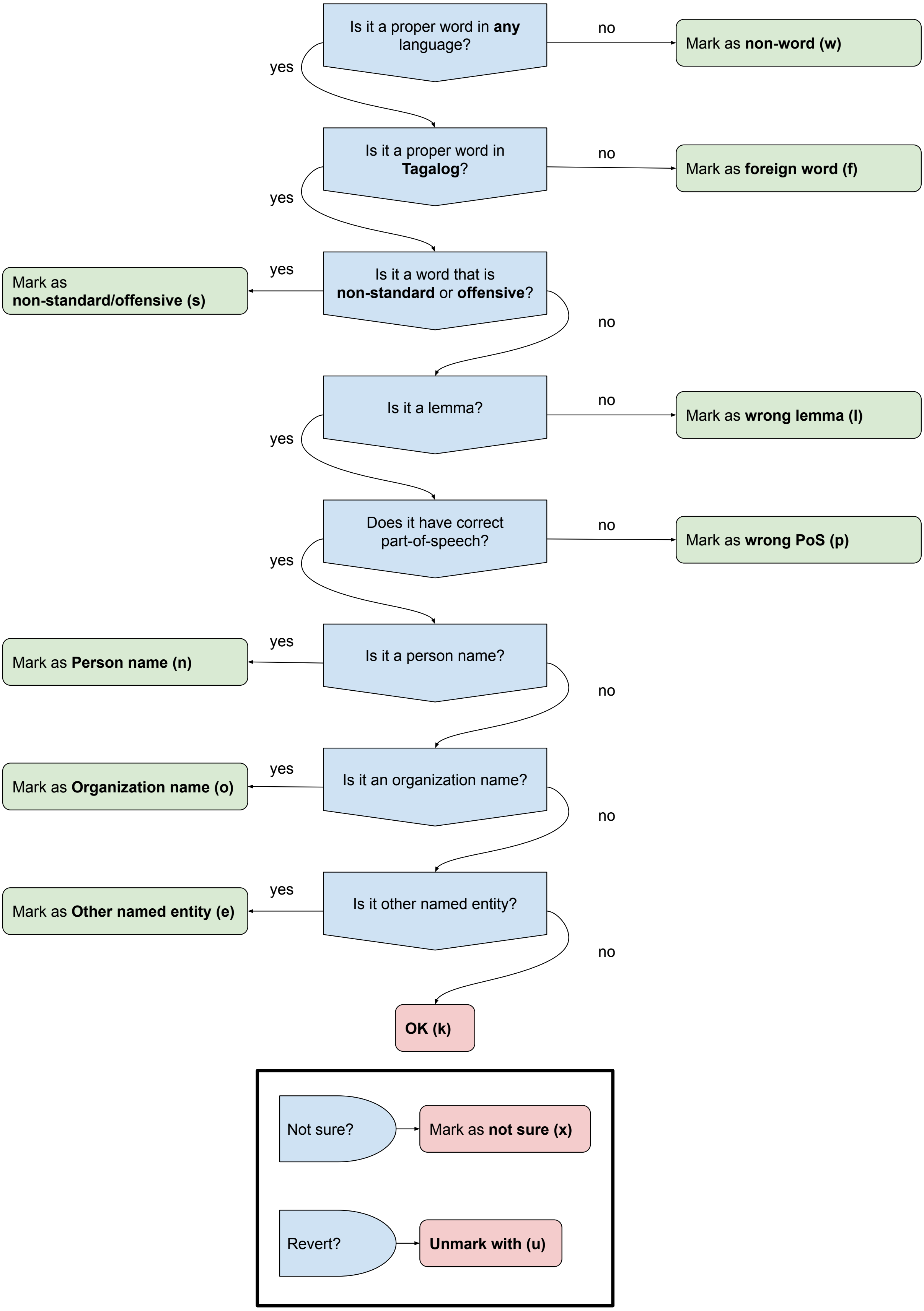

Is it a proper word in **any** language?
no
Mark as **non-word (w)**
yes
Is it a proper word in **Tagalog**?
no
Mark as **foreign word (f)**
yes
Mark as **non-standard/offensive (s)**
yes
Is it a word that is **non-standard** or **offensive**?
no
Is it a lemma?
no
Mark as **wrong lemma (l)**
yes
Does it have correct part-of-speech?
no
Mark as **wrong PoS (p)**
yes
Mark as **Person name (n)**
yes
Is it a person name?
no
Mark as **Organization name (o)**
yes
Is it an organization name?
no
Mark as **Other named entity (e)**
yes
Is it other named entity?
no
OK (k)
Not sure?
Mark as **not sure (x)**
Revert?
Unmark with (u)

Figure 3: Decision scheme for post-editing Tagalog headwords

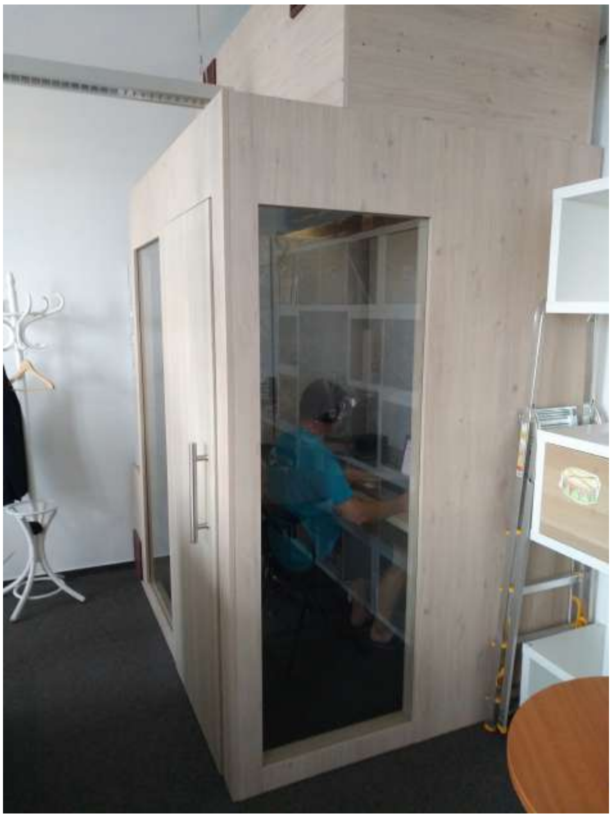


Figure 4: A sound-proof recording booth.

In our setting, native speakers recorded the pronunciation in a small recording booth (see Figure 4). They used a simple tool that displayed the next word to be recorded. A keyboard key press started the recording for a fixed amount of time (3 seconds); afterwards, the recorded sound was replayed to the speaker for confirmation (who then proceeded to the next word) or rejection (and the re-recording of the same word). This workflow allowed the speaker to record about 1,000 words during a working day of 8 hours, including inevitable rest breaks.

3.5 Word Sense Induction

We used a hybrid approach for identifying word senses by clustering the word sketch contexts according to the embedding vectors. We used skip-gram embeddings of dimension 300 using the fastText package (Joulin et al., 2016) and for every word sketch collocation of the examined word, we averaged collocation occurrence embeddings and used the HDBSCAN (McInnes et al., 2017) algorithm to cluster these vectors. The method is shown in Figure 5. HDBSCAN can determine the number of clusters automatically which is important because there is no reliable estimate for the number of word senses that we could use beforehand.

Editors were presented with the identified word sense clusters. Each cluster contained a set of collocations selected by the clustering algorithm. The editors reassigned

[1] The model was obtained from https://github.com/matthewgo/FilipinoStanfordPOSTagger

[2] Available at: https://github.com/crlwingen/TagalogStemmerPython

[3] http://ltrc.iiit.ac.in/showfile.php?filename=downloads/shallow_parser.php

collocations across the clusters/senses or created new senses. Along with of that process, disambiguators were created and translated into English. The editorial interface for this task can be seen in Figure 6.

After this post-editing step was finished, the corpus was sense-annotated by the senses/clusters as post-edited and this annotation was further used to generate a sense-based thesaurus, sense-based example sentences, and sense-based images.

3.6 Thesaurus

We obtained thesaurus items from the distributional thesaurus in Sketch Engine (Rychlý & Kilgarriff, 2007). The task for the editor during the post-editing phase was to validate each item for the particular word sense and classify them into synonyms and antonyms. Distinguishing between synonyms and antonyms is not yet automated and is a good candidate for further research.

3.7 Examples

Example sentences were obtained using GDEX (Kilgarriff et al., 2008) from Sketch Engine and validated and translated into English in the post-editing phase. This turned out to be one of the most tedious tasks in the end, owing to the very modest corpus sizes.

3.8 Images

We downloaded images from copyright-free online sources (Wikimedia projects, Pixabay, targeted Google Custom Search) and had the editors choose the best image (if any) for the particular word sense. The editing interface is demonstrated in Figure 7.

3.9 Translations

As the last step, disambiguators and example sentences were translated into Korean. Disambiguators were pre-translated using both Google Translate and Microsoft Bing; the latter was used mainly because it offers multiple translation candidates as part of its API. Unfortunately, it turned out that the alternative translation candidates given by Bing are just alternative word forms or spellings, so it did not help much to increase the diversity of the translation candidates before a human translator was validated them. Example sentences were translated using just Google Translate.

4. Data Management

We started the project with the idea of separated XML files, "batches" containing a few dozens of entries, which would fall through the annotation process as atomic units – and in the end we would just put them together into a dictionary. However, errors and disagreements in annotation (e.g. the example annotator refused to process the word previously accepted) led to a shrinking of the batches, complicated dependencies among them, the overall complexity of the data, and massive delays in processing.

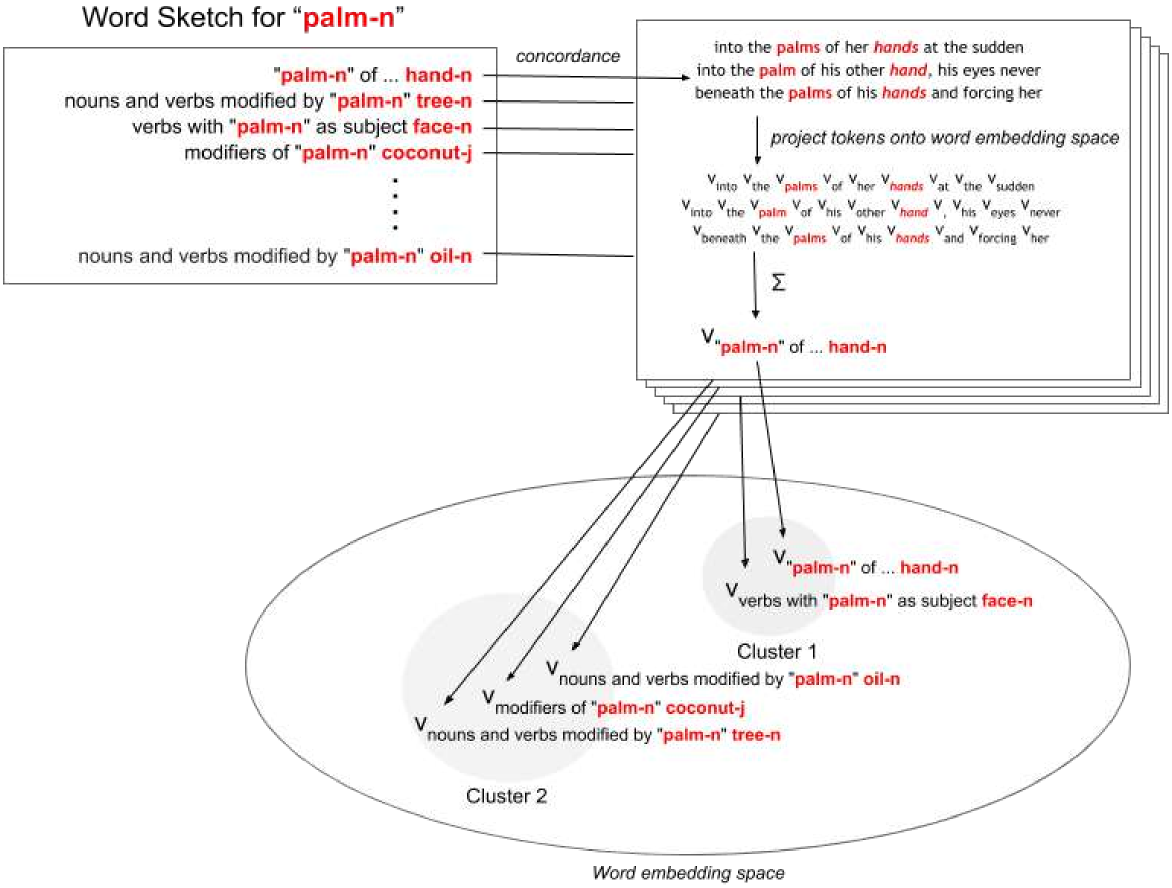

Word Sketch for "palm-n"
"palm-n" of ... hand-n
nouns and verbs modified by "palm-n" tree-n
verbs with "palm-n" as subject face-n
modifiers of "palm-n" coconut-j
nouns and verbs modified by "palm-n" oil-n
concordance
into the palms of her hands at the sudden
into the palm of his other hand, his eyes never
beneath the palms of his hands and forcing her
project tokens onto word embedding space
Σ
V"palm-n" of ... hand-n
V"palm-n" of ... hand-n
Vverbs with "palm-n" as subject face-n
Cluster 1
Vnouns and verbs modified by "palm-n" oil-n
Vmodifiers of "palm-n" coconut-j
Vnouns and verbs modified by "palm-n" tree-n
Cluster 2
Word embedding space

Figure 5: Using HDBSCAN over word sketch collocation embeddings

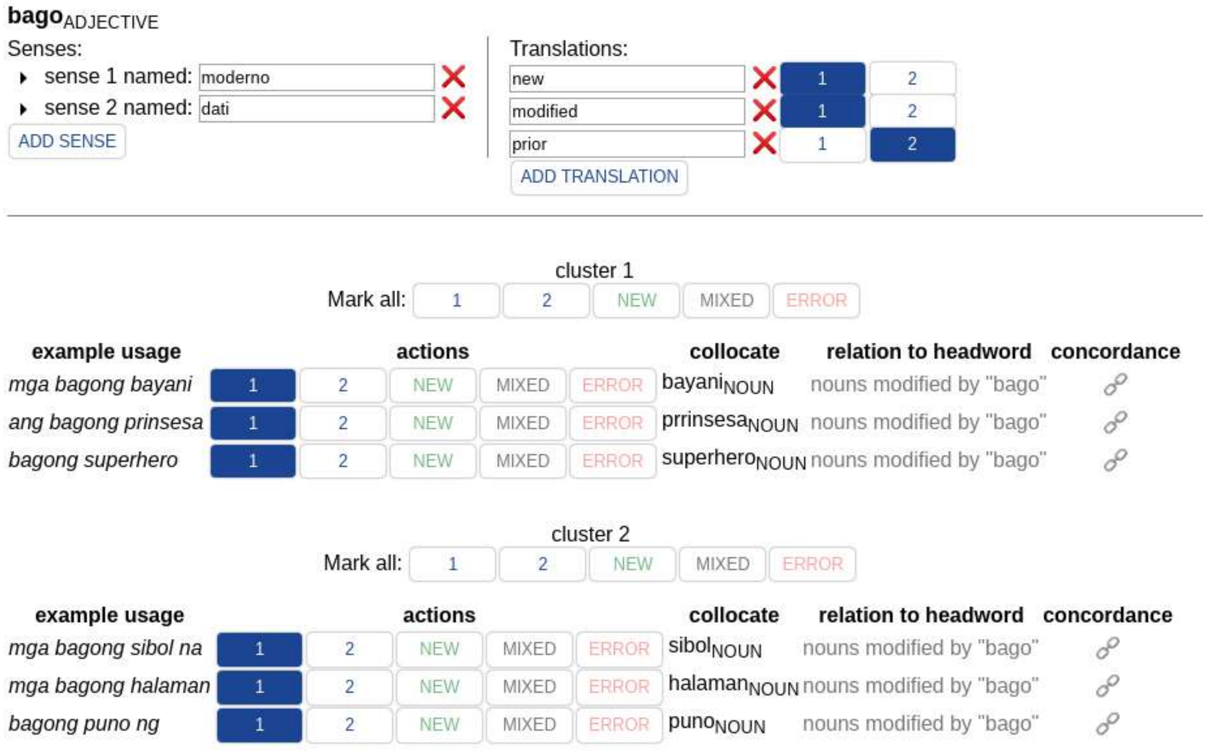
Therefore, we switched to a central database, stored in a novel textual format called NVH (name-value hierarchy)[4]. The batches for annotation were created as XML exports from this database, and finished annotations were processed as imports into the database. For each phase of entry processing, we implemented automatic import and export procedures that ensured consistency. The Git version control system was used so that it was subsequently possible to inspect, track, and fix problematic imports.

This mechanism worked much better and we managed to complete the dictionaries with it. However, there were still significant drawbacks:

- some of the annotation errors propagated and were only discovered when it was too hard to fix them (many of the fixes were done manually in the last phase of the project); and,
- there were errors in the original corpus annotation, so it was necessary to correct many of the headwords and propagate their correct form back to the corpus (so that the subsequent phases could be carried out correctly).

We find it crucial to understand that data management needs to be designed to take into account the inevitable human errors (and have a mechanism to handle them easily) and the fact that a source corpus is a noisy resource that the can be improved using

[4] http://www.namevaluehierarchy.org



bago$_{\text{ADJECTIVE}}$

Senses:

▸ sense 1 named: moderno ✕

▸ sense 2 named: dati ✕

ADD SENSE

Translations:

new	✕	1	2
modified	✕	1	2
prior	✕	1	2

ADD TRANSLATION

cluster 1

Mark all: 1 2 NEW MIXED ERROR

example usage	actions	collocate	relation to headword	concordance
mga bagong bayani	1 2 NEW MIXED ERROR	bayani$_{\text{NOUN}}$	nouns modified by "bago"	🔗
ang bagong prinsesa	1 2 NEW MIXED ERROR	prrinsesa$_{\text{NOUN}}$	nouns modified by "bago"	🔗
bagong superhero	1 2 NEW MIXED ERROR	superhero$_{\text{NOUN}}$	nouns modified by "bago"	🔗

cluster 2

Mark all: 1 2 NEW MIXED ERROR

example usage	actions	collocate	relation to headword	concordance
mga bagong sibol na	1 2 NEW MIXED ERROR	sibol$_{\text{NOUN}}$	nouns modified by "bago"	🔗
mga bagong halaman	1 2 NEW MIXED ERROR	halaman$_{\text{NOUN}}$	nouns modified by "bago"	🔗
bagong puno ng	1 2 NEW MIXED ERROR	puno$_{\text{NOUN}}$	nouns modified by "bago"	🔗

Figure 6: Post-editing interface for word sense identification.

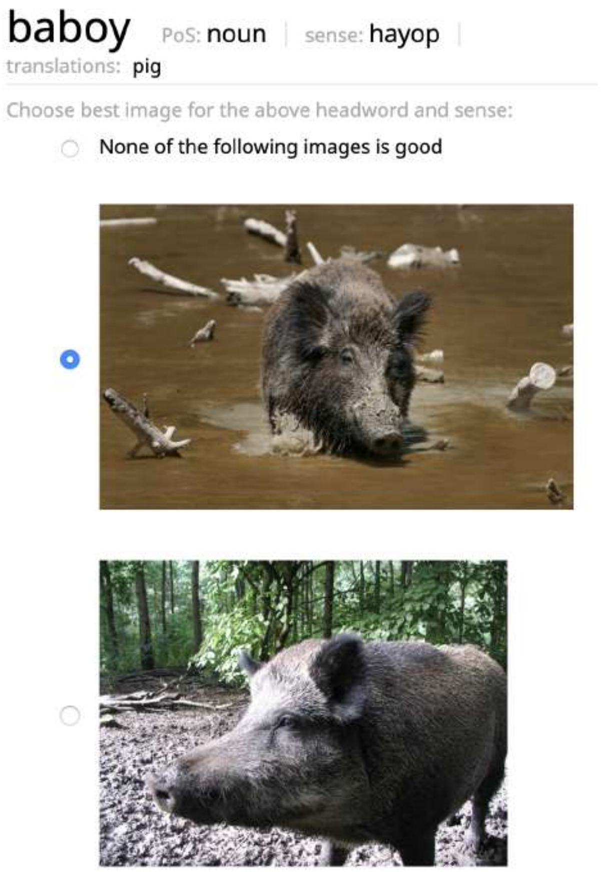
the annotations obtained in the post-editing phase. In our case, the automatic corpus annotation was improved whenever the annotators submitted corrections to part-of-speech tagging, lemmatisation or sense-identification. Updating the corpus and using its best version for further work was speeding up further post-editing tasks as well as created a better corpus, which for us was an important by-product in itself.

5. Time effort

The overall time effort for the Tagalog dictionary is available in Table 1. All three projects were started with approximately a 6-month lag and we managed to utilise the experiences gained as well as reuse many of the tools (such as the custom editing widgets in Lexonomy) so that the time effort for the Urdu dictionary (which started second) was about 20% less than for Tagalog, and for Lao (which started third) it was again about 20% less than for Urdu.

6. Conclusion

Before the start of the project execution, we mainly anticipated problems with the automatic algorithms generating the data – our main concerns were the possible low quality of the automatically generated data and therefore the low efficiency of the post-editing process. The reality was quite different. The output from these algorithms was mostly sufficient and the post-editing process was effective. We experienced the largest challenges in the management part of the project, and especially regarding the data management: keeping the data consistent, keeping the corpus consistent with the corrected data, keeping the annotation process running smoothly, and avoiding repeated cycles.



baboy PoS: noun | sense: hayop |

translations: pig

Choose best image for the above headword and sense:

○ None of the following images is good

●

○

Figure 7: Post-editing interface for images in Lexonomy.

Annotation phase	PH
Headwords	396
Revisions	464
Inflections	478
Audio (recording)	100
Senses + En translation	669
Collocations	204
Images	313
Thesaurus	617
Examples + En translation	1,938
Examples proofreading	135
Examples corrections	373
Translation into Korean	772
Final review	591
Final manual changes	87
Training, communication	64
Total	7,199

Table 1: Person-hours spent on annotation for the different phases of the Tagalog dictionary

In further projects, this is the part that needs to be focused on in the first place: solving this successfully is key to the overall success of the project.

The projects have also reconfirmed the importance of corpus size for quality lexicographic work. The sizes of the corpora used should be seen as the necessary minimum and many of the issues we faced would not be present if the corpora had been, for example, 10 times bigger, which would easily be the case for many better-resourced languages.

Overall, the projects clearly showed the vitality of the post-editing workflow in lexicography as well as the technological readiness of the lexicographic tools that we used. We are confident that further improvements in the management of the whole process can bring further significant savings as regards the in time effort required.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. This work has been partly supported by the Ministry of Education of CR within the LINDAT-CLARIAH-CZ project LM2018101.

7. References

Baisa, V., Blahuš, M., Cukr, M., Herman, O., Jakubíček, M., Kovář, V., Medveď, M., Měchura, M., Rychlý, P., Suchomel, V. (2019). Automating dictionary production: a Tagalog-English-Korean dictionary from scratch. In *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference. 1-3 October 2019, Sintra, Portugal.* Lexical Computing, pp. 805–818.

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P. & Suchomel, V. (2013). The TenTen Corpus Family. *International Conference on Corpus Linguistics, Lancaster.*

Jakubíček, M., Kovář, V. & Rychlý, P. (2020). Million-Click Dictionary. In *Proceedings of XIX EURALEX Congress: Lexicography for Inclusion, Vol. II [to be published].*

Joulin, A., Grave, E., Bojanowski, P. & Mikolov, T. (2016). Bag of Tricks for Efficient Text Classification. *CoRR*, abs/1607.01759. URL http://arxiv.org/abs/1607.01759. 1607.01759.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, 1. URL http://dx.doi.org/10.1007/s40607-014-0009-9.

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M. & Rychlỳ, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus. In *Proceedings of the XIII EURALEX international congress.* Documenta Universitaria Barcelona, Spain, pp. 425–432.

McInnes, L., Healy, J. & Astels, S. (2017). hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, 2(11).

Měchura, M.B. (2017). Introducing Lexonomy: an open-source dictionary writing and publishing system. In *Electronic Lexicography in the 21st Century: Lexicography from Scratch. Proceedings of the eLex 2017 conference, 19-21 September 2017.*

Rychlý, P. & Kilgarriff, A. (2007). An efficient algorithm for building a distributional thesaurus (and other Sketch Engine developments). In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pp. 41–44.

Schmid, H. & Laws, F. (2008). Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pp. 777–784.

Toutanova, K., Klein, D., Manning, C.D. & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*. Association for Computational Linguistics, pp. 173–180.

Word-embedding based bilingual terminology alignment

Andraž Repar[1], **Matej Martinc**[2], **Matej Ulčar**[3], **Senja Pollak**[4]

[1]International Postgraduate School, Institut Jozef Stefan, Jamova 39, Ljubljana, Slovenia
[2] Institut Jozef Stefan, Jamova 39, Ljubljana, Slovenia
[3] Faculty of Computer and Information Science, University of Ljubljana, Večna pot 113, Ljubljana, Slovenia
[4] Institut Jozef Stefan, Jamova 39, Ljubljana, Slovenia
E-mail: andraz.repar@ijs.si, matej.martinc@ijs.si, matej.ulcar@fri.uni-lj.si, senja.pollak@ijs.si

Abstract

The ability to accurately align concepts between languages can provide significant benefits in many practical applications. In this paper, we extend a machine learning approach using dictionary and cognate-based features with novel cross-lingual embedding features using pretrained fastText embeddings. We use the tool VecMap to align the embeddings between Slovenian and English and then for every word calculate the top 3 closest word embeddings in the opposite language based on cosine distance. These alignments are then used as features for the machine learning algorithm. With one configuration of the input parameters, we managed to improve the overall F-score compared to previous work, while another configuration yielded improved precision (96%) at a cost of lower recall. Using embedding-based features as a replacement for dictionary-based features provides a significant benefit: while a large bilingual parallel corpus is required to generate the Giza++ word alignment lists, no such data is required for embedding-based features where the only required inputs are two unrelated monolingual corpora and a small bilingual dictionary from which the embedding alignments are calculated.

Keywords: terminology alignment; word embeddings; embeddings alignment; machine learning

1. Introduction

The ability to accurately align concepts between languages can provide significant benefits in many practical applications. For example, in terminology, terms can be aligned between languages to provide bilingual terminological resources, while in the news industry, keywords can be aligned to provide better news clustering or search in another language. Accurate bilingual resources can also serve as seed data for various other NLP tasks, such as multilingual vector space alignment.

Bilingual terminology alignment[1] is the process of aligning terms between two candidate term lists in two languages. The primary purpose of bilingual terminology extraction is to build a term bank - i.e. a list of terms in one language along with their equivalents in the other language. With regard to the input text, we can distinguish between alignment on the basis of a parallel corpus and alignment on the basis of a comparable corpus. For the translation industry, bilingual terminology extraction from parallel corpora is extremely relevant due to the large amounts of sentence-aligned parallel corpora available in the form of translation memories. Consequently, initial attempts at bilingual terminology extraction involved parallel input data (Kupiec, 1993; Daille et al., 1994; Gaussier, 1998), and the interest of the community has continued until today. However, most parallel corpora are owned by private companies[2], such as language service providers, who consider them to be their intellectual property and are reluctant to share them publicly. For this reason (and in particular for language pairs not involving English) considerable efforts have also been invested into researching bilingual terminology extraction from comparable corpora (Fung & Yee, 1998; Rapp, 1999; Chiao & Zweigenbaum, 2002; Cao & Li, 2002; Daille &

[1] Note that bilingual terminology alignment has a narrower focus than *bilingual terminology extraction*, but the two terms are often used interchangeably in various papers. The latter covers extraction and alignment of terms between languages.

[2] However, some publicly available parallel corpora do exist. A good overview can be found at the OPUS web portal (Tiedemann, 2012).

Morin, 2005; Morin et al., 2008; Vintar, 2010; Bouamor et al., 2013; Hazem & Morin, 2016, 2017).

The approach designed by Aker et al. (2013) and replicated and adapted in Repar et al. (2019) served as the basis of our work. It was developed to align terminology between languages with the help of parallel corpora using machine-learning techniques. They use terms from the Eurovoc (Steinberger et al., 2002) thesaurus and train an SVM binary classifier (Joachims, 2002) (with a linear kernel and the trade-off between training error and margin parameter c = 10). The task of bilingual alignment is treated as a binary classification task – each term from the source language S is paired with each term from the target language T and the classifier then decides whether the aligned pair is correct or incorrect. Aker et al. (2013) run their experiments on the 21 official EU languages covered by Eurovoc with English always being the source language (20 language pairs altogether). They evaluate the performance on a held-out term pair list from Eurovoc using recall, precision and F-measure for all 21 languages. Next, they propose an experimental setting for a simulation of a real-world scenario where they collect English-German comparable corpora of two domains (IT, automotive) from Wikipedia, perform monolingual term extraction using the system by Pinnis et al. (2012) followed by the bilingual alignment procedure described above and manually evaluate the results (using two evaluators). They report excellent performance on the held-out term list with many language pairs reaching 100% precision and the lowest recall being 65%. For Slovenian, which is our main interest, the reported results were excellent with perfect or nearly perfect precision and good recall. The reported results of the manual evaluation phase were also good, with two evaluators agreeing that at least 81% of the extracted term pairs in the IT domain and at least 60% of the extracted term pairs in the automotive domain can be considered exact translations. Repar et al. (2019) tried to reproduce their approach and after initially having little success they were at the end able to achieve comparable results with precision exceeding 90% and recall over 50%.

Despite the problem of bilingual term alignment lending itself well to the binary classification task, there have been relatively few approaches utilising machine learning. Similar to Aker et al. (2013), Baldwin & Tanaka (2004) generate corpus-based, dictionary-based and translation-based features and train an SVM classifier to rank the translation candidates. Note that they only focus on multi-word noun phrases (noun + noun). A similar approach, again focusing on noun phrases, is also described by Cao & Li (2002). Finally, Nassirudin & Purwarianti (2015) also reimplement Aker et al. (2013) for the Indonesian-Japanese language pair and further expand it with additional statistical features.

This paper is organised as follows: the present section introduces the problem and related work, Section 2 describes the datasets used for the experiments, Section 3 lists the features used in the machine learning process, Section 4 contains a description of the experiments and lists their results and Section 5 provides the conclusion.

2. Resources

The approach described in this paper requires four types of resources. The first two are the same as in Aker et al. (2013) and Repar et al. (2019), whereas the third and fourth resources are required for the additional experiments conducted for this paper:

- aligned term pairs in two languages that serve as training data
- a parallel corpus to generate a Giza++ word alignment list
- pretrained embeddings in two languages
- a (small) bilingual dictionary

We create term pairs from the Eurovoc (Steinberger et al., 2002) thesaurus, which at the time of Repar et al. (2019) consisted of 7,083[3] terms, by pairing Slovenian terms with English ones. The test set consisted of 600 positive (correct) term pairs — taken randomly out of the total 7,083 Eurovoc term pairs — and around 1.3 million negative pairs which were created by pairing each source term with 200 distinct incorrect random target terms. Aker et al. (2013) argue that this was done to simulate real-world conditions where the classifier would be faced with a larger number of negative pairs and a comparably small number of positive ones. The 600 positive term pairs were further divided into 200 pairs where both (i.e. source and target) terms were single words, 200 pairs with a single word only on one side and 200 pairs with multiple-word terms on both sides. The remaining positive term pairs (approximately 6,200) were used as training data along with additional 6,200 negative pairs. These were constructed by taking the source side terms and pairing each source term with one target term (other than the correct one). Using Giza++, we created source-to-target and target-to-source word alignment dictionaries based on the DGT translation memory (Steinberger et al., 2013). The resulting dictionary entries consist of the source word *s*, its translation *t* and the number indicating the probability that *t* is an actual translation of *s*. To improve the performance of the dictionary-based features, the following entries were removed from the dictionaries:

- entries where probability is lower then 0.05
- entries where the source word was less than 4 characters and the target word more than 5 characters long and vice versa in order to avoid translations of stop word to content words)

In addition to the resources described above, we used fastText (Bojanowski et al., 2016) pre-trained word embedding vectors to calculate distances (or similarities) between terms. We aligned monolingual fastText embeddings using the VecMap (Artetxe et al., 2018) tool which can align embeddings with the help of a small bilingual dictionary. We used a bilingual dictionary compiled from two sources: single word terms from Eurovoc and Wiktionary entries extracted using the wikt2dict tool (Acs, 2014). Using the aligned embedding vectors, we then calculated cosine distances between all words present in Eurovoc terms in one language and all words present in Eurovoc terms in the other language.

Using the fastText-based lists of aligned words, we created 3-tuples[4] of most similar — based on cosine similarity — source-to-target and target-to-source words, such as:

- ksenofobija [‘xenophobia’, ‘0.744’], [‘racism’, ‘0.6797’], [‘anti-semitism’, ‘0.654’]
- ženska [‘woman’, ‘0.7896’], [‘women’, ‘0.73’], [‘female’, ‘0.722’]

[3] While new terms are constantly added to Eurovoc, we decided not to use them to allow for better comparison between the approaches

[4] This number was determined experimentally.

where the tuple contains the source language word along with their three most likely corresponding words in the target language and their cosine similarities. The 3-tuples of most similar words were used to construct additional features for the machine learning algorithm.

3. Feature construction

The updated approach in this paper uses three types of features that express correspondences between the words (composing a term) in the target and source language. The dictionary and cognate-based features are same as in Repar et al. (2019), while embedding-based features are newly developed. The three feature types are as follows (for a detailed description see Table 1):

- 7 dictionary-based (using Giza++) features which take advantage of dictionaries created from large parallel corpora of which 6 are direction-dependent (source-to-target or target-to-source) and 1 direction-independent — resulting in altogether 13 features
- 7 cognate-based features (on the basis of Gaizauskas et al. (2012)) which utilize string-based word similarity between languages
- 5 cognate-based features using specific transliteration rules which take into account the differences in writing systems between two languages: e.g. Slovenian and English. Transliteration rules were created for both directions (source-to-target and target-to-source) separately and cognate-based features were constructed for both directions — resulting in an additional 10 cognate-based features with transliteration rules. The following transliteration rules were used: *x:ks, y:j, w:v, q:k* for English to Slovenian and *č:ch, š:sh, ž:zh* for Slovenian to English
- 5 direction-dependent combined[5] features where the term pair alignment is correct if either the dictionary or the cognate-based method returns a positive result — resulting in a total of 10 combined features
- 12 novel direction-dependent embedding-based features utilising fastText embeddings — resulting in a total of 24 features
- 5 novel combined features constructed in the same manner as the existing combined features but replacing Giza++ word lists with fastText-based lists of top 3 aligned words - resulting in a total of 10 novel combined features
- 3 term length features: sourceTargetLengthMatch, sourceTermLength, targetTermLength

To match words with morphological differences, we do not perform direct string matching but utilise Levenshtein Distance. Two words were considered equal if the Levenshtein Distance Levenshtein (1966) was equal or higher than 0.95.

[5] For combined features, a word is considered as covered if it can be found in the corresponding set of Giza++ translations or if one of the cognate-based measures (Longest Common Subsequence, Longest Common Substring, Levensthein Distance, Needleman-Wunsch Distance, Dice) is 0.70 or higher (set experimentally by Aker et al. (2013))

Feature	Cat	Description
isFirstWordTranslated	Dict	Checks whether the first word of the source term is a translation of the first word in the target term (based on the Giza++ dictionary)
isLastWordTranslated	Dict	Checks whether the last word of the source term is a translation of the last word in the target term
percentageOfTranslatedWords	Dict	Ratio of source words that have a translation in the target term
percentageOfNotTranslatedWords	Dict	Ratio of source words that do not have a translation in the target term
longestTranslatedUnitInPercentage	Dict	Ratio of the longest contiguous sequence of source words which has a translation in the target term (compared to the source term length)
longestNotTranslatedUnitInPercentage	Dict	Ratio of the longest contiguous sequence of source words which do not have a translation in the target term (compared to the source term length)
Longest Common Subsequence Ratio	Cogn	Measures the longest common non-consecutive sequence of characters between two strings
Longest Common Substring Ratio	Cogn	Measures the longest common consecutive string (LCST) of characters that two strings have in common
Dice similarity	Cogn	2*LCST / (len(source) + len(target))
Needlemann-Wunsch distance	Cogn	LCST / min(len(source), len(target))
isFirstWordCognate	Cogn	A binary feature which returns True if the longest common consecutive string (LCST) of the first words in the source and target terms divided by the length of the longest of the two words is greater than or equal to a threshold value of 0.7 and both words are longer than 3 characters
isLastWordCognate	Cogn	A binary feature which returns True if the longest common consecutive string (LCST) of the last words in the source and target terms divided by the length of longest of the two words is greater than or equal to a threshold value of 0.7 and both words are longer than 3 characters
Normalized Levensthein distance (LD)	Cogn	1 - LD / max(len(source), len(target))
isFirstWordCovered	Comb	A binary feature indicating whether the first word in the source term has a translation or transliteration in the target term
isLastWordCovered	Comb	A binary feature indicating whether the last word in the source term has a translation or transliteration in the target term
percentageOfCoverage	Comb	Returns the percentage of source term words which have a translation or transliteration in the target term
percentageOfNonCoverage	Comb	Returns the percentage of source term words which have neither a translation nor transliteration in the target term
difBetweenCoverageAndNonCoverage	Comb	Returns the difference between the last two features
isFirstWordMatch	Emd	Checks whether the first word of the source term is the most likely translation of the first word in the target term (based on the aligned embeddings)
isLastWordMatch	Emd	Checks whether the last word of the source term is the most likely translation of the last word in the target term (based on the aligned embeddings)
percentageOfFirstMatchWords	Emb	Ratio of source words that have a first match (i.e. first position in the 3-tuple) in the target term
percentageOfNotFirstMatchWords	Emb	Ratio of source words that do not have a first match (i.e. first position in the 3-tuple) in the target term
longestFirstMatchUnitInPercentage	Emb	Ratio of the longest contiguous sequence of source words which has a first match (first position in the 3-tuple) in the target term (compared to the source term length)
longestNotFirstMatchUnitInPercentage	Emb	Ratio of the longest contiguous sequence of source words which do not have a first match (first position in the 3-tuple) in the target term (compared to the source term length)
isFirstWordTopnMatch	Emd	Checks whether the first word of the source term is in the 3-tuple of most likely translations of the first word in the target term (based on the aligned embeddings)

isLastWordTopnMatch	Emd	Checks whether the first word of the source term is not in the 3-tuple of most likely translations of the first word in the target term (based on the aligned embeddings)
percentageOfTopnMatchWords	Emb	Ratio of source words that have a match (i.e. any position in the 3-tuple) in the target term
percentageOfNotTopnMatchWords	Emb	Ratio of source words that do not have a match (i.e. any position in the 3-tuple) in the target term
longestTopnMatchUnitInPercentage	Emb	Ratio of the longest contiguous sequence of source words which has a match (any position in the 3-tuple) in the target term (compared to the source term length)
longestNotTopnMatchUnitInPercentage	Emb	Ratio of the longest contiguous sequence of source words which do not have a match (any position in the 3-tuple) in the target term (compared to the source term length)
isFirstWordCoveredEmbeddings	Comb	A binary feature indicating whether the first word in the source term has a match (any position in the 3-tuple) or transliteration in the target term
isLastWordCoveredEmbeddings	Comb	A binary feature indicating whether the last word in the source term has a match (any position in the 3-tuple) or transliteration in the target term
percentageOfCoverageEmbeddings	Comb	Returns the percentage of source term words which have a match (any position in the 3-tuple) or transliteration in the target term
percentageOfNonCoverageEmbeddings	Comb	Returns the percentage of source term words which do not have a match (any position in the 3-tuple) or transliteration in the target term
diffBetweenCoverageAnd-NonCoverageEmbeddings	Comb	Returns the difference between the last two features

Figure 1: Features used in the experiments. Note that some features are used more than once because they are direction-dependent.

4. Experimental setup and results

The constructed features were then used to train an SVM binary classifier (Joachims, 2002) (with a linear kernel and the trade-off between training error and margin parameter c = 10). We selected three configurations from Repar et al. (2019) for comparison:

- Training set 1:200: a very unbalanced training set (ratio of 1:200 between positive and negative examples [6]) greatly improves the precision of the classifier at a cost of somewhat lower recall, when compared to a balanced train set or a less unbalanced train set (e.g., ratio of 1:10 between positive and negative examples).
- Training set filtering 3: In Repar et al. (2019), we have performed an error analysis and found that many incorrectly classified term pairs are cases of partial translation where one unit in a multi-word term has a correct Giza++ dictionary translation in the corresponding term in the other language. Based on the problem of partial translations, leading to false positive examples, we focused on the features that would eliminate such partial translations from the training set. After a systematic experimentation, we noticed that we can drastically improve precision if we only keep positive term pairs with the following feature values: isFirstWordTranslated = True, isLastWordTranslated = True, percentageOfCoverage > 0.66, isFirstWordTranslated-reversed = True, isLastWordTranslated-reversed = True, percentageOfCoverage-reversed > 0.66.

[6] 1:200 imbalance ratio was the largest imbalance we tried, since the testing results indicated that no further gains could be achieved by further increasing the imbalance.

- Cognates: the dataset is additionally filtered according to the following criteria: isFirstWordCognate = True and isLastWordCognate = True, isFirstWordTranslated = True and isLastWordCognate = True, isFirstWordCognate = True and isLastWordTranslated = True and we also use a Gaussian kernel instead of the linear one, since this new dataset structure represents a classic "exclusive or" (XOR) problem which a linear classifier is unable to solve.

The selection was made based on our experience and previous work with this approach. The three selected configurations were among the best performing in previous experiments and we believed they had the highest potential for improvement. For a complete description of the decisions that led to these configurations, please refer to Repar et al. (2019).

No.	Config EN-SL	Training set size	Pos/Neg ratio	Precision	Recall	F-score
Dictionary-based and cognate-based features						
1	Training set 1:200	1,303,083	1:200	0.4299	**0.7617**	0.5496
2	Training set filtering 3	645,813	1:200	0.9342	0.4966	0.6485
3	Cognates approach	672,345	1:200	0.8732	0.5167	0.6492
Dictionary-based, embedding-based and cognate-based features						
1	Training set 1:200	1,303,083	1:200	0.5375	0.680	0.6004
2	Training set filtering 3	695,058	1:200	0.8170	0.5133	0.6305
3	Cognates approach	706,113	1:200	0.8991	0.5200	**0.6589**
Embedding-based and cognate-based features only						
1	Training set 1:200	1,303,083	1:200	0.3232	0.4967	0.3916
2	Training set filtering 3	322,605	1:200	0.9545	0.2450	0.3899
3	Cognates approach	394,362	1:200	**0.9618**	0.3617	0.5242

Table 2: Results on the English-Slovenian term pair.

First, we simply added the new embedding-based features to the dataset to see if they improved the overall performance. Later, we removed the dictionary-based features from the dataset to see whether the novel embedding-based features could replace them without a major impact on the performance. As can be observed from Table 2, the results are a mixed bag when using all available features. Without any training set filtering, the new features improve precision at the expense of recall, but are less effective when filtering is applied. Nevertheless, when we use additional trainset filters for the Cognates approach, we can observe a slight increase in both precision and recall resulting in the overall highest F-score. When we use only embedding-based and cognate-based features, which would be beneficial for language pairs without access to large parallel corpora needed to create Giza++ word alignments, there is a significant drop in recall in all cases, but precision actually increases when trainset filtering is applied and the Cognates approach achieves the overall best precision.

5. Conclusion

In this paper, we continued our experiments on bilingual terminology alignment using a machine learning approach by adding new features based on fastText word embedding vectors. We took advantage of the availability of large pre-trained datasets by Bojanowski et al. (2016), and a cross-lingual word embedding mapping tool Vecmap by Artetxe et al. (2018) to create word alignment dictionaries similar to the output of traditional word alignment tools, such as Giza++ (Och & Ney, 2003). The single most important advantage of this approach is that while Giza++ requires a large parallel corpus, fastText vectors are trained on monolingual data and Vecmap needs only a (much smaller) bilingual dictionary. Bilingual dictionaries are readily available for many language pairs via Wiktionary (Acs, 2014).

The experiments showed that the new features can have a positive impact on the F-score (depending on the configuration), but precision was somewhat lower compared to when we were using only Giza++ features. When we removed Giza++ features and using only the new embedding-based features (alongside cognate features which are based on word similarity and require no pre-existing bilingual data), we observed somewhat lower recall and slightly higher precision. This means that the embedding-based features can be used instead of Giza++ features for language pairs where no large parallel bilingual corpora are available.

In terms of future work, we plan on creating additional features using contextual embeddings, such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2018), which could potentially help us improve recall, and explore more granular and detailed training set filtering techniques. We also plan to expand the experiments and test other configurations in a more systematic way.

6. Acknowledgements

The work was supported by European Union's Horizon 2020 research and innovation programme under grant agreement No 825153, project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media). We also acknowledge the project the financial support from the Slovenian Research Agency for research core funding for the programme Knowledge Technologies (No. P2-0103) and the project TermFrame - Terminology and Knowledge Frames across Languages (No. J6-9372).

7. References

Acs, J. (2014). Pivot-based multilingual dictionary building using Wiktionary. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation LREC*.

Aker, A., Paramita, M. & Gaizauskas, R. (2013). Extracting bilingual terminologies from comparable corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1. pp. 402–411.

Artetxe, M., Labaka, G. & Agirre, E. (2018). Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Baldwin, T. & Tanaka, T. (2004). Translation by machine of complex nominals: Getting it right. In *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*. pp. 24–31.

Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. (2016). Enriching Word Vectors with Subword Information. *arXiv preprint arXiv:1607.04606*.

Bouamor, D., Semmar, N. & Zweigenbaum, P. (2013). Context Vector Disambiguation for Bilingual Lexicon Extraction from Comparable Corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. pp. 759–764.

Cao, Y. & Li, H. (2002). Base Noun Phrase Translation Using Web Data and the EM Algorithm. In *Proceedings of the 19th International Conference on Computational Linguistics - Volume 1*. pp. 1–7.

Chiao, Y.C. & Zweigenbaum, P. (2002). Looking for Candidate Translational Equivalents in Specialized, Comparable Corpora. In *Proceedings of the 19th International Conference on Computational Linguistics - Volume 2*. pp. 1–5.

Daille, B., Gaussier, E. & Langé, J.M. (1994). Towards Automatic Extraction of Monolingual and Bilingual Terminology. In *Proceedings of the 15th Conference on Computational Linguistics - Volume 1*. pp. 515–521.

Daille, B. & Morin, E. (2005). French-English Terminology Extraction from Comparable Corpora. In *Natural Language Processing – IJCNLP 2005*. pp. 707–718.

Devlin, J., Chang, M.W., Lee, K. & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

Fung, P. & Yee, L.Y. (1998). An IR Approach for Translating New Words from Nonparallel, Comparable Texts. In *Proceedings of the 17th International Conference on Computational Linguistics - Volume 1*. pp. 414–420.

Gaizauskas, R., Aker, A. & Yang Feng, R. (2012). Automatic bilingual phrase extraction from comparable corpora. In *24th International Conference on Computational Linguistics*. pp. 23–32.

Gaussier, E. (1998). Flow Network Models for Word Alignment and Terminology Extraction from Bilingual Corpora. In *Proceedings of the 17th International Conference on Computational Linguistics - Volume 1*. pp. 444–450.

Hazem, A. & Morin, E. (2016). Efficient Data Selection for Bilingual Terminology Extraction from Comparable Corpora. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. pp. 3401–3411.

Hazem, A. & Morin, E. (2017). Bilingual Word Embeddings for Bilingual Terminology Extraction from Specialized Comparable Corpora. In *Proceedings of the 8th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. pp. 685–693.

Joachims, T. (2002). *Learning to classify text using support vector machines: Methods, theory and algorithms*. Kluwer Academic Publishers.

Kupiec, J. (1993). An algorithm for finding noun phrase correspondences in bilingual corpora. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics*. pp. 17–22.

Levenshtein, V.I. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10, p. 707.

Morin, E., Daille, B., Takeuchi, K. & Kageura, K. (2008). Brains, Not Brawn: The Use of Smart Comparable Corpora in Bilingual Terminology Mining. *ACM Trans. Speech Lang. Process.*, 7(1), pp. 1:1–1:23.

Nassirudin, M. & Purwarianti, A. (2015). Indonesian-Japanese term extraction from bilingual corpora using machine learning. In *Advanced Computer Science and Information Systems (ICACSIS), 2015 International Conference on*. pp. 111–116.

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

Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. & Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proc. of NAACL*.

Pinnis, M., Ljubešić, N., Stefanescu, D., Skadina, I., Tadić, M. & Gornostaya, T. (2012). Term extraction, tagging, and mapping tools for under-resourced languages. In *Proceedings of the 10th Conference on Terminology and Knowledge Engineering (TKE 2012), June*. pp. 20–21.

Rapp, R. (1999). Automatic Identification of Word Translations from Unrelated English and German Corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*. pp. 519–526.

Repar, A., Martinc, M. & Pollak, S. (2019). Reproduction, replication, analysis and adaptation of a term alignment approach. *Language Resources and Evaluation*, pp. 1–34.

Steinberger, R., Eisele, A., Klocek, S., Pilos, S. & Schlüter, P. (2013). DGT-TM: A freely available translation memory in 22 languages. In *Proceedings of the 8th international conference on Language Resources and Evaluation (LREC'2012)*.

Steinberger, R., Pouliquen, B. & Hagman, J. (2002). Cross-lingual document similarity calculation using the multilingual thesaurus eurovoc. *Computational Linguistics and Intelligent Text Processing*, pp. 101–121.

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In N.C.C. Chair), K. Choukri, T. Declerck, M.U. Dogan, B. Maegaard, J. Mariani, J. Odijk & S. Piperidis (eds.) *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA).

Vintar, Š. (2010). Bilingual term recognition revisited: The bag-of-equivalents term alignment approach and its evaluation. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 16(2), pp. 141–158.

Identifying Metadata-Specific Collocations in Text Corpora

Ondřej Herman[1,2], Miloš Jakubíček[1,2], Vojtěch Kovář[1,2]

[1]Natural Language Processing Centre
Faculty of Informatics, Masaryk University
Brno, Czech Republic
E-mail: xherman1@fi.muni.cz, jak@fi.muni.cz, xkovar3@fi.muni.cz

[2]Lexical Computing
Brno, Czech Republic
E-mail: ondrej.herman@sketchengine.eu, milos.jakubicek@sketchengine.eu, vojtech.kovar@sketchengine.eu

Abstract

Statistical corpus analysis of collocations is one of the important steps in creating a dictionary entry: collocations may distinguish senses, describe typical phrasemes and idioms and outline the whole picture of a word's behaviour. However, some collocations are domain-specific, typical only in particular contexts, and thus far there has been no easy way to distinguish "general" collocations from those that are predominantly typical in particular domains.
In this paper, we present a tool which allows lexicographers to see typical domains in which a particular collocation occurs. We introduce a statistical procedure based on corpus metadata to identify domain-specific collocations in an intuitive way, and we also present a user interface connected to the word sketch feature of the Sketch Engine corpus interface (Kilgarriff et al., 2014a).
The new feature can be used in the manual inspection of collocation lists, as well as when using the API or in a semi-automatic post-editing scenario of building a dictionary.

Keywords: collocations; word sketch; meta-data; text types; corpus

1. Introduction

Word sketches (Kilgarriff et al., 2014a) are an intuitive and intelligible summary of a word's collocational behaviour; they have been used in lexicography for nearly 20 years. However, additional information for some of the collocations is sometimes needed.

One of the missing pieces of information is whether a particular collocation is evenly distributed within the corpus, or somehow specific to a particular text type, or even found exclusively in a particular text type. By text type, we understand any type of metadata annotation available within the corpus: web domain, genre, topic, year of publication, author of the text, etc.

This paper addresses the possibilities of adding text type information into lists of collocations such as word sketches. After a discussion of various possible approaches, we select two types of information that may be beneficial for users and show how it can be presented to the users in the Sketch Engine interface and in the API.

We also describe the practical implementation of this new feature within Sketch Engine and discuss some particular advantages and potential problems. Finally, we introduce the compilation of new word sketch indexes that enable this feature and briefly discuss its efficiency.

2. Related Work

Corpus meta-data, as well as collocations, have been used in countless projects and it would make no sense to try to list them all. For example, (Sharoff et al., 2014) used log-likelihood statistics to extract candidates for multiword dictionary entries. The Word sketch itself, with its default *logDice* score (Rychlý, 2008), has been intensively used since its introduction in 2004 (Kilgarriff et al., 2014a).

Corpus meta-data information has also been used widely. Corpora and subcorpora of different domains have been compared (Kilgarriff, 2009; Kilgarriff et al., 2014b) to obtain domain-specific headword lists suitable for specialised dictionaries, and the automatic generation of dictionary labels using corpus meta-data has been proposed and implemented (Rundell & Kilgarriff, 2011).[1] However, all of this has only been suggested on the word (or term) level. Similar computations have, to the best of our knowledge, never been suggested on the level of collocations, which is what we propose in this paper. The statistics for collocations need to be different from single-word meta-data usage, as the expected usage will be different – we do not need a list of most domain-specific collocations, but we do need to mark all collocations that are likely to be domain-specific.

2.1 Meta-Data and Collocations

To the best of our knowledge, there is no corpus tool capable of adding meta-data information into lists of collocations. However, the statistics presented in the folling sections more or less just play with relative frequencies within particular text types, and specify conditions under which observation of these relative frequencies is interesting.

Of course, finding the frequency distribution of a given collocation across text types was possible before: for example in Sketch Engine it was possible to create a concordance for a specific item in word sketch, and to create a text type frequency distribution for this collocation that contains relative frequencies in particular text types, as illustrated in Figure 1. In that case it reveals that "oil spill" is more than 3x more frequent in *W_misc* and *W_non_ac_polit_law_edu*, than in the rest of the corpus – which may be an interesting item of information.

However, this process is very time-consuming and we cannot expect anyone to investigate such a frequency distribution for all collocations in a word sketch. Instead, we let the computer do it, and we set conditions under which a collocation is highlighted as specific for a particular text type. That will give lexicographers easy access to information they probably did not access previously.

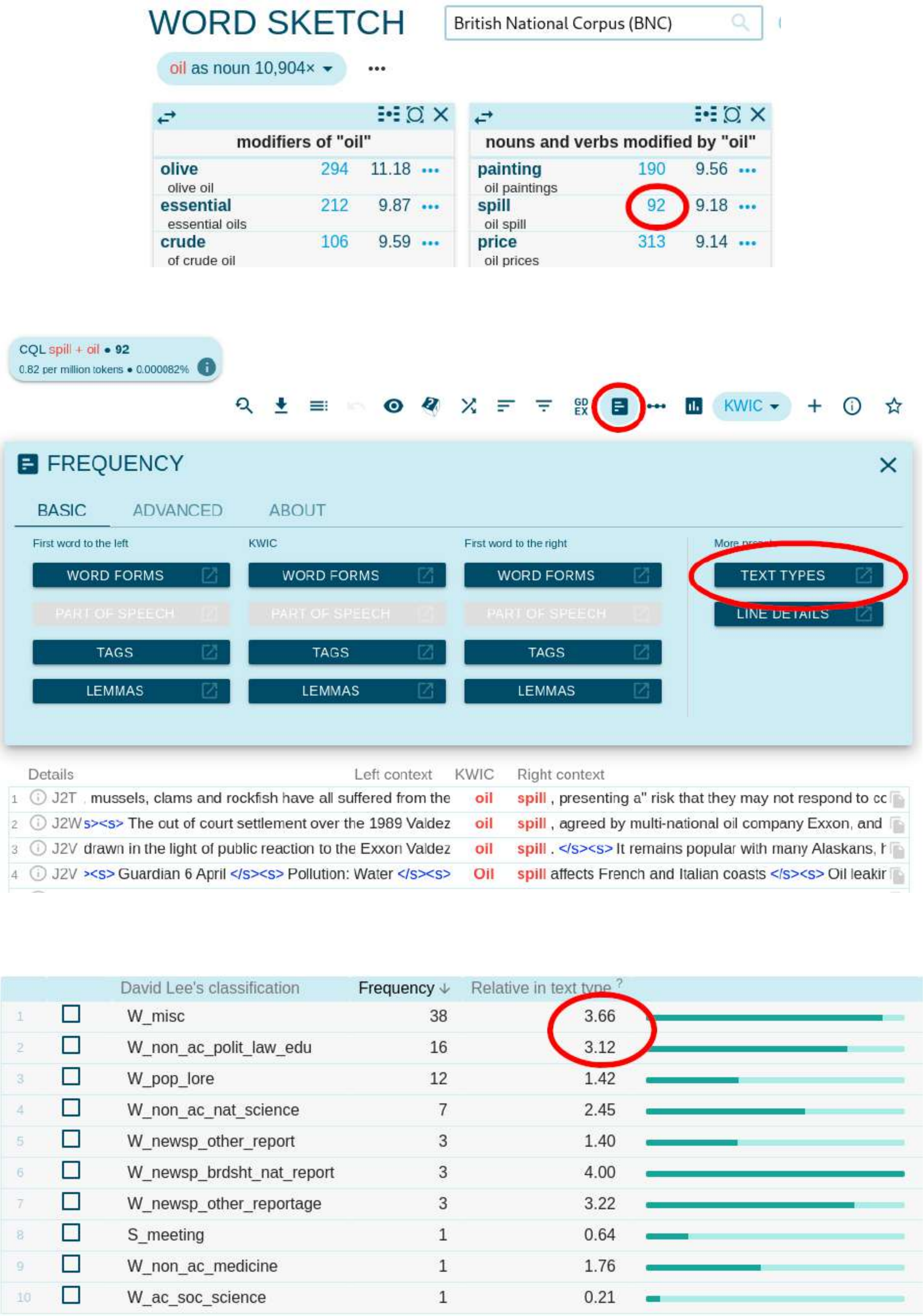
3. The Evolution of the Idea

In the following text, let us think about a particular collocation **C** (e.g. *good news*), and a particular text type **T** (e.g. *genre: newspaper*). Let us suppose that **C** occurs **N** times in the whole corpus, and **M** times in the text type **T**.

3.1 Initial Idea

We started with a very rough simple idea: if a substantial majority of collocation **C** occurs in **T**, we should report it to the user. For example, if 70% of **C** falls into **T** (or $M/N \geq 0.7$),

[1] However, the automation of dictionary labels does not seem to be intensively used, perhaps due to the lack of useful corpus meta-data, no clear general conception of dictionary labels, or the low accessibility of the related features in the corpus tools.



WORD SKETCH British National Corpus (BNC)

oil as noun 10,904×

modifiers of "oil"		
olive olive oil	294	11.18
essential essential oils	212	9.87
crude of crude oil	106	9.59

nouns and verbs modified by "oil"		
painting oil paintings	190	9.56
spill oil spill	92	9.18
price oil prices	313	9.14

CQL spill + oil • 92
0.82 per million tokens • 0.000082%

FREQUENCY

	Details	Left context	KWIC	Right context
1	J2T	mussels, clams and rockfish have all suffered from the	oil	spill , presenting a" risk that they may not respond to cc
2	J2W	s><s> The out of court settlement over the 1989 Valdez	oil	spill , agreed by multi-national oil company Exxon, and
3	J2V	drawn in the light of public reaction to the Exxon Valdez	oil	spill . </s><s> It remains popular with many Alaskans, h
4	J2V	><s> Guardian 6 April </s><s> Pollution: Water </s><s>	Oil	spill affects French and Italian coasts </s><s> Oil leakir

		David Lee's classification	Frequency ↓	Relative in text type ?
1	□	W_misc	38	3.66
2	□	W_non_ac_polit_law_edu	16	3.12
3	□	W_pop_lore	12	1.42
4	□	W_non_ac_nat_science	7	2.45
5	□	W_newsp_other_report	3	1.40
6	□	W_newsp_brdsht_nat_report	3	4.00
7	□	W_newsp_other_reportage	3	3.22
8	□	S_meeting	1	0.64
9	□	W_non_ac_medicine	1	1.76
10	□	W_ac_soc_science	1	0.21

Figure 1: Finding the relative meta-data frequencies of a collocation.

we would say that **C** "usually occurs in" **T**. Or, if 99% of **C** belongs to **T** ($M/N \geq 0.99$), we would tell the user that **C** "only occurs in" **T**. Actually, this method has been built into Sketch Engine for years, it was just not directly visible in the interface.

However, there are significant problems with this simple approach.

It would work well if all of the text types in the corpus were the same size. But if **T** covers a substantial part of the corpus — e.g. 90%, like *Publication date: 1985-1993* in the British National Corpus, BNC (Leech, 1992) — then it is absolutely normal and expectable that the majority of the occurrences of **C** will fall into this text type. The vast majority of all the collocations would probably exceed some 70% threshold and we would report that almost all the collocations "usually occur in **T**". Such information is more or less useless.

On the other hand, if, e.g. half of the occurrences of **C** fall into a small text type (such as *Publication date: 1960-1974* in the BNC, covering only 1.2% of the corpus), it is definitely something interesting and users will want to know. However, our simple method would miss it.

3.2 Including the Text Type Size

It is clear that we need to include the text type size into the computation. Let us suppose that text type **T** covers **P** percent of the corpus text.

As the naive approach from the previous section works well if all the text types are the same size, we thought about a statistical correction that would use a weighting of the occurrences within particular text types, in order to virtually make all of them the same size. We normalised the raw number of hits using the percentage of the corpus covered by the text type, and compared these normalised numbers with their sum. In other words, we used $\mathbf{M/P}$ for all the text types instead of **M**, and the sum of all these fractions instead of **N**. Let us call this sum $\mathbf{N}_{corrected}$.

This approach, however, is problematic in another set of cases, as we noticed shortly. If **T** is small (such as regarding the *Publication date: 1960-1974* in the BNC, $P = 1.2\%$), the normalisation will end up with an unwanted result: imagine two text types **T1** and **T2**, the first covering 99% of the text and containing 45 out of 50 occurrences of **C**. Then $P1 = 99\%, P2 = 1\%, M1 = 45, M2 = 5$. The normalised frequencies are $M1/P1 = 45, M2/P2 = 500$. $N_{corrected} = 545$, so **T2** contains 500/545 = 92% of the corrected occurrences and we would report that **C** "usually occurs in **T2**". But this does not correspond to the real distribution; **T2** contains only 5 of 50 occurrences and "**C** usually occurs in **T2**" is very misleading information.

Another problematic case is when we have two small text types, **T1** and **T2**, both covering e.g. 5% of the corpus ($P1 = P2 = 5\%$). Collocation **C** occurs in both of them with the same frequency (e.g. 30), and never outside these two text types — i.e. $M1 = M2 = 30, N = 60$. Then $M1/P1 = M2/P2 = 30/0.05 = 600, N_{corrected} = 1{,}200$. Neither of the two text types will be mentioned because the corrected ratio for both of them is 50%, which will not exceed the threshold. We will not say anything but that the initial situation is very interesting — **C** only occurs in 10% of the corpus! — so not saying anything is clearly wrong.

3.3 Expected vs. Observed

The last mentioned situation made us rethink the idea of saying "usually in **T**" or "only in **T**": sometimes we have two or more significant text types to report, and none of these messages describes the situation correctly. We came to the conclusion that, in specified cases, we need to say "especially in **T**" which would mean that the collocation is *more often found in this text type than in the others.*

What does this mean? To avoid the problematic results mentioned in the previous section, we used the concept of *expected* and *observed* occurrences of collocation **C**. The expected number of occurrences means, how many hits we would expect in this text type, according to the number of hits in the whole corpus. In other words, $M_{exp} = N * P$. Then we contrast this number with the *observed* **M**. If the observed **M** is significantly higher, we would say "**C** occurs especially in **T**".

3.4 Statistical Significance

Significantly higher in the previous sentence should definitely incorporate statistical significance. For our purposes, however, it is crucial that the information provided to users can be explained easily. And in pure hypothesis testing, we usually do not get easily explainable numbers: How to communicate to the user that e.g. an increase 1,000→1,100 (i.e. 10%) is statistically significant, whereas 40→60 (i.e. 50%) may not be? Especially when we only want to provide an extremely simple message "**C** occurs especially in **T**" – we want users to have some clear idea behind this message.

In addition to that, it has recently been argued (Kilgarriff, 2005; Koplenig, 2019) that statistical significance is not the right measure in corpus linguistics, because

- language is not random and therefore does not fulfil the assumptions of statistical hypotheses testing,
- therefore, if we have enough data, almost everything becomes statistically significant,
- therefore measuring statistical significance means only measuring if we have enough data, and it is not a good base for estimating what is linguistically interesting.

For these two reasons, we decided to employ a simple, explainable criterion: if observed **M** is at least twice as big as the expected $\mathbf{M}_{exp}$, we will show that "**C** occurs especially in **T**". To avoid reporting random noise, we added the following thresholds that must be met in order to display the message:

- the minimum total frequency of the collocation (**N**) is 20
- the minimum $\mathbf{M}_{exp}$ is 5

The minimum thresholds still ensure statistical significance with $p < 0.05$, using the binomial test.

3.5 Usually and Only

In the previous two sections, we specified some notable criteria and decided to mark them by telling the user “**C** occurs especially in **T**”. However, we did not abandon the idea of marking “usually” and “only” along with “especially”. We just returned back to their original, naive meaning.

For “usually” and “only”, we use absolute frequencies, the uncorrected number of hits, to ensure that the words really mean the same to the system and to the user. If absolute frequency in text type **T** stands for more than 70% of the occurrences of the collocation’s overall frequency, we indicate “**C** occurs usually in **T**”. If it is more than 97%, we show “**C** occurs only in **T**”. (These two thresholds are arbitrary, as agreed with initial users of this new feature.)

However, we will show the message under this condition only if **T** is not a dominant text type, i.e. only if it covers less than 50% of the corpus – this is to avoid the problematic scenario with *Publication date: 1985-1993* described above. For dominant text types (covering more than 50% of the corpus), we can still show “usually” and “only” but the conditions are different:

- absolute frequency in text type **T** stands for more than 70% (97%) of the occurrences of the collocation’s overall frequency,
- the minimum expected frequency $\mathbf{M}_{exp}$ in the rest of the corpus is 20,
- the observed frequency in the rest of the corpus is less than 20% of $\mathbf{M}_{exp}$.

In other words, we report “usually” and “only” for the dominant text type only if the frequency in the rest of the corpus is much lower than expected.

4. Specification

In less detail, we want to inform word sketch users about three types of the collocation’s specificity:

1. The collocation is *only* present in a particular text type, and (nearly) not at all in the others. We show “only **T**” if more than 97% of the collocation’s occurrences (in absolute numbers) falls into text type **T**.
2. Most of the collocation occurrences fall into a particular text type, i.e. the text type is dominant for the collocation but not for the whole corpus. We show “usually **T**” if more than 70% (but less than 97%) of the collocation’s occurrences falls into text type **T**. (There are separate rules for the dominant text type, see the previous section.)
3. The relative frequency of a particular collocation in a particular text type is much higher than the relative frequency of that collocation in the whole corpus. We show “especially **T**” if the collocation’s *relative* frequency in text type **T** is at least twice as high as its relative frequency in the whole corpus.

These three characteristics are now part of the word sketch interface, if compiled. We describe the compilation procedure and the user interface in the following sections.

5. Implementation

5.1 Compilation

The statistics are computed at the time of corpus compilation and are instantly available in the word sketch database indexes. To save the numbers for each collocation, we had to change the format of the word sketch indexes. The resulting data are slightly larger, for the BNC with 3 different text types ("Text type", "Publication date" and "David Lee's classification") the increase was 22% (1.03GB→1.25GB). The additional compilation time was 13 minutes.

Of course, these numbers depend on various details (sketch grammar, the number of text types included, the distribution of text types within the corpus etc.) and cannot be generalised; they are rather illustrative.

The compilation program is written in the Go programming language.

5.2 User Interface

The notes "only", "usually", and "especially" are displayed in the standard word sketch interface under the particular collocations. Depending on the sketch grammar, the number of text types and their distribution in the corpus, they can take up a lot of space on user's screen – therefore they can be turned off. We have also considered an option where they are displayed on mouseover or after clicking a small icon, but this is so far only a matter for future development.

Another idea for future development is the option to filter the word sketch by the metadata labels, or by *always/usually/especially.* This is likely to appear in the interface soon.

The notes are also available in the Sketch Engine REST API, so that external tools can benefit from this new feature.

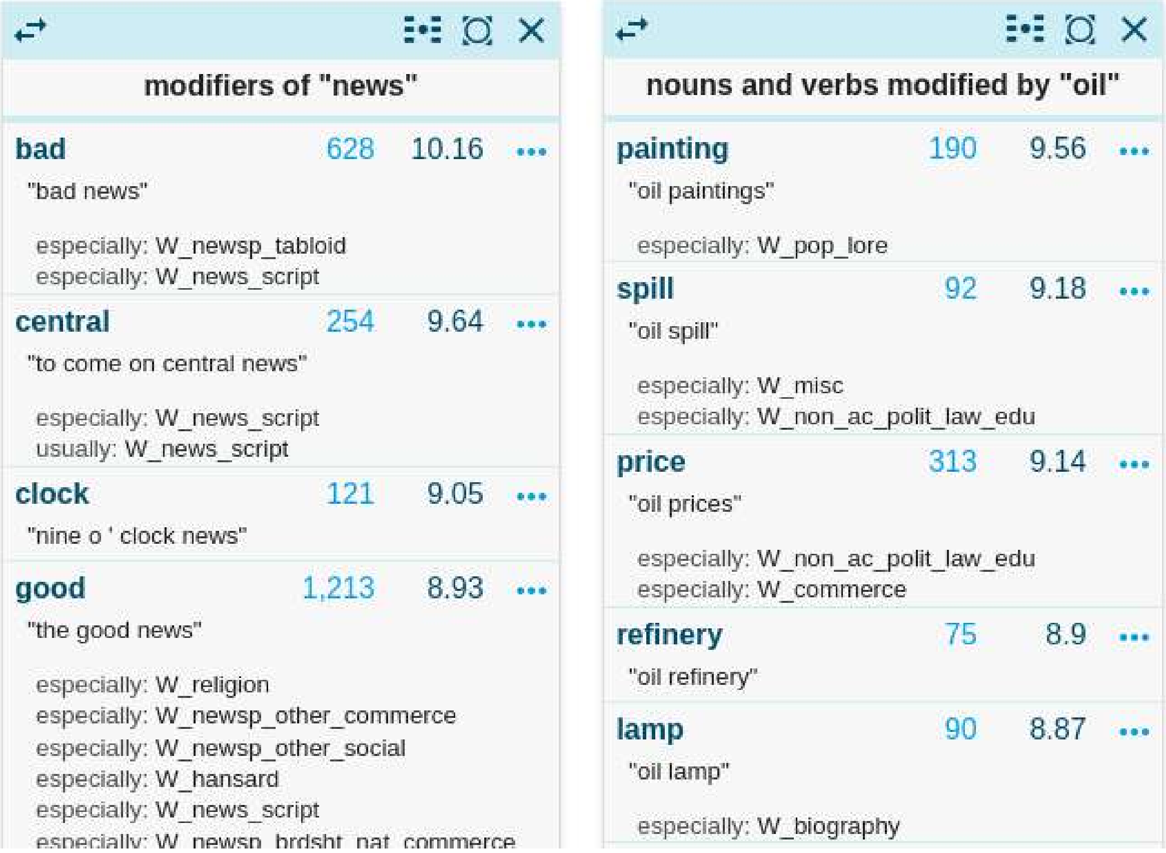
6. Lexicographic Potential

Of course, the new feature can be used in lexicographical work – the text types in the corpus may provide useful insights leading to dictionary labels for particular collocations, or even for whole entries:

- **Revealing metadata-specific senses.** Collocations are often used to describe different senses of the headword. If we notify the lexicographer that a particular collocation is domain-specific, it may lead to a useful dictionary label for the particular sense (e.g. *American English* or *legal texts*, depending on the available meta-data).
- **Richer information on collocations.** Dictionaries often include typical collocations and examples of the headword. Now it is easy to add more information to these particular collocations, e.g. *black hole (astronomy).*
- **Pre-generating label candidates.** In post-editing lexicography, which is becoming increasingly popular, it can be used directly for suggesting the labels. The collocations can be exported from the corpus into a dictionary writing system, together with the meta-data information, and a lexicographer can only edit the collocations and the labels – which will result in richer dictionaries with less work.

7. Examples

Figure 2 shows two examples of metadata-specific collocations, as can be newly identified in word sketches. Both examples use the British National Corpus and David Lee's classification (Lee, 2002).



modifiers of "news"

bad 628 10.16 •••
"bad news"
especially: W_newsp_tabloid
especially: W_news_script

central 254 9.64 •••
"to come on central news"
especially: W_news_script
usually: W_news_script

clock 121 9.05 •••
"nine o ' clock news"

good 1,213 8.93 •••
"the good news"
especially: W_religion
especially: W_newsp_other_commerce
especially: W_newsp_other_social
especially: W_hansard
especially: W_news_script
especially: W_newsp_brdsht_nat_commerce

nouns and verbs modified by "oil"

painting 190 9.56 •••
"oil paintings"
especially: W_pop_lore

spill 92 9.18 •••
"oil spill"
especially: W_misc
especially: W_non_ac_polit_law_edu

price 313 9.14 •••
"oil prices"
especially: W_non_ac_polit_law_edu
especially: W_commerce

refinery 75 8.9 •••
"oil refinery"

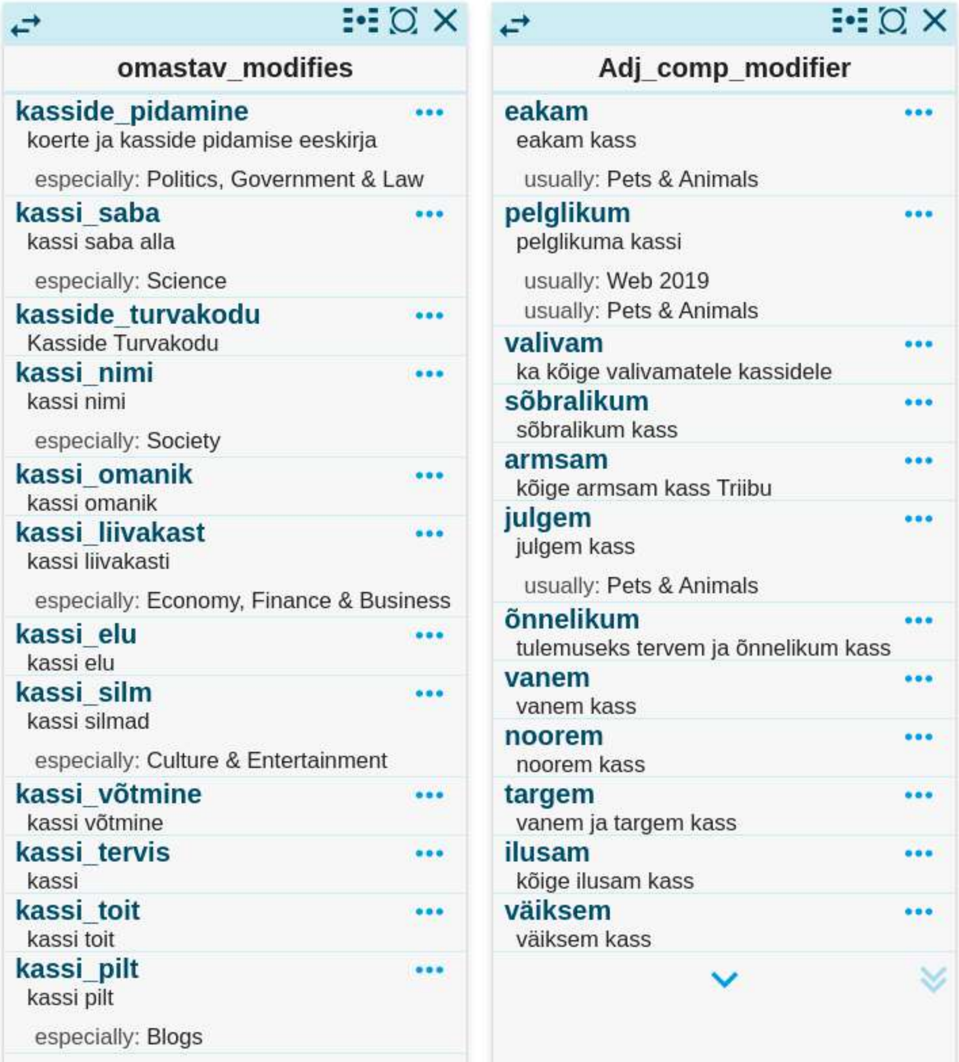
lamp 90 8.87 •••
"oil lamp"
especially: W_biography

Figure 2: Examples of metadata-specific collocations in the British National Corpus

The first one is a fragment of a word sketch for "news" and shows that *bad news* is specific to tabloid newspapers and TV autocue scripts, whereas *good news* occurs mostly in religious and commercial texts and a variety of other genres.

The second fragment shows the genre-specific collocations of the word "oil": *oil paintings* occurs most frequently in popular magazines, *oil lamps* in biographies, *oil prices* and *oil spills* are political topic,s and *oil prices* is also important in financial texts (*oil spills* is not). *Oil refineries* is covered evenly within all the text types.

Figure 3 shows another example and different text types in the Estonian National Corpus. The example is a fragment of a word sketch for "*kass*" (cat) and shows, for example "*koerte ja kasside pidamise eeskiri*" (rules for keeping dogs and cats) being typical in *Politics, Government & Law*, "*kassi silmad*" (cat eyes) being typical in *Culture & Entertainment* or "*julgem kass*" (braver cat) being predominantly present in *Pets & Animals*.



omastav_modifies

kasside_pidamine
koerte ja kasside pidamise eeskirja
especially: Politics, Government & Law

kassi_saba
kassi saba alla
especially: Science

kasside_turvakodu
Kasside Turvakodu

kassi_nimi
kassi nimi
especially: Society

kassi_omanik
kassi omanik

kassi_liivakast
kassi liivakasti
especially: Economy, Finance & Business

kassi_elu
kassi elu

kassi_silm
kassi silmad
especially: Culture & Entertainment

kassi_võtmine
kassi võtmine

kassi_tervis
kassi

kassi_toit
kassi toit

kassi_pilt
kassi pilt
especially: Blogs

Adj_comp_modifier

eakam
eakam kass
usually: Pets & Animals

pelglikum
pelglikuma kassi
usually: Web 2019
usually: Pets & Animals

valivam
ka kõige valivamatele kassidele

sõbralikum
sõbralikum kass

armsam
kõige armsam kass Triibu

julgem
julgem kass
usually: Pets & Animals

õnnelikum
tulemuseks tervem ja õnnelikum kass

vanem
vanem kass

noorem
noorem kass

targem
vanem ja targem kass

ilusam
kõige ilusam kass

väiksem
väiksem kass

Figure 3: Examples of metadata-specific collocations in the Estonian National Corpus (Estonian NC 2019)

8. Conclusion

In this paper, we introduced a procedure for including text type information into collocation summaries, such as word sketches. We explained the mental process that ended up with the current specification, then we outlined the implementation, described the user interface and illustrated the output with examples.

The newly introduced functionality is still in its early stage of existence; so far it has only limited production use and has not yet been tested on a large scale. Therefore, some of the parameters may change slightly in the future.

However, we can say that – as in most of the cases concerning corpus data – the future usability of the new feature depends on the quality of the data: the text type annotation, the selection of the right text types to be shown in the word sketch, the corpus having a decent size, as well as the size of particular text types. The quality of the language data in general is one of the biggest challenges for computational linguistics and semi-automatic lexicography in the coming years.

9. Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

This work has been partly supported by the Ministry of Education of CR within the LINDAT-CLARIAH-CZ project LM2018101.

10. References

Kilgarriff, A. (2005). Language is never, ever, ever, random. *Corpus Linguistics and Linguistic Theory*, 1(2), pp. 263–276.

Kilgarriff, A. (2009). Simple maths for keywords. In *Proceedings of the Corpus Linguistics Conference. Liverpool, UK.*

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. (2014a). The Sketch Engine: ten years on. *Lexicography*, 1. URL http://dx.doi.org/10.1007/s40607-014-0009-9.

Kilgarriff, A., Jakubíček, M., Kovář, V., Rychlý, P. & Suchomel, V. (2014b). Finding terms in corpora for many languages with the Sketch Engine. *EACL 2014*, p. 53.

Koplenig, A. (2019). Against statistical significance testing in corpus linguistics. *Corpus Linguistics and Linguistic Theory*, 15(2), pp. 321–346. URL https://doi.org/10.1515/cllt-2016-0036.

Lee, D. (2002). Genres, Registers, Text Types, Domains, and Styles: Clarifying the Concepts and Navigating a Path through the BNC Jungle. *Language Learning and Technology*, 5.

Leech, G. (1992). 100 million words of English: the British National Corpus (BNC). *Language Research*, 28(1), pp. 1–13.

Rundell, M. & Kilgarriff, A. (2011). Automating the creation of dictionaries: Where will it all end? *A Taste for Corpora. In Honour of Sylviane Granger*, pp. 257–282.

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

Sharoff, S., Umanskaya, E. & Wilson, J. (2014). *A frequency dictionary of Russian: Core vocabulary for learners.* Routledge.

Porting the Latin WordNet onto OntoLex-Lemon

Stefania Racioppa[1], Thierry Declerck[1,2]

[1]German Research Center for Artificial Intelligence (DFKI)
Saarland Informatics Campus D3 2, Saarbrücken, Germany
[2]Austrian Centre for Digital Humanities and Cultural Heritage
Sonnenfelsgasse 19, Wien 1010, Austria
E-mail: stefania.racioppa@dfki.de, declerck@dfki.de

Abstract

In this paper we describe the porting of the Latin WordNet data available at the University of Exeter onto the OntoLex-Lemon model, focusing on the representation of both morphological and conceptual information. In the longer term, we aim at integrating the resulting data set in the Linguistic Linked Open Data (LLOD) infrastructure, linking (or even merging) it to the Latin data sets already published in the LOD framework by the ERC "Linking Latin" (LILA) project. We discuss some lessons learned, as it turned out that such a transformation and linking exercise can lead to an improved consistency and accuracy of the original data.

Keywords: Latin; WordNet; Morphology; OntoLex-Lemon

1. Introduction

In our work, we are concerned with the transformation of heterogeneous digital lexical resources, available in a multitude of formats, into a harmonised representation in the context of the OntoLex-Lemon model,[1] which is briefly introduced in Section 3 of this paper. Besides mainstream language resources, we are also dealing with ancient and low-resourced languages, as we are aiming at contributing to the improved access to such resources, which could further support the deployment of language technologies in the broader field of digital humanities.

Our first steps in dealing with Latin language data consisted in mapping the Latin WordNet available at the University of Exeter[2] onto the OntoLex-Lemon model. We set the main focus on the semantic representation of both the morphological and conceptual information encoded in the Latin WordNet. In this context, we are also starting a cooperation with the ERC project "LiLa" (*Linking Latin Building a Knowledge Base of Linguistic Resources for Latin*)[3] on the harmonisation of the semantic representation of Latin language data for their optimal publication on the Linguistic Linked Open Data cloud.

In the following sections, we introduce first the Latin WordNet data of the University of Exeter, and describe then briefly the Linguistic Linked Open Data cloud as well as the OntoLex-Lemon representation model for lexical data. We continue with the presentation of the first results of the mapping of the Latin WordNet data onto OntoLex-Lemon, comparing them with the Latin data already ported to the Linked Open Data by the LILA project. We close with the discussions of some lessons learned.

2. Latin WordNet at the University of Exeter

The Latin WordNet initiative at the University of Exeter "builds on, and extends, the original Latin WordNet developed as part of the Fondazione Bruno Kessler's

[1] https://www.w3.org/2016/05/ontolex/. See also (Cimiano et al., 2016).

[2] https://latinwordnet.exeter.ac.uk/. See also (Fedriani et al., 2020).

[3] https://lila-erc.eu/. See also (Mambrini & Passarotti, 2019) and the Latin Lemma Bank Query Interface of the LiLa project, available at https://lila-erc.eu/query/.

MultiWordNet project"[4] and is developed in the context of a cooperation, among others, with the University of Genoa[5] and the LiLa project.[6]

Periodically updated versions of the Latin WordNet are available in two formats (JSON and CSV) in a GitHub repository.[7] The data is distributed over distinct files for different categories, from which we considered the files displaying information on the lemmas, literal senses and synsets.

After working on the CSV data set of January 2020,[8] we communicated some issues we found in the source data to the resource developer. In a second step, we worked on the CSV data set of October 2020.[9] Also in this case, the communication with the developer was essential to solve the remaining open issues.

Concerning the lemma information associated with the synsets, both data sets differ slightly in their layout. The January version included the following columns in the files containing the lemmas: *ID*, *URI*, the lemma itself, *part of speech*, *morphological information*, *principal parts*, *irregular forms*, *alternative forms*, *IPA phonetic representation*, *prosody*, and *validation id*. In the October data set, the irregular and alternative forms, as well as the phonetic representation have been dropped, and the column order was changed. The information included in the distinct files are described in detail in Listings 2.1, 2.2 and 2.3.

In Listing 2.1, displaying the lemma "abdicatio" (in the version of 2020.10.10), we can see the lexical and morphological information associated with the lemma, that needs to be represented in OntoLex-Lemon. The lemma is included in the file "lemma_0.csv" with the ID "19117". This ID is present three times in the file "literalsense_0.csv", indicating that the lemma has 3 senses pointing to the synsets "136508", "136706" and "104057", which are included in the file "synset_10.csv", and "synset_0.csv".

Listing 2.3 displays the information associated with the synsets, where the glosses are from the Princeton WordNet,[10] while in Listing 2.2 we can see how the synsets are related to the lemmas by the use of their respective IDs.

```
id,uri,lemma,pos,validated,morpho,principal_parts,prosody
19117,a0031,abdicatio,n,1,n–s——fn3–,abdication,abdicatio
```

Listing 2.1: The entry *abdicatio* in the 2020-10-10 data set

```
id,lemma,synset,period,genre,notes
2,19117,136508,,,
3,19117,136706,,,
4,19117,104057,,,
```

Listing 2.2: The literal senses for the lemma *abdicatio* in the 2020-10-10 data set

[4] Quoted from https://latinwordnet.exeter.ac.uk/. See also (Fedriani et al., 2020).

[5] This cooperation is documented, for example, in (Fedriani et al., 2020).

[6] The "Linked Latin" (LILA) is a project funded by the European Research Council (ERC). See https://lila-erc.eu/ for more details. See also (Passarotti et al., 2019; Mambrini & Passarotti, 2019).

[7] https://github.com/latinwordnet/latinwordnet-archive/tree/master/csv/.

[8] https://github.com/latinwordnet/latinwordnet-archive/tree/master/csv/2020-01-31.

[9] https://github.com/latinwordnet/latinwordnet-archive/tree/master/csv/2020-10-10.

[10] https://wordnet.princeton.edu/. See also (Fellbaum, 1998).

```
id,offset,pos,language,gloss,semfield
136508,05385235,n,10,refusal to acknowledge as one's own,
136706,05414335,n,10,a verbal act of renouncing,
104057,00134568,n,10,the act of renouncing,
```

Listing 2.3: The synsets to which the literal senses for the lemma *abdicatio* are pointing to (in the 2020-10-10 data set)

3. OntoLex-Lemon

The OntoLex-Lemon model, which results from a W3C Community Group,[11] was originally developed with the aim to provide a rich linguistic grounding for ontologies, meaning that the natural language expressions used in the labels, definitions or comments of ontology elements are equipped with an extensive linguistic description.[12] This rich linguistic grounding includes the representation of morphological and syntactic properties of lexical entries as well as their syntax-semantics interface, i.e. the meaning of these lexical entries with respect to an ontology or to specialised vocabularies.

The main organising unit for those linguistic descriptions is the *LexicalEntry* class, which enables, among other things, the representation of morphological patterns for each entry (a multi-word expression, a word or an affix). The connection of a lexical entry to an ontological entity is marked mainly by the *ontolex:denotes* property or is mediated by the *LexicalSense* or the *LexicalConcept* classes, as this is represented in Figure 1, which displays the core module of the model.

OntoLex-Lemon builds on and extends the preceding *lemon* model (McCrae et al., 2012). A major difference is that OntoLex-Lemon includes an explicit way to encode conceptual hierarchies, using the SKOS[13] standard. As can be seen in Figure 1, lexical entries can be linked, via the *ontolex:evokes* property, to such SKOS concepts, which can represent WordNet synsets. This structure aligns the relation between lexical entries and ontological resources, which is implemented either directly by the *ontolex:reference* property or mediated by the instances of the *ontolex:LexicalSense* class.

More recently, OntoLex-Lemon has been used also as a de facto standard in the field of digital lexicography and is being applied for example in the European infrastructure project ELEXIS (European Lexicographic Infrastructure).[14]

4. Representation of the Latin WordNet Lemmas in OntoLex-Lemon

The modelling of the linguistic information from the Latin WordNet data within OntoLex-Lemon took into consideration the recent morphology module, currently under (advanced) discussion within the W3C "Ontology-Lexica" Community Group[15], in which

[11] See https://www.w3.org/2016/05/ontolex/.

[12] See (Cimiano et al., 2016).

[13] SKOS stands for "Simple Knowledge Organization System". SKOS provides "a model for expressing the basic structure and content of concept schemes such as thesauri, classification schemes, subject heading lists, taxonomies, folksonomies, and other similar types of controlled vocabulary" (https://www.w3.org/TR/skos-primer/).

[14] See http://www.elex.is/ for more detail.

[15] See (Klimek et al. (2019))

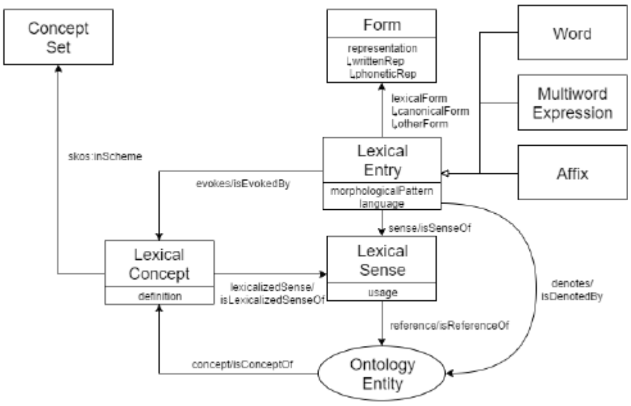


Figure 1: The core modules of OntoLex-Lemon. Graphic taken from https://www.w3.org/2016/05/ontolex/

(among others) also members of the LiLa developer team are actively involved. However, as the discussion is still ongoing, we cannot exclude discrepancies with the most recent model definition.

In a first step, we performed a light data clean-up, i.e. merging split entries, and separating the elements in the column *principal parts* consistently with blanks. In the "cleaned" files, we looked for potential duplicates. While identical entries can just be dropped in the generated OntoLex-Lemon compliant output, in some cases we detected lemmas with the same part of speech, but different genders, or declension groups. As we cannot decide if these are actually errors, homographs with different senses, or if the lemmas really allow for different inflections, we shared our findings with the developers and are currently iteratively adapting and running our transformation process from the updated CSV data onto OntoLex-Lemon.

Analysing the available data in the lemma category of both January and October data sets, we found out that the required morphological features were represented in a quite structured form for each lemma, which includes in the "morpho" column an abbreviated information, i.e. `n-s---fn3-`. This indicates the values for, respectively, part of speech (here: *noun*), adjective degree, number (*singular*), verb tense, mood, and voice, gender (*feminine*), case (*nominative*), declension group (*3rd*), and stem variations (where applicable, e.g. in *abnept-abneptis*: `n-s---fn3i`).

The morphological information was not changed in the latest Latin WordNet version, so that we were able to map the *morpho* value of both January and October data sets into an OntoLex-compatible format using a simple Python script. As a side effect, the script also helped us to highlight and remove the very few errors in the original data.

For the further processing, we split the data by part of speech and converted it in a "readable" CSV/Pandas format, as shown in Listing 4.1 below:[16]

```
,base ,forms ,pos ,number ,gender ,case ,group ,fonipa ,stem ,degree ,person ,tense ,
mood,voice
,abdicatio ,abdication ,noun ,singular ,feminine ,nominative ,3 ,– ,,,,,,

,base ,forms ,pos ,number ,gender ,case ,group ,stem ,degree ,person ,tense ,mood,
voice
```

[16] The phonetic transcription (value *fonipa*) of the January dataset is not displayed in this example.

```
,abdicatio ,abdication ,noun ,singular ,feminine ,nominative ,3 ,,,,,,
```

Listing 4.1: CSV/Pandas output for the entry *abdicatio*
from the 2020-01-31 and 2020-10-10 data set

As the readers can notice, some values — as well as the meaning of the used "codes" — depend partially on the part of speech of the corresponding lemma, and the part of speech is listed separately in a dedicated column. Above this, the Latin declensions can (mostly) be recognised by the ending of the lemma. All these factors helped us in this phase to detect several inconsistencies in the original data, such as a wrong gender or declension groups, or even inconsistent part of speech information.

After processing the January data set, we forwarded our findings to the Latin WordNet developers, and some corrections were implemented in the following versions. The inconsistencies found in the October data set are currently under revision. A first feedback from the developer confirmed that some items were indeed mis-tagged, although the "morpho" fields are mostly correct. However, examining the apparently "duplicated" entries might be more complicated. Some highlighted items seem to be morphological variants, which need to be checked also with respect to the semantic distance between the items. While "real" variants can be merged, it is possible that others mean something different, in which case it would be reasonable to keep them as distinct lemmas.

Also, the *prosody* column plays a relevant role in the lemma disambiguation (e.g. *scŏpa* vs. *scōpa*), and it might be worth including this piece of information in a future OntoLex version of this resource. In general, the Latin WordNet can be seen as "work in progress", so that besides this, further changes might be made in the future.

As the Latin WordNet does not include full forms or the declension tables corresponding to the defined groups, we decided to represent the lemma inflection not as full-form reference, but using the morphological patterns principle described in the OntoLex Morphology Module, which explicitly recommends linking to external sources for such purposes. We found a detailed description of the Latin declension groups in Wikipedia[17] and mapped the declension tables listed there into the OntoLex-Lemon format.

This work resulted in the generation of 73,949 entries (19,999 adjectives, 38,135 nouns, 60 prepositions, 4902 adverbs, 10,854 verbs) from the January data set, and 73,945 entries (19,999 adjectives, 38,130 nouns, 60 prepositions, 4,901 adverbs, 10,855 verbs) from the October data set, as well as 1,219 morphological patterns (192 for nouns, 192 for adjectives and 835 for verbs). However, as the possible inconsistencies we mentioned above are currently under review, the final figures might change in the future.

Listing 4.2 displays the OntoLex-Lemon lemma for "abdicatio" and its forms. The representation is the same for both data sets. However, the IPA phonetic representation was dropped in the latest version.

```
:lex_abdicatio a ontolex:LexicalEntry ;
    lexinfo:gender lexinfo:feminine ;
    lexinfo:partOfSpeech lexinfo:noun ;
    morph:inflects :la-noun_3 ;
    ontolex:canonicalForm :form_abdicatio ;
    ontolex:evokes :a0031 ;
```

[17] https://en.wikipedia.org/wiki/Latin_declension

```
    ontolex:otherForm :form_abdicatio_root .

:form_abdicatio a ontolex:Form ;
    lexinfo:case lexinfo:nominative ;
    lexinfo:number lexinfo:singular ;
    ontolex:writtenRep "abdicatio"@la .

:form_abdicatio_root a ontolex:Form ;
    ontolex:writtenRep "abdication"@la .
```

Listing 4.2: The OntoLex-Lemon representation for *abdicatio* including the "canonical" and the "other" forms

The corresponding morphological pattern and some associated "rules" are displayed in Listing 4.3. In the examples, we can see the entries for the accusative forms, singular (*abdicationem*) and plural (*abdicationes*). The inflections are represented in the "replacement" value as a pattern, using the syntax of regular expressions.

The feature *generates* lists the morphological information related to each inflection. Alternative values (*lexinfo:feminine*, *lexinfo:masculine*) indicate the allowed morphological information, which is disambiguated by the corresponding value in the "main" entry.

```
:la-noun_3 a morph:paradigm ;
    rdfs:comment "Latin 3rd noun declension" .

:la-noun_3_acc_m-f_pl a morph:rule ;
    morph:generates [ lexinfo:case lexinfo:accusative ;
            lexinfo:gender lexinfo:feminine ,
                lexinfo:masculine ;
            lexinfo:number lexinfo:plural ] ;
    morph:paradigm :la-noun_3 ;
    morph:replacement [ morph:source "$" ;
            morph:target "es" ] .

:la-noun_3_acc_m-f_sg a morph:rule ;
    morph:generates [ lexinfo:case lexinfo:accusative ;
            lexinfo:gender lexinfo:feminine ,
                lexinfo:masculine ;
            lexinfo:number lexinfo:singular ] ;
    morph:paradigm :la-noun_3 ;
    morph:replacement [ morph:source "$" ;
            morph:target "em" ] .
```

Listing 4.3: The *la-noun_3* paradigm and some of the associated rules

This way, we are making the morphological information available in a declarative manner.

5. The OntoLex-Lemon Representation of the Synsets of Latin WordNet and their Relations to the Lemmas

The original Latin WordNet corpus includes 107,687 synsets, which are taken from Princeton WordNet. The mapping from the original conceptual data in CSV format onto OntoLex-Lemon was simpler to achieve as for the lexical and morphological data, as there was no need to design paradigms or rules to be included in the target representation format.

Listing 5.1 displays an example of a synset, encoded as an instance of the *LexicalConcept* class, and the way it is related to the instances of the *LexicalEntry* class that "evokes" it.

```
:LexicalConcept_134535
  skox:definition "a line drawn on a map connecting points of equal height" ;
  ontolex:isEvokedBy :lex_conputatio ;
  ontolex:isEvokedBy :lex_configuratio ;
  ontolex:isEvokedBy :lex_computatio ;
  ontolex:isEvokedBy :lex_idolon ;
  ontolex:isEvokedBy :lex_efformatio ;
  ontolex:isEvokedBy :lex_sinus ;
  ontolex:isEvokedBy :lex_circumcaesura ;
  ontolex:isEvokedBy :lex_spectrum ;
.
```

Listing 5.1: The OntoLex-Lemon representation for the original synset with id *134535* – including the Princeton WordNet definition and the links to the lexical entries realising the lexical concept

We noticed that many synsets have not yet been related to a Latin word (or lemma). We also discovered that some synsets are on the contrary linked to a multitude of lemmas, like the example in Listing 5.2, which clearly points to an issue in the granularity of the relations between synsets and lemmas in the current version of the data set.

```
:LexicalConcept_134565
  skox:definition "a symbol used to represent a number:
    'he learned to write the numerals before he went to school'" ;
  ontolex:isEvokedBy :lex_numerus ;
  ontolex:isEvokedBy :lex_dessignatio ;
  ontolex:isEvokedBy :lex_plurimus ;
  ontolex:isEvokedBy :lex_auditus ;
  ontolex:isEvokedBy :lex_simplum ;
  ontolex:isEvokedBy :lex_conplus ;
  ontolex:isEvokedBy :lex_carnuficina ;
  ontolex:isEvokedBy :lex_caudex ;
  ontolex:isEvokedBy :lex_compactura ;
  ontolex:isEvokedBy :lex_penecostas ;
  ontolex:isEvokedBy :lex_connubium ;
  ontolex:isEvokedBy :lex_flexio ;
  ontolex:isEvokedBy :lex_quoteni ;
  ontolex:isEvokedBy :lex_reuolutio ;
  ontolex:isEvokedBy :lex_chilias ;
  ontolex:isEvokedBy :lex_aditio ;
  ontolex:isEvokedBy :lex_offa ;
  ontolex:isEvokedBy :lex_cybus ;
  ontolex:isEvokedBy :lex_simulacrum ;
  ontolex:isEvokedBy :lex_infrequentia ;
  ontolex:isEvokedBy :lex_plurimum ;
  ontolex:isEvokedBy :lex_frenus ;
  ontolex:isEvokedBy :lex_binio ;
  ontolex:isEvokedBy :lex_trias ;
  ontolex:isEvokedBy :lex_compar ;
  ....
  ....
```

Listing 5.2: The OntoLex-Lemon representation for the original synset with id *134565* (in the January version) – with a very high number of lemmas that are referred to

6. Linguistic Linked Open Data Cloud

The Linguistic Linked Open Data (LLOD) cloud is an initiative started in 2012 by a working group of the Open Knowledge Foundation.[18] The aim of the initiative was to break the data silos of linguistic data and thus encourage Natural Language Processing (NLP) applications that make use of data from multiple languages and modalities (e.g., lexicon, corpora, etc.). Technologies for representing language data in the LLOD include tools for the discovery, transformation and linking of language data sets which can be applied to both data and metadata, in order to provide multi-portal access to heterogeneous data repositories.

Looking at the current state of the LLOD, displayed in Figure 2, the reader can see that the data sets published in this cloud are classified along the lines of six categories:

- Corpora
- Terminologies, Thesauri and Knowledge Bases
- Lexicons and Dictionaries
- Linguistic Resource Metadata
- Linguistic Data Categories
- Typological Databases

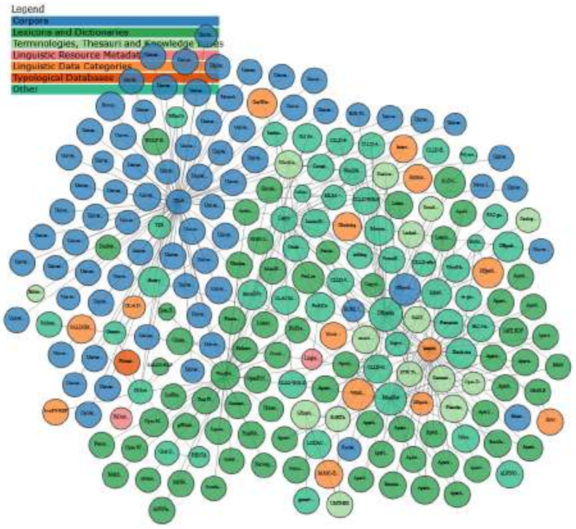

Legend
Corpora
Lexicons and Dictionaries
Terminologies, Thesauri and Knowledge Bases
Linguistic Resource Metadata
Linguistic Data Categories
Typological Databases
Other

Figure 2: The Linguistic Linked Data Cloud

The final goal of our work is to publish as many language data as possible in the LLOD cloud, and to do this, a representation of the data with the Resource Description Framework (RDF) is a prerequisite.

The research community involved in the development of the LLOD cloud aims at increasing the uptake of language technologies also in the broader field of digital humanities and cultural heritage. Dealing with historical languages and porting them to RDF is therefore an important achievement.

[18] See https://linguistic-lod.org/llod-cloud and (McCrae et al., 2016).

The encoding of the Latin in WordNet in RDF and OntoLex-Lemon also allows to establish more precise comparisons with the Latin data already available in the Linked Data framework, resulting from the work pursued by the "Linked Latin" (LiLa) project.[19]

Apart from the different naming of the single features and values, the OntoLex-Lemon representation of our example "abdicatio" (displayed above in Listing 2.1) and the corresponding LiLa lemma (https://lila-erc.eu/data/id/lemma/86857, displayed below in *turtle* format) indeed show a large degree of compatibility: Both have in the "main" entry dedicated values for part of speech and gender definition, as well as a written representation of the lemma itself and a reference to the inflection class.

The main difference between both corpora is how the inflected forms of the lemma are handled. While the OntoLex-Lemon representation just builds a plain reference to the canonical and the "other" form(s) (*abdicatio*, *abdication*), the LiLa representation offers a better analysis of the lemma, because it labels its constituent elements - prefix, radix, and suffix (*ab-*, [base], *-(t)io(n)*). Above this, LiLa adds a reference to the lemma group "dico", which is inflected similarly. Finally, the synset value is a specific feature of the Latin WordNet corpus.

```
<data/id/lemma/86857> a lila:Lemma ;
     rdfs:label "abdicatio" ;
     lila:hasBase <data/id/base/8> ;
     lila:hasGender lila:feminine ;
     lila:hasInflectionType lila:n3 ;
     lila:hasPOS lila:noun ;
     lila:hasPrefix <data/id/prefix/1> ;
     lila:hasSuffix <data/id/suffix/2> ;
     ontolex:writtenRep "abdicatio" .

<data/id/base/8> a lila:Base ;
     rdfs:label "Base of dico" .

<data/id/prefix/1> a lila:Prefix ;
     rdfs:label "a(b)-" .

<data/id/suffix/2> a lila:Suffix ;
     rdfs:label "-(t)io(n)" .
```

Listing 6.1: LiLa lemma representation for "abdicatio" in *turtle* format

For this reason, both data sets could be put in relation by using the OntoLex-Lemon element they have in common: the value of the *ontolex:writtenRep* property. It would also be straightforward to establish a mapping between the LiLa properties expressing the morphological information and the corresponding properties of the LexInfo vocabulary,[20] which are used in OntoLex-Lemon. This way, we could detect which elements are only in one of the data sets, or if inconsistencies are present in describing one and the same phenomenon.

Last but not least, we could suggest the merging of (compatible) pieces of information. Just to mention a few examples, we could share the value of the associated synset from the

[19] Repeating a former footnote for the convenience of the reader: https://lila-erc.eu/. See also (Mambrini & Passarotti, 2019) and the Latin Lemma Bank Query Interface of the LiLa project, available at https://lila-erc.eu/query/.

[20] See https://lexinfo.net/ontology/3.0/lexinfo.

OntoLex-Lemon entry (expressed in the property *evokes*) with the LiLa representation of the same lemma. On the other hand, as mentioned above, LiLa offers a more detailed analysis of the lemma decomposition (i.e. the values *hasPrefix* and *hasSuffix*), which would complete the shallow representation of alternative forms in OntoLex-Lemon (i.e. the simple value *otherForm* and its written representation).

While this is work we have ahead of us, it shows perspective for cross-linked or event unified resources for the Latin language.

7. Lessons Learned

Our work on porting the Latin WordNet onto a Linked Data-compliant format has reinforced our conviction that the encoding in such a format is an added value, as the information contained in the original data set is made available in a declarative way, which supports its linking to other sources of information. Here we see particularly the possibility to cooperate with the LiLa project, as the data encoding is really interoperable.

Another added value lies in the fact that such (automated) transformation work helps to detect potential inconsistencies in the original data. We experienced this in both morphological and conceptual aspects of the CSV data we were working with. The new versions of the Latin WordNet could also benefit of the feedback given to the developer. A simple example of small errors in the conceptual domain is the missing of correct data in a column of the CSV file. Something very difficult to find manually, but which causes an error message when running the Python code to generate the OntoLex-Lemon representation.

8. Conclusions

We presented the current state of our work consisting in mapping the Latin WordNet data onto the OntoLex-Lemon model, in order to support its publication in the Linguistic Linked Open Data cloud. This way this type of language resources can be made directly accessible to NLP applications in the field of eLexicography and digital humanities.

The next steps in our work will be directed at a close cooperation with the LiLa project, towards the best possible semantic representation of Latin language data for their consumption on the Web of Linguistic Linked Data. Thereby we will aim at linking to both encyclopaedic resources, DBpedia[21] and Wikidata,[22] in order to link the Latin language data to additional extra-linguistic information.

Our data set and the algorithms for generating the OntoLex-Lemon representation will be made freely available, either at the GitHub repository of the Latin WordNet or within the LOD presence of the LiLa project.

9. Acknowledgements

Contributions by the German Research Center for Artificial Intelligence (DFKI GmbH) were supported in part by the H2020 project Prêt-à-LLOD with Grant Agreement number

[21] https://wiki.dbpedia.org/.

[22] https://www.wikidata.org/wiki/Wikidata:Main_Page.

825182. Contributions by the Austrian Centre for Digital Humanities and Cultural Heritage at the Austrian Academy of Sciences were supported in part by the H2020 project "ELEXIS" with Grant Agreement number 731015. The work described in this paper was also pursued in part in the larger context of the COST Action CA18209 - NexusLinguarum "European network for Web-centred linguistic data science". We also thank the anonymous reviewers for their insightful comments.

10. References

Cimiano, P., McCrae, J.P. & Buitelaar, P. (2016). Lexicon Model for Ontologies: Community Report.

Fedriani, C., Felice, I.D. & Shorth, W.M. (2020). The Digital Lexicon Translaticium Latinum: Theoretical and Methodological Issues. In C. Marras, M. Passarotti, G. Franzini & E. Litta (eds.) *Atti del IX Convegno Annuale AIUCD. La svolta inevitabile: sfide e prospettive per l'Informatica Umanistica.* Associazione per l'Informatica Umanistica e la Cultura Digitale, pp. 106–113.

Fellbaum, C. (ed.) (1998). *WordNet: An Electronic Lexical Database.* Language, Speech, and Communication. Cambridge, MA: MIT Press.

Klimek, B., McCrae, J., Bosque-Gil, J., Ionov, M., Tauber, J. & Chiarcos, C. (2019). Challenges for the Representation of Morphology in Ontology Lexicons. In *Proceedings of eLex 2019.* URL https://elex.link/elex2019/wp-content/uploads/2019/09/eLex_20.

Mambrini, F. & Passarotti, M. (2019). Harmonizing Different Lemmatization Strategies for Building a Knowledge Base of Linguistic Resources for Latin. In *Proceedings of the 13th Linguistic Annotation Workshop.* Florence, Italy: Association for Computational Linguistics, pp. 71–80. URL https://www.aclweb.org/anthology/W19-4009.

McCrae, J., de Cea, G.A., Buitelaar, P., Cimiano, P., Declerck, T., Gómez-Pérez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D. & Wunner, T. (2012). Interchanging lexical resources on the Semantic Web. *Language Resources and Evaluation*, 46(6), pp. 701–709.

McCrae, J.P., Chiarcos, C., Bond, F., Cimiano, P., Declerck, T., de Melo, G., Gracia, J., Hellmann, S., Klimek, B., Moran, S., Osenova, P., Pareja-Lora, A. & Pool, J. (2016). The Open Linguistics Working Group: Developing the Linguistic Linked Open Data Cloud. In N.C.C. Chair), K. Choukri, T. Declerck, M. Grobelnik, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis (eds.) *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16).* ELRA, 9, rue des Cordelières, 75013 Paris: ELRA.

Passarotti, M.C., Cecchini, F.M., Franzini, G., Litta, E., Mambrini, F. & Ruffolo, P. (2019). The LiLa Knowledge Base of Linguistic Resources and NLP Tools for Latin. In *LDK.*

Automatic induction of a multilingual taxonomy of discourse markers

Rogelio Nazar

Instituto de Literatura y Ciencias del Lenguaje
Pontificia Universidad Católica de Valparaíso
rogelio.nazar@pucv.cl

Abstract

This paper describes a proposed method for the identification and classification of discourse markers (e.g., *however, therefore, by the way*) by applying statistical analysis to large parallel corpora. The objective is to build a lexical resource consisting of a multilingual taxonomy, so far in English, Spanish, German and French. A method is proposed that first separates discourse markers from the rest of the lexical units in the corpus using a measure of entropy, and then classifies them in groups by function using a clustering procedure especially designed for massive data processing. From that point onwards, the system is used to recursively identify and classify more units. Experimental evaluation shows that, in terms of precision, the automated method is able to perform as well as a team of human annotators (undergraduate students of linguistics), and it outperforms them in terms of recall.

Keywords: automatic creation of dictionary content; connectives; discourse markers; taxonomy induction; natural language processing

1. Introduction

This paper presents the first results of a lexicographic research project aimed at cataloging discourse markers (DMs) by means of statistical analysis of large parallel corpora. It describes a newly developed algorithm for the automatic induction of a multilingual taxonomy of DMs, which is then used to recursively identify and classify more units. The objective of the research is to obtain an exhaustive inventory of DMs of different languages. Some preliminary results are described, including a classifier of DMs and a first version of the multilingual taxonomy, so far in English, Spanish, German and French.

The method is solely based on the exploitation of parallel corpora by statistical algorithms. There is no human intervention in the process chain, and no external resources are used, such as POS-taggers or dictionaries. The reason for disregarding external resources, even when such resources are available for the languages considered in the present research, is in part for scientific parsimony but also to facilitate replication of experiments in other, possibly less resourced, languages. One has to take into consideration, too, that one of the outcomes of a purely corpus-based approach is that it may lead to the detection of new units, those that are currently in use in the texts but have not yet been added to dictionaries.

The method uses only co-occurrence association measures and an entropy model to identify DMs according to their distribution in the corpus. As DMs are independent of the content of the texts in which they appear, their occurrence in texts cannot be used to predict the occurrence of other units. Once they are separated from the set of vocabulary units, they are then grouped together using a clustering method which uses their shared equivalence in other languages as a similarity measure. The algorithm will classify new candidates by language, will then decide if they are effectively DMs and, if that is the case, it will assign them to a category.

The identification and subsequent classification of DMs is an extremely difficult task due to various factors. Even for humans (and, indeed, for specialists) it is not always clear where the distinction between DMs and the rest of the lexical units lies, and the definition of the concept varies according to authors and theories. This is due to several reasons. Among them, there is the polyfunctionality of DMs (Pons Bordería & Fischer, 2021), i.e. the fact that the same unit can have a DM function in some contexts but not in others, and even that the same unit can have different DM functions depending on the context. Other factors that further complicate any attempt to determine a clear-cut distinction is that, while some of them operate at the discourse level (one of their characteristic features), others instead seem to be more integrated into the syntactic structure. In part, this is one of the reasons why it is important to conduct empirical research on the subject, especially when the field is dominated by theoretical approaches that rely heavily on introspection or with corpus-based research but with hand-picked examples.

The method's performance varies by language. It is fairly successful in English, Spanish and French, but less so in German, where it has been only moderately successful. On the whole, however, the results of the approach are promising, especially when a preliminary evaluation with Spanish results shows that the method outperforms a group of human annotators. This is a remarkable achievement considering that it is an extremely minimalist approach, one which is computationally inexpensive and has no dependency on linguistic resources other than a parallel corpus. In its current form, the method could be of interest to lexicographers working on DMs, for researchers applying algorithms to automate some levels of discourse analysis, and also for final users, such as translators or people writing in a first or a second language.

2. Related work

In recent years, linguistic theorists have turned their attention to DMs, with an increasing number of publications being devoted to the subject (Fraser, 1999; Pons Bordería, 2001; Schiffrin, 2001). The topic, however, is by no means new in linguistics, and appears in some early grammars, especially of the Spanish tradition. For instance, grammarians such as Antonio de Nebrija, Gregorio Garcés o Andrés Bello in the 15th, 18th and 19th centuries, respectively (Casado Velarde, 1993; Pons Bordería, 2001) all make reference to DMs in their works; more recently there is Gili Gaya (1943), who discusses DMs, albeit using different terminology.

Greater interest in the subject began to appear much later, with the advent of discourse analysis, and more specifically in the field of text grammars. Early work by van Dijk (1973), for instance, presents the main functions of what he then called connectives, which mark the logical relations between propositions, such as conjunction, disjunction, causality, condition, concession, contrast, purpose and so on. A few years later, Halliday & Hasan (1976) presented a developed categorisation of what they call conjunctive relations, with additive, adversative, causal and temporal markers, as well as other continuative or conversational units. A final important historical precedent in the study of DMs is the analysis of connectives in the field of argumentation theory by Anscombre & Ducrot (1976). They notably pointed out that the absurdity of an example such as (1) is a consequence of the use of the expression *même* ('even'):

(1) # *Une mule vaut mieux qu'un âne, même mauvais.*
(A mule is better than a donkey, even a bad donkey).

DMs are perceived to be a driving force behind the proliferation of text grammars, as they were a subject for which earlier linguistic theories proved inadequate. As Stubbs (1983: 77) puts it, DMs "provide problems for sentence based grammars, but are of great interest in a study of discourse sequences, since their functions are largely to do with the organization of connected discourse, and with the interpretation of functional categories of speech acts".

The following years saw a profusion of publications dealing with DM's defining properties and attempting to delineate their boundaries and categorisations. DMs are, probably, a universal feature of language, but they are not easily defined as a single class. They have been defined as particles that facilitate the interpretation of coherence relations in texts (Fraser, 1999; Pons Bordería, 2001). That is to say, they are instructions on how to connect propositions and organise argumentation. It must be noticed, however, that coherence relations between propositions can be inferred even in the absence of DMs, and therefore they are considered optional. However, their presence facilitates comprehension and reduces the chances of ambiguity. They also have an important function in facilitating the interaction between participants, so they have an interpersonal value beyond their textual one, by signalling changes of subject or turn taking (Mosegaard Hansen, 1998). In this sense, one must consider DMs in the context of other pragmatic particles with an interpersonal function, such as interjections, modal particles, focus particles, conjunctions, etc.

In terms of their morphology, they are formally mostly invariable. They have no inflection, do not admit modifiers and cannot be negated or coordinated (Martín Zorraquino & Portolés, 1999). They can pertain to different categories, such as conjunctions, adverbs, prepositional phrases, idioms, and so on.

Regarding their syntactic nature, Schiffrin (2001) describes them as utterance-initial and syntactically independent, although this is perhaps a too restrictive characterisation that would leave out many valid DMs. But it is true that they often are parenthetical and seem to be outside of the syntactic structure of the sentence. More critically, they do not participate directly in the sentence's propositional content, but rather affect the whole sentence or the relation between the sentence and other chunks of text. Their scope varies across different levels of discourse (Pons Bordería, 2001; Brinton, 2010).

In terms of their semantics, they have procedural rather than semantic content, i.e., no referential, propositional or truth value. Historically, though, they derive from lexical units that did have these properties (Traugott & Dasher, 2002), but lost them due to a process of grammaticalisation. It is therefore said that their propositional content has been gradually 'bleached' (Wichmann & Chanet, 2009).

DMs can be organised according to function. One of the most common classifications is counter-argumentation, with expressions such as *however* or *nevertheless*, among others. These are intended to alert the reader/listener that the following propositions will not be what might be expected based on what came before it. Other common functions are to make a cause-consequence relation explicit, such as *consequently* or *therefore*. In their well-known taxonomy, Martín Zorraquino & Portolés (1999) describe a series of broad categories that then divide into branches. Among the main classes we find the structuring type (e.g. *on the one hand, on the other, finally*), connectives (e.g. *moreover, furthermore, in the same way*), reformulatives (e.g. *in other words, better said*), and others. This categorisation has been extremely influential not only in the Spanish tradition, but in other languages as well, e.g. in German (Blühdorn et al., 2017).

The vast majority of the literature on DMs has been devoted to the qualitative study of individual cases, e.g. Urgelles-Coll (2010) in the case of the English DM *anyway* or Llopis-Cardona (2014) in the case of several DMs in Spanish. Fewer are the attempts to compile extended lists of DMs. Two exceptions are Knott (1996) and Stede (2002) who took on this task in English and German, respectively. More work was carried out later in the case of Spanish, for instance dictionaries such as those by Santos Río (2003) or Briz et al. (2008). Recent years have seen an increase in activity in this area. For instance, the material provided by Roze et al. (2012) for French, Feltracco et al. (2016) for Italian, Mírovský et al. (2017) for Czech and Mendes et al. (2018) for Portuguese. Special mention must be made of the contribution by Stede et al. (2019), who are centralising a multilingual taxonomy of DMs in a single database: http://connective-lex.info/.

The computational linguistics community that deals with discourse analysis has paid comparatively less attention to the topic of DMs, Stubbs (1996) being among the exceptions. When these researchers do mention DMs, they use different terminology to refer to them, for instance "discourse cues" (Moore & Wiemer-Hastings, 2003). The field has seen a renewed interest in DMs as of late, in part motivated by recent progress in the field of discourse parsing (Xue et al., 2016), but there is still much to be done. Lopes et al. (2015: 1), for instance, note that "little has been said on their cross-language behavior and, subsequently, on building an inventory of multilingual lexica of discourse markers".

A driving force in this renewed interest seems to be the application of parallel corpora and machine translation. Versley (2010) used an English-German parallel corpus to transfer linguistic annotations from English to German. In a similar way, Lopes et al. (2015) used machine translation to obtain a list of equivalent DMs in different languages from an original list of 427 markers in English.

Also using parallel corpora, but taking a different approach, one similar to that being presented in this study, Robledo & Nazar (2018) described a method based on clustering to offer a bottom-up taxonomy of Spanish DMs. There, as in the current paper, the functional equivalence of different DMs is based on their shared translation as shown in the corpus alignment. Using that method, 587 Spanish DMs were obtained, with evaluation figures showing 0.93 precision and 0.78 recall in the task of identifying false DMs in a list with mixed genuine and false items. A limitation is that the method requires a variety of language-dependent resources, such as POS-taggers, syntax-based rules to filter out improbable candidates and a *gazetteer* used as a stoplist for the same purpose. The main drawback, however, is the hierarchic clustering method that is used. Based on a distance matrix, it entails great computational expense when dealing with large datasets.

More recently, Sileo et al. (2019) used a curated list of 174 markers for English in order to discover sentence initial, parenthetical, high-frequency DMs using contextual cues (word ngrams). After a complex and computationally expensive machine learning procedure involving sentence selection, tokenising, tagging and finally classification with the Fasttext library, they discovered 243 DM candidates, but their results are modest in terms of accuracy.

This study continues in the same vein as the aforementioned ones in that it is an empirical method, based on the statistical analysis of large corpora. The difference is that the present one is comparatively a very simple method, and with a focus on a multilingual and language-agnostic approach. With regards to earlier qualitative studies on DMs, the main difference is that the present one is an empirical method, i.e., a bottom-up rather than a top-down approach. This is important for practical reasons, as the automation saves a lot of effort, but also, and most importantly, for scientific reasons, as the quantitative method favours objectivity. Also, in contrast with the manually compiled DM lexicons existing today, which comprise only a few hundred entries, in this project thousands of them are discovered, which are offered to the public in an open database online. All these are reasons to believe that the present paper represents a substantial contribution to the state of the art on DM research methodology.

3. Methodology

As already anticipated, the methodology consists of first identifying DMs in corpora by separating them from the rest of the vocabulary and then classifying them in a bottom-up functional taxonomy. It is a minimalist approach based solely on statistical measures and without any type of external resource apart from a parallel corpus. Section 3.1 explains how DMs are identified according to their distribution in the corpus by exploiting one of their characteristics, which is to be independent of the content of the texts in which they appear. In operational terms, this means that their occurrence cannot be used to predict the occurrence of other lexical units. Section 3.2 describes the subsequent step, i.e. their classification, which is performed using an original clustering algorithm. Section 3.3 shows how the clusters are tagged and organised. Finally, section 3.4 explains how, once this core taxonomy is built, it is then used to further populate it by classifying new DMs obtained from corpora in a recursive manner.

3.1 Separating DMs from the rest of the vocabulary

The same parallel corpus was used for all steps of the procedure: the Opus Corpus (Tiedemann, 2012), a large collection of parallel corpora in different languages, freely available and organised by corpus in different TMX files, a standard format in the field of translation. The number of corpora varies according to the language pairs, but is close to 30 files per pair. Each corpus presents a different specialised technical domain and/or discourse genre. It is aligned at ‘translation units’, which generally correspond to sentences but sometimes larger segments, like paragraphs. The corpus does not include lemmatisation or POS-tagging annotations but that is not a problem since such data is not needed for the method presented here.

For the first step, only the target language segment is used, ignoring the alignments. An initial set of vocabulary units is obtained from the corpus by sorting ngrams, defined as sequences of one, two and three words not including punctuation marks. These are not used as a means to determine the boundaries of the ngrams because doing so would lead to the obtainment of only parenthetical DMs, which are merely a subset of all existing DMs. Moreover, DMs do not behave in this way in all languages. For instance, German DMs are not used parenthetically as frequently as in the other languages.

The result is a very large initial vocabulary set, denoted as $InVoc$, which is then reduced in size in subsequent steps by filtering units according to their distribution in the corpus and according to a measure of information. As DMs are procedural instead of semantic, that means that their appearance in a text is not related to the semantic content and they cannot be used to predict the co-occurrence of other vocabulary units. Thus, a subset of $InVoc$ called $FiVoc$ contains units that appear in at least seven of the 30 TMX files with a minimum frequency of 50 occurrences, all thresholds being arbitrary but empirically motivated.

This first operation results in a dramatic decrease in the size of the vocabulary lists, from an average of half a million units per language to fewer than 5,000. Yet, fewer than a third of the latter are genuine DMs, as the majority of these are words or sequences of words bearing a very general semantic content. In the case of English, these would be high frequency words such as *property* or *language* as well as names of places like cities or countries (e.g. *Paris, the Netherlands*), among others. As a consequence, a more refined procedure is then applied, which is computationally more expensive but justifiable considering that it is applied to only a few thousand units.

The second filtering operation consists of determining a measure of information of the candidates. This measure aims to tell how informative a word is in relation to its ability to predict the appearance of other words. A word with a clear semantic content, e.g. *Paris*, should exhibit a tendency to co-occur in large numbers of contexts with other units that are semantically related, e.g. *France.* The contrary would be the case of the units we are interested in, the DMs, which should score very low with this type of measure. Therefore, given a target unit x, it is possible to obtain a set $M(x)$ consisting of a sample of contexts of occurrence of x from the corpus and then sort all the vocabulary units[1] in a ranking R_x by decreasing order of frequency. One can then use the relation between this frequency and the sample size in order to obtain a distinction between semantic and procedural units. The coefficient used to calculate this is shown in (1). The parameter n is arbitrary, but experimentally fixed at 20. The decision to accept or reject x as a member of the candidate set C is based on another empirically parameter t, as shown in (2). Alternatively, one could also keep the best k candidates in C.

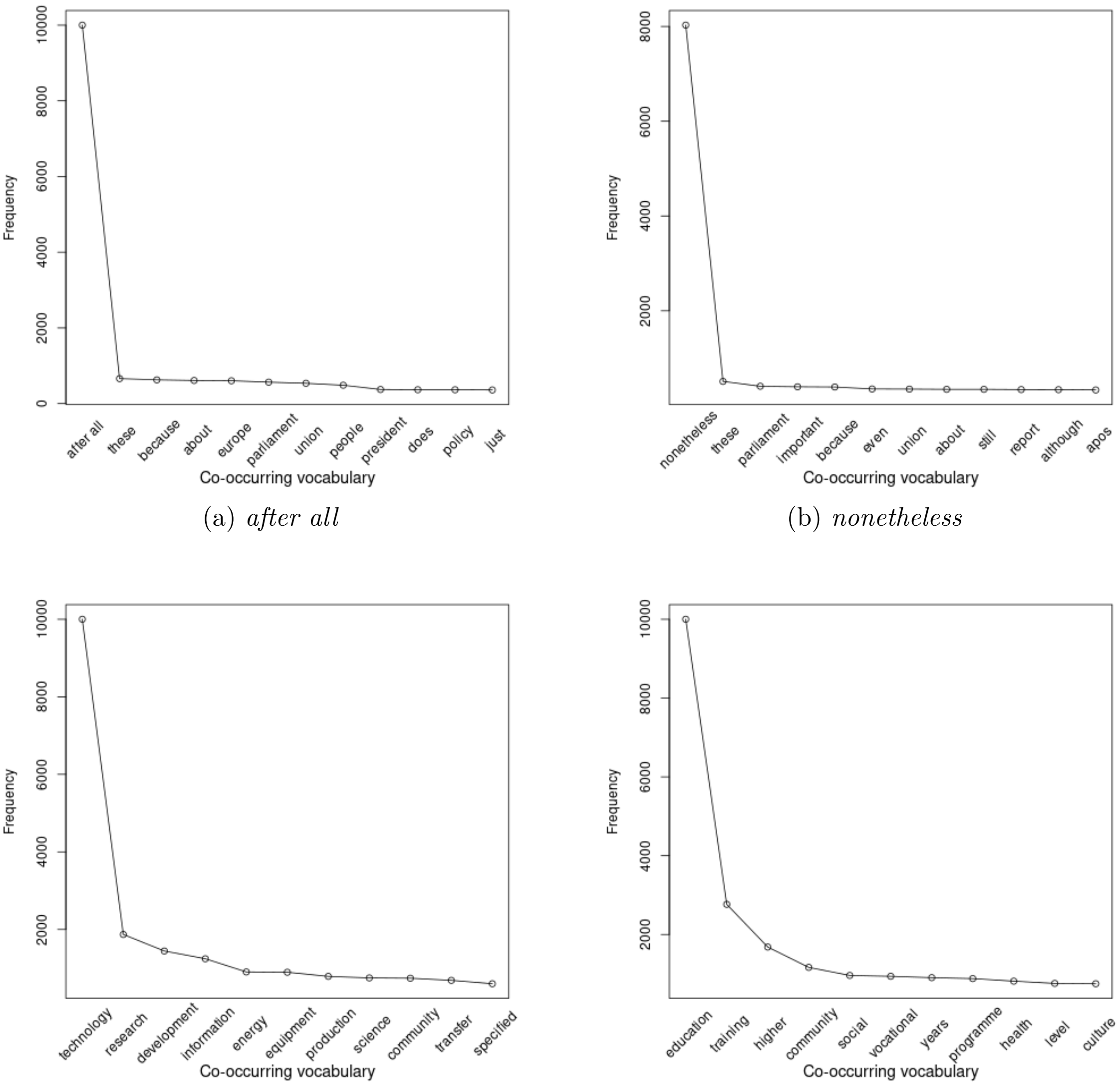
$$I(x) = \frac{\log_2 \sum_{i=1}^{n} R_{x,i}}{\log_2 |M(x)|} \tag{1}$$

$$x \in C = \begin{cases} true & I(x) < t \\ false & \text{otherwise} \end{cases} \tag{2}$$

For illustration, Figure 1 presents how it is possible to obtain an almost clear-cut separation between the two classes. Functional units such as *after all* (Panel a) or *nonetheless* (Panel b) are very different from semantically-charged vocabulary units such as *technology* (Panel c) or *education* (Panel d), and the difference is revealed by their co-occurrence pattern. E.g., in the case of *technology,* one can say that if this word is found in a sentence, then there is a relatively high probability of finding other words[2],

[1] The units considered here are only single-words instead of word-ngrams. This is done this way for simplicity and to reduce computational cost, but the possibility of using larger-than-word units is worth exploring in future research.

[2] Function words (i.e., those that would appear in any random sentence such as *with, that, from, this,* etc.) are also ignored precisely because they are themselves very uninformative



(a) *after all*

(b) *nonetheless*

(c) *technology*

(d) *education*

Figure 1: The shape of the co-occurrence frequency curve is used to predict the semantic or procedural nature of lexical units

such as *research, development, information* and so on. This does not happen in the case of DMs. An item like *after all* shows an extremely low frequency of co-occurrence with other units. Thus, finding this phrase in a sentence does not make it possible to predict the occurrence of any other lexical item.

3.2 Induction of a functional taxonomy of DMs

The previous phase yielded a set $C(l)$ of DM candidates for each language l (en, fr, es, de). In this phase, in turn, for each l, a functional taxonomy of DMs will be created in the form of a hierarchic clustering, for which the parallel corpus is used. At this point, languages are paired together. It is irrelevant which languages are used in each pair, but for practical reasons English is used as one of the languages for each pair, as it is usually the language for which more material is available. Thus, with the English-French pair, for instance, the algorithm produces an alignment of sets C_{en} and C_{fr}. The alignment of the units in both lists can be achieved with the use of a co-occurrence measure such as $A(i, j)$, shown in (3).

$$A(C_{en,i}, C_{fr,j}) = \frac{f(C_{en,i}, C_{fr,j})}{\sqrt{f(C_{en,i})}.\sqrt{f(C_{fr,j})}} \quad (3)$$

This coefficient compares the frequency of co-occurrence of the vocabulary units in the aligned segments with their independent frequency in the whole corpus. Thus, if, for instance, $C_{en,i}$ is *nonetheless* and $C_{fr,j}$ is *néanmoins*, the algorithm contrasts the number of times they appear in translated sentences with the number of times they appear in general, that is, alone or together. For each unit in C_{en} there will be a limited number of equivalent candidates in C_{fr}. The top three candidates, as long as they have a score greater than 0.20, are kept. This parameter is again arbitrary but empirically defined.

The purpose of aligning the DM candidates in this fashion is only to allow for their organisation in a taxonomy, a result that is achieved by means of a clustering procedure. This procedure is conducted using the aligned pairs as a similarity measure, i.e., two units are considered similar for the clustering if they share the same equivalent markers in the parallel corpus. To continue with the same example, English items like *nonetheless* and *nevertheless* are considered similar because they share the same equivalence in a second language, such as *néanmoins* in the case of French.

The exact procedure of the clustering is as follows. It consists of a greedy-matching, graph-based clustering algorithm that has the property of being very efficient in comparison with regular hierarchic clustering algorithms such as those used in previous studies (Robledo & Nazar, 2018), which suffer from quadratic complexity and are not scalable to many thousands of objects. The option applied here is simpler, and is called 'the cocktail-party algorithm'. One often sees, at conference cocktail parties or coffee-breaks, that people tend to cluster together as they arrive on the basis, at least initially, of their mutual acquaintance. If the DM candidates have been aligned, one can imagine them as people coming to the cocktail in pairs. For instance, first Paul (*nonetheless*) and Eva (*néanmoins*) arrive together, followed by Robert (*of course*) and María (*évidemment*), who also arrive together. The two pairs do not know each other, so they stay apart and keep to themselves. Then, however, Eva sees that Michael (*nevertheless*) just arrived, and since she already knows him (*néanmoins* and *nevertheless* were also found to be equivalent according to the parallel corpus), she introduces him to Paul. Now, Paul, Eva and Michael form a single cluster, as depicted in Figure 2.

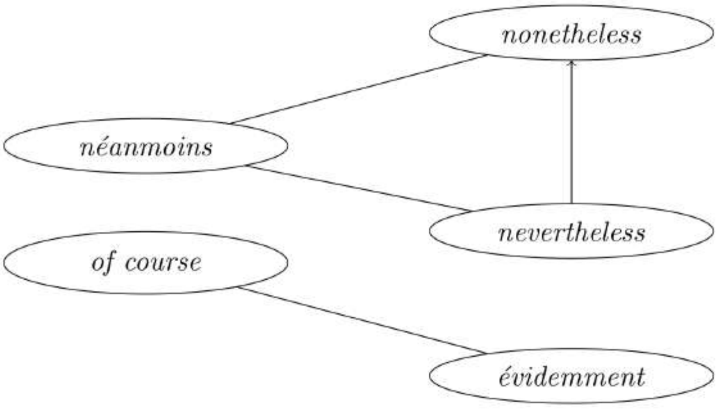


Figure 2: Illustration of a moment of the graph-based clustering process

If someone else arrives and knows at least one of these three in the cluster, she will also join the group, unless she finds another group with more acquaintances. This process goes on, and more clusters are produced during the event as more people/DMs arrive, and the result will be a bilingual taxonomy.

3.3 Tagging the clusters

One limitation of the taxonomy created so far is that clusters have no meaningful names. They are identified by numeric codes that bear no relation to their content. Also, there is the problem that some of these clusters should be grouped in order to form larger categories. Since it would be too laborious to manually tag each cluster with a name, it was decided to resort to an automatic tagging procedure based on the taxonomy originally proposed by Martín Zorraquino & Portolés (1999) because, as already mentioned, it has been extensively used, even in languages other than Spanish.

Using the examples provided by these authors, a matching algorithm was developed to tag a given cluster from the induced taxonomy with the names of the categories they provide. For example, if there is a cluster that consists of contrastive connectors, it will probably include some of the examples mentioned by those authors, such as *sin embargo, no obstante*, etc. Thanks to these shared examples, the tagging algorithm can recognise the relationship between the cluster and said category and confidently assign a meaningful name to each cluster.

As the examples are in Spanish, the Spanish side of the taxonomy has to be used to do the tagging. But, since all the taxonomy is multilingually aligned, a tag assigned to a cluster in one side of the taxonomy is inherited by the other sides as well. The tagging also has the effect of aggregating similar clusters in larger categories.

In any case, the content of the clusters is kept separate, although hierarchically organised. For example, there is one broad category in the terminology of Martín Zorraquino & Portolés (1999) called *Estructuradores de la información*, referring to DMs used for information structuring, and within this category there is a subcategory called *Comentadores*, referring to DMs used to introduce commentary. It happens that this algorithm finds new divisions within this category, and there are different clusters under the tag of *Comentadores*. For example, one of these clusters contains DMs like *arguably, certainly, presumably, probably*, among other units, while another contains DMs such as *at this point, at this stage, at this time*, etc. Keeping them separate allows one to obtain a layered categorisation, which in turn can be used as the basis for the further categorisation of new DMs.

3.4 Further population of the taxonomy

Once a basic or core multilingual taxonomy of DMs has been obtained (hereinafter *Dismark*), it is then possible to use such material as the basis for the categorisation of new DMs, done recursively. For this final part of the procedure, an input candidate x is needed ($x \notin Dismark$) for the algorithm to perform the following three subtasks:

1. Classify x by language
2. Decide if x is effectively a DM
3. If 2 is true, assign x to a category in *Dismark*

For subtask 1, one is of course limited to the available languages. The algorithm will retrieve contexts of occurrence of x in the corpora of the different languages and select the one with the highest number of hits. For subtask 2 it will use the parallel corpora. If x appears in the aligned sentences with other DMs already registered in the taxonomy, then this is taken as indication that x is a true DM. Once this has been decided, the algorithm has to find the best matching category for x, and this is done in a way reminiscent of the method explained in Section 3.3. That is, using the equivalences for x in a different language that were just obtained from the parallel corpus, the best category is selected on the basis of their matching. For instance, if x is *in that sense* and is not already in *Dismark*, its analysis in the parallel corpus will reveal that valid French equivalents are, among others, units like *à cet égard* and *dans ce sens*, which are already in the taxonomy. If this is the case, then the algorithm can safely place x on the English side of this cluster.

This taxonomy operates automatically and without supervision. Moreover, the larger the taxonomy becomes, the better the result of its predictions because it has a better knowledge base. Thus we can see how, from nothing more than a parallel corpus and a set of category names for the clusters, it is possible to obtain a taxonomy of DMs thanks to a system that is characterised by a virtuous cycle and that can incrementally improve in precision and thoroughness.

4. Evaluation

At the time of writing, the database contains a total of 2,463 different DMs classified in 20 different categories and 71 subcategories. Tables 1 and 2 show examples of two clusters belonging to different categories. These are meant to be read as groups of DMs that are functionally equivalent, and no correspondence is implied in their horizontal alignment. They share the same cluster simply because they can be used with the same function.

English	**Spanish**	**French**	**German**
• in a manner similar • in a similar manner • in the same manner • in the same way • likewise • similarly	• de forma similar • de la misma forma • de la misma manera • de manera similar • de modo similar • del mismo modo • forma similar • manera similar	• de la même façon • de la même manière • de même • même façon • même manière	• auf dieselbe Weise • desgleichen • dieselbe Weise • ebenso • gleiche Weise • gleichen Weise • in ähnlicher Weise • ähnlicher Weise

Table 1: An example of a subcategory (cluster) of the category ‘additive connectives’

English	**Spanish**	**French**	**German**
• after all • at last • at some point • at some time • at the end • but after all • eventually • in a few words • in a word • in brief • in short • in sum • in summary • in the end • on balance • sooner or later • to sum up • to summarise • ultimately • upon the whole	• a fin de cuentas • a la larga • al final • así pues • de forma resumida • después de todo • en algún momento • en definitiva • en fin • en pocas palabras • en resolución • en resumen • en resumidas cuentas • en suma • en una palabra • en última instancia • en último término • eventualmente • tarde o temprano	• après tout • au bout du compte • au final • en bref • en définitive • en fin de compte • en résumé • en somme • enfin • finalement • forme résumée • forme résumée ou agrégée	• am Ende • erweitert • irgendwann • kurz gefasst • kurz gesagt • kurzum • letzten Endes • letztendlich • letztlich • schließlich

Table 2: Another example of subcategory (cluster) of the category ‘recapitulation connectives’

The first look on the results reveals that there is a considerable mismatch in quality between languages. While the results on English, Spanish and French seem very impressive (on average 95% of the DMs are correct), in German, instead, one can claim only that there has been moderate success, with 84% of the DMs being correct. A worse performance in German was in part to be expected, as this language presents more challenges for automatic processing. This is due to the fact that the syntactic behaviour of DMs in German is different from the other languages regarding position, punctuation and the use of cases (e.g. nominative, accusative, dative). Many of the problems were also related to segmentation faults (e.g., the system retrieves *solchen Fällen* instead of the correct form *in solchen Fällen*).

In order to offer a more precise evaluation, we conducted a small experiment to compare the performance of the algorithm with a group of human annotators in the task of identifying DMs. After a university semester course on Text Grammar which deals extensively on the subject of DMs, seven of the best performing students were selected to participate in the task. Their training consisted of both theoretical lessons on the subject and practical exercises in which they had to identify and classify DMs using the taxonomy by Martín Zorraquino & Portolés (1999).

For the task, the annotators received a list of 709 expressions, roughly two thirds of which were mixed DMs and one third of which were lexical units of other types, in alphabetical order. The students, unaware of the composition of the list, were asked to place a number one beside every unit that they considered not to be a DM. They were asked to perform the task alone, without asking their classmates, and to refrain from using corpora, dictionaries or any other type of lexicographic resource. It was emphasised to them that they should follow their intuition. Table 3 shows the results.

Annotator	**Precision**	**Recall**	**F1**
Dismark	97	94	95
Student 1	96	51	66
Student 2	95	61	74
Student 3	95	41	57
Student 4	94	59	72
Student 5	93	66	77
Student 6	92	32	47
Student 7	91	75	82

Table 3: Comparing the performance of algorithm vs. humans in the task of identifying DMs

In general, they all performed fairly well in terms of precision, and as the table shows, when they selected something as a DM, they were almost always correct. They tended, however, to be more conservative. A series of follow-up interviews with the students revealed that they were unwilling to select something as a DM unless they were very sure it was one. That is, the students marked DMs that were prototypical, meaning highly grammaticalised and showing no sign of morphological variation. They tended to reject genuine cases such as *en estas circunstancias* (‘in these circumstances’) or *en términos más generales* (‘in broader terms’).

Another reason for them to reject genuine DMs was the fact that they found them too polysemous or polyfunctional, in the sense that they were elements that could function as DMs but only in certain contexts. In this regard, the lack of lines of context certainly put humans at a disadvantage. An interesting direction for future research would be to present the participants with the task of detecting DMs in a particular text. This, however, would be a different type of research, because it would not be about classifying DMs in abstract. Instead, its focus would be the classification of particular instances of DMs. That would require totally different sets of measures, such as contextual cues, to determine in which contexts something is used as a DM and in which not. Such an endeavour would be out of the scope of a lexicography project and closer to the area of discourse analysis.

At any rate, what is to be learned from this experiment is that distinguishing between a DM and a non-DM element is not an easy task and that, perhaps, the way forward would be to follow the same criterion as Rysová & Rysová (2018) with the Prague Discourse Bank. This would be to establish a distinction between primary DMs, with those more prototypical or grammaticalised units, and other categories with secondary and free DMs, to accommodate those units that fulfil the same function but are less prototypical.

5. Conclusions

This paper presented a newly developed method for the automatic induction of a multilingual taxonomy of DMs, including a description of its first results. The method is simple and effective. It is also computationally inexpensive and easy to replicate in different languages. The method is, in fact, robust to language varieties, as it could provide useful results even in German, which is, morphologically speaking, a language very different from the others.

Also, in comparison with manually curated classifications of DMs, which in most cases offer a few hundred items, the multilingual taxonomy already offers thousands of them, including items of medium to low frequency in the corpus. The results of the project, including the full database of DMs and a demo for the DM classifier, are offered at the project's website[3]. Even though this is still work in progress, the results currently available can be useful for lexicographers interested in DM projects as well as NLP professionals working on text understanding or text generation. Final users, such as writers or translators, can benefit from this collection in order to improve vocabulary richness.

With respect to future research, the priorities would be the following: 1) to continue evaluating and exploring variations in the method; 2) to continue populating the taxonomy with new, maybe less frequent items and 3) to incorporate new languages, first from Europe and later from other language typologies, taking advantage of the fact that no external resources are needed.

6. Acknowledgments

This paper has been made possible thanks to funding from a research grant by the Government of Chile: *Proyecto Fondecyt Regular 1191481. Inducción automática de taxonomías de marcadores discursivos a partir de corpus multilingües* (Automatic induction of taxonomies of discourse markers from multilingual corpora). I would like

[3] http://www.tecling.com/dismark

to thank the anonymous reviewers for pointing out different ways to improve the paper. I would also like to express my gratitude to Maureen Noble for proofreading and for her valuable comments.

7. References

Anscombre, J.C. & Ducrot, O. (1976). L'argumentation dans la langue. *Langages*, 42, pp. 5–27.

Blühdorn, H., Foolen, A. & Loureda, O. (2017). Diskursmarker: Begriffsgeschichte – Theorie – Beschreibung. Ein bibliographischer Überblick. In H. Blühdorn, A. Deppermann, H. Helmer & T. Spranz-Fogasy (eds.) *Diskursmarker im Deutschen. Reflexionen und Analysen.* Göttingen: Verlag für Gesprächsforschung.

Brinton, L. (2010). Discourse Markers. In A. Jucker & I. Taavitsainen (eds.) *Historical Pragmatics.* Berlin: Gruyter Mouton.

Briz, A., Pons, S. & Portolés, J. (2008). Diccionario de partículas discursivas del español. URL http://www.dpde.es.

Feltracco, A., Jezek, E., Magnini, B. & Stede, M. (2016). LICO: A Lexicon of Italian Connectives. In *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016), Napoli, Italy, December 5-7, 2016*, volume 1749 of *CEUR Workshop Proceedings.* CEUR-WS.org.

Fraser, B. (1999). What are discourse markers? *Journal of Pragmatics*, (31), pp. 931–952.

Halliday, M. & Hasan, R. (1976). *Cohesion in English.* London: Longman.

Knott, A. (1996). *A data-driven methodology for motivating a set of coherence relations.* Ph.D. thesis, University of Edinburgh, UK. British Library, EThOS.

Llopis-Cardona, A. (2014). *Aproximación funcional a los marcadores discursivos. Análisis y aplicación lexicográfica.* Frankfurt am Main: Peter Lang.

Lopes, A., de Matos, D.M., Cabarrão, V., Ribeiro, R., Moniz, H., Trancoso, I. & Mata, A.I. (2015). Towards Using Machine Translation Techniques to Induce Multilingual Lexica of Discourse Markers. arXiv 1503.09144.

Martín Zorraquino, M.A. & Portolés, J. (1999). Los marcadores del discurso. In I. Bosque & V. Demonte (eds.) *Gramática Descriptiva de la Lengua Española.* Madrid: Espasa, pp. 4051–4214.

Mendes, A., del Rio, I., Stede, M. & Dombek, F. (2018). A Lexicon of Discourse Markers for Portuguese – LDM-PT. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).* Miyazaki, Japan: European Language Resources Association (ELRA).

Moore, J.D. & Wiemer-Hastings, P. (2003). Discourse in Computational Linguistics and Artificial Intelligence. In A.C. Graesser, M.A. Gernsbacher & S.R. Goldman (eds.) *Handbook of Discourse Processes.* Routledge.

Mosegaard Hansen, M.B. (1998). *The Function of Discourse Particles : A study with special reference to spoken standard French.* Amsterdam/Philadelphia: John Benjamins.

Mírovský, J., Synková, P., Rysová, M. & Poláková, L. (2017). CzeDLex – A Lexicon of Czech Discourse Connectives. *The Prague Bulletin of Mathematical Linguistics*, (109), pp. 61–91.

Pons Bordería, S. (2001). Connectives/Discourse markers. An Overview. *Quaderns de Filologia. Estudis Literaris*, (6), pp. 219–243.

Pons Bordería, S. & Fischer, K. (2021). Using discourse segmentation to account for the polyfunctionality of discourse markers: The case of well. *Journal of Pragmatics*, 173, pp. 101–118.

Robledo, H. & Nazar, R. (2018). Clasificación automatizada de marcadores discursivos. *Procesamiento del Lenguaje Natural*, (61), pp. 109–116.

Roze, C., Danlos, L. & Muller, P. (2012). LEXCONN: a French lexicon of discourse connectives. *Discours - Revue de linguistique, psycholinguistique et informatique.* URL https://hal.inria.fr/hal-00702542.

Rysová, M. & Rysová, K. (2018). Primary and secondary discourse connectives: Constraints and preferences. *Journal of Pragmatics*, 130, pp. 16–32.

Santos Río, L. (2003). *Diccionario de partículas.* Salamanca: Luso-española de ediciones.

Schiffrin, D. (2001). Discourse Markers: Language, Meaning, and Context. In D. Schiffrin, D. Tannen & H. Hamilton (eds.) *The Handbook of Discourse Analysis.* Oxford: Blackwell, pp. 54–75.

Sileo, D., van de Cruys, T., Pradel, C. & Muller, P. (2019). Mining Discourse Markers for Unsupervised Sentence Representation Learning. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers).* Minneapolis, Minnesota: Association for Computational Linguistics, pp. 3477–3486.

Stede, M. (2002). DiMLex: A Lexical Approach to Discourse Markers. In A. Lenci & V.D. Tomaso (eds.) *Exploring the Lexicon - Theory and Computation.* Alessandria: Edizioni dell'Orso.

Stede, M., Scheffler, T. & Mendes, A. (2019). Connective-Lex: A Web-Based Multilingual Lexical Resource for Connectives. *Discours.* URL https://journals.openedition.org/discours/10098.

Stubbs, M. (1983). *Discourse Analysis. The Sociolinguistic Analysis of Natural Language.* Chicago: University of Chicago Press.

Stubbs, M. (1996). *Text and Corpus Analysis.* Oxford: Blackwell.

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12).* Istanbul, Turkey: European Language Resources Association (ELRA), pp. 2214–2218.

Traugott, E. & Dasher, R. (2002). *Regularity in semantic change.* New York: Cambridge University Press.

Urgelles-Coll, M. (2010). *The Syntax and Semantics of Discourse Markers.* London: Continuum.

van Dijk, T. (1973). Text Grammar and Text Logic. In J. Petöfi & H. Rieser (eds.) *Studies in Text Grammar.* Dordrecht: Reidel, pp. 17–78.

Versley, Y. (2010). Discovery of Ambiguous and Unambiguous Discourse Connectives via Annotation Projection. In L. Ahrenberg, J. Tiedemann & M. Volk (eds.) *Proceedings of the Workshop on Annotation and Exploitation of Parallel Corpora (AEPC).* Tartu: Northern European Association for Language Technology, pp. 83–92.

Wichmann, A. & Chanet, C. (2009). Discourse markers: A challenge for linguists and teachers. *Nouveaux cahiers de linguistique française*, 29(4), pp. 23–40.

Xue, N., Ng, H.T., Pradhan, S., Rutherford, A., Webber, B., Wang, C. & Wang, H. (2016). CoNLL 2016 Shared Task on Multilingual Shallow Discourse Parsing. In *Proceedings of the CoNLL-16 shared task.* Berlin, Germany: Association for Computational Linguistics, pp. 1–19.

New developments in Lexonomy

Adam Rambousek[1,2,4], Miloš Jakubíček[1,2], Iztok Kosem[3,4]

[1]Faculty of Informatics, Masaryk University, Brno, Czech Republic
[2]Lexical Computing, Brno, Czech Republic
[3]Centre for Language Resources and Technologies, University of Ljubljana, Slovenia
[4]Jožef Stefan Institute, Ljubljana, Slovenia
E-mail: rambousek@fi.muni.cz, milos.jakubicek@sketchengine.eu, iztok.kosem@cjvt.si

Abstract

This article describes new developments and enhanced features in the open-source web application for dictionary writing, Lexonomy. Since its introduction in 2017, a growing number of users and organisations have chosen Lexonomy to edit their dictionaries. We describe the motivation and process of the source code refactoring to Python programming language. Next, we provide details on integration with the Sketch Engine corpus manager. We also cover the completely new feature of dictionary linking, both as a graphical interface for users, and API to include Lexonomy in the process of automatic dictionary linking. Finally, the article describes the new functionality needed for Lexonomy integration within the ELEXIS project processes. Furthermore, we provide usage statistics on users and dictionaries they create.

Keywords: Dictionary editing; Dictionary writing system; Lexicographic tools; XML; Corpora connection

1. Introduction

Lexonomy (Měchura et al., 2017) is a free, open-source, web-based dictionary writing system. Since its introduction in 2017, it is used by a growing number of users and organisations. The publicly available installation at www.lexonomy.eu is currently used by over 2,700 users who created over 5,000 dictionaries.

Lexonomy was selected to be part of the ELEXIS (Krek et al., 2018) project infrastructure, providing the primary tool for dictionary creation, storage, and browsing. Thanks to this, the number of users and their dictionaries increased significantly, which led to two groups of updates to Lexonomy. Integration into ELEXIS brought new feature requests from various project partners. Furthermore, we had to address performance issues for a larger amount of data and users.

The following chapters present new updates and features in Lexonomy since 2018.

2. Improved scalability

Originally, Lexonomy was developed in Node.js[1] at the backend side and HTML+JavaScript on the client-side. While the Node.js server provided a connection to the database, core functionalities, and application interface, HTML webpages enriched with JavaScript provided a graphical user interface. To store metadata about users and dictionaries, and dictionary entries, Lexonomy uses the SQLite database[2]. Each dictionary is stored in a separate database file. One of the benefits is working directly with the database file, e.g., using dictionary templates for various projects or backup.

As the number of users and dictionaries in the system grew, we experienced performance issues and long response times when users searched in their dictionaries. After profiling all parts of the application, we identified the handling of concurrent database access requests

[1] https://nodejs.org/en/about/
[2] https://www.sqlite.org/

to be the main cause of the issue. When many users at once searched for entries or imported dictionary data, Node.js server kept database queries in a queue and processed them one by one. This means that one complex database search or import of extensive data into a dictionary may slow down the response time for other users.

At the same time, more developers wanted to participate in Lexonomy, and the issues with Node.js meant that they had to wait before they were able to join the team.

We thus decided to refactor the code of the backend part of Lexonomy. After considering the pros and cons of several programming languages, we selected Python as the best option. From the beginning, we addressed performance by using a multi-threaded environment and running time-consuming tasks (e.g., dictionary import) as background jobs.

After we deployed the refactored backend on the Lexonomy server, Lexonomy could smoothly handle dictionaries of millions of entries. Users only noticed the better performance of Lexonomy, as the graphical user interface was not changed and it still uses the same HTML templates with JavaScript. For developers, the Lexonomy source code is now smaller and more transparent, and they do not need to repeat the same code several times (e.g., checking user access rights).

3. Closer integration with Sketch Engine

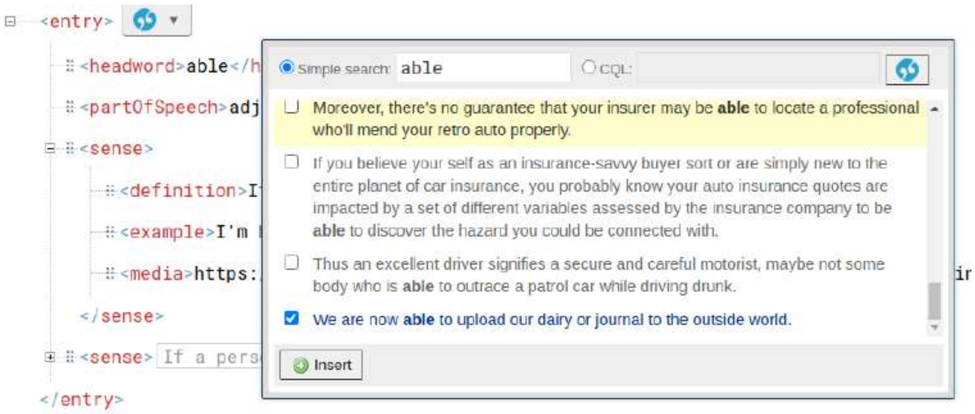
Lexonomy may still work as a standalone tool that can be installed locally on anybody's desktop. It can also be easily coupled with the (No)Sketch Engine corpus management system (Kilgarriff et al., 2014) to get access to corpus content. Connection with the Sketch Engine was extended to provide more options and a better user interface.

The first option is to retrieve the corpus data directly while working in the dictionary editor. For each dictionary, users can select which corpus to use and which elements in the entry structure correspond with different corpus data (examples, collocations, thesaurus items, or definitions). When editing an entry, users will see the Sketch Engine icon on the right elements. After clicking the icon, they may run a CQL query and select which results to include, see Figure 1 for example. The data will be copied to the dictionary entry structure where users can post-edit them. As a default, sketchengine.eu is accessed. However, users may specify their own installation of (No)Sketch Engine.

And from the other side, it is possible to create a new dictionary and fill it with data from the Sketch Engine interface. Users will start in the Sketch Engine and its OneClick Dictionary tool (Kilgarriff & Jackson, 2013). Depending on language support and user selection, the process generates a headword list with part-of-speech labels, provides candidates for example sentences, collocations, synonyms, or definitions. Subsequently, all the data are pushed into Lexonomy, where the new dictionary is created. Users are able to extend or edit the dictionary during the post-editing phase, thus saving time.

4. Single sign-on

To make registration and authentication more comfortable for users, Lexonomy provides the option to log in via the Sketch Engine application. Thanks to this integration, users



```
<entry>
  <headword>able</h
  <partOfSpeech>adj
  <sense>
    <definition>I
    <example>I'm
    <media>https:                ir
  </sense>
  <sense> If a pers
</entry>
```

● Simple search: able ○ CQL:

□ Moreover, there's no guarantee that your insurer may be **able** to locate a professional who'll mend your retro auto properly.

□ If you believe your self as an insurance-savvy buyer sort or are simply new to the entire planet of car insurance, you probably know your auto insurance quotes are impacted by a set of different variables assessed by the insurance company to be **able** to discover the hazard you could be connected with.

□ Thus an excellent driver signifies a secure and careful motorist, maybe not some body who is **able** to outrace a patrol car while driving drunk.

☑ We are now **able** to upload our dairy or journal to the outside world.

Figure 1: Connection with the Sketch Engine, searching for example sentences.

are able to log in to Lexonomy with easy single sign-on through the worldwide eduGAIN research network[3] and other institutions.

5. Integration with Elexifier

Lexonomy was selected as a primary tool for dictionary creation and editing in the ELEXIS project. Apart from dictionary editing, Lexonomy is the base for Elexifier (McCrae et al., 2019), a tool that is designed to digitise printed dictionaries in PDF or XML format. Utilising the option to change the default Lexonomy entry editor with custom JavaScript and XSLT code, Elexifier developers created their own entry editor for annotation of dictionary data in PDF files.

6. Dictionary linking

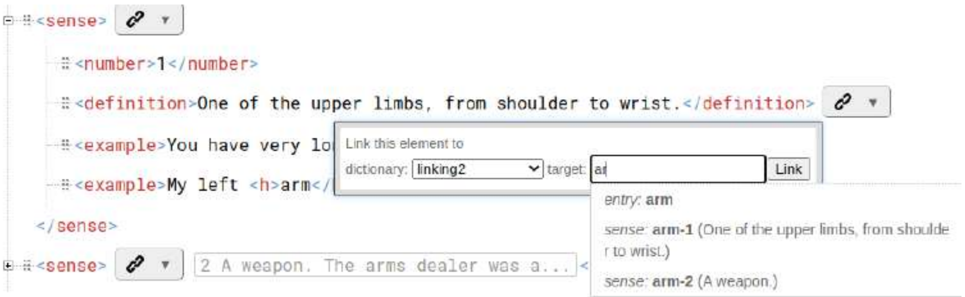
Lexonomy was selected as the dictionary storage in the ELEXIS project, where available dictionaries will be interlinked. To support this task and other scenarios where users need to connect dictionaries, Lexonomy was extended with the general mechanism for dictionary linking.

6.1 Manual linking

The linking mechanism in Lexonomy supports links between any entry elements in any dictionary. As a first step, users have to specify which entry elements should serve as the link point and how each element is identified. For example, *entry* may serve as a link point and each entry is uniquely identified with *(headword + PoS)*, or *definition* may be used as a link point and each definition is uniquely identified with *(headword + PoS + sense number)*.

When users are editing an entry, they have the option to add or view links at corresponding entry elements. When they want to add a new link, they select the target dictionary, choose which element to use in the target entry, and search for a particular link target. Source and target elements may be on a different level in an entry structure. For example, it is

[3] https://technical.edugain.org/status



```
<sense>
    <number>1</number>
    <definition>One of the upper limbs, from shoulder to wrist.</definition>
    <example>You have very lo
    <example>My left <h>arm</
</sense>
<sense>  2 A weapon. The arms dealer was a...
```

Link this element to

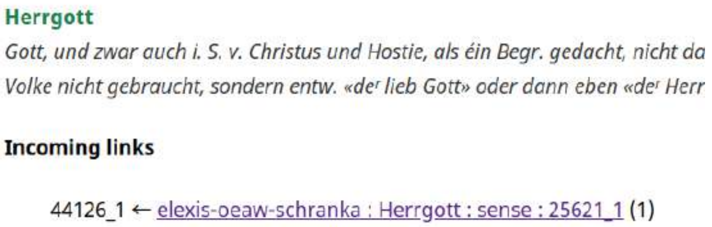
dictionary: linking2 target: ar Link

entry: arm

sense: arm-1 (One of the upper limbs, from shoulde r to wrist.)

sense: arm-2 (A weapon.)

Figure 2: Creating link between *definition* and searching for target element (*entry* or *sense*).



Herrgott

Gott, und zwar auch i. S. v. Christus und Hostie, als éin Begr. gedacht, nicht da Volke nicht gebraucht, sondern entw. «deʳ lieb Gott» oder dann eben «deʳ Herr

Incoming links

44126_1 ← elexis-oeaw-schranka : Herrgott : sense : 25621_1 (1)

Figure 3: Example of link information in an entry preview.

possible to create a link between full entry and one definition. See Figure 2 for an example of link creation and searching for the target of the link.

When browsing the dictionary, links are also displayed in the entry preview (see Figure 3). To provide a general overview, Lexonomy also displays the complete list of links for the dictionary (see Figure 4).

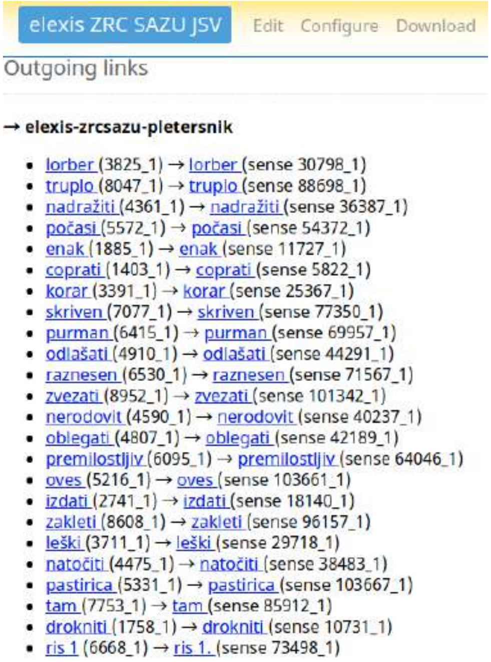
6.2 Automatic linking

For integration with automatic linking tools, Lexonomy provides API interface to work with the cross-links. As of now, the NAISC tool (McCrae & Buitelaar, 2018) is available for automatic linking directly from Lexonomy. Although the process was developed with the NAISC tool, it may be easily extended to work with other tools.

The process of automated link creation uses the following steps:

- user selects source and target dictionary,
- both dictionaries are converted to the OntoLex RDF format required by NAISC,
- NAISC detects the links,
- output from NAISC is converted to the internal Lexonomy format and stored in the database,
- links are available, and users may post-edit the results in Lexonomy editor.

As an input, NAISC requires files in the OntoLex RDF containing headword, part-of-speech, and definitions texts for each entry. Since we anticipate many dictionaries with various entry structures, users may not be able to configure linking elements



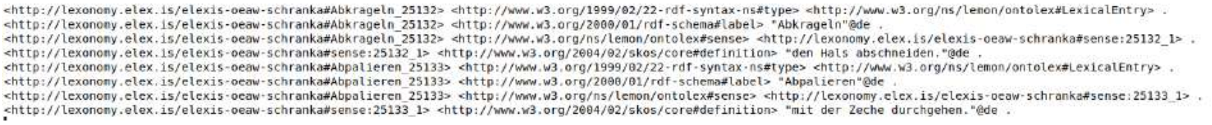
Outgoing links

→ elexis-zrcsazu-pletersnik

- lorber (3825_1) → lorber (sense 30798_1)
- truplo (8047_1) → truplo (sense 88698_1)
- nadražiti (4361_1) → nadražiti (sense 36387_1)
- počasi (5572_1) → počasi (sense 54372_1)
- enak (1885_1) → enak (sense 11727_1)
- coprati (1403_1) → coprati (sense 5822_1)
- korar (3391_1) → korar (sense 25367_1)
- skriven (7077_1) → skriven (sense 77350_1)
- purman (6415_1) → purman (sense 69957_1)
- odlašati (4910_1) → odlašati (sense 44291_1)
- raznesen (6530_1) → raznesen (sense 71567_1)
- zvezati (8952_1) → zvezati (sense 101342_1)
- nerodovit (4590_1) → nerodovit (sense 40237_1)
- oblegati (4807_1) → oblegati (sense 42189_1)
- premilostljiv (6095_1) → premilostljiv (sense 64046_1)
- oves (5216_1) → oves (sense 103661_1)
- izdati (2741_1) → izdati (sense 18140_1)
- zakleti (8608_1) → zakleti (sense 96157_1)
- leški (3711_1) → leški (sense 29718_1)
- natočiti (4475_1) → natočiti (sense 38483_1)
- pastirica (5331_1) → pastirica (sense 103667_1)
- tam (7753_1) → tam (sense 85912_1)
- drokniti (1758_1) → drokniti (sense 10731_1)
- ris 1 (6668_1) → ris 1. (sense 73498_1)

Figure 4: Example of dictionary overview of all available links (from the *JSV* dictionary to the *Pleteršnik's* dictionary, linking between senses of both dictionaries.

beforehand for each dictionary. In such a case, Lexonomy tries to guess the entry structure to provide all the data for NAISC – starting with the TEI-Lex0 entry structure, followed by several common elements for headwords and definitions (see Figure 5 for an example of OntoLex RDF export).



```
<http://lexonomy.elex.is/elexis-oeaw-schranka#Abkrageln_25132> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://www.w3.org/ns/lemon/ontolex#LexicalEntry> .
<http://lexonomy.elex.is/elexis-oeaw-schranka#Abkrageln_25132> <http://www.w3.org/2000/01/rdf-schema#label> "Abkrageln"@de .
<http://lexonomy.elex.is/elexis-oeaw-schranka#Abkrageln_25132> <http://www.w3.org/ns/lemon/ontolex#sense> <http://lexonomy.elex.is/elexis-oeaw-schranka#sense:25132_1> .
<http://lexonomy.elex.is/elexis-oeaw-schranka#sense:25132_1> <http://www.w3.org/2004/02/skos/core#definition> "den Hals abschneiden."@de .
<http://lexonomy.elex.is/elexis-oeaw-schranka#Abpalieren_25133> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://www.w3.org/ns/lemon/ontolex#LexicalEntry> .
<http://lexonomy.elex.is/elexis-oeaw-schranka#Abpalieren_25133> <http://www.w3.org/2000/01/rdf-schema#label> "Abpalieren"@de .
<http://lexonomy.elex.is/elexis-oeaw-schranka#Abpalieren_25133> <http://www.w3.org/ns/lemon/ontolex#sense> <http://lexonomy.elex.is/elexis-oeaw-schranka#sense:25133_1> .
<http://lexonomy.elex.is/elexis-oeaw-schranka#sense:25133_1> <http://www.w3.org/2004/02/skos/core#definition> "mit der Zeche durchgehen."@de .
```

Figure 5: Example of dictionary export into OntoLex RDF format.

7. Standardisation

The ELEXIS project develops a standardised data model for digitally-born dictionaries as part of the OASIS LEXIDMA technical committee[4]. When the standardised format is published, Lexonomy will switch to the LEXIDMA data model as a default template for dictionaries. Keeping the complete configurability of custom user formats, of course.

In the meantime, Lexonomy supports TEI-Lex0 (Romary & Tasovac, 2018) and the OnotoLex RDF format for ontologies (McCrae et al., 2017) as temporary formats. Lexonomy was updated to support both formats in API interfaces and to be integrated into automated lexicographic pipelines in the ELEXIS.

[4] https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=lexidma

8. Usage analysis

As of April 2021, over 2,700 users are working with Lexonomy. Altogether, they created more than 5,400 dictionaries containing over 34 million entries.

8.1 *OneClick Dictionary* dictionaries

Thanks to the connection with the Sketch Engine and its OneClick Dictionary tool, it is possible to create a new dictionary with the data from the corpus (e.g., headwords, examples). Utilising the OneClick Dictionary tool, users created 798 dictionaries in Lexonomy, which shows the popularity of automatic dictionary creation and post-editing. Most dictionaries cover a particular topic, e.g., terms from sports, medical science, or computer science. The most popular language with OneClick Dictionaries is English, followed by Czech, Italian and Latvian. Users created dictionaries in 30 different languages.

8.2 ELEXIS lexical resources

We have obtained 75 lexical resources from ELEXIS partners and observers (coming from 25 different institutions). The lexical resources range from different types of dictionaries, e.g., large general dictionaries, bilingual dictionaries, thesauri, specialised dictionaries (terminology, dialects), to lemma lists. Resources that were available in the XML format were directly uploaded to Lexonomy in their original format. Several resources were provided in different file formats, e.g., CSV or JSON. They were converted to the XML format before uploading to Lexonomy. Several dictionaries were provided in the PDF format, and these were converted to the XML format using the Elexifier tool. We list the largest resources (in terms of number of entries) in Table 1. Lexical resources provided by partners and observers are not publicly available, until licences are settled and exact access rights are specified. The Lexonomy application takes care of user accounts and access setting.

9. Conclusion

This paper summarises about two years of Lexonomy development. We introduced several features for a better user experience that attracted many new users to work with Lexonomy. Currently, over 2,700 users edit their dictionaries with Lexonomy, and we hope this number will grow even more. Other important updates include features that are integrating Lexonomy in various automated lexicographic pipelines. These integrations highlight the post-editing aspect of dictionary editing, and Lexonomy provides cutting-edge technologies even for small lexicographic teams or even one-person dictionary projects.

9.1 Future work

We are aware that the graphical user interface of Lexonomy is getting more cluttered with new features over time, and is also not suitable for work on mobile devices. On the developer side, currently used HTML templates are getting harder to maintain and extend. We decided to redesign and also refactor the user interface completely. The new

Lexical resource	**Institution**	**Licence**	**Number of entries**
Nova beseda frequency lexicon	ZRC SAZU Scientific Research Centre of Slovenian Academy of Sciences and Arts	CC BY 4.0	2,251,151
Svenska Akademiens Ordlista	Swedish Academy	open access license	984,823
Swedish Academy Dictionary	Swedish Academy	open access license	550,424
The Dictionary of Standard Estonian 2013	Institute of the Estonian Language	academic	425,766
Monier-Williams Sanskrit-English Dictionary	Cologne Center for Humanities	CC BY 3.0	398,412
Tezaurs Latvian	Institute of Mathematics and Computer Science, University of Latvia	CC BY-SA 4.0	320,869
The lemma list of the German dictionary "elexiko"	Leibniz Institute for the German Language	open access	275,756
Czech lemma lists	Institute of the Czech National Corpus	CC BY-SA 4.0	169,934
Dictionary of the Danish Language - ODS lemmas	The Society for Danish Language and Literature	restricted	163,012
Finnish dialect dictionary	Institute for the Languages of Finland	CC BY 4.0	161,148
Schweizerisehes Idiotikon	Schweizerisehes Idiotikon	CC BY-SA	160,254
Nords Ordbank - Bokmal	University of Bergen Library	CC-BY	153,939

Table 1: Selected lexical resources from ELEXIS partners, with more than 150,000 entries, sorted by the number of entries

user interface is currently in development, utilising the JavaScript component library Riot.js[5] and interface design framework Materialize[6].

As we keep including more dictionaries with an increasing number of entries (tens of thousands of entries is getting more common) in Lexonomy, we are constantly monitoring the database performance. If we notice a decrease in speed with large amounts of data, we will evaluate other databases to select the best storage for big lexicographic data.

10. Acknowledgements

The research received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. This work has been partly supported by the Ministry of Education of ČR within the LINDAT-CLARIAH-CZ project LM2018101.

11. References

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, 1. URL http://dx.doi.org/10.1007/s40607-014-0009-9.

Kilgarriff, A. & Jackson, H. (2013). Using corpora as data sources for dictionaries. *The Bloomsbury Companion to Lexicography. London: Bloomsbury*, pp. 77–96.

Krek, S., Kosem, I., McCrae, J.P., Navigli, R., Pedersen, B.S., Tiberius, C. & Wissik, T. (2018). European lexicographic infrastructure (elexis). In *Proceedings of the XVIII EURALEX International Congress on Lexicography in Global Contexts*. pp. 881–892.

McCrae, J.P., Bosque-Gil, J., Gracia, J., Buitelaar, P. & Cimiano, P. (2017). The Ontolex-Lemon model: development and applications. In *Proceedings of eLex 2017 conference*. pp. 19–21.

McCrae, J.P. & Buitelaar, P. (2018). Linking datasets using semantic textual similarity. *Cybernetics and information technologies*, 18(1), pp. 109–123.

McCrae, J.P., Tiberius, C., Khan, A.F., Kernerman, I., Declerck, T., Krek, S., Monachini, M. & Ahmadi, S. (2019). The ELEXIS interface for interoperable lexical resources. In *Proceedings of the sixth biennial conference on electronic lexicography (eLex)*. eLex 2019.

Měchura, M.B. et al. (2017). Introducing Lexonomy: an open-source dictionary writing and publishing system. In *Electronic Lexicography in the 21st Century: Lexicography from Scratch. Proceedings of the eLex 2017 conference*. pp. 19–21.

Romary, L. & Tasovac, T. (2018). TEI Lex-0: A target format for TEI-encoded dictionaries and lexical resources. In *TEI Conference and Members' Meeting.*

[5] https://riot.js.org/

[6] https://materializecss.com/

Lemmatisation, etymology and information overload on English and Swedish editions of Wiktionary

Allahverdi Verdizade

Uppsala University, P.O. Box 256, SE-751 05 Uppsala
E-mail: allahverdi.verdizade@lingfil.uu.se

Abstract

Wiktionary is a user-generated wiki-project with the goal of building a universal dictionary covering all words in all languages. Various language editions of Wiktionary have community-specific policies regulating concrete lexicographic questions. The distinct entry structures of English and Swedish Wiktionaries are examined in the context of the relation between headword and etymological information, under special consideration of the user-friendliness of the respective approach. The English Wiktionary applies the etymological approach in setting the headword, which splits identical forms into parts of speech, but also into headwords based on word origin. Additionally, the semantic information is separated from non-semantic more rigorously than is done in the Swedish Wiktionary, placing lists of related and derived terms below the headword rather than under each definition. The Swedish Wiktionary applies the formal-grammatical approach, where division into headwords is made strictly based on identical form and part of speech. In this approach, homonymy is disregarded. The etymological information is nested under each definition rather than having a separate section above the headword. The analysis of the two language editions suggests that the different approaches lead to different amounts of information overload in users, depending on the extent of non-semantic information. Equally extensive entries are handled better within the layout structure of the English Wiktionary.

Keywords: Wiktionary; information overload; etymology

1. Wiktionary, the universal dictionary

Wiktionary is a collaborative project aiming at creating a copyright-free, universal dictionary. The project declares as its goal nothing less than “describing all words in all languages”, including all living and extinct natural languages, as well as a selection of constructed languages. Wiktionary is currently available in 171 language editions. Each edition is characterised by information about the word, be it definitions, word etymologies, labels informing about the word’s register and usage, etc., provided in one meta-language[1]. Each language edition housed under a domain prefix (en., sv., de. etc) thus has only one meta-language, but contains entries and definitions of words in (potentially) all languages.

Language editions vary strongly in coverage, quality and growth rate. It is hardly surprising that the large languages have the highest number of entries: the English Wiktionary, hereafter referred to as *en.wikt*, has as of now 3.6 million definitions distributed over 2.6 million entries in 4,500 target languages, out of which English is the largest, with 550,000 entries (21% of all entries). Three other languages – Chinese, Finnish and Italian – are also particularly well-represented on en.wikt, having over 100,000 entries each, whereas some 3000 other languages are represented by fewer than 10 definitions each. The Swedish edition, sv.wikt, is much smaller, at 356,000 entries, out of which 83,000 are entries on words in Swedish. The ratio between entries in the meta-language and other languages is approximately the same (23% of all entries in sv.wikt are entries of Swedish words).

Size and quality do not always go together, and one of the largest editions was until recently that in Malagasy. Wiktionary in Malagasy was able to keep up with en.wikt for a long time in terms of amount of entries, but the key to success was not the cumulative work

[1] This is referred to as "native language" in Meyer & Gurevych (2012), which provides an excellent and well-informed introduction to Wiktionary

of an active community, but machine-translation coupled with bot-assisted mass-creation of entries entirely without subsequent human involvement with a low accuracy of glosses and generally poor quality of entries as a result. Therefore, even if the size of the lexical stock covered and growth rate are not always associated with the size of the active community, the overall quality tends to be. As such, en.wikt has 6,000 active editors, sv.wikt has 170, whereas the Malagasy edition has 14. An “active editor” is defined broadly as a user with at least one edit in the past month. As has been noted in the literature, a collaborative project needs to reach a “critical mass” of active editors in order for the lexicographical work to take off in earnest (Törnqvist, 2015).

2. Target audience and functions

Svensén (2009: p. 482-3) lists criteria that can be used to assess a dictionary. Some of the aspects to take into account when critically reviewing a dictionary are: 1. the amount of information provided by a dictionary, 2. the quality of the provided information, and 3. the way it is presented. It is emphasised that every dictionary review must depart from the dictionary’s own idea of the target audience and functions it intends to fill. Neither the quality (1) nor the quantity (2) of the word-stock provided by any edition of Wiktionary is within the scope of this paper: only the various approaches chosen to present it in the relation to lemmatisation (3) are examined. Fuertes-Olivera (2009) evaluates and compares the quantity and the quality of the coverage of English and Spanish lemmas on en.wikt at the time, although findings of a qualitative analysis of Wiktionary like this quickly become outdated in view of the high growth ratio of the project.

Compared to printed dictionaries, the aspects listed above can be somewhat hard to apply when dealing with web-based collaborative projects. Wiktionary is, strictly speaking, not a dictionary, but a dictionary project, which unlike most products developed by private companies or other organizations (referred to as "institutional internet reference works" by Fuertes-Olivera (2009) is not intended to be complete within a certain time framework. This is partly due to the declared goal of “describing all words in all languages”, partly because human languages are in a constant state of change, with new words and senses emerging by the day, while others fall out of use or change their meaning. Seen from this perspective, all Wiktionary editions have the same, next to indefinite, potential to grow and to be reworked. This is only limited by the number of active editors and their interest in different aspects of lexicographic work.

The formal absence of a target audience must therefore be addressed for a meaningful analysis to be possible. I will therefore exclude from the following groups of users: 1. language learners, typically benefiting from information about a word’s formal, semantic and pragmatic aspects. The core vocabulary, i.e. 2,000 of the most frequent words or so, is of primary interest for this group. Examples of usage and collocations are also of uttermost importance. 2. Users looking up words in their native language, such as less frequent words, specialist vocabulary, neologisms, controversial terms or usage prescription. The needs of both above-mentioned groups may include both reception and production; semantic relations (synonyms, antonyms) are thus important. 3. Users interested in linguistic history: here, word etymologies are of primary interest. The potential of Wiktionary is perhaps greatest precisely in this area, and its importance (at least that of en.wikt) in academic contexts as a resource for both finding etymological information and data for novel etymological research becomes increasingly salient (see, for example, Meyer &

Gurevych (2012); Khoury & Sapsford (2016); Sagot (2017) to name a few). It has at times even been proposed that Wiktionary *is* above all an etymological dictionary[2][3], constituting a secondary source, which, unlike tertiary sources, not only accounts for and summarises published research, but also evaluates its adequacy, comments, and complements it[4]. In view of this, group 3 is perhaps as important as the first two, which usually are the main target audience of a dictionary.

Finally, a fourth group of users can be discerned, since Wiktionary is a project run by unpaid enthusiasts: the editors themselves. They may be representatives of groups 2 and 3, too, and in addition to that native speakers of a project's meta-language, and thus might not have the language learner's perspective in mind. Paradoxically, absence of formally stated target audience can make the editors a target audience in themselves: the unpaid community of hobby lexicographers compiles entries (first of all) for their own community, constituting the primary readership and critics.

This may also be the reason why en.wikt can be perceived as less helpful for learners of English: if the main bulk of the editors are native speakers of English, they might not be interested in contributing information that would help learners of their language, disambiguating definitions, adding synonyms and example usages etc. This is hardly unique for Wiktionary, as monolingual dictionaries are normally written by native speakers regardless of medium. In the case of Wiktionary, however, there is no commissioner to set "production goals" regarding content and time framework. One could argue that en.wikt is not intended for learners of English: however, making English entries more elaborate and user-friendly is of course a legitimate way of contributing, and it also makes it more useful for learners of English. Thus, en.wikt being less suitable for learners of English is not a result of a specific policy, but a consequence of most editors' backgrounds and fields of interests.

The functions filled by Wiktionary can be inferred from the target groups listed above. Another function, that can be hard to tie to any of the above, is that which can be inferred from the slogan "all words of all languages" – that of documentation. A potential target audience benefiting from this is possibly researchers, enthusiasts and activists of linguistic revitalisation and language technology developers.

If the assumption put forward by Gouws & Tarp (2017) regarding too much information being at odds with the needs of users to the same extent as too little is to be accepted, it is easy to see that there is a potential conflict between the will to document everything and degree of user-friendliness. As they note: "In many consultation procedures where problems are experienced there is little doubt that the provision of less lexicographic data would have raised the success rate" (ibid.: 896). Removing valid lexicographic data from

[2] User Widsith, 2018.11.14, in Beer Parlour, internal discussion page: "I think earliest senses should be first, including when they're obsolete, as in any historical dictionary (which Wiktionary is, like it or not)".

[3] User KevinUp, 2019.05.10, BP, "Since Wiktionary is an etymological dictionary, I would prefer to see native Japanese words being lemmatized at their kana forms and Sino-Japanese terms lemmatized at their kanji forms".

[4] User Rua, 2015.09.1, BP, "Hence, the question that still remains to be answered is whether Wiktionary is an etymological dictionary (secondary source with its own interpretations) or an encyclopedia/compendium of etymological research (tertiary source). Currently, Wiktionary is an etymological dictionary/secondary source as it contains its own interpretations of the data."

Wiktionary is, however, disallowed. One can only seek to relieve the information overload that occurs in the reader, i.e. by reorganising the content visually.

The user groups listed in this section may seem a case of unnecessary coinage of novel terminology, considering the well-established concept of consulting situations, such as reception, perception, translation, etc. However, these would be more relevant for an investigation of the *contents* of Wiktionary, rather than its layout structure. The relationship between entry e.g. entry structure and the etymology affects readers in all these consulting situations to the same degree. It does not mean, however, that we cannot draw between the user typology proposed here and a traditional typology of dictionaries, as suggested for example by Tarp (2017: p. 247):

Adapted from Tarp (2017)		Target groups proposed in current paper
communicative	assist users in solving problems related to written and oral communication, such as text reception, text production, translation and text revision	language learners; users looking up words in their native language
cognitive	transmit knowledge to their users	readers interested in language history; Wiktionary editors; researchers
operative	assist users in performing specific types of action	language learners
interpretive	assist users in interpreting non-linguistic signs	-

3. The overall structure

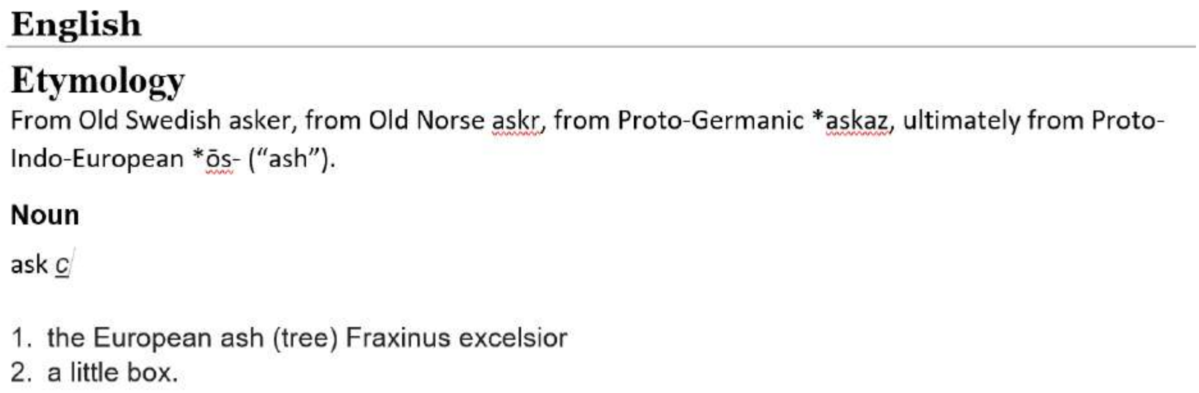
The starting point in the access structure at Wiktionary is spelling, which means that words in different languages are displayed alphabetically on the same page[5]. The entry layout is originally not developed for the purposes of a dictionary, but for encyclopaedic articles, whence it has been "inherited" and subsequently adjusted to a certain degree, making it radically different from a printed dictionary in several ways. The alphabetical order of entries within a language is not visible for the reader: although the sought entry can be reached by consulting the alphabetical index, the usual way is by using the search function. In order to compensate for the absence of a natural connection with other relevant entries (which can often be found on the same or adjacent pages in printed dictionaries), hyperlinks are used to refer to derived terms, compounds or otherwise related terms. Entries interconnected through semantic relationships (synonymy, antonymy etc.), that are normally not found next to each other in printed dictionaries, are also connected via hyperlinking.

Except for some very general principles applying over the edition boundaries (such as criteria for inclusion[6]), specific lexicographic policies are decided over by the local communities of each edition. One such policy is the question of lemmatisation, or "how lexical units with identical citation forms be presented" Svensén (2004). The differences in how this affects the entry layout in each edition can be exemplified with two constructed entries from the focal editions.

[5] However, entries in meta-language are displayed at the top regardless of the language name's initial letter. English always comes first on the page on en.wikt, Swedish on sv.wikt, etc.

[6] Some differences regarding which words may be included do exist, too: i.e. given names as well as surnames may be included on en.wikt but are not permitted on sv.wikt.

Figure 1: A simplified basic entry on en.wikt



English

Etymology

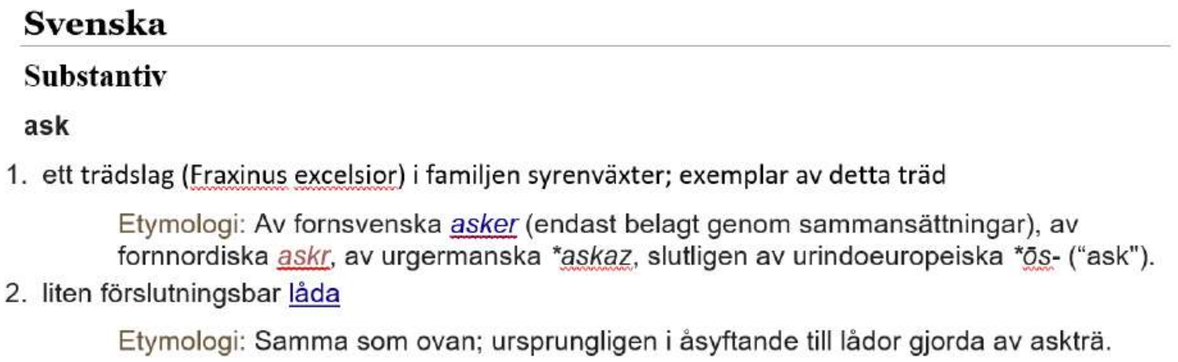
From Old Swedish asker, from Old Norse askr, from Proto-Germanic *askaz, ultimately from Proto-Indo-European *ōs- ("ash").

Noun

ask c

1. the European ash (tree) Fraxinus excelsior
2. a little box.

Figure 2: A simplified basic entry on sv.wikt



Svenska

Substantiv

ask

1. ett trädslag (Fraxinus excelsior) i familjen syrenväxter; exemplar av detta träd

   Etymologi: Av fornsvenska *asker* (endast belagt genom sammansättningar), av fornnordiska *askr*, av urgermanska **askaz*, slutligen av urindoeuropeiska **ōs-* ("ask").
2. liten förslutningsbar låda

   Etymologi: Samma som ovan; ursprungligen i åsyftande till lådor gjorda av askträ.

The main difference lies in how the entry is organised in relation to etymology: while en.wikt structures the content (primarily) around individual etymologies, it is organised (primarily) around the part of speech on sv.wikt. The contrast is most visible in the order of headers: the etymology section constitutes a higher-order section on en.wikt, and the etymological information is given above the definitions, at its own top-level on the page. On sv.wikt, the etymological information is provided inside the lexeme, under each definition. As can be seen, the division into parts of speech constitutes the higher-order hierarchy on sv.wikt, whereas it is subordinate to etymologies on en.wikt. It could be argued that sv.wikt has moved further away from the encyclopaedic entry layout inherited from Wikipedia and done away with the level in the page structure hierarchy, which on en.wikt is made up by the etymology section. The etymology has ceased to be central part of the macro-structure and is demoted to the micro-structure, under the individual definitions. The contrast can be presented schematically, and compared with printed dictionaries in Table (1).

The Swedish word *ask* featured in the constructed entries above presents a case of polysemy: the sense 'a little box' developed from the primary sense 'ash (tree)'. This simplistic example does therefore not fully reflect the contrast in entry structure brought about by the different approaches to lemmatisation adopted by each edition, which is most evident with regard to homonyms. The constructed entries below exemplify each edition's approach to the homonymous English word *bore* (figures 6 and 7), belonging to

Table 1: Comparison of layout structure in printed dictionaries, en.wikt and sv.wikt

Printed dictionaries	English Wiktionary	Swedish Wiktionary
Page (all lemmata sorted alphabetically, fitting in a single paper page)	Page (all lemmata with identical spelling)	Page
↓	↓	↓
↓	Etymologies (lemmata which can be derived from the same source)	↓
↓	↓	↓
Lemmata (independent entries with identical formal properties: spelling, part of speech, declension/conjugation)	Lemmata	Lemmata
↓	↓	↓
Definitions	Definitions	Definitions

several parts of speech[7]. Note the striking difference in the amount of screenspace used by the entries: the entry on en.wikt is visually much larger than the one on sv.wikt.

4. Lemmatisation

The principles that can be discerned behind the organisation of the entry structure can and should be contextualised within the ones traditionally applied in printed dictionaries. A central reason for the different appearance of the entries on en.wikt and sv.wikt is lemmatisation. Below follows a short review of how it is approached in paper dictionaries, and, by extension, how the question of polysemy vs. homonymy is resolved there. Svensén (2004) lists four approaches: the etymological, the semantic, the morpho-semantic and the formal-grammatical. These four approaches can also be seen as four ways of answering the question “what is a word, in the lexicographical sense?” (and, by extension, “what is another?”).

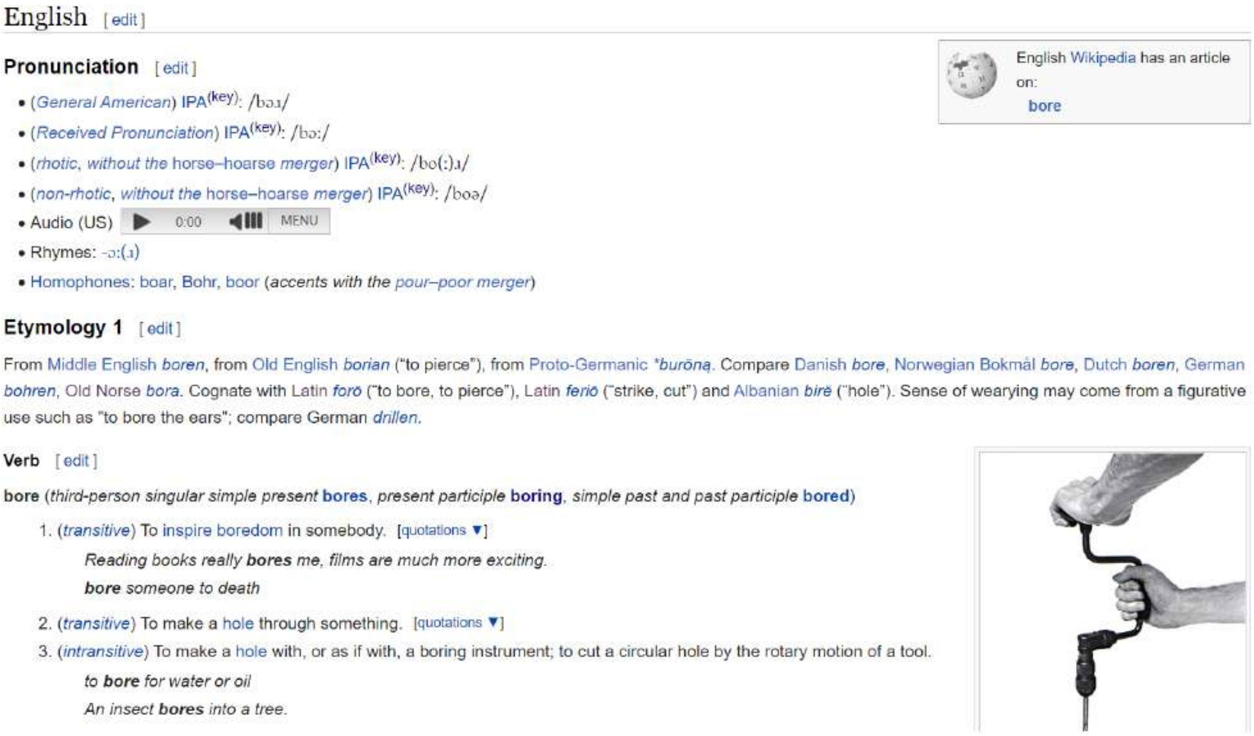
4.1 Approaches to lemmatisation in printed dictionaries

The etymological method[8] in its strict application departs from wordhood based on forms of shared origin. Such lexical units are treated as polysemous and lemmatised under the same entry. The readers’ intuitions regarding which forms belong together

[7] Certain departures were made from the actual entries in order to secure the same amount of information in both constructed entries. For example, the entry on en.wikt is in reality much larger and the one on sv.wikt is smaller. Some lemmata belonging to other parts of speech have been left out. The translation section in the Swedish entry is given merely for comparability, as translations to other target languages are only allowed from the entry in the meta-language. Figures (3) and (4) show parts of the actual entries

[8] The English-language edition of Svensén (2009) does not include the etymological approach as a distinct way of organizing the entries and concludes further that ”the place of etymology in the micro-structure is usually uncomplicated“. Since our analysis suggests that etymology is far from uncomplicated in the context of Wiktionaries, we will utilise the original analysis proposed in Svensén (2004).

Figure 3: Parts of actual entries for the word *bore*



English [edit]

Pronunciation [edit]

- (*General American*) IPA(key): /bɔɹ/
- (*Received Pronunciation*) IPA(key): /bɔː/
- (*rhotic, without the horse–hoarse merger*) IPA(key): /bo(ː)ɹ/
- (*non-rhotic, without the horse–hoarse merger*) IPA(key): /boə/
- Audio (US) ▶ 0:00 MENU
- Rhymes: -ɔː(ɹ)
- Homophones: boar, Bohr, boor (*accents with the pour–poor merger*)

Etymology 1 [edit]

From Middle English *boren*, from Old English *borian* ("to pierce"), from Proto-Germanic **burōną*. Compare Danish *bore*, Norwegian Bokmål *bore*, Dutch *boren*, German *bohren*, Old Norse *bora*. Cognate with Latin *forō* ("to bore, to pierce"), Latin *feriō* ("strike, cut") and Albanian *birë* ("hole"). Sense of wearying may come from a figurative use such as "to bore the ears"; compare German *drillen*.

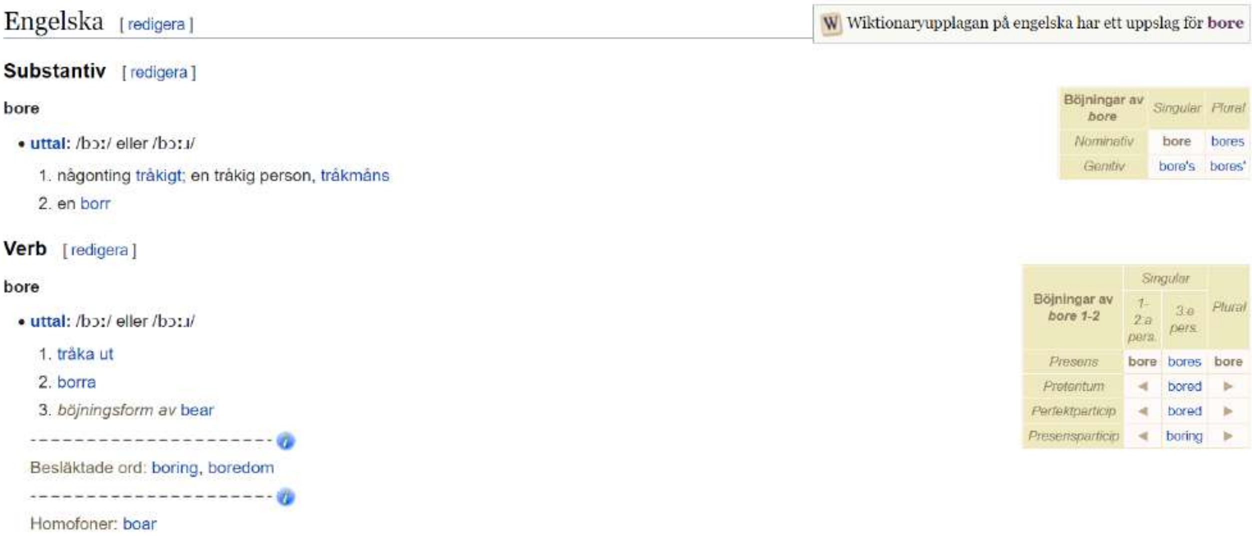
English Wikipedia has an article on: bore

Verb [edit]

bore (*third-person singular simple present* **bores**, *present participle* **boring**, *simple past and past participle* **bored**)

1. (*transitive*) To inspire boredom in somebody. [quotations ▼]
   *Reading books really **bores** me, films are much more exciting.*
   ***bore** someone to death*
2. (*transitive*) To make a hole through something. [quotations ▼]
3. (*intransitive*) To make a hole with, or as if with, a boring instrument; to cut a circular hole by the rotary motion of a tool.
   *to **bore** for water or oil*
   *An insect **bores** into a tree.*

Figure 4: Parts of actual entries for the word *bore*



Engelska [redigera]

Wiktionaryupplagan på engelska har ett uppslag för **bore**

Substantiv [redigera]

bore

- **uttal:** /bɔː/ eller /bɔːɹ/
  1. någonting tråkigt; en tråkig person, tråkmåns
  2. en borr

Böjningar av *bore*	*Singular*	*Plural*
Nominativ	bore	bores
Genitiv	bore's	bores'

Verb [redigera]

bore

- **uttal:** /bɔː/ eller /bɔːɹ/
  1. tråka ut
  2. borra
  3. *böjningsform av* bear

Besläktade ord: boring, boredom

Homofoner: boar

Böjningar av *bore 1-2*	*Singular*		*Plural*
	1-2:a pers.	*3:e pers.*	
Presens	bore	bores	bore
Preteritum	◀	bored	▶
Perfektparticip	◀	bored	▶
Presensparticip	◀	boring	▶

are thereby of no importance. Words demonstrating identical formal properties (part of speech, inflection, pronunciation) but unrelated historically are seen as homonymous, unrelated forms merely coinciding on the surface and are treated under separate entries. As the name suggests, this approach is best suited for etymological dictionaries, but the principle has been adopted in general-purpose dictionaries too, such as the *Concise Oxford English Dictionary* (2011), which groups lemmata by word origin.

The semantic approach[9], on the other hand, disregards etymology and groups words by (groups of) meaning. Words are treated as homonymous when their senses are deemed to be too divergent. Etymologically related words like the Swedish *ask* (1. ‘a kind of tree; 2. ‘a small box’) are divided between two entries, whereas i.e. English *crown* (‘a royal headdress; the top of a tree’) is viewed as polysemous and lemmatised under one entry. This approach is well suited for general-purpose dictionaries.

The morpho-semantic approach[10] has the same view on the relationship between etymology and semantics as the previous approach but implies a more learner-friendly macro-structure since semantically related groups of lemmas are given under one “super-lemma” chosen to represent the word-family. This model deviates from formal properties (alphabetical sorting, part of speech and inflection) as a base for the access structure to a larger extent and lemmatises all members of the word-family under the main lemma (cf. Swedish *basal* ‘basal’ ADJ, *basning* ‘steaming’ VN, *basera* ‘to base’ V, under the superlemma *bas* ‘base’ N.). In addition to the learner-friendliness of this approach, it is also a natural choice for languages relying heavily on prefixation for word-formation, such as Indonesian (e.g. in Korigodskiy et al. (1990)), as it allows us to quickly find derived forms which otherwise would end up in another part of the volume. At the same time, Svensén (2004: p. 124) puts forward the argument that it can be harder, not easier, for the reader to arrive at the sought word if he isn’t able to identify the super-lemma[11]. However, since Wiktionaries only have one form per language and page (entries with distinct spellings are not listed on under the same page and can be accessed via the search function), this weakness does not really apply.

The fourth logical way of handling homonymy and polysemy is the formal-grammatical approach[12], which bases lemmatisation entirely on formal properties of a word without any reference to either etymology or semantics. All forms with identical spelling, part of speech and inflection are treated under the same entry.

Although the formal-grammatical approach eliminates the need for making decisions on how to divide formally identical words into several entries based on extra-linguistic (historical) and semantic grounds, thus speeding up the compilation, it does have drawbacks, too. It is not well-suited for an etymological dictionary and, at the same time, can be somewhat counter-intuitive for readers looking up polysemous/homonymous words in their native language, where it can be assumed that semantic groupings and sub-groupings would facilitate successful look-up. The approach is fully implemented in the latest edition of the printed *Swedish Academic Word List* (SAWL, 2015), which focuses on listing the vocabulary of the Swedish language and attaches less importance to definitions.

[9] Svensén (2009) provides a more clear-cut typology and calls this ”macro-structure oriented homonymization of core senses“ (p.366)

[10] ”Homonymization of individual senses“ (Svensén, 2009: p. 365) and ”non-strict-alphabetical macro-structure“ (pp. 374-276)

[11] As such, it can be challenging for the learner to recognise that the Indonesian *menyerahkan* ‘to hand over’ should be looked up under *serah* ‘to give up’ unless the former is referring to the latter in the overall alphabetic structure in addition to being placed under the base-form; providing such reference for all derived forms easily becomes exceedingly space-consuming in a printed dictionary, since all verbs have a derived form prefixed with *me-*

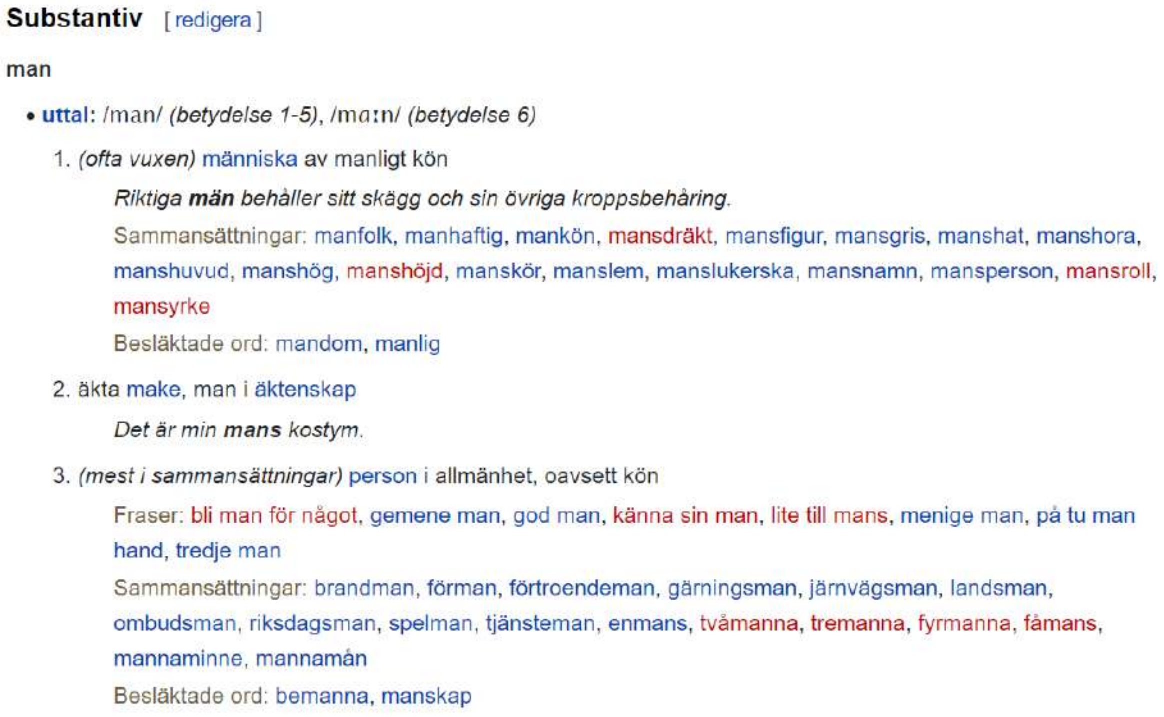
[12] “Strict alphabetical macrostructure” (Svensén, 2009: p. 371-374)

4.2 Approaches to lemmatisation on Wiktionary

The universality in respect to target groups and purpose reflects the relation to lemmatisation described above: elements pertaining to all three methods can be identified. En.wikt is organised almost entirely according to the etymological approach, but its lemmatisation strategy is in a sense even more radically etymological compared to printed dictionaries: all etymologically related lexemes with identical forms are treated under the same lemma. Several lexemes with distinct formal properties are organised under the same etymology section or divided between several sections if they have different etymologies. The English term *base* is divided between four etymologies: etymology 1 contains subsections both for the noun and the verb *base*, etymology 2 only the adjective *base* etc.

The fundamental structure of sv.wikt is, first and foremost, in line with the formal-grammatical approach, part of speech and inflection are central for lemmatisation. The etymological information is nested under one or several senses by means of so called templates, which automatise the formatting (the position, font size and colour) of different elements. Nesting of links to related terms, such as compounded forms, can be viewed as incorporation of elements of the morpho-semantic approach.

Figure 5: Elements of the morpho-semantic approach implemented on sv.wikt: compounded forms (*sammansättningar*), related terms (*besläktade ord*) and phrases (*fraser*) linked to from the relevant senses of the lemma *man* '1. male 2. husband 3. person'.



Substantiv [redigera]

man

- **uttal**: /man/ *(betydelse 1-5)*, /mɑːn/ *(betydelse 6)*

1. *(ofta vuxen)* människa av manligt kön

   Riktiga ***män*** *behåller sitt skägg och sin övriga kroppsbehåring.*

   Sammansättningar: manfolk, manhaftig, mankön, mansdräkt, mansfigur, mansgris, manshat, manshora, manshuvud, manshög, manshöjd, manskör, manslem, manslukerska, mansnamn, mansperson, mansroll, mansyrke

   Besläktade ord: mandom, manlig

2. äkta make, man i äktenskap

   Det är min ***mans*** *kostym.*

3. *(mest i sammansättningar)* person i allmänhet, oavsett kön

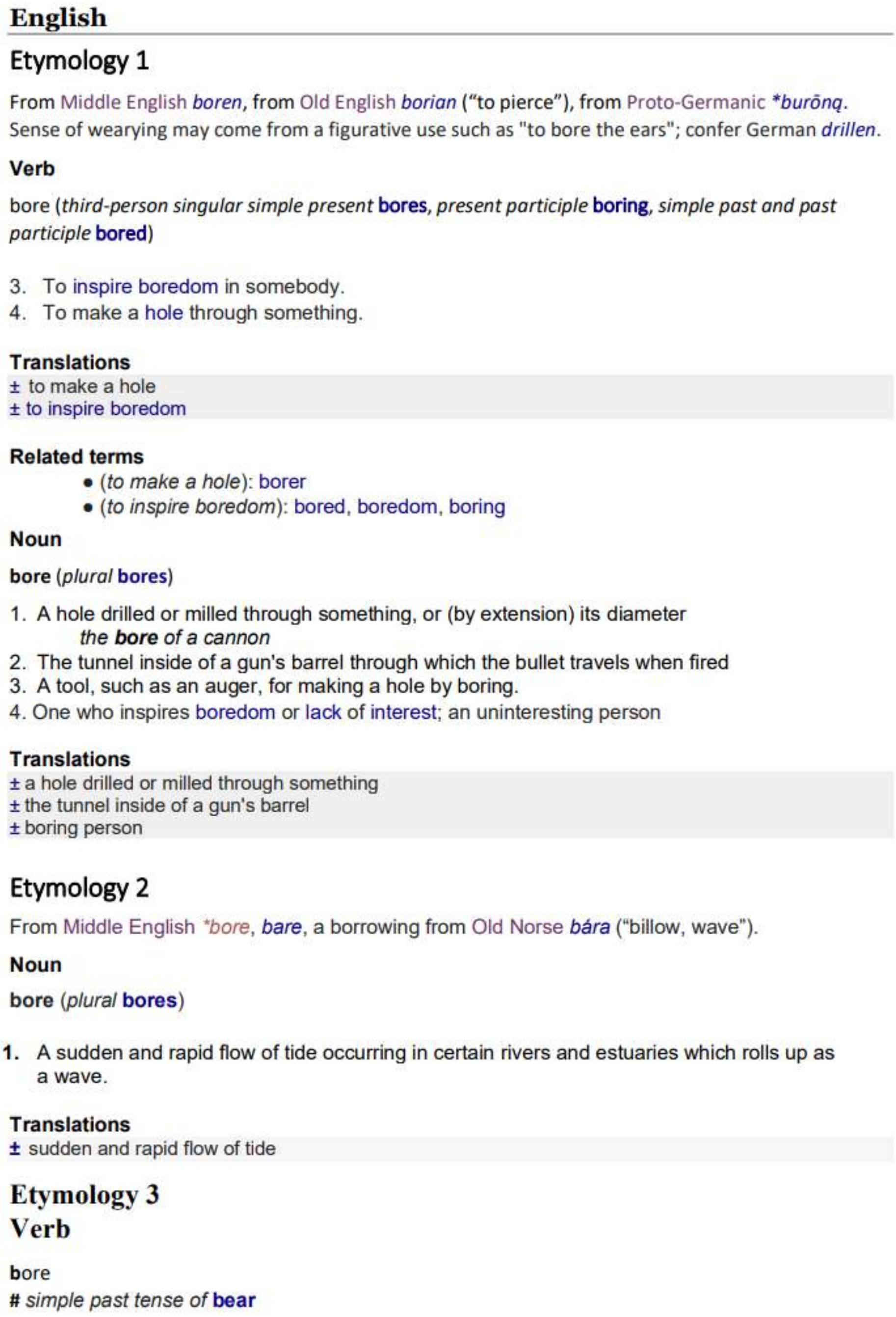
   Fraser: bli man för något, gemene man, god man, känna sin man, lite till mans, menige man, på tu man hand, tredje man

   Sammansättningar: brandman, förman, förtroendeman, gärningsman, järnvägsman, landsman, ombudsman, riksdagsman, spelman, tjänsteman, enmans, tvåmanna, tremanna, fyrmanna, fåmans, mannaminne, mannamån

   Besläktade ord: bemanna, manskap

In sum, the community of sv.wikt decided to move away from the structure inherited from Wikipedia to a further extent in order to get closer to the formal-grammatical method. Remnants of the original layout can still be found in some entries: i.e. the entry *person* has etymology as a separate section under the noun rather than having a template inside the definitions. Sv.wikt's layout policy page, *Stilguiden*, is states that this way of including

Figure 6: A simplified homonymous entry on en.wikt



English

Etymology 1

From Middle English *boren*, from Old English *borian* ("to pierce"), from Proto-Germanic **burōną*. Sense of wearying may come from a figurative use such as "to bore the ears"; confer German *drillen*.

Verb

bore (*third-person singular simple present* **bores**, *present participle* **boring**, *simple past and past participle* **bored**)

3. To inspire boredom in somebody.
4. To make a hole through something.

Translations

± to make a hole
± to inspire boredom

Related terms

- (*to make a hole*): borer
- (*to inspire boredom*): bored, boredom, boring

Noun

bore (*plural* **bores**)

1. A hole drilled or milled through something, or (by extension) its diameter
   *the **bore** of a cannon*
2. The tunnel inside of a gun's barrel through which the bullet travels when fired
3. A tool, such as an auger, for making a hole by boring.
4. One who inspires boredom or lack of interest; an uninteresting person

Translations

± a hole drilled or milled through something
± the tunnel inside of a gun's barrel
± boring person

Etymology 2

From Middle English **bore*, *bare*, a borrowing from Old Norse *bára* ("billow, wave").

Noun

bore (*plural* **bores**)

1. A sudden and rapid flow of tide occurring in certain rivers and estuaries which rolls up as a wave.

Translations

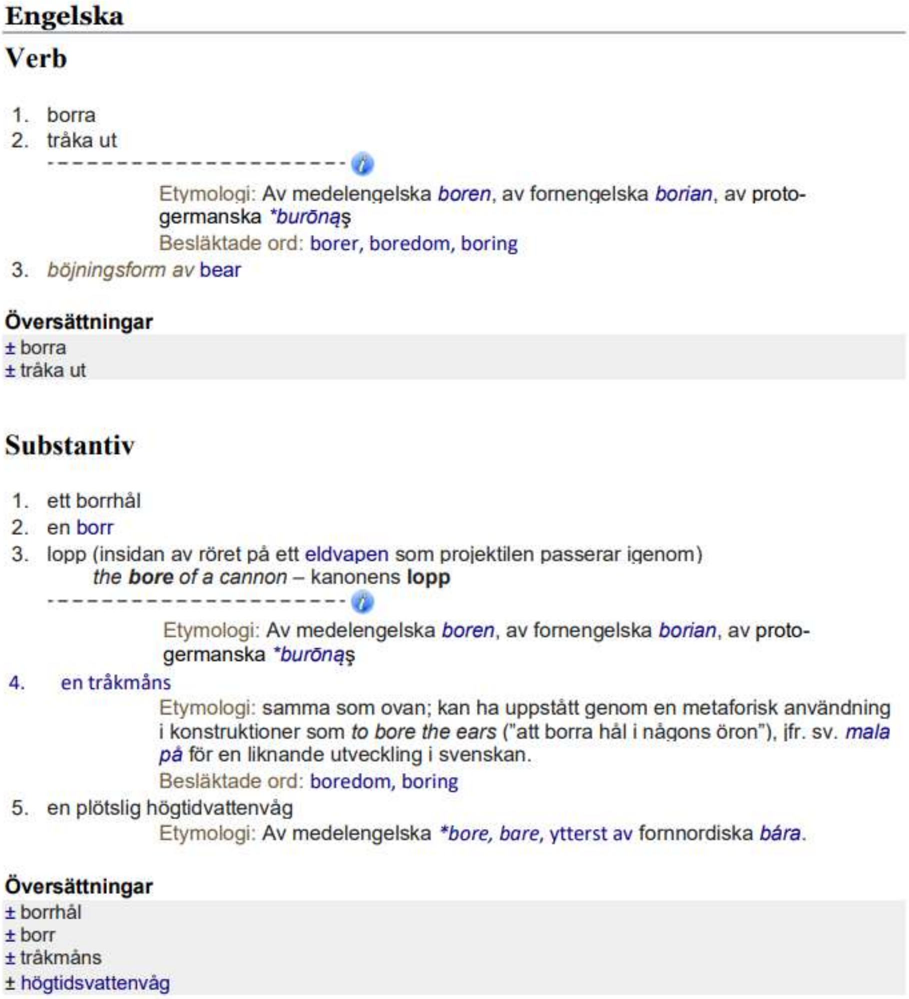
± sudden and rapid flow of tide

Etymology 3
Verb

bore
# simple past tense of **bear**

etymological information is being phased out. En.wikt, on the other hand, has retained a more encyclopedic layout in order to structure entries around shared origin and, by keeping the screenspace intended for formal and semantic properties of the word visually apart from the screenspace intended for etymologies, created a solid groundwork for inclusion of elaborate etymological information. Indeed, insufficient space has historically limited proper etymologisation in printed dictionaries, e.g. when it comes to derived terms (Buchi, 2016: p. 345), and in order to fully utilise the advantages of the paperless format, access to enough (screen)space for the etymology section must be assured in one form or another.

Figure 7: A simplified homonymous entry on sv.wikt



Engelska

Verb

1. borra
2. tråka ut

   Etymologi: Av medelengelska *boren*, av fornengelska *borian*, av proto-germanska **burōnąȿ*

   Besläktade ord: borer, boredom, boring

3. *böjningsform av* bear

Översättningar

± borra
± tråka ut

Substantiv

1. ett borrhål
2. en borr
3. lopp (insidan av röret på ett eldvapen som projektilen passerar igenom)

   *the **bore** of a cannon* – kanonens **lopp**

   Etymologi: Av medelengelska *boren*, av fornengelska *borian*, av proto-germanska **burōnąȿ*

4. en tråkmåns

   Etymologi: samma som ovan; kan ha uppstått genom en metaforisk användning i konstruktioner som *to bore the ears* ("att borra hål i någons öron"), jfr. sv. *mala på* för en liknande utveckling i svenskan.

   Besläktade ord: boredom, boring

5. en plötslig högtidvattenvåg

   Etymologi: Av medelengelska **bore, bare*, ytterst av fornnordiska *bára*.

Översättningar

± borrhål
± borr
± tråkmåns
± högtidsvattenvåg

4.3 Implications for target groups

Taking apart definitions of a word and placing them in several sections based on origin (as done on en.wikt) can cause inconvenience for the casual reader uninterested in linguistic history and potentially impede a successful look-up. At the same time, it clears the micro-structure of all non-semantic information: no etymological information is given in the visual vicinity of definitions, being placed in a specially designated section. The part of speech section is reserved for definitions and language samples in the form of user-constructed example sentences, collocations and quotations. Lexical relations, such as synonyms and antonyms, which are deemed to be valuable for comprehension of the sense, are allowed too.

The etymology section is often made up of a short list of attested or reconstructed historical word-forms ancestral to the word in question and cognates in related languages, but there are also many instances of elaborate and sourced inquiries of a words history, including discussion of possible directions of borrowing, semantic shifts and typological parallels. Such inquiries often have a very high academic standard. In view of the very large number of contributors at en.wikt (as compared to sv.witk), often with special interest in language history, it is not uncommon to see etymology sections of rather extensive size.

Having them nested among definitions would make the latter very hard to navigate, and likely reduce the editors' disposition to compile the often space-demanding review of the existing research, which should ideally be the basis of every etymology.

Derivations, otherwise related terms and translations, links to descendants in other languages all have their own sections visually separated from definitions. This results in an overall page structure with many sections and subsections. This might not be a problem for the seasoned readers of en.wikt, but it should be kept in mind that it was originally developed for encyclopaedic articles with relatively large amount of text in each section. Therefore, navigating a page with many sections, several of which only contain lists of links to other entries, can be challenging to first-time visitors, as it requires a lot of screenspace.

The question is, however, whether the overload of etymological information in the micro-structure (nested under the definitions) would not imply an even more severe impediment to successful look-up than a messy macro-structure. Compare the Swedish noun *bas*, mentioned by (Svensén, 2004: §52) as an example of a polysemous/homonymous word. The *Swedish Academic Dictionary* (SAD, the standard reference work for Swedish etymologies) lists five distinct homonyms belonging to the form. At present, the word encompasses 16 senses unsorted for etymology on sv.wikt. These senses could probably be derived from more than five etymologies provided by SAD at the time of the entry's compilation in the year 1900, as novel senses have emerged since. If fairly complete etymological information would be added under the definitions of the word on sv.wikt, the navigation and possibility for successful look-up would deteriorate for historically interested readers and learners alike.

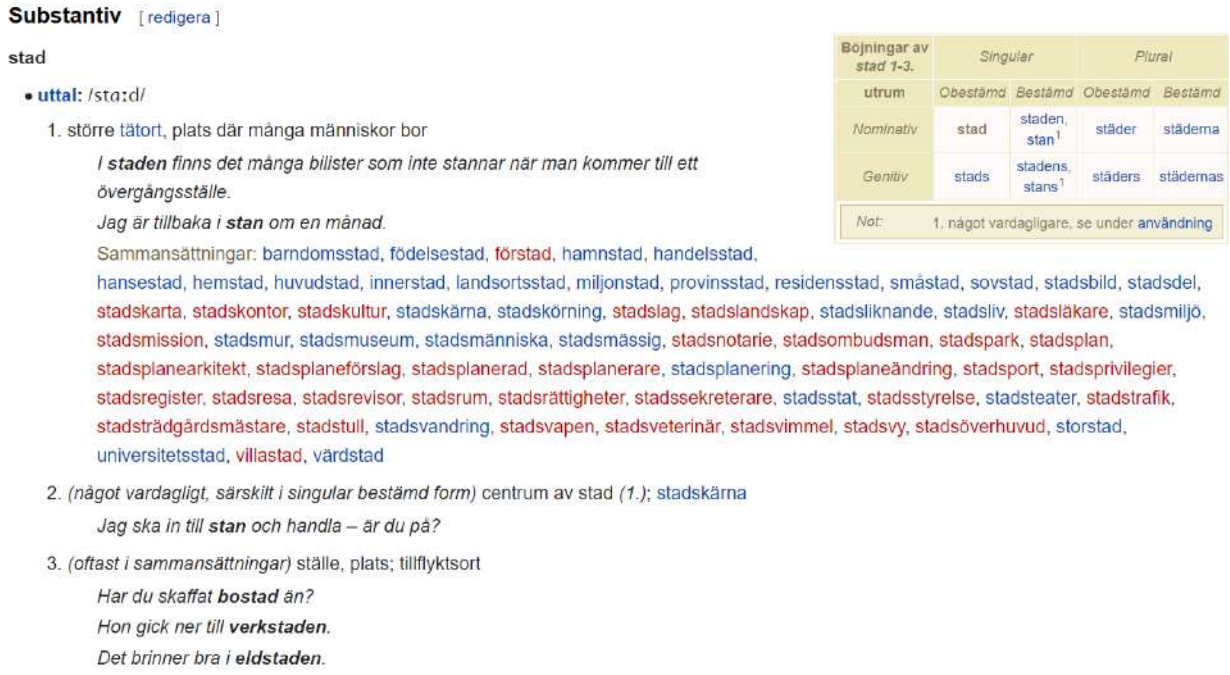
However, this is in reality not much of a problem for sv.wikt in view of the fact that elaborate etymology sections are at present rare in homonymous words. It is not clear whether this depends on the entry layout reducing the willingness to compile elaborate etymologies, the small number of active editors, or a combination of both factors.

Considering the groups of users outlined at the beginning of this paper, it can safely be assumed that native speakers without interest in etymology and advanced learners benefit from this state of affairs at sv.wikt, as they are unlikely to look up highly frequent words. The latter are precisely the type of words that tend to be polysemous, homonymous and serve as bases of derivation for a great number of terms. Learners and readers who take interest in etymologies are more likely to look up frequent words with a potential for overloaded micro-structure. In particular, the decision to rely on templates nested under definitions for etymological information could discourage potential editors with interest in language history from making elaborate contributions.

5. Information overload

As indicated above, the extensive amount of etymological information on en.wikt results in slower look-up due to the definitions being split between several etymology sections, whereas sv.wikt is spared from this side-effect due to comparatively low amount of etymological information. The incorporation of elements of the morpho-semantic approach into sv.wikt, however, has a potential to slow-down the look-up, too. As such, compare the entry *stad* at sv.wikt (fig. 8) and en.wikt (fig. 9), where compounded terms are visually separated from the defintions to a greater extent. Both the messy macro-structure, caused by splitting of the definitions between several etymology sections and the messy micro-structure caused by piling up of elements irrelevant for understanding the sense of the word in question are ultimately the results of the goal of including everything there is to say about a word ("all words of all languages").

Figure 8: Excessive nesting of compounded terms into the micro-structure of the sv.wikt entry *stad*



Substantiv [redigera]

stad

- **uttal:** /stɑːd/

1. större tätort, plats där många människor bor

   *I **staden** finns det många bilister som inte stannar när man kommer till ett övergångsställe.*

   *Jag är tillbaka i **stan** om en månad.*

   Sammansättningar: barndomsstad, födelsestad, förstad, hamnstad, handelsstad, hansestad, hemstad, huvudstad, innerstad, landsortsstad, miljonstad, provinsstad, residensstad, småstad, sovstad, stadsbild, stadsdel, stadskarta, stadskontor, stadskultur, stadskärna, stadskörning, stadslag, stadslandskap, stadsliknande, stadsliv, stadsläkare, stadsmiljö, stadsmission, stadsmur, stadsmuseum, stadsmänniska, stadsmässig, stadsnotarie, stadsombudsman, stadspark, stadsplan, stadsplanearkitekt, stadsplaneförslag, stadsplanerad, stadsplanerare, stadsplanering, stadsplaneändring, stadsport, stadsprivilegier, stadsregister, stadsresa, stadsrevisor, stadsrum, stadsrättigheter, stadssekreterare, stadsstat, stadsstyrelse, stadsteater, stadstrafik, stadsträdgårdsmästare, stadstull, stadsvandring, stadsvapen, stadsveterinär, stadsvimmel, stadsvy, stadsöverhuvud, storstad, universitetsstad, villastad, värdstad

2. *(något vardagligt, särskilt i singular bestämd form)* centrum av stad *(1.)*; stadskärna

   *Jag ska in till **stan** och handla – är du på?*

3. *(oftast i sammansättningar)* ställe, plats; tillflyktsort

   *Har du skaffat **bostad** än?*

   *Hon gick ner till **verkstaden**.*

   *Det brinner bra i **eldstaden**.*

Böjningar av *stad* 1-3.	*Singular*		*Plural*	
utrum	*Obestämd*	*Bestämd*	*Obestämd*	*Bestämd*
Nominativ	stad	staden, stan[1]	städer	städerna
Genitiv	stads	stadens, stans[1]	städers	städernas

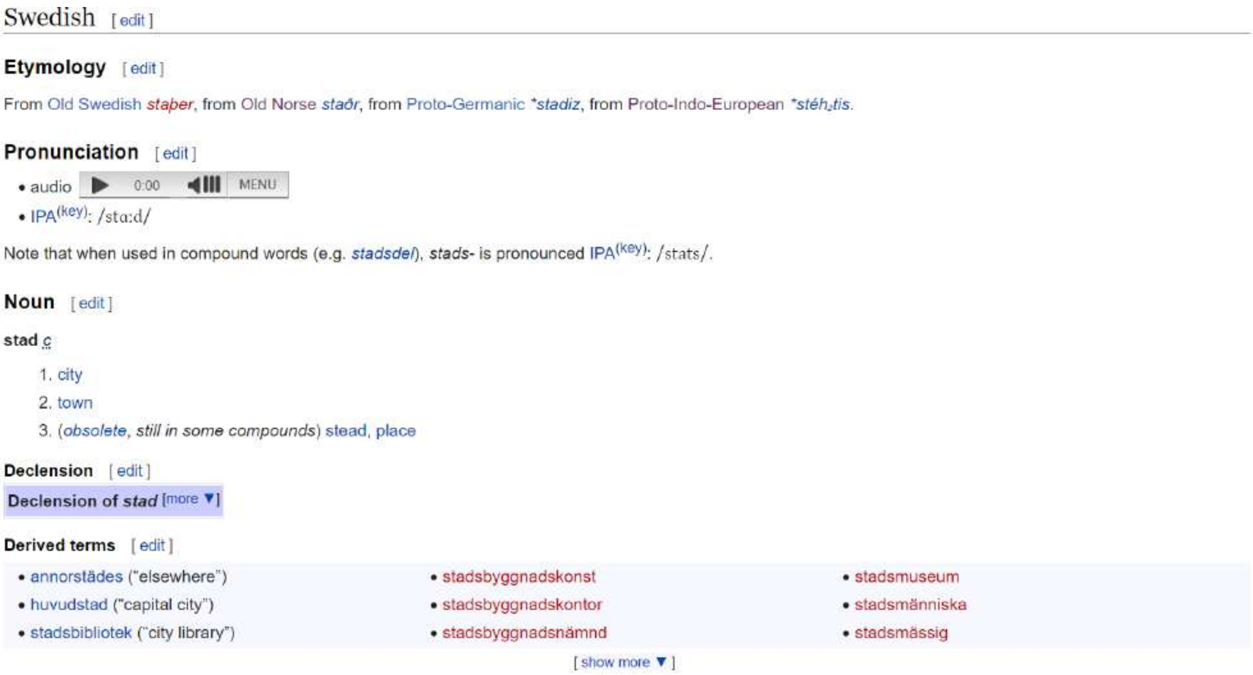
Not: 1. något vardagligare, se under användning

The information overload on Wiktionary is, in the typology of Gouws & Tarp (2017) a form of *concrete data overload*, where the formal properties of a word, formal lexical relations (derivations, compounded terms) and etymologies are incorporated into the micro-structure although not necessary demanded by the reader. The main bulk of readers are here assumed to be primarily interested in semantic and pragmatic information rather than etymology or formal lexical relations. Reducing the amount of information to remedy this kind of overload cannot be done, since Wiktionary strives to be as complete as possible.

The *perceptive data overload* (not presenting information optimally), however, can be dealt with. A perceptive data overload emerges when screen space is not used optimally. This is the case, for instance, with the pile up of compounded terms in the entry *stad*. Another example of this are translation sections at en.wikt: some translation sections of frequent terms grow so large that in order to navigate them meaningfully they must be moved to a separate page.[13] The potential for (almost) infinite growth of entry contents, only limited by the number of active editors, makes this type of overload ever more pressing. The way it is dealt with (moving contents to separate pages or hiding them under "spoilers") relieves some of the problems, but creates new ones, such as the need for more clicks to arrive at the sought content.

[13] This is the case for example with the entry *hand*, for which there are 340 translations just for the primary, literal sense. Considering the fact that translations to any language (for which there are many more than 340) are allowed and welcome to be added, this constitutes a clear conflict between the ambition to include everything and reader-friendliness.

Figure 9: Compounded terms under a separate section, partially under a “spoiler”, at the en.wikt entry *stad*



Swedish [edit]

Etymology [edit]

From Old Swedish *staþer*, from Old Norse *staðr*, from Proto-Germanic **stadiz*, from Proto-Indo-European **stéh₂tis*.

Pronunciation [edit]

- audio
- IPA(key): /stɑːd/

Note that when used in compound words (e.g. *stadsdel*), *stads-* is pronounced IPA(key): /stats/.

Noun [edit]

stad c

1. city
2. town
3. (*obsolete*, *still in some compounds*) stead, place

Declension [edit]

Declension of *stad* [more ▼]

Derived terms [edit]

- annorstädes (“elsewhere”)
- huvudstad (“capital city”)
- stadsbibliotek (“city library”)
- stadsbyggnadskonst
- stadsbyggnadskontor
- stadsbyggnadsnämnd
- stadsmuseum
- stadsmänniska
- stadsmässig

[show more ▼]

To sum up, the perceptive data overload on en.wikt arises from the large number of sections and from fragmentation of definitions in homonymous words over multiple etymologies. The perceptive data overload on sv.wikt varies greatly with the amount of content at each individual entry, and arises from the too tight integration of the semantic and non-semantic information. The micro-structure becomes overloaded, since many different types of non-semantic information are placed under the definitions, impeding the chances of successful look-up for casual readers. While it is true that *some* related terms belonging to a definition would be beneficial for quick comprehension, a pile-up of the kind seen in the entry *stad* hardly serves the reader well.

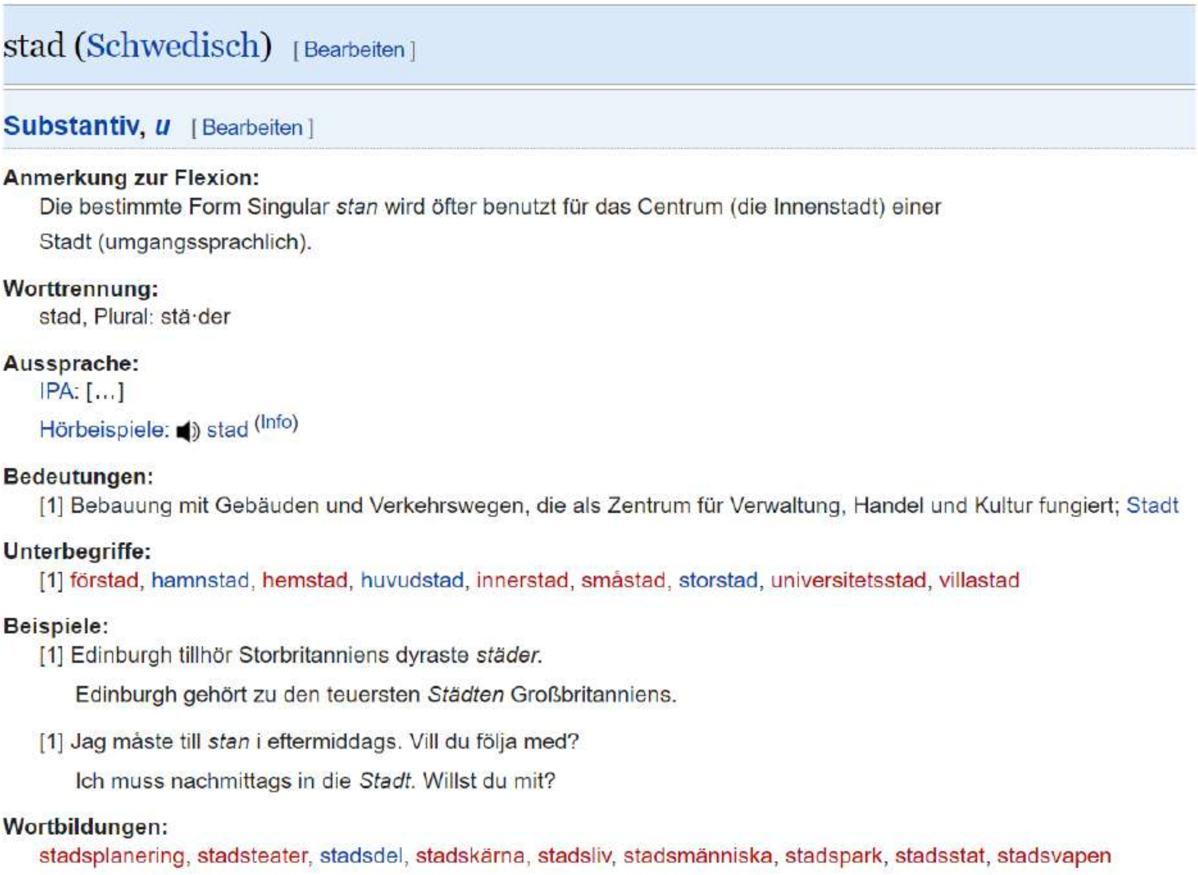
These two degrees of integration could be contrasted with a third solution, presented by the German edition of Wiktionary. It is quite extreme and obviously suffers from too large *dis*integration of different types of information instead: here, every type of information is given under a separate section (see figure 10 for an example of this).

6. Concluding discussion

Every decision on entry layout, lemmatisation and visual integration of different types of information has its own (dis)advantageous effects. As such, the decisions made by the community of sv.wikt to move further away from the encyclopaedic layout of Wikipedia, abolishing separate sections for, for example etymologies, and adherence to the formal-grammatical approach made the screenspace of an entry much smaller (see figures 6 and 7), which is undoubtedly beneficial for the visual grasp of the contents. But this advantage lasts only as long as non-semantic information is held to a minimum. This is also the case for the majority of entries at sv.wikt[14], which is why the decision can

[14] As such, out of 303,373 pages on sv.wikt containing lexical entries, only 18,142 pages contain entries with compounded terms, 43,497 with otherwise related terms and 14,007 with etymologies. A page may include several entries in more than one language. For comparison: there are 2,594,263 lexical

Figure 10: Example sentences at the German Wiktionary entry *stad* separated from the definitions by a list of compounded terms.



stad (Schwedisch) [Bearbeiten]

Substantiv, *u* [Bearbeiten]

Anmerkung zur Flexion:
Die bestimmte Form Singular *stan* wird öfter benutzt für das Centrum (die Innenstadt) einer Stadt (umgangssprachlich).

Worttrennung:
stad, Plural: stä·der

Aussprache:
IPA: […]
Hörbeispiele: stad (Info)

Bedeutungen:
[1] Bebauung mit Gebäuden und Verkehrswegen, die als Zentrum für Verwaltung, Handel und Kultur fungiert; Stadt

Unterbegriffe:
[1] förstad, hamnstad, hemstad, huvudstad, innerstad, småstad, storstad, universitetsstad, villastad

Beispiele:
[1] Edinburgh tillhör Storbritanniens dyraste *städer.*
Edinburgh gehört zu den teuersten *Städten* Großbritanniens.
[1] Jag måste till *stan* i eftermiddags. Vill du följa med?
Ich muss nachmittags in die *Stadt*. Willst du mit?

Wortbildungen:
stadsplanering, stadsteater, stadsdel, stadskärna, stadsliv, stadsmänniska, stadspark, stadsstat, stadsvapen

be seen as justified. Since completeness of included information is the absolute ideal for Wiktionary, it would be beneficial for sv.wikt to find a way to sustain increasing depth of its content without increasing the concrete overload and exacerbating the user experience. One such way could be relying less on the use of micro-structure templates and establishing separate sections at least for some types of information. An alternative solution would be to introduce a so called "spoiler", or "fold/unfold" function, where the non-semantic information remains structurally subordinate to the definitions, but is hidden under a spoiler by default. This way, etymologies and lists of related terms would still be one click away without impeding the look-up for users who don't need them.

The comparatively large size of the editor community on en.wikt makes the vision of completeness, especially with regards to etymological information, much closer to the reality. As a result, the entry layout had to undergo a larger separation between semantic and non-semantic information, including fragmentation of definitions in homonymous entries between several etymologies. This has increased the concrete information overload in such entries for readers uninterested in language history, but enabled continued growth of high-quality content, such as elaborate etymology sections.

entries (distributed over a smaller number of pages) on en.wikt, 1,410,582 pages contain entries with etymologies, 69,194 pages contain homonymous entries with at least two etymologies. A total of 254,547 pages contain entries with derived terms and 267,975 pages contain entries with related terms.

Since Wiktionary is a project with enormous potential and increasing relevance to lexicography, it would be desirable to address some issues outlined but not examined in this paper. An in-depth study of the contents (in addition to the structure), its quality and adequacy in meeting the needs of target groups (both suggested here and derived from traditional consulting situations) are some of the topics for future research. Empirical verification of the findings of current paper using online user surveys or eye-trackers would also shed more light on the relation between the entry layout and various types of information overload.

7. References

Buchi, É. (2016). *The Oxford Handbook of Lexicography*, chapter Etymological dictionaries. Oxford University Press.

Fuertes-Olivera, P.A. (2009). The Function Theory of Lexicography and Electronic Dictionaries: Wiktionary as Prototype of Collective Multiple-Language Internet Dictionary. *Lexicography at a Crossroads. Dictionaries and Encyclopedias Today, Lexicographic Tools Tomorrow*, pp. 99–134.

Gouws, R.H. & Tarp, S. (2017). Information overload and data overload in lexicography. *International Journal of Lexicography*, 30(4), pp. 389–415.

Khoury, R. & Sapsford, F. (2016). Latin word stemming using Wiktionary. *Digital Scholarship in the Humanities*, 31(2), pp. 368–373.

Korigodskiy, R., Kondrasykin, O.N., Zinowyev, B.I. & Losyagin, W. (1990). *Kamus Besar Bahasa Indonesia-Rusia.* Moscow: Russkiy Yazik.

Meyer, C.M. & Gurevych, I. (2012). *Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography.* na.

Oxford Dictionaries (2011). *Concise Oxford English Dictionary: Main edition.* OUP Oxford. URL https://books.google.se/books?id=DneZcQAACAAJ.

Sagot, B. (2017). Extracting an etymological database from wiktionary. In *Electronic Lexicography in the 21st century (eLex 2017).* pp. 716–728.

Sköldberg, E. & Wenner, L. (2020). Folkmun. se: A Study of a User-Generated Dictionary of Swedish. *International Journal of Lexicography*, 33(1), pp. 1–16.

Svensén, B. (2004). *Handbok i lexikografi : Ordböcker i teori och praktik.* Norstedts.

Svensén, B. (2009). *A handbook of lexicography: The theory and practice of dictionary-making.* Cambridge University Press Cambridge.

Tarp, S. (2017). *The Routledge handbook of lexicography*, chapter The concept of dictionary. Routledge.

Törnqvist, L. (2015). Nordiska dialekt-och slangordböcker på Internet. *LexicoNordica*, 22, pp. 57–75.

Wolfer, S. & Müller-Spitzer, C. (2016). How many people constitute a crowd and what do they do? Quantitative analyses of revisions in the English and German Wiktionary editions. *Lexikos*, 26, pp. 347–371.

Creating an Electronic Lexicon for the Under-resourced Southern Varieties of the Kurdish Language

Zahra Azin[1], **Sina Ahmadi**[2]

[1]Geomatics and Cartographic Research Center, Carleton University
[2]National University of Ireland Galway, Ireland
E-mail: zahraazin@cmail.carleton.ca, ahmadi.sina@outlook.com

Abstract

Thanks to the advances in information technology and communication, many endangered, vulnerable and under-represented language communities have a chance to revitalise and document their languages. In comparison to other Kurdish variants such as central Kurdish (also known as Sorani) and northern Kurdish (also known as Kurmanji), southern Kurdish has received little attention, making it an under-documented and under-resourced language that is spoken primarily in the Kurdish regions of Iran, particularly Kermanshah and Ilam provinces. As the case of our study, we focus on creating an electronic monolingual lexicon of significant size for the southern variants of Kurdish in the OntoLex-Lemon ontology by converting a bilingual and monolingual dictionary. In addition, we report our efforts in using a semi-automatic pivot-based translation inference approach to align the current resource with other resources in Kurdish and Gorani. We believe that this resource increases inter-operability across various natural language processing systems and facilitates many tasks in computational linguistics for Kurdish. Our resource is publicly available under a Creative Commons Attribution-ShareAlike 4.0 International License[1].

Keywords: southern Kurdish; electronic lexicography; less-resourced languages; machine-readable dictionary

1. Introduction

Given the increasing importance of information technology and accessibility in our era, language communities around the globe are experiencing a momentous period to consolidate their languages with technology. As an initial step in documenting and processing natural languages, electronic resources, particularly lexicons, are of significance to pave the way for gradual and more advanced progress. That being said, many endangered and under-documented languages face further challenges due to the scarcity of language and linguistic resources.

In this paper, we focus on one of the under-represented variants of the Kurdish language, southern Kurdish. Kurdish is an Indo-European language spoken by 20-30 million people in the Kurdish regions of Iraq, Iran, Turkey and Syria, and also among the Kurdish diaspora around the world. Generally, the language is categorised as a less-resourced one with few linguistic resources and sparse documentation (Abdulrahman et al., 2019). Among the three main dialects of Kurdish, namely northern Kurdish or Kurmanji, central Kurdish or Sorani and southern Kurdish, the latter lacks resources to a greater extent than the other two variants (Ahmadi, 2020). To remedy this, in this paper we discuss our efforts in creating an electronic lexicon for southern Kurdish.

Aware of the advances in the Semantic Web and Linked Data technologies, we focus on converting a printed dictionary, which is provided to us by a native lexicographer, into the Ontolex-Lemon ontology (McCrae et al., 2017). The dictionary is compiled based on lemmata of southern Kurdish and provides translations in Persian and Sorani Kurdish. In this regard, our methodology is based on Ahmadi et al. (2019), where the printed dictionary is semi-automatically converted into OntoLex-Lemon (McCrae et al., 2017). OntoLex-Lemon aims at modelling existing lexicographic resources as linked data and

[1] https://github.com/sinaahmadi/SKurdishLexicon

providing a conceptual model of language and linguistic objects to increase the re-usability of lexicographic content by following Semantic Web standards. Thanks to the current advances in linguistic linked open data (LLOD), lexicographic resources are now widely used in OntoLex-Lemon, and one compelling example is Wikidata,[2] which openly provides access to data regarding lexemes, senses and lexical forms (Nielsen, 2020).

Moreover, as a preliminary study, we carry out a translation inference task where our southern Kurdish lexicon is aligned with the Sorani dictionary produced by (Ahmadi et al., 2019) at the sense level. In addition to lexicons which are crucial resources in many natural language processing (NLP) tasks, such as word-sense disambiguation and spelling-error correction, alignment of lexical resources has proved to be beneficial in many natural language processing tasks (Ahmadi et al., 2020).

2. Southern Kurdish

There are different approaches proposed by linguists and dialectologists to classify Kurdish. All these classifications contain a group representing a bundle of familiar varieties including Kalhori, Feyli, Kermashani, and Laki. In the classification of Kurdish varieties, Hassanpour (1992) names this group Kermashani and identifies it as one of the main varieties of Kurdish alongside Kurmanji, Sorani, and Hawrami. At the same time, (Izady, 1992: p. 169) identifies two main groups of Kurdish language: Kurmanji, which includes north Kurmanji and south Kurmanji, which refers to Sorani; and Pahlawani, which consists of Zaza and Gorani (also written as Gurani). According to this classification, dialects spoken in the southern areas of Kurdish speaking settlements (starting from eastern Turkey to western Iran) are considered as varieties of Gorani. Later, Fattah (2000) provided a clearer picture of the Kurdish language based on a detailed fieldwork proposing a plausible classification of Kurdish into five groups of northern Kurdish (or Kurmanji), central Kurdish (or Sorani), southern Kurdish, Zazaki, and Hawrami (also referred to as Gorani). In this section, partly complying with Fattah's classification (Fattah, 2000), we discuss southern Kurdish (also called SK) and some of its issues.

Southern Kurdish is a variety of the language consisting of a group of vernaculars spoken by almost three million people across an extensive region of western Iran, including Ilam, a large area of Kermanshah, and some parts of Lorestan and Kurdistan provinces (Fattah, 2000). As shown in Figure 1, this variety is also spoken in eastern Iraq in Khanaqin and Mandali, very close to the borders with Iran. Due to the geography of the areas where southern Kurdish varieties are spoken, the population of southern Kurdish speakers is quite dispersed. On the other hand, as shown on the map, the presence of other languages such as Lori, mainly spoken in Lorestan, Chaharmahal and Bakhtiari, and parts of Ilam has resulted in a linguistic continuum between Kurdish and Lori in those areas (Aliakbari et al., 2015).

The existence of other languages and varieties such as Gorani, Lori, Persian, Turkic, and Arabic has resulted in a complex linguistic situation, with language contact and multilingualism slowing down the progress of studies on southern Kurdish and its vernaculars, and this poses challenges to the classification of southern Kurdish dialects. In such a linguistic context, ethnic affiliation directly affects the categorisation of language

[2] https://www.wikidata.org

varieties. Therefore, the study of southern Kurdish is closely related to ethnogeography which means the names of the vernaculars refer to specific ethnic groups or villages where their speakers reside. For instance, the names of some major variants such as “Kalhor” or “Kordali” have been taken from large tribes whereas variants with a smaller group of speakers are named after the geographical district, such as “Malikshahi”, which is spoken in a district called Malikshah. The variations of southern Kurdish are mainly mutually intelligible, but as the geographical distance between the speakers increases, more effort is required to understand other varieties (Fattah, 2000).

Nevertheless, the language shift towards Persian among the southern Kurdish speakers due to sociolinguistic factors has been considered a threat to the native languages over since the past few decades (Yarahmadi, 2021). This language shift, according to (Yarahmadi, 2021), has not changed the vocabulary of the language in such a way that Kurdish words are replaced by Persian equivalents, but the syntax and phonology of the language have also undergone many changes.

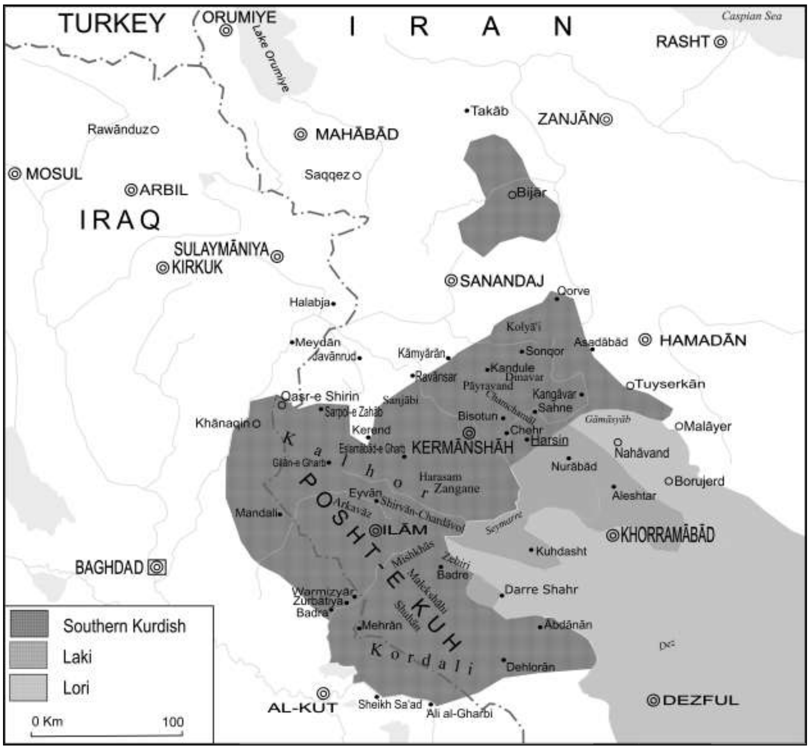


Figure 1: Revised map of the distribution of southern Kurdish dialects (Fattah, 2000) from (Belelli, 2019: p. 3)

2.1 Dialects

Classification of southern Kurdish varieties is not easy mainly due to the lack of descriptive studies regarding the nuances among them. Fattah (2000) was the first who outlined an initial classification of southern Kurdish vernaculars. He identified 27 sub-groups of southern Kurdish in Iran and eight in Iraq based on which he proposed seven main dialects: Bijari (also known as Garusi), Kolya’i (called Chardawri in the Kurdistan Province of

Iran), Laki spoken in the city of Kermanshah[3], Kalhori (including Sanjabi and Zangana), Malekshahi, Badre'i, and Kordali (Belelli, 2019).

There are different factors affecting dialectal variations including geographic, social, and individual properties. The linguistic complexity of the area where southern Kurdish is spoken makes the study of varieties even more challenging due to the extensive population mobility, language contact, and complexity of intersecting some dialects. Extensive fieldwork is thus required to better understand the linguistic characteristics of this dialect and its varieties.

2.2 Scripts

It is not until recently that southern Kurdish has been used in writing, except for some literary and religious works. Historically, literary Gorani was the primary means for writing literary works in southern Kurdish (Kreyenbroek & Chamanara, 2013). Unlike other Kurdish varieties, such as Sorani, southern Kurdish has never gone through language standardisation, mainly because of the dominance of official languages in regions where it is spoken, and the speakers mainly use languages such as Persian, Arabic, and Sorani Kurdish for writing. Today, southern Kurdish varieties are rarely found in written form or they simply follow other existing scripts. It seems that using such writing systems does not prevent this variety from expressing itself properly. However, some phonological features which distinguish southern Kurdish from other varieties have not been represented in such writing systems. Table 1 represents Kurdish scripts[4] and tentatively illustrates the place of southern Kurdish in this system. As the variety have not been standardised, different existing forms used in existing resources are shown in the table (Fattah, 2000; Jalilian, 2006). The examples in this manuscript are provided in the Latin-based script of Kurmanji Kurdish.

The phonological variations among southern Kurdish dialects depend on the region they are spoken in. The pharyngeal consonants [ħ], [ɣ] and [ʕ] are absent in southern Kurdish varieties spoken in Iran, unlike Sorani Kurdish. The voiced velar nasal [ŋ] seems to be missing in variations spoken in Khanaqin of Iraq and Qasr-e Shirin in Iran; however, it is common in Kalhori and Mandali dialects as in *řeη* 'colour' (Fattah, 2000). In some Kalhori dialects (and along the border with Iraq) [g^j] replaces word final [g] or [k], e.g. *seg^j* 'dog' (Belelli, 2019).

Unlike the consonants, vowels in southern Kurdish vary extensively among its dialects in different regions. Among the vowels represented in Table 1, [ə] and [ü] are southern Kurdish specific and do not appear in other Kurdish varieties. The mid central /ə/ is used as an Ezafe (also known as Izafe) marker in southern Kurdish to distinguish between /î/ in Sorani and /i/ in its varieties, e.g. *kuř-i xas* "(the) good boy" vs. Sorani *kuř-î xas* "a good boy" (Karimpour, 2003).

One of the main challenges before southern Kurdish standardisation is phonological variations among its dialects. Language contact and multilingualism lead to gradual

[3] Whether Laki is a variety of Kurdish or Lori is still an open question, and we avoid this discussion as it is not in the scope of the present study.

[4] In this table, following the Unified Kurdish Alphabet introduced by the Kurdish Academy of Language, the Latin based Yekgirtú has been used to represent Kurdish script. See http://www.kurdishacademy.org

Kurdish Phonemes (IPA)	Latin-based	Yekgirtû	SK (existing resources)	Arabo-Persian			
				initial	middle	final	single
[ɑː]	A a	A a	A a	ئاـ	ـاـ	ـا	ا
[b]	B b	B b	B b	بـ	ـبـ	ـب	ب
[ʧ]	Ç ç	C c	Ç ç Č č Ch ch	چـ	ـچـ	ـچ	چ
[ʤ]	C c	J j	C c J j Ĵ ĵ	جـ	ـجـ	ـج	ج
[d]	D d	D d	D d	دـ	ـدـ	ـد	د
[æ]	E e	E e	E e A a	ئه	ـه	ـه	ه
[eː]	Ê ê	É é	Ê ê	ئێـ	ـێـ	ـێ	ێ
[f]	F f	F f	F f	فـ	ـفـ	ـف	ف
[g]	G g	G g	G g	گـ	ـگـ	ـگ	گ
[h]	H h	H h	H h	هـ	ـهـ	ـه	ه
[ħ]	H h	H' h'	H h H' h'	حـ	ـحـ	ـح	ح
[ə]	I i	I i	I i				
[iː]	Î î	Í í	Î î	ئیـ	ـیـ	ئی	ی
[ʒ]	J j	Jh jh	J j Ž ž	ژ	ـژـ	ـژ	ژ
[k]	K k	K k	K k	کـ	ـکـ	ـک	ک
[l]	L l	L l	L l	لـ	ـلـ	ـل	ل
[ɫ]	Ł ł	Ll ll	Ḷ ḷ Ł ł Lˆ lˆ	ڵـ	ـڵـ	ـڵ	ڵ
[m]	M m	M m	M m	مـ	ـمـ	ـم	م
[n]	N n	N n	N n	نـ	ـنـ	ـن	ن
[oː]	O o	O o	O o	ئۆ	ـۆـ	ـۆ	ۆ
[p]	P p	P p	P p	پـ	ـپـ	ـپ	پ
[q]	Q q	Q q	Q q	قـ	ـقـ	ـق	ق
[ɾ]	R r	R r	R r	ر	ـرـ	ـر	ر
[r]	Ř ř	Rr rr	Ř ř Ṛ ṛ	ڕ	ـڕـ	ـڕ	ڕ
[s]	S s	S s	S s	سـ	ـسـ	ـس	س
[ʃ]	Ş ş	Sh sh	Ş ş Š š Sh sh	شـ	ـشـ	ـش	ش
[t]	T t	T t	T t	تـ	ـتـ	ـت	ت
[ʊ]	U u	U u	U u	ئو	ـوـ	ـو	و
[uː]	Û û	Ú ú	Û û	ئوو	ـووـ	ـوو	وو
[v]	V v	V v	V v	ڤـ	ـڤـ	ـڤ	ڤ
[w]	W w	W w	W w	و	ـوـ	ـو	و
[x]	X x	X x	X x	خـ	ـخـ	ـخ	خ
[j]	Y y	Y y	Y y	ـی	ـیـ	یـ	ی
[z]	Z z	Z z	Z z	ز	ـزـ	ـز	ز
[ʔ]	ʻ		ʻ	ئـ	-	-	-
[ʕ]		'	É é	عـ	ـعـ	ـع	ع
[ɣ]	X x	X' x'	X' x'	غـ	ـغـ	ـغ	غ
[ŋ]	ŋ		ŋ	-	ـنْـ	ـنْ	نْ
[gʲ]			Ğ ğ	گْـ	ـگْـ	ـگْ	گْ
[y]	Ü ü		Ü ü	ۊ	ـۊـ	ـۊ	ۊ

Table 1: Current forms of Kurdish alphabets found in existing resources

.

changes in the phonological system which, in turn, lead to a more complex linguistic situation in regions where this dialect is spoken. However, as the first step towards the standardisation of this variety, a close investigation of the scripts used in existing resources is indispensable.

2.3 Vocabulary

Lexical differences in southern Kurdish varieties are found in dictionaries and everyday conversation of the speakers. However, using various lexical items does not obscure the intelligibility of the dialects for listeners who speak different varieties. Moreover, neighboring dialects and inter-dialectal means of communication result in the gradual change of the lexicon (Belelli, 2019). Speakers try to approximate their dialects to the highest prestige one and sometimes they do this to build solidarity across their differences.

Despite all those similarities, there still exist nuances in the lexical items used by the speakers of southern Kurdish dialects. As the provincial border of south-west Kermanshah is crossed to Ilam and eastern Iraq, lexical differences become more salient. The difference might be in the form of a simple shift of the vowels (e.g. *çö* vs. *çaw* ‘eye’), or by using different words for a same concept (e.g. *keřemye* vs. *tem* ‘fog’).

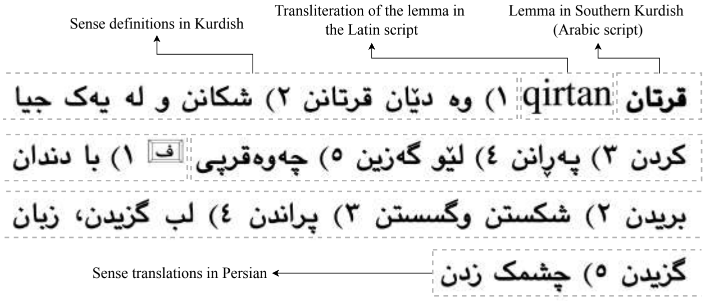
One way to study dialect variation is the lexicon, or the vocabulary used by the speakers. Varieties might either use different words or same words with different meanings for instance, Badre’i speakers use *xwazî* ‘want’ to ask for something, while in Kalhori, speakers use the same verb when proposing to a woman. Although such differences might lead to misunderstanding, it does not interrupt the communication.

3. Approach

3.1 Dictionary Compilation

In this study, we use “*Ferhengî Başur*” (literally meaning “South Dictionary”), a southern Kurdish-central Kurdish-Persian dictionary compiled and edited by Jalilian (2006) with the purpose of codification of southern Kurdish. Initially, the dictionary was created to be a part of “*Henbane Borîne*”, a Kurdish-Kurdish-Persian dictionary written by Abdurrahman Sharafkandi known as Hazhar in 1990 (Sharafkandi, 1991). “*Henbane Borîne*” contains around 60,000 entries with lexemes from different Kurdish varieties. In addition, the Persian equivalent of entries along with a few examples are provided. “*Ferhengî Başur*” maintains the same structure but with a focus on the southern Kurdish varieties, particularly Kalhori and Laki.

Despite the attempt to document the general and folkloric vocabulary of southern variants of Kurdish in this resource, there is a lack of coverage of topics due to the scarcity of terminologies for Kurdish in general, and for these variants in particular. Similar to the majority of Kurdish dictionaries, our resource lacks consistent definition of entries in such a way that sense glosses are provided for only a few lemmata. The same issue can be observed with respect to idioms, examples and pronunciation. Among the words in various varieties of Kurdish, only those in Laki are specified by the lexicographer. Figure 2 illustrates the entry *qirtan* ‘to cut’ in southern Kurdish in the printed dictionary.



قرتان qirtan ١) وه دێان قرتانن ٢) شکانن و له یەک جیا
کردن ٣) پەڕانن ٤) لێو گەزین ٥) چەوەقرپی ف ١) با دندان
بریدن ٢) شکستن وگسستن ٣) پراندن ٤) لب گزیدن، زبان
گزیدن ٥) چشمک زدن

Sense definitions in Kurdish
Transliteration of the lemma in the Latin script
Lemma in Southern Kurdish (Arabic script)
Sense translations in Persian

Figure 2: "*qirtan*" 'to cut', an entry in the printed version of Jalilian (2006: p. 534) Southern Kurdish dictionary

3.2 Conversion into OntoLex-Lemon

Ahmadi et al. (2019) propose an approach to create electronic lexicons for Sorani Kurdish, Kurmanji Kurdish and Gorani. Following the same approach, we use a semi-automatic technique to extract entries from the printed dictionary using regular expressions. The extracted information is followed by a manual verification regarding the lemma in the Arabic-Persian script, its transliteration in the Latin script of Kurdish, glosses in Kurdish and their translations in Persian. In addition, a few entries are provided with additional information, such as the sub-dialect where the word is used, which are similarly included in the conversion process.

In order to increase the interoperability and accessibility of this resource, we use the electronic dictionary in OntoLex-Lemon in the Resource Description Framework (RDF). The OntoLex-Lemon standard provides rich linguistic grounding for ontologies, such as representation of morphological and syntactic properties of lexical entries (McCrae et al., 2017). The core of the Ontolex-Lemon model is shown in Figure 3. In addition, we also use the lexicography module Lexicog, which provides a conceptual model of language and linguistic objects in lexicography (Bosque-Gil et al., 2017).

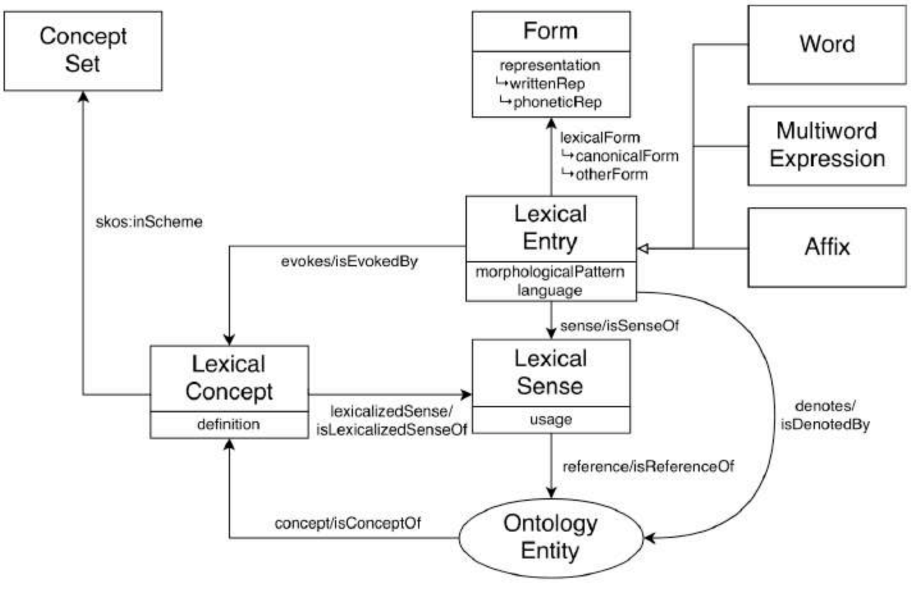


Figure 3: Lemon-OntoLex Core (McCrae et al., 2017)

Figure 4 shows the same entry in Figure 2 where the lemmas in the Arabic-based and Latin scripts are provided along with the senses and their translations into Persian. It should be noted that morphosyntactic information, such as part-of-speech tags, are not provided in the current version of the electronic dictionary. Due to the inconsistency in differentiating glosses and senses in the microstructure of the printed dictionary, we only include senses which are composed of at most two space-separated words. This measure was taken to only include senses rather than glosses in the converted dictionary. Overall, 14,326 entries are extracted from the printed lexicon.

```
@prefix ontolex: <http://www.w3.org/ns/lemon/ontolex#> .
@prefix vartrans: <http://www.w3.org/ns/lemon/vartrans#> .
@prefix lime: <http://www.w3.org/ns/lemon/lime#> .
:lexicon a lime:Lexicon;
    lime:language <www.lexvo.org/page/iso639-3/sdh> ;
    lime:entry :lex_qirtan .

:lex_qirtan a ontolex:LexicalEntry, ontolex:Word ;
    dct:language <www.lexvo.org/page/iso639-3/sdh> ;
    rdfs:label "qirtan"@sdh-latn ;
    rdfs:label "قرتان"@sdh-arab .
    ontolex:sense :qirtan_sense_1, qirtan_sense_2, qirtan_sense_3, qirtan_sense_4, qirtan_sense_5 .

:qirtan_sense_1 rdfs:label "پەڕانن"@sdh-arab .
:qirtan_sense_2 rdfs:label "چەوەقرپی"@sdh-arab .

:fa_lex_1 a ontolex:LexicalEntry ;
    rdfs:label "پراندن"@fa ; ontolex:sense :fa_lex_1_sense .
:fa_lex_2 a ontolex:LexicalEntry ;
    rdfs:label "چشمک زدن"@fa ; ontolex:sense :fa_lex_2_sense .

:trans_qirtan_sense_1_fa_lex_1 a vartrans:Translation ;
   vartrans:source :qirtan_sense_1 ; vartrans:target :fa_lex_1 .
:trans_qirtan_sense_2_fa_lex_2 a vartrans:Translation ;
   vartrans:source :qirtan_sense_2 ; vartrans:target :fa_lex_2 .
```

Figure 4: An example entry from our Southern Kurdish dictionary. The original printed entry in the left and the equivalent in RDF Turtle based on the OntoLex-Lemon model

3.3 Alignment with a Sorani Kurdish dictionary

Some of the senses in the southern Kurdish dictionary are provided in Sorani Kurdish. Such cases allow us to align the current resource with other existing ones, particularly (Ahmadi et al., 2019) Sorani dictionary in OntoLex-Lemon[5]. The latter provides translations in English for Sorani which can also used as a pivot language to align with other lexical resources. More precisely, we first align the entries in the southern Kurdish dictionary by matching senses that appear in the Sorani Kurdish dictionary. Therefore, the initial headwords can be aligned with the English translations of the Sorani lemmas. The alignment of the southern Kurdish dictionary with the Sorani one yielded 1,047 cross-dialect links.

[5] https://github.com/KurdishBLARK/KurdishLex/blob/master/Sorani.ttl

4. Conclusion and Future Work

In this paper, we present a preliminary study to create an electronic lexicon in southern Kurdish using the OntoLex-Lemon ontology. Our primary goal is to shed light on the current state of the southern varieties of Kurdish. We believe that this resource helps the southern varieties of the Kurdish language, which are under-represented, to be documented. Moreover, it will pave the way for further developments for southern Kurdish, in particular in language technology and natural language processing for tasks such as spelling error detection and correction, part-of-speech tagging and syntactic analysis.

A major limitation of this work is due to the limited coverage of the dictionary, and also the lack of glosses, examples, pronunciations and morphosyntactic properties. The current dictionary can be further completed by adding morphosyntactic information, etymological and usage examples. In order to increase inter-operability among Kurdish resources, it is also suggested to align the resource with other lexical and semantic resources such as KurdNet, the Kurdish WordNet Aliabadi et al. (2014) and more dialects of Kurdish.

5. Acknowledgments

The authors would like to thank Ako Jalilian who graciously provided his lexicographical content for this project. His efforts in documenting different varieties of southern Kurdish are greatly appreciated.

6. References

Abdulrahman, R.O., Hassani, H. & Ahmadi, S. (2019). Developing a fine-grained corpus for a less-resourced language: the case of Kurdish. *arXiv preprint arXiv:1909.11467*.

Ahmadi, S. (2020). Building a Corpus for the Zaza–Gorani Language Family. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*. pp. 70–78.

Ahmadi, S., Hassani, H. & McCrae, J.P. (2019). Towards electronic lexicography for the Kurdish language. In *Proceedings of the sixth biennial conference on electronic lexicography (eLex)*. eLex 2019.

Ahmadi, S., McCrae, J.P., Nimb, S., Khan, F., Monachini, M., Pedersen, B., Declerck, T., Wissik, T., Bellandi, A., Pisani, I., Troelsgård, T., Olsen, S., Krek, S., Lipp, V., Váradi, T., Simon, L., Gyorffy, A., Tiberius, C., Schoonheim, T., Ben Moshe, Y., Rudich, M., Abu Ahmad, R., Lonke, D., Kovalenko, K., Langemets, M., Kallas, J., Dereza, O., Fransen, T., Cillessen, D., Lindemann, D., Alonso, M., Salgado, A., Luis Sancho, J., Ureña-Ruiz, R.J., Porta Zamorano, J., Simov, K., Osenova, P., Kancheva, Z., Radev, I., Stanković, R., Perdih, A. & Gabrovsek, D. (2020). A Multilingual Evaluation Dataset for Monolingual Word Sense Alignment. In *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 3232–3242. URL https://www.aclweb.org/anthology/2020.lrec-1.395.

Aliabadi, P., Ahmadi, M.S., Salavati, S. & Esmaili, K.S. (2014). Towards building kurdnet, the kurdish wordnet. In *Proceedings of the Seventh Global Wordnet Conference*. pp. 1–6.

Aliakbari, M., Gheitasi, M. & Anonby, E. (2015). On language distribution in Ilam province. *Iranian studies*, 48(6), pp. 835–850.

Belelli, S. (2019). Towards a Dialectology of Southern Kurdish: Where to begin? In *Current issues in Kurdish linguistics*. University of Bamberg Press, pp. 73–92.

Bosque-Gil, J., Gracia, J. & Montiel-Ponsoda, E. (2017). Towards a Module for Lexicography in OntoLex. In *LDK Workshops*. pp. 74–84.

Fattah, I.K. (2000). *Les dialectes kurdes méridionaux: Étude linguistique et dialectologique (Acta Iranica 37)*. Leuven: Peeters.

Hassanpour, A. (1992). *Nationalism and language in Kurdistan*. New York: Edwin Mellen Pr.

Izady, M. (1992). *The Kurds: A concise handbook*. London: Taylor & Francis.

Jalilian, A. (2006). *Farhang-ī bāšūr: Tāībat ba nāwčakānī kirmāšān-u īlām-u luřistān (Bashur dictionary: Specific for the regions of Kemanshah, Ilam and Lorestan)*. Tehran: Porseman.

Karimpour, K. (2003). *Khovar Halat: Farhang-e guyeš-e kordi-e kalhori (kermānšāhi), kordi-fārsi (Dictionary of the Kalhori Kurdish (Kermanshahi) dialect, Kurdish-Persian)*. Kermanshah: Sobh-e Roshan.

Kreyenbroek, P. & Chamanara, B. (2013). Literary Gurāni: Koinè or Continuum? *Chez les Kurdes*, pp. 151–169.

McCrae, J.P., Bosque-Gil, J., Gracia, J., Buitelaar, P. & Cimiano, P. (2017). The Ontolex-Lemon model: development and applications. In *Proceedings of eLex 2017 conference*. pp. 19–21.

Nielsen, F. (2020). Lexemes in Wikidata: 2020 status. In *Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020)*. pp. 82–86.

Sharafkandi, A. (1991). *Henbane Borîne Farhang-e Kordi-Farsi (Henbane Borine Kurdish-Persian dictionary)*. Tehran: Soroush Press.

Yarahmadi, J. (2021). Language Shift Among Speakers of Kalhuri Kurdish in Iran. *International Journal of Kurdish Studies*, 7(1), pp. 82–102.

Encoding semantic phenomena in verb-argument combinations

Elisabetta Jezek[1], Costanza Marini[1,2], Emma Romani[1]

[1]University of Pavia, Dipartimento di Studi Umanistici, Strada Nuova 65, 27100 Pavia
[2]University of Bergamo, Dipartimento di Lingue, Letterature e Culture Moderne, via Salvecchio 19, 24129 Bergamo
E-mail: e.jezek@unipv.it, costanza.marini01@universitadipavia.it, emma.romani01@universitadipavia.it

Abstract

In this paper, we report the classification we adopted in two electronic resources of corpus-derived verbal patterns for Italian and Croatian (T-PAS and CROATPAS) to account for three different semantic phenomena that we observed occurring between nouns and verbs in valency structure contexts: Semantic Type alternation, Semantic Type shift (metonymy), and Complex Type exploitation. After presenting the two resources in the context of similar projects (Section 2), in Sections 3, 4, and 5 we examine the three phenomena in detail and show how we registered them in the editor we developed for this purpose, called Skema. The encoding of these phenomena in the editor is of paramount importance for being able to query them in the interface of the two resources, which will soon be publicly available online. In Section 5, we draw our conclusions and suggest possible ways to use the annotated data.

Keywords: pattern resource; verb argument structure; semantic type; corpus analysis; word sense

1. Introduction

Lexical resources traditionally rely on lists of word senses, although several studies have long shown that word senses are very slippery entities (Kilgarriff, 1993), and that sense inventories fail to capture the large spectrum of meanings words acquire in their context of use. From a theoretical perspective, the variation in the senses of a word stems from the fact that natural languages are semantically flexible, that is, the meaning of a word varies from occurrence to occurrence as a function of the interaction with the other words it combines with, and with the context of utterance (Pustejovsky, 1995; Recanati, 2002). Within this framework, in lexicography word senses are then better conceived as abstractions from clusters of corpus citations (Kilgarriff, 1993: 91).

In this paper, we present two resources of verbal patterns that take this background into account, and address the problem of encoding the sense variation that can be observed in the nouns filling the argument positions in the pattern, which we assume are triggered by the verb the nouns combine with. Specifically, we report the classification we adopted in two inventories of predicate-argument structures – namely, T-PAS for Italian (Ježek et al., 2014) and its sister project CROATPAS for Croatian (Marini & Ježek, 2019) – to account for three different semantic phenomena that may affect nouns within a valency structure context: Semantic Type Alternation, Semantic Type Shift (Metonymy), and Complex Type Exploitation. This is possible thanks to a shared System of Semantic Types used to classify the semantics of arguments (Ježek, 2019), to the compositional principles of type coercion and type exploitation inspired by the Generative Lexicon (Pustejovsky & Ježek, 2008), to the methodological framework of corpus analysis adopted from Hanks (2013), and, last but not least, thanks to the editor that was developed to encode the phenomena at play (Baisa et al., 2020).

The structure of the paper is as follows: in Section 2 we introduce the two resources; in Section 3 we provide examples of Semantic Type Alternation occurring in different syntactic positions; in Section 4 we discuss Metonymy; in Section 5 we illustrate Complex

Type Exploitation. Finally, in Section 6 we highlight the usefulness of encoding these phenomena in electronic resources.

2. The resources: T-PAS and CROATPAS

T-PAS (Ježek et al., 2014) and CROATPAS (Marini & Ježek, 2019) are two corpus-derived resources consisting of repositories of Typed Predicate-Argument Structures for Italian (T-PAS) and Croatian (CROATPAS) verbs. Both projects are being developed at the University of Pavia with the technical support of *Lexical Computing Ltd.* and are intended to be used for linguistic analysis, language teaching, and computational applications. The resources share their organisation as regards four fundamental components:

1. a repository of corpus-derived predicate argument structures (called *patterns*) with semantic specification of their argument slots, e.g. [Human] drinks [Beverage];
2. an inventory of ca. 200 corpus-derived semantic classes (called *Semantic Types*) organised in a hierarchy (called *System of Semantic Types*), used for the semantic specification of the arguments;
3. a corpus of annotated sentences that instantiate the different patterns of the verbs in the inventory. Corpus lines are tagged with their respective pattern numbers and anchored to the verb they feature, which is the lexical unit of analysis;[1]
4. an editing system called Skema (Baisa et al., 2020), which allows the registration of patterns and all the syntactic and semantic information associated therewith, and facilitates the manual annotation of corpus instances (directly linked to the patterns).[2]

Typed predicate-argument structures are patterns that display the semantic properties of verbs: for each meaning of a verb, a specific pattern is provided. As referenced above, the patterns are corpus-derived, i.e. they are acquired through the manual clustering and annotation of corpus instances, following the CPA methodology (Hanks, 2013). Currently, T-PAS contains 1160 implemented verbs, 5,529 patterns, and ca. 200,000 annotated corpus instances, while CROATPAS contains 180 verb entries, 683 patterns and ca. 23,000 annotated corpus lines.

In the resources, each pattern is labelled with a pattern number and connected to a list of corpus instances realising that specific verb meaning. The Skema editor (see Figure 1) enables the registration of different semantic and lexical information in each pattern, more specifically:

1. the *verb*, which in T-PAS is generally in its infinitive form - e.g. *bere* (Eng., ‘to drink’);

[1] The reference corpora for the resources are two web corpora, namely *ItWac* (reduced) for T-PAS and *hrWac 2.2* for CROATPAS. *ItWac* (reduced) contains around 935 million tokens, while *hrWac 2.2* contains roughly 1.2 billion tokens.

[2] Skema (Baisa et al., 2020) is a corpus pattern editor system implemented to facilitate the management of manual annotation of concordance lines with user-defined labels and the editing of the corresponding patterns in terms of slots, attributes and other features following the lexicographic technique of CPA (Hanks, 2013).

2. the *Semantic Types* (e.g. [Human], [Beverage], always portrayed within square brackets), specifying the semantics of the arguments selected by the verb. Semantic Types can be found on six different arguments positions: *subjects* (portrayed in red), *direct objects* (green), *adverbials* (grey), *clausals* (violet), *predicative complements* (blue), *prepositional complements* (orange, only in T-PAS), and *indirect complements* (light blue, only in CROATPAS).
3. the *sense description*, i.e. a brief definition of the meaning of the verb in that specific pattern, which usually features the same Semantic Types registered in the pattern in question;
4. a *lexical set* (optional) for each Semantic Type in the pattern, i.e. a selection of the most representative lexical items instantiating that Semantic Type (e.g. *vino* = 'wine' | *birra* = 'beer' | *aranciata* = 'orange juice' are good candidates for the lexical set of [Beverage]);[3]
5. the *roles* (optional) played by some specific Semantic Types in certain contexts: in particular, the Semantic Type [Human] can acquire the role of Athlete, Doctor, Musician, Host, Guest, Writer, etc., depending on the verb selecting it as an argument;[4]
6. the *features* (optional) associated with the Semantic Types, i.e. certain semantic characteristics required by the pattern syntax (e.g. Plural) or by the specific verb meaning (e.g. Female, Negative, Visible);[5]
7. *prepositions* (for prepositional and indirect complements), *particles* (for adverbials), *complementisers* (for clausals), *quantifiers*, and *determiners* (for lexical sets), which can be implemented according to the specific argument position in question.

The System of Semantic Types used to classify the semantics of arguments (Pustejovsky et al., 2004; Ježek, 2019) is a hierarchy of general semantic categories obtained by manual clustering of the lexical items found in the argument positions of corpus-derived valency structures. The System currently contains ca. 200 Semantic Types that are hierarchically organised on the basis of the 'is a' (subsumption) relation (e.g. [Human] is an [Animate]).[6] The System of Semantic Types is shared by both resources.

Figure 1 shows the general organisation of both resources in the Skema editor (using patterns and corpus examples from the Italian T-PAS resource) with its four components used by the annotators to compile the patterns:[7]

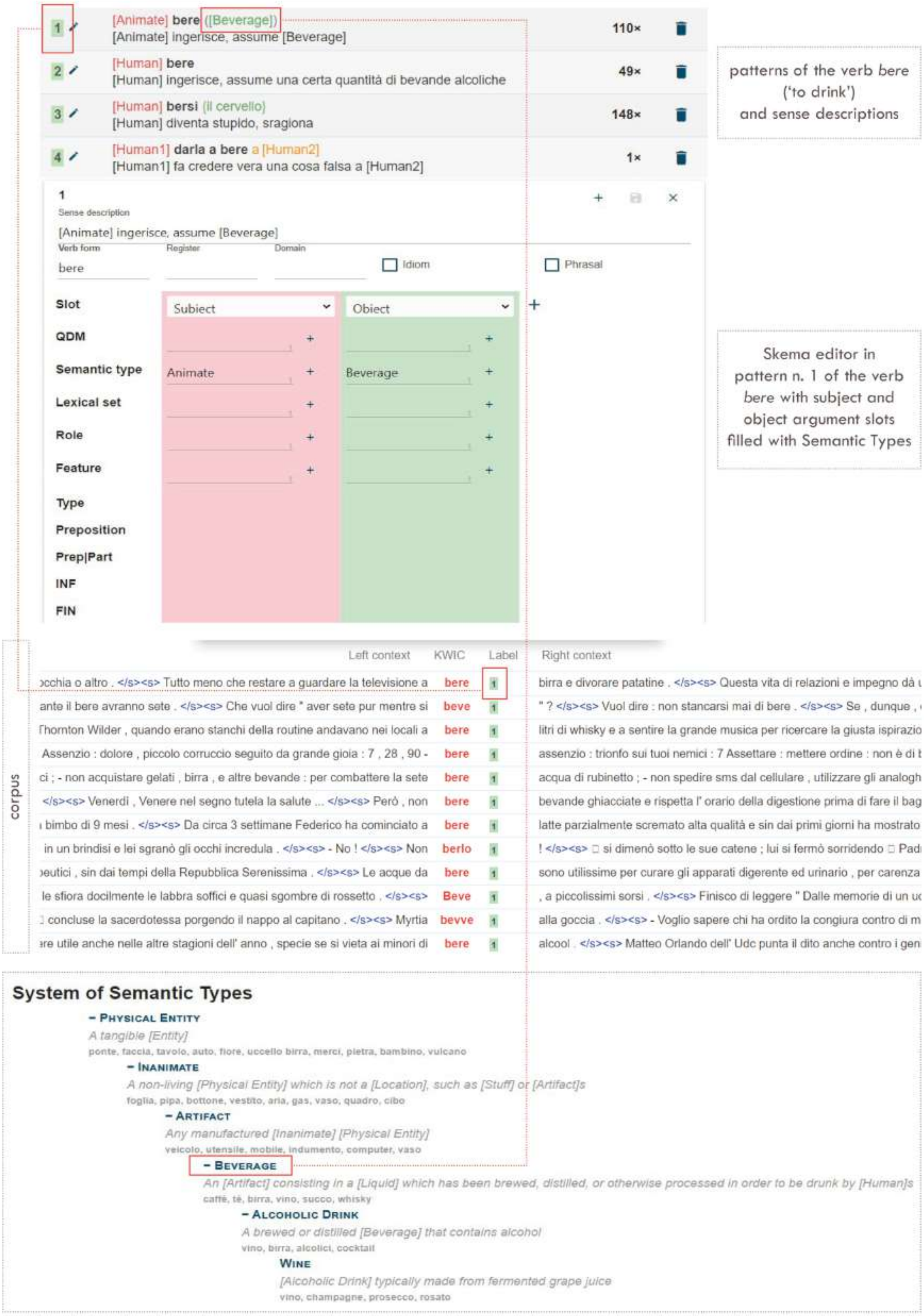
[3] Lexical sets appear next to their respective Semantic Types, in curly brackets.

[4] In Skema, Roles appear within square brackets, next to the Semantic Types they apply to, and preceded by '=', e.g. [Human = Doctor].

[5] In Skema, Features appear within square brackets, after the Semantic Types they apply to, and preceded by ':', e.g. [Human : Plural].

[6] The System of Semantic Types, together with definitions and examples for each Type, is made accessible to lexicographers through a customised function of Skema, so that it can be readily consulted while editing the patterns.

[7] The Skema editor is only accessible to the annotators working on the projects; the online public version based on Skema will display the patterns in a graphical interface that can be browsed.



1 [Animate] **bere** ([Beverage])
[Animate] ingerisce, assume [Beverage] — 110×

2 [Human] **bere**
[Human] ingerisce, assume una certa quantità di bevande alcoliche — 49×

3 [Human] **bersi** (il cervello)
[Human] diventa stupido, sragiona — 148×

4 [Human1] **darla a bere** a [Human2]
[Human1] fa credere vera una cosa falsa a [Human2] — 1×

patterns of the verb *bere*
('to drink')
and sense descriptions

1
Sense description
[Animate] ingerisce, assume [Beverage]

Verb form: bere
Register:
Domain:
□ Idiom
□ Phrasal

Slot	Subiect	Obiect
QDM		
Semantic type	Animate	Beverage
Lexical set		
Role		
Feature		
Type		
Preposition		
Prep|Part		
INF		
FIN		

Skema editor in pattern n. 1 of the verb *bere* with subject and object argument slots filled with Semantic Types

corpus

Left context	KWIC	Label	Right context
ɔcchia o altro . </s><s> Tutto meno che restare a guardare la televisione a	bere	1	birra e divorare patatine . </s><s> Questa vita di relazioni e impegno dà ι
ante il bere avranno sete . </s><s> Che vuol dire " aver sete pur mentre si	beve	1	" ? </s><s> Vuol dire : non stancarsi mai di bere . </s><s> Se , dunque , ‹
Thornton Wilder , quando erano stanchi della routine andavano nei locali a	bere	1	litri di whisky e a sentire la grande musica per ricercare la giusta ispirazio
Assenzio : dolore , piccolo corruccio seguito da grande gioia : 7 , 28 , 90 -	bere	1	assenzio : trionfo sui tuoi nemici : 7 Assettare : mettere ordine : non è di t
ci ; - non acquistare gelati , birra , e altre bevande : per combattere la sete	bere	1	acqua di rubinetto ; - non spedire sms dal cellulare , utilizzare gli analogh
</s><s> Venerdì , Venere nel segno tutela la salute ... </s><s> Però , non	bere	1	bevande ghiacciate e rispetta l' orario della digestione prima di fare il bag
ı bimbo di 9 mesi . </s><s> Da circa 3 settimane Federico ha cominciato a	bere	1	latte parzialmente scremato alta qualità e sin dai primi giorni ha mostrato
in un brindisi e lei sgranò gli occhi incredula . </s><s> - No ! </s><s> Non	berlo	1	! </s><s> □ si dimenò sotto le sue catene ; lui si fermò sorridendo □ Padı
ɔeutici , sin dai tempi della Repubblica Serenissima . </s><s> Le acque da	bere	1	sono utilissime per curare gli apparati digerente ed urinario , per carenza
le sfiora docilmente le labbra soffici e quasi sgombre di rossetto . </s><s>	Beve	1	, a piccolissimi sorsi . </s><s> Finisco di leggere " Dalle memorie di un uc
ː concluse la sacerdotessa porgendo il nappo al capitano . </s><s> Myrtia	bevve	1	alla goccia . </s><s> - Voglio sapere chi ha ordito la congiura contro di m
ɜre utile anche nelle altre stagioni dell' anno , specie se si vieta ai minori di	bere	1	alcool . </s><s> Matteo Orlando dell' Udc punta il dito anche contro i gen

System of Semantic Types

- Physical Entity
  A tangible [Entity]
  ponte, faccia, tavolo, auto, fiore, uccello birra, merci, pietra, bambino, vulcano
  - Inanimate
    A non-living [Physical Entity] which is not a [Location], such as [Stuff] or [Artifact]s
    foglia, pipa, bottone, vestito, aria, gas, vaso, quadro, cibo
    - Artifact
      Any manufactured [Inanimate] [Physical Entity]
      veicolo, utensile, mobile, indumento, computer, vaso
      - Beverage
        An [Artifact] consisting in a [Liquid] which has been brewed, distilled, or otherwise processed in order to be drunk by [Human]s
        caffè, tè, birra, vino, succo, whisky
        - Alcoholic Drink
          A brewed or distilled [Beverage] that contains alcohol
          vino, birra, alcolici, cocktail
          Wine
          [Alcoholic Drink] typically made from fermented grape juice
          vino, champagne, prosecco, rosato

Figure 1: The general structure of the resources (based on T-PAS) with the four main components as encoded in the Skema pattern editor (from the top of the image): patterns, pattern editor, corpus, System of Semantic Types

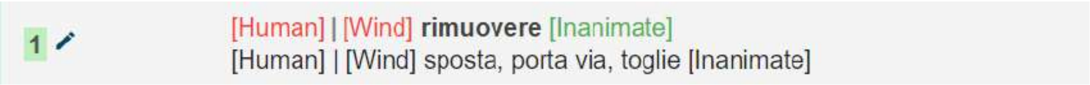
When it comes to pattern resources, it is necessary to mention some noteworthy projects revolving around several different languages. Chronologically, the first project where Corpus Pattern Analysis was applied was the *Pattern Dictionary of English Verbs* (PDEV) (Hanks & Pustejovsky, 2005), which is being developed at the Research Institute for Information and Language Processing of the University of Wolverhampton. An equivalent Spanish project is *Verbario* (Renau & Nazar, 2021), developed at the Pontifical Catholic University of Valparaíso (Chile). As for Dutch, a recent tool combining verb patterns, collocations and idioms is *Woordcombinaties* (Colman & Tiberius, 2018), which is being developed in Leiden at the *Instituut voor de Nederlandse taal.* Last but not least, another Italian pattern dictionary is currently being designed at the University of Heidelberg (Germany). The project is aimed at creating a learner's dictionary with phraseological disambiguators (Di Muccio-Failla & Giacomini, 2017).

In the rest of the paper, we will focus on the encoding of the three semantic phenomena that we have detected while building the pattern resources, and encoded in Skema. They are: Semantic Type Alternation, Semantic Type Shift (Metonymy), and Complex Type Exploitation.

3. Semantic Type Alternation

Let us start with the most frequent phenomenon, Semantic Type Alternation. When different Semantic Types alternate on the same argument slot within the same verb sense – i.e. within the same *pattern* – a Semantic Type alternation is at play. Semantic Type alternations are a pervasive phenomenon in both the T-PAS and CROATPAS resources and are graphically encoded by adding vertical bars "|" (which stand for the OR operator) between the alternating Semantic Types.

An example of Semantic Type Alternation on the subject position is the one between [Human] and [Wind] in the context of pattern 1 of the Italian verb *rimuovere* 'to remove' (Figure 2).

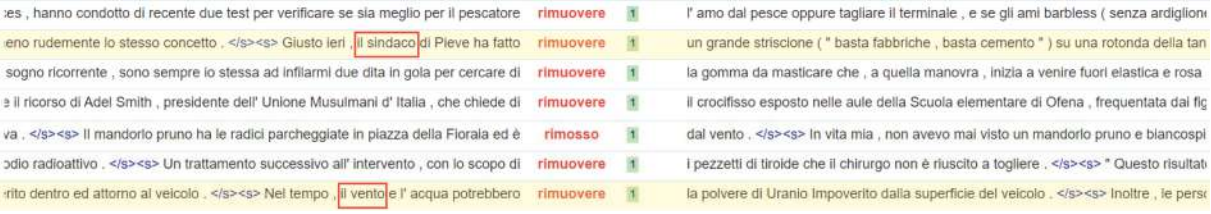


1

[Human] | [Wind] **rimuovere** [Inanimate]
[Human] | [Wind] sposta, porta via, toglie [Inanimate]

Figure 2: Pattern 1 of the Italian verb *rimuovere* 'to remove'

The following corpus lines (Figure 3) can be considered to be instantiations of the pattern:

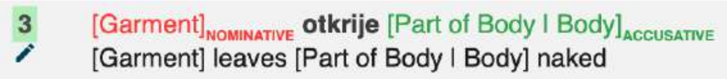


:es , hanno condotto di recente due test per verificare se sia meglio per il pescatore	rimuovere	1	l' amo dal pesce oppure tagliare il terminale , e se gli ami barbless (senza ardiglion
eno rudemente lo stesso concetto . </s><s> Giusto ieri , il sindaco di Pieve ha fatto	rimuovere	1	un grande striscione (" basta fabbriche , basta cemento ") su una rotonda della tan
sogno ricorrente , sono sempre io stessa ad infilarmi due dita in gola per cercare di	rimuovere	1	la gomma da masticare che , a quella manovra , inizia a venire fuori elastica e rosa
e il ricorso di Adel Smith , presidente dell' Unione Musulmani d' Italia , che chiede di	rimuovere	1	il crocifisso esposto nelle aule della Scuola elementare di Ofena , frequentata dai fig
va . </s><s> Il mandorlo pruno ha le radici parcheggiate in piazza della Fioraia ed è	rimosso	1	dal vento . </s><s> In vita mia , non avevo mai visto un mandorlo pruno e biancospi
odio radioattivo . </s><s> Un trattamento successivo all' intervento , con lo scopo di	rimuovere	1	i pezzetti di tiroide che il chirurgo non è riuscito a togliere . </s><s> " Questo risultat
rito dentro ed attorno al veicolo . </s><s> Nel tempo , il vento e l' acqua potrebbero	rimuovere	1	la polvere di Uranio Impoverito dalla superficie del veicolo . </s><s> Inoltre , le pers

Figure 3: Corpus lines linked to pattern 1 of the Italian verb *rimuovere* 'to remove' with subjects in red

Let us compare the two highlighted sentences: *Il sindaco di Pieve ha fatto rimuovere un grande striscione*, 'The major of Pieve had a big banner removed', in which the word *sindaco* is an instance of the Semantic Type [Human], and *Il vento e l'acqua potrebbero rimuovere la polvere di Uranio impoverito dalla superficie del veicolo*, 'Wind and water may remove uranium dust from the vehicle's surface', in which *vento* instantiates the Semantic Type [Wind]. In both cases, the meaning of the verb is the same, that is, 'removing' something. In this case, the two Semantic Types are not linked by any kind of relation. This is not true for all Semantic Type Alternations, as we will show below.

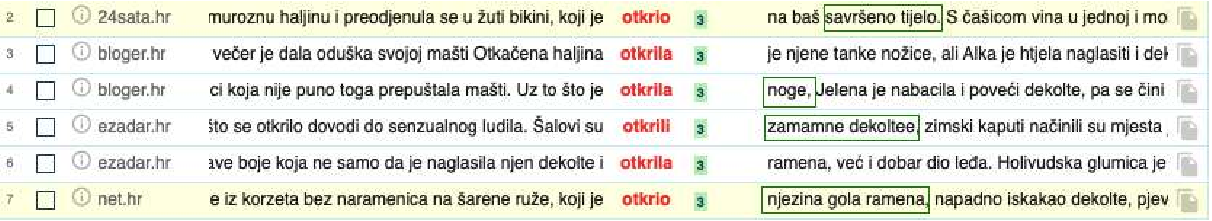
Turning now to the object position, an interesting alternation taking place on the object slot of pattern 3 of the Croatian verb *otkriti* 'to reveal' is [Part of Body | Body] (Figure 4).[8]



3 [Garment]$_{\text{NOMINATIVE}}$ **otkrije** [Part of Body I Body]$_{\text{ACCUSATIVE}}$
[Garment] leaves [Part of Body I Body] naked

Figure 4: Pattern 1 of the Croatian verb *otkriti* 'to reveal'

Unlike the previous case, in this case the two alternating Semantic Types are clearly linked by a meronymic relationship of Part/Whole. For this reason, it is all the more obvious that their alternation does not imply any meaning shift in the verb, as is testified by the highlighted sentences from Figure 5: *(Korzet) je otkrio njezina gola ramena*, 'The corset revealed her bare shoulders', and *Skinula je glamuroznu haljinu i preodjenula se u žuti bikini, koji je otkrio na baš savršeno tijelo*, 'She took off the glamourous dress and changed into a yellow bikini, which revealed a truly perfect body'.



2	□	24sata.hr	muroznu haljinu i preodjenula se u žuti bikini, koji je	otkrio	3	na baš savršeno tijelo. S čašicom vina u jednoj i mo
3	□	bloger.hr	večer je dala oduška svojoj mašti Otkačena haljina	otkrila	3	je njene tanke nožice, ali Alka je htjela naglasiti i del
4	□	bloger.hr	ci koja nije puno toga prepuštala mašti. Uz to što je	otkrila	3	noge, Jelena je nabacila i poveći dekolte, pa se čini
5	□	ezadar.hr	što se otkrilo dovodi do senzualnog ludila. Šalovi su	otkrili	3	zamamne dekoltee, zimski kaputi načinili su mjesta
6	□	ezadar.hr	ave boje koja ne samo da je naglasila njen dekolte i	otkrila	3	ramena, već i dobar dio leđa. Holivudska glumica je
7	□	net.hr	e iz korzeta bez naramenica na šarene ruže, koji je	otkrio	3	njezina gola ramena, napadno iskakao dekolte, pjev

Figure 5: Corpus lines linked to pattern 3 of the Croatian verb *otkriti* 'to reveal', with objects in green

To provide an idea of the frequency of Semantic Type alternations, we report some raw figures from T-PAS. For each argument position (column 1), we provide the number of patterns that include that argument slot in their valency structure (column 2) and the number of patterns featuring at least one Semantic Type Alternation in that position (column 3).

The final line of Table 1 displays the overall number of T-PAS patterns (column 2) and the overall number of T-PAS patterns with at least one alternation on any argument position (column 3). Note that these numbers are lower than the sum of the elements in each column, since the same pattern can encompass more than one argument slot (e.g. a subject and an object), each potentially bearing a Semantic Type Alternation. However, we can still state that nearly 45 percent (2,468 out of 5,529) of the patterns

[8] Since Croatian is a Slavic language equipped with its own case system to express the relationships between sentence components, the Croatian version of the Skema editor has been enriched by adding explicit bottom-right case markings on each argument slot, such as *nominative* or *accusative*.

Argument	No. of patterns	No. of patterns with Semantic Type Alternation
Subject	5,503	1,687
Object	3,184	1,097
Prepositional complement	1,668	450
Adverbial	379	0
Clausal	435	9
Predicative complement	108	16
Overall	5,529	2,468

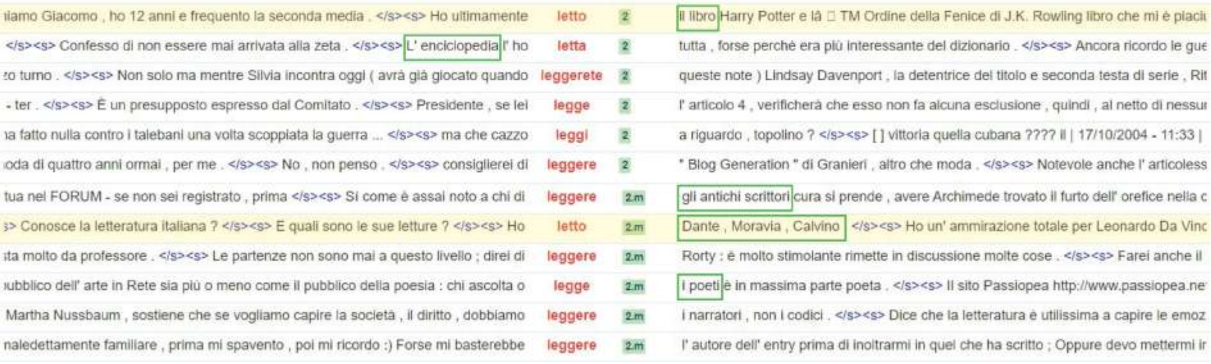
Table 1: T-PAS patterns featuring Semantic Type Alternations for each argument position

in the inventory feature a Semantic Type alternation on at least one of their argument positions.

4. Semantic Type Shift

In both T-PAS and CROATPAS, the changes in meaning of an argument caused by metonymic displacements are not encoded as Semantic Type Alternations but as Semantic Type Shifts. Following Pustejovsky (1995), we assumed that such shifts take place when a Semantic Type is forced by the verb to be understood as a different one (which satisfies its semantic selectional requirements or preferences).

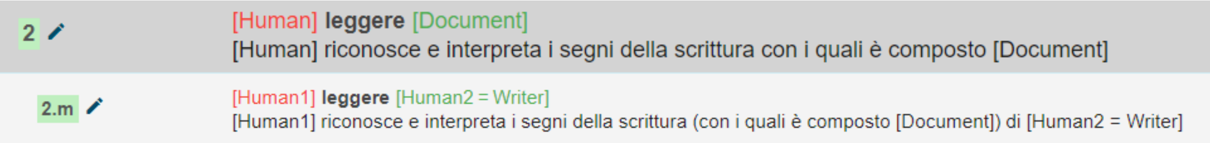
Three clear-cut cases of metonymy are offered by the sentence *Ho letto Dante, Moravia, Calvino*, 'I have read Dante, Moravia, Calvino, ' from Figure 6, where the Italian verb *leggere*, 'to read', triggers a shift from [Human = Writer] to [Document]. Unlike in the first highlighted sentence - *Ho ultimamente letto il libro di Harry Potter*, 'I have recently read the Harry Potter book' – each time the verb *leggere* combines with the name of an author on the object position, the well-known Author/Work metonymy takes place, forcing that person to be interpreted as the *document he or she has written.*



ıiamo Giacomo , ho 12 anni e frequento la seconda media . </s><s> Ho ultimamente	letto	2	il libro Harry Potter e là □ TM Ordine della Fenice di J.K. Rowling libro che mi è piaciı
</s><s> Confesso di non essere mai arrivata alla zeta . </s><s> L' enciclopedia l' ho	letta	2	tutta , forse perchè era più interessante del dizionario . </s><s> Ancora ricordo le gue
zo turno . </s><s> Non solo ma mentre Silvia incontra oggi (avrà già giocato quando	leggerete	2	queste note) Lindsay Davenport , la detentrice del titolo e seconda testa di serie , Rit
- ter . </s><s> È un presupposto espresso dal Comitato . </s><s> Presidente , se lei	legge	2	l' articolo 4 , verificherà che esso non fa alcuna esclusione , quindi , al netto di nessuı
ıa fatto nulla contro i talebani una volta scoppiata la guerra ... </s><s> ma che cazzo	leggi	2	a riguardo , topolino ? </s><s> [ ] vittoria quella cubana ???? il | 17/10/2004 - 11:33 |
ıoda di quattro anni ormai , per me . </s><s> No , non penso . </s><s> consiglierei di	leggere	2	" Blog Generation " di Granieri , altro che moda . </s><s> Notevole anche l' articoless
tua nel FORUM - se non sei registrato , prima </s><s> Si come è assai noto a chi di	leggere	2.m	gli antichi scrittori cura si prende , avere Archimede trovato il furto dell' orefice nella c
s> Conosce la letteratura italiana ? </s><s> E quali sono le sue letture ? </s><s> Ho	letto	2.m	Dante , Moravia , Calvino </s><s> Ho un' ammirazione totale per Leonardo Da Vinc
ıta molto da professore . </s><s> Le partenze non sono mai a questo livello ; direi di	leggere	2.m	Rorty : è molto stimolante rimette in discussione molte cose . </s><s> Farei anche il
ıubblico dell' arte in Rete sia più o meno come il pubblico della poesia : chi ascolta o	legge	2.m	i poeti è in massima parte poeta . </s><s> Il sito Passiopea http://www.passiopea.ne
Martha Nussbaum , sostiene che se vogliamo capire la società , il diritto , dobbiamo	leggere	2.m	i narratori , non i codici . </s><s> Dice che la letteratura è utilissima a capire le emoz
naledettamente familiare , prima mi spavento , poi mi ricordo :) Forse mi basterebbe	leggere	2.m	l' autore dell' entry prima di inoltrarmi in quel che ha scritto ; Oppure devo mettermi ir

Figure 6: Corpus lines linked to pattern 2 and subpattern 2.m of the Italian verb *leggere* 'to read'

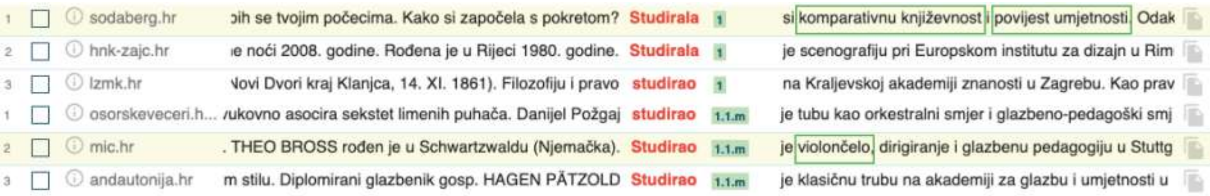
As shown in Figure 7, the metonymy at play is encoded in Skema as a sub-pattern of the main pattern [Human] reads [Document] (Romani & Ježek, 2020; Marini & Ježek, 2020). Note that the labels of the subpattern and of the metonymic corpus lines linked to it are the same: they start with the same number as the main pattern label and end in '.m', which stands for *metonymic*.



2	[Human] **leggere** [Document] [Human] riconosce e interpreta i segni della scrittura con i quali è composto [Document]
2.m	[Human1] **leggere** [Human2 = Writer] [Human1] riconosce e interpreta i segni della scrittura (con i quali è composto [Document]) di [Human2 = Writer]

Figure 7: Pattern 2 and metonymic subpattern 2.m of the Italian verb *leggere* 'to read'

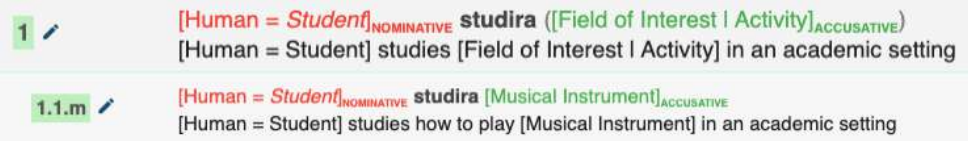
Let us now consider the Semantic Type shift taking place in the last three corpus instances from Figure 8 – one of them being *Studirao je violončelo*, 'He studied cello' – and compare them to the first three non-metonymic examples – e.g. *Studirala si komparativnu književnost i povijest umjetnosti*, 'You studied comparative literature and art history'.



1	□	sodaberg.hr	ɔih se tvojim počecima. Kako si započela s pokretom?	Studirala	1	si komparativnu književnost i povijest umjetnosti. Odak
2	□	hnk-zajc.hr	ıe noći 2008. godine. Rođena je u Rijeci 1980. godine.	Studirala	1	je scenografiju pri Europskom institutu za dizajn u Rim
3	□	lzmk.hr	Novi Dvori kraj Klanjca, 14. XI. 1861). Filozofiju i pravo	studirao	1	na Kraljevskoj akademiji znanosti u Zagrebu. Kao prav
1	□	osorskeveceri.h...	/ukovno asocira sekstet limenih puhača. Danijel Požgaj	studirao	1.1.m	je tubu kao orkestralni smjer i glazbeno-pedagoški smj
2	□	mic.hr	. THEO BROSS rođen je u Schwartzwaldu (Njemačka).	Studirao	1.1.m	je violončelo, dirigiranje i glazbenu pedagogiju u Stuttg
3	□	andautonija.hr	m stilu. Diplomirani glazbenik gosp. HAGEN PÄTZOLD	Studirao	1.1.m	je klasičnu trubu na akademiji za glazbu i umjetnosti u

Figure 8: Corpus lines linked to pattern 1 and subpattern 1.m of the Croatian verb *studirati* 'to study'

Indeed, having studied comparative literature and art history implies having acquired a deep knowledge of those theoretical fields, whereas having studied a [Musical Instrument] means "having studied how to play it". This last piece of information is not explicitly stated, but is the result of a metonymic shift triggered by the verb *studirati*, 'to study', which requires either a theoretical [Field of Interest] or an [Activity] in the direct object slot (Figure 9), thus forcing [Musical Instrument] to be understood as the [Activity] of playing it.



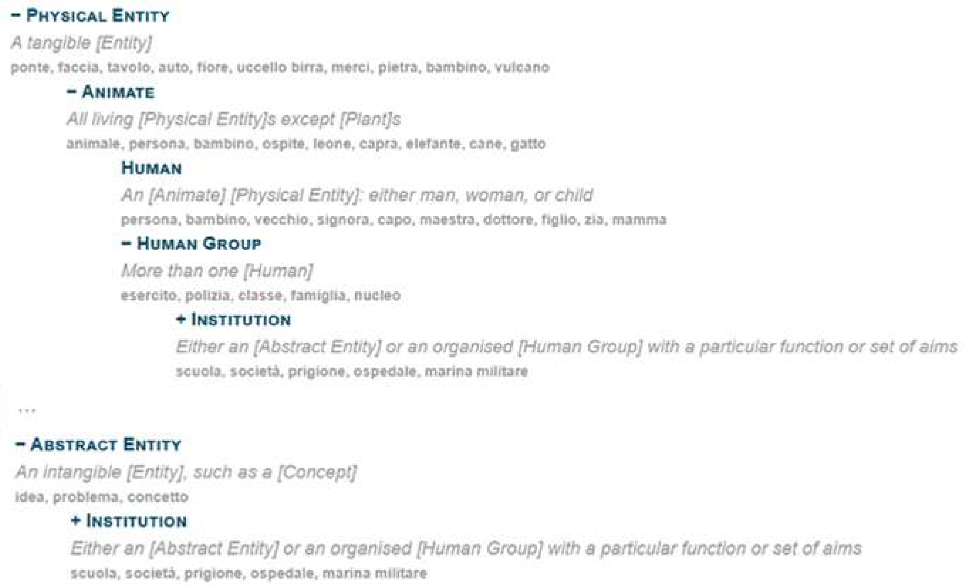
1	[Human = *Student*]$_{\text{NOMINATIVE}}$ **studira** ([Field of Interest I Activity]$_{\text{ACCUSATIVE}}$) [Human = Student] studies [Field of Interest I Activity] in an academic setting
1.1.m	[Human = *Student*]$_{\text{NOMINATIVE}}$ **studira** [Musical Instrument]$_{\text{ACCUSATIVE}}$ [Human = Student] studies how to play [Musical Instrument] in an academic setting

Figure 9: Pattern 1 and metonymic subpattern 1.1.m of the Croatian verb *studirati*, 'to study'.

5. Complex Type Exploitation

In our System of Semantic Types, we acknowledge the existence of Complex Types. Complex Types are unique Semantic Types "made up" of two (or more) components (Pustejovsky & Ježek, 2008): for example, [Institution] is a Complex Type made up of

[Abstract Entity] and [Human Group]. In the Skema editor, we currently encode Complex Types as “simple” Semantic Types (e.g. [Institution]). However, we keep track of their internal complexity by locating them in multiple places in the System of Semantic Types, as sub-types of their components: for example, as one can see in Figure 10, the Complex Type [Institution] is located both under [Human Group], which is a kind of [Animate] entity, and under [Abstract Entity].[9] We call this phenomenon multiple inheritance, meaning that a Complex Semantic Type inherits from different Types of the hierarchy.



− PHYSICAL ENTITY
A tangible [Entity]
ponte, faccia, tavolo, auto, fiore, uccello birra, merci, pietra, bambino, vulcano

− ANIMATE
All living [Physical Entity]s except [Plant]s
animale, persona, bambino, ospite, leone, capra, elefante, cane, gatto

HUMAN
An [Animate] [Physical Entity]: either man, woman, or child
persona, bambino, vecchio, signora, capo, maestra, dottore, figlio, zia, mamma

− HUMAN GROUP
More than one [Human]
esercito, polizia, classe, famiglia, nucleo

+ INSTITUTION
Either an [Abstract Entity] or an organised [Human Group] with a particular function or set of aims
scuola, società, prigione, ospedale, marina militare

...

− ABSTRACT ENTITY
An intangible [Entity], such as a [Concept]
idea, problema, concetto

+ INSTITUTION
Either an [Abstract Entity] or an organised [Human Group] with a particular function or set of aims
scuola, società, prigione, ospedale, marina militare

Figure 10: The System of Semantic Types used in T-PAS and CROATPAS, where the Complex Type [Institution] is registered both as a type of [Human Group] and [Abstract Entity]

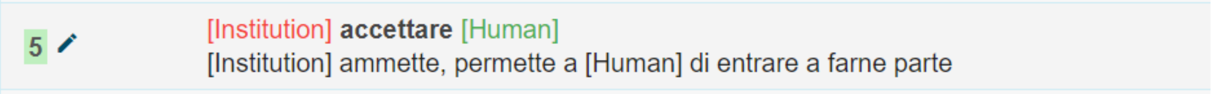
That having been said, we encode a Complex Type Exploitation when a verb exploits only one of the components of a Complex Semantic Type associated with an argument. In this case, no metonymy occurs because there is no change of referent, as is the case in the examples in Section 4. In the following, we provide some examples of Complex Type Exploitation in the patterns of T-PAS and CROATPAS, focusing on two Complex Types, and highlighting which component is exploited. We also include instances of co-predication, i.e. contexts in which both components are simultaneously exploited.[10]

The first Complex Type we examine is [Institution], whose components are [Abstract Entity] and [Human Group]. In pattern 5 of the T-PAS verb *accettare*, ‘to accept’, for example, the verb only exploits the human component of the Complex Type [Institution] of its subject (Figure 11), as the act of accepting someone is typical of a [Human] or a [Human Group]:[11]

[9] Each component of a Complex Type is a “real” Semantic Type, which can also be used independently of the Complex Type.

[10] Recall that co-predication is the test traditionally used in linguistic and ontological studies to identify Complex Types (Pustejovsky, 1995).

[11] Even though the Semantic Types [Human] and [Human Group] are connected by the Whole/Part relationship (given that a [Human Group] is a group of more than one [Human]), they are not related in our System as the only relation that we consider is the relation of subsumption, e.g. ‘is-a-type-of’. They are listed at the same level and subtypes of [Animate]s

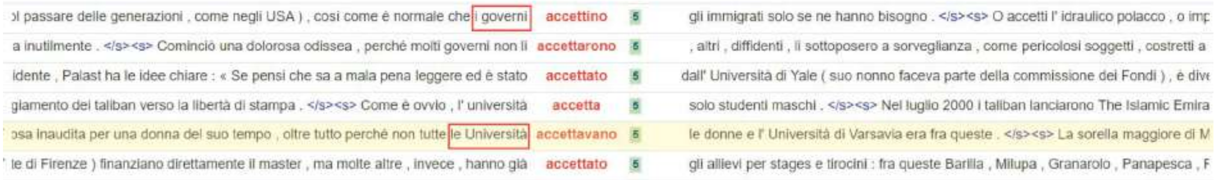


5

[Institution] **accettare** [Human]
[Institution] ammette, permette a [Human] di entrare a farne parte

Figure 11: Pattern 5 of the Italian verb *accettare*, ‘to accept’, featuring the Complex Type [Institution] in the subject position, exploited in its [Human Group] component.

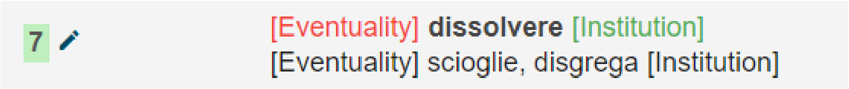
This pattern is instantiated by corpus examples such as *Non tutte le università accettavano le donne e l’Università di Varsavia era tra queste* ‘Not all universities accepted women, and the University of Warsaw was one of them’, as highlighted in Figure 12.



ol passare delle generazioni , come negli USA) , così come è normale che i governi	accettino	5	gli immigrati solo se ne hanno bisogno . </s><s> O accetti l' idraulico polacco , o imp
a inutilmente . </s><s> Cominciò una dolorosa odissea , perché molti governi non li	accettarono	5	, altri , diffidenti , li sottoposero a sorveglianza , come pericolosi soggetti , costretti a
idente , Palast ha le idee chiare : « Se pensi che sa a mala pena leggere ed è stato	accettato	5	dall' Università di Yale (suo nonno faceva parte della commissione dei Fondi) , è dive
glamento dei taliban verso la libertà di stampa . </s><s> Come è ovvio , l' università	accetta	5	solo studenti maschi . </s><s> Nel luglio 2000 i taliban lanciarono The Islamic Emira
osa inaudita per una donna del suo tempo , oltre tutto perché non tutte le Università	accettavano	5	le donne e l' Università di Varsavia era fra queste . </s><s> La sorella maggiore di M
le di Firenze) finanziano direttamente il master , ma molte altre , invece , hanno già	accettato	5	gli allievi per stages e tirocini : fra queste Barilla , Milupa , Granarolo , Panapesca , F

Figure 12: Corpus instances for the verb *accettare* ‘to accept’ and instantiations of the Semantic Type [Institution]

Conversely, in pattern 7 from Figure 13, the verb *dissolvere*, ‘to dissolve’, only selects the [Abstract Entity] component of [Institution]. The meaning of the verb in this case is figurative:

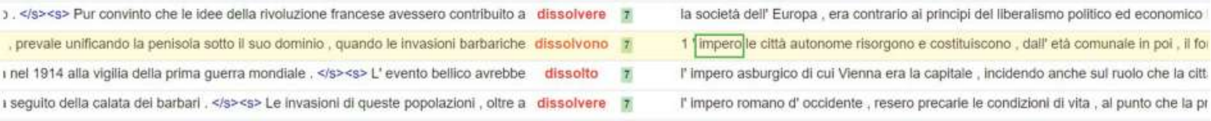


7

[Eventuality] **dissolvere** [Institution]
[Eventuality] scioglie, disgrega [Institution]

Figure 13: Pattern 7 of the verb *dissolvere*, ‘to dissolve’, featuring the Complex Type [Institution] in the object position, exploited in its [Abstract Entity] component

An example of this kind of exploitation is *Le invasioni barbariche dissolvono l’Impero* ‘Barbaric invasions disintegrate the Empire’, as highlighted in Figure 14:

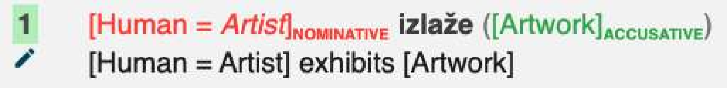


ɔ . </s><s> Pur convinto che le idee della rivoluzione francese avessero contribuito a	dissolvere	7	la società dell' Europa , era contrario ai principi del liberalismo politico ed economico
, prevale unificando la penisola sotto il suo dominio , quando le invasioni barbariche	dissolvono	7	1 impero le città autonome risorgono e costituiscono , dall' età comunale in poi , il fo
ı nel 1914 alla vigilia della prima guerra mondiale . </s><s> L' evento bellico avrebbe	dissolto	7	l' impero asburgico di cui Vienna era la capitale , incidendo anche sul ruolo che la citt
ı seguito della calata dei barbari . </s><s> Le invasioni di queste popolazioni , oltre a	dissolvere	7	l' impero romano d' occidente , resero precarie le condizioni di vita , al punto che la pr

Figure 14: Corpus instances of the verb *dissolvere*, ‘to dissolve’, and instantiations of the Semantic Type [Institution]

We may also consider corpus sentences that display co-predication, that is, cases in which both components are exploited with regard to the same argument, as in *L’Università di Padova fu fondata nel 1222 ed è stata la prima al mondo ad accettare studenti ebrei*, ‘The University of Padova was founded in 1222, and it was the first in the world to accept Jewish students.’ In this case, the verb *fondare*, ‘to found’, taking [Institution] as an object, exploits the [Abstract Entity] component, whereas the verb *accettare*, as in the previous sentence, exploits the [Human Group] component.

As a second case, we consider examples of the exploitation of the Complex Type [Artwork], whose components are [Artifact] and [Concept]. For example, the Croatian verb *izlagati*, ‘to exhibit’, exploits only the Artifact component of this Complex Type, which we can find in the object position of pattern 1 in Figure 15 from CROATPAS.

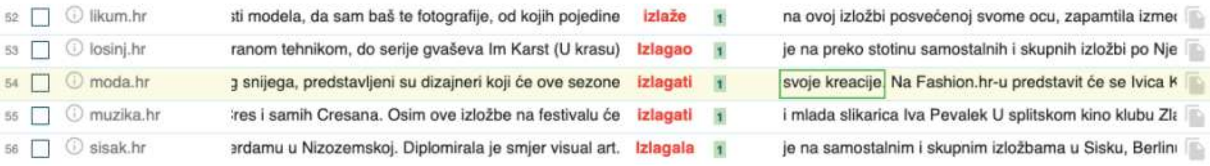


1 [Human = *Artist*]$_{\text{NOMINATIVE}}$ **izlaže** ([Artwork]$_{\text{ACCUSATIVE}}$)

✎ [Human = Artist] exhibits [Artwork]

Figure 15: Pattern 1 of the Croatian verb *izlagati*, ‘to exhibit’, and the Complex Type [Artwork] in the object position, exploited in its [Artifact] component

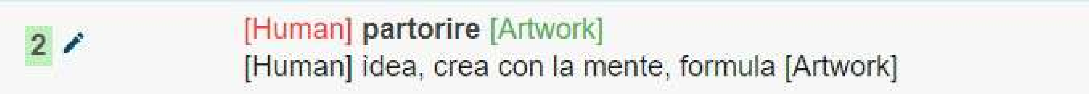
After all – as we can see from sentences such as *Predstavljeni su dizajneri koji će ove sezone izlagati svoje kreacije*, ‘The designers that will exhibit their creations this season have been presented’, from Figure 16 – artistic creations need to be physical entities in order to be exhibited.



52	□	likum.hr	ɜti modela, da sam baš te fotografije, od kojih pojedine	izlaže	1	na ovoj izložbi posvećenoj svome ocu, zapamtila izme
53	□	losinj.hr	ranom tehnikom, do serije gvaševa Im Karst (U krasu)	Izlagao	1	je na preko stotinu samostalnih i skupnih izložbi po Nje
54	□	moda.hr	ȝ snijega, predstavljeni su dizajneri koji će ove sezone	izlagati	1	svoje kreacije Na Fashion.hr-u predstavit će se Ivica K
55	□	muzika.hr	ːres i samih Cresana. Osim ove izložbe na festivalu će	izlagati	1	i mlada slikarica Iva Pevalek U splitskom kino klubu Zl
56	□	sisak.hr	ɘrdamu u Nizozemskoj. Diplomirala je smjer visual art.	Izlagala	1	je na samostalnim i skupnim izložbama u Sisku, Berlin

Figure 16: Corpus lines linked to pattern 1 of the Croatian verb *izlagati*, ‘to exhibit’.

Conversely, we can consider pattern 2 of the Italian verb *partorire*, ‘to give birth’, for the exploitation of the [Concept] component of the Complex Type [Artwork] (Figure 17). Note that *partorire* does not select the physical component of [Artwork], since its meaning is figurative: one cannot literally ‘give birth to an [Artwork]’, but rather we can talk of conceiving it in our mind, which is why we are only exploiting its conceptual component.

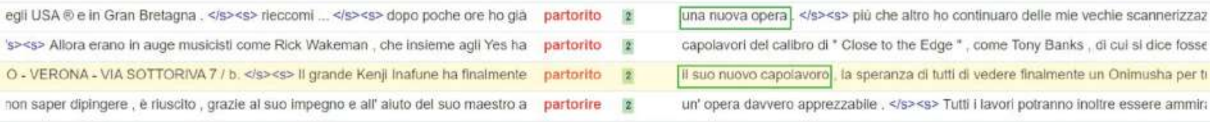


2

[Human] partorire [Artwork]
[Human] idea, crea con la mente, formula [Artwork]

Figure 17: Pattern 2 of the Italian verb *partorire*, 'to give birth', featuring the Complex Type [Artwork] in the object position, exploited in its [Concept] component

As we can see from the following corpus instances (Figure 18), the meaning is clearly 'to mentally conceive something', as in the sentence *Il grande Kenji Inafune ha finalmente partorito il suo nuovo capolavoro*, 'The great Kenji Inafune has finally given birth to its new masterpiece'.



egli USA ® e in Gran Bretagna . </s><s> rieccomi ... </s><s> dopo poche ore ho già	partorito	2	una nuova opera . </s><s> più che altro ho continuaro delle mie vechie scannerizzaz
's><s> Allora erano in auge musicisti come Rick Wakeman , che insieme agli Yes ha	partorito	2	capolavori del calibro di " Close to the Edge " , come Tony Banks , di cui si dice fosse
O - VERONA - VIA SOTTORIVA 7 / b. </s><s> Il grande Kenji Inafune ha finalmente	partorito	2	il suo nuovo capolavoro , la speranza di tutti di vedere finalmente un Onimusha per t
non saper dipingere , è riuscito , grazie al suo impegno e all' aiuto del suo maestro a	partorire	2	un' opera davvero apprezzabile . </s><s> Tutti i lavori potranno inoltre essere ammir

Figure 18: Corpus instances for *partorire*, 'to give birth', and instantiations of the [Artwork] Semantic Type

Finally, consider this instance of co-predication in which both components ([Concept] with *partorire*, 'to give birth' and [Artifact] with *presentare*, 'to present') are present: *Nel 1501 Leonardo da Vinci partorì un'opera di straordinaria importanza, che fu presentata al sultano Bezajet II: si trattava di un ponte ...* (Eng.: 'In 1501 Leonardo da Vinci gave birth to an artwork of extraordinary importance, which was presented to Sultan Bezajet II: it was a bridge ...').

6. Conclusions and future prospects

In this paper, we have shown how the semantic phenomena that take place in verb-argument combinations are encoded in two electronic resources dedicated to the description of corpus-derived verb-argument structures present in Italian and Croatian. In particular, we have discussed cases in which there is an alternation of Semantic Types on the same argument position within the same verb sense, cases where there is a Semantic Type Shift, and, finally, cases where a single component of a Complex Type denoted by a noun is exploited in the composition. We have shown how these data are currently stored in the off-line editor that we developed, called Skema.

In the near future, the data will be made public through a graphical interface, where users will be able to search for the three phenomena by browsing for the pattern and subpattern inventory (accompanied by Good Dictionary EXamples selected from the annotated corpus for each pattern (Kilgarriff et al., 2008)). Users will also be able to query the system of Semantic Types and the different argument positions (subject, object, prepositional complement, indirect complement, clausal, and predicative complement), both one at a time, as well as in combination.

The data in the two resources can be useful for linguistic research in syntax and semantics, for example, for studies aiming at classifying verbs based on the semantic selection of their arguments. Moreover, they can be useful for corpus-based approaches to language teaching, and possibly as a gold standard in natural language processing tasks involving figurative language recognition in accordance with Shutova et al. (2013), who used a combination of corpus analysis and knowledge base extraction to predict classes of words in order to identify instances of logical metonymy.

7. Acknowledgments

We would like to thank the three anonymous reviewers for their useful comments and suggestions.

8. References

Baisa, V., Tiberius, C., Ježek, E., Colman, L., Marini, C. & Romani, E. (2020). Skema: A New Tool for Corpus-driven Lexicography. In *Proceedings of the 19th EURALEX International Congress.*

Colman, L. & Tiberius, C. (2018). A Good Match: a Dutch Collocation, Idiom and Pattern Dictionary Combined. In *Proceedings of the 18th EURALEX International Congress.* Ljubljana, Slovenia.

Di Muccio-Failla, P. & Giacomini, L. (2017). Designing a Learner's Dictionary with Phraseological Disambiguators. In *Proceedings of the 2nd International Conference on Computational and Corpus-Based Phraseology (Europhras)*. London, UK.

Hanks, P. (2013). *Lexical Analysis: Norms and Exploitations.* Cambridge: The MIT Press.

Hanks, P. & Pustejovsky, J. (2005). A Pattern Dictionary for Natural Language Processing. *Revue Française de Linguistique Appliquée*, 11(2), pp. 63–82.

Ježek, E. (2019). Sweetening Ontologies Cont'd: Aligning bottom-up with top-down ontologies. In A. Barton, S. Seppälä & D. Porello (eds.) *Proceedings of the Joint Ontology Workshops 2019.* Graz, Austria.

Ježek, E., Magnini, B., Feltracco, A., Bianchini, A. & Popescu, O. (2014). T-PAS: A resource of corpus-derived Types Predicate-Argument Structures for linguistic analysis and semantic processing. In *Proceedings of LREC.* pp. 890–895.

Kilgarriff, A. (1993). I Don't Believe in Word Senses. *Computers and the Humanities*, 31(2), pp. 91–113.

Kilgarriff, A., Husák, M., Mcadam, K., Rundell, M. & Rychlý, P. (2008). GDEX: automatically finding good dictionary examples in a corpus. In *Proceedings of the 13th EURALEX International Congress.* Barcelona, Spain.

Marini, C. & Ježek, E. (2019). CROATPAS: A Resource of Corpus-derived Typed Predicate Argument Structures for Croatian. In *Proceedings of the 6th Italian Conference on Computational Linguistics.* Bari, Italy.

Marini, C. & Ježek, E. (2020). Annotating Croatian Semantic Type Coercions in CROATPAS. In *Proceedings of the 16th Joint ACL-ISO Workshop on Interoperable Semantic Annotation.* Marseille, France.

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

Pustejovsky, J., Hanks, P. & Rumshisky, A. (2004). Automated Induction of Sense in Context. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING).* Geneva, Switzerland.

Pustejovsky, J. & Ježek, E. (2008). Semantic Coercion in Language: Beyond Distributional Analysis. *Italian Journal of Linguistics/ Rivista Italiana di Linguistica*, 20(1), pp. 123–138.

Recanati, F. (2002). *Literal Meaning.* Cambridge: Cambridge University Press.

Renau, I. & Nazar, R. (2021). Verbario. URL: http://www.verbario.com. Last accessed: 1 June, 2021.

Romani, E. & Ježek, E. (2020). Tracing Metonymic Relations in T-PAS: An Annotation Exercise on a Corpus-based Resource for Italian. In *Proceedings of the 7th Italian Conference on Computational Linguistics*. Bologna, Italy.

Shutova, E., Kaplan, J., Teufel, S. & Korhonen, A. (2013). A computational model of logical metonymy. *ACM Transactions on Speech and Language Processing*, 11.

Heteronym Sense Linking

Lenka Bajčetić[1], Thierry Declerck[1,2], John P. McCrae[3]

[1]Austrian Centre for Digital Humanities and Cultural Heritage
Sonnenfelsgasse 19, Wien 1010, Austria
[2]German Research Center for Artificial Intelligence (DFKI)
Saarland Informatics Campus D3 2, Saarbrücken, Germany
[3] Data Science Insitute, NUI Galway, Ireland
E-mail: lenka.bajcetic@oeaw.ac.at, declerck@dfki.de, john@mccr.ae

Abstract

In this paper we present ongoing work which aims to semi-automatically connect pronunciation information to lexical semantic resources which currently lack such information, with a focus on WordNet. This is particularly relevant for the cases of heteronyms — homographs that have different meanings associated with different pronunciations — as this is a factor that implies a re-design and adaptation of the formal representation of the targeted lexical semantic resources: in the case of heteronyms it is not enough to just add a slot for pronunciation information to each WordNet entry. Also, there are numerous tools and resources which rely on WordNet, so we hope that enriching WordNet with valuable pronunciation information can prove beneficial for many applications in the future. Our work consists of compiling a small gold standard dataset of heteronymous words, which contains short documents created for each WordNet sense, in total 136 senses matched with their pronunciation from Wiktionary. For the task of matching WordNet senses with their corresponding Wiktionary entries, we train several supervised classifiers which rely on various similarity metrics, and we explore whether these metrics can serve as useful features as well as the quality of the different classifiers tested on our dataset. Finally, we explain in what way these results could be stored in OntoLex-Lemon and integrated to the Open English WordNet.

Keywords: Sense Linking; Heteronyms; Wordnets; Wiktionary

1. Introduction

There are many types of ambiguity in language, and one interesting example are homographs. These are words that are spelled the same, but they have different pronunciations. Specifically, homographs that have different meanings associated with different pronunciations, are called heteronyms (Martin et al., 1981).

Heteronyms can cause great challenges for speech-to-text and text-to-speech systems. They also provide an interesting use-case for our endeavour to enrich WordNet with pronunciation information.

Recently, the Global WordNet Association (GWA) updated its Global Wordnet Formats (McCrae et al., 2021)[1], which have been introduced to enable wordnets to have a common representation. One of the updates performed by GWA concerns the possibility to add pronunciation information to the entries of wordnets. GWA decided to “support the use of IETF language tags to indicate dialect”. This update is a great step towards integrating pronunciation information in wordnets.

As a complementary task to the representation of heteronymy in wordnets, we start with the task of supporting an automated linking between the senses of the heteronyms we extracted from Wiktionary and those included in the Open English WordNet (McCrae et al., 2020). While the sense linking task is in itself interesting Ahmadi & McCrae (2021), it can lead to an automated addition of the pronunciation information to the heteronyms included in English WordNet. Since English WordNet is a manually curated gold standard resource, this would lead to the possibility to get an evaluation of the linking work for this

[1] https://globalwordnet.github.io/schemas/

specific type of phenomenon and also to the building of a training set for an extension of the linking work.

2. Related Work

In order to be valuable, language resources should be accessible and legal to use, sufficient in terms of quality and size, and ideally with a documented interface (Ishida, 2006). According to these aspects, both WordNet and Wiktionary are language resources of the highest value, and it is no surprise there are many endeavours aimed at connecting the two. For instance, the work of Meyer & Gurevych (2011) shows that automatic alignments between Wiktionary senses and Princeton WordNet can be established by combining several text similarity scores to compare a bag of words based on several pieces of information linked to a WordNet sense with another bag of words obtained from a Wiktionary entry. This is quite similar to the approach we have followed also, as explained in the Method section. A large part of this work is also harnessing the multilingualism of the two resources, in an attempt to create very large multilingual corpora by aligning several Wiktionaries and WordNets.

Our previous work on heteronyms is presented in (Declerck & Bajčetić, 2021). This work consisted of extracting entries from Wiktionary that carry pronunciation information (following suggestions made by Schlippe et al. (2010)), for the four categories that are relevant for WordNet: nouns, verbs, adjectives, and adverbs. The result of this procedure consisted of listing each heteronymous word, together with its pronunciations, associated with their respective meanings and related example sentences. Declerck & Bajčetić (2021) propose a first representation of such entries using the OntoLex-Lemon model (Cimiano et al., 2016), suggesting a deduplication of lexical entries on the base of their different pronunciations, if those are related with specific meanings.

3. Method

When designing our linking approach[2], we have decided to pose our task as a classification problem. First we have created a dataset which consisted of the correct matches from the gold standard, and added their incorrect counterparts. In the end, we are left with a dataset of 272 examples labelled ‘True’ or ‘False’ depending on the matching. This means we have effectively transformed our sentence similarity task into a binary classification problem. While binary classification can be tackled in many ways, we have decided to experiment with supervised classifiers which were trained using various sentence similarity metrics.

3.1 Gold standard

In order to test and train our classifiers, we have compiled a small dataset which covers 10 examples of heteronymous words. The dataset consists of 136 WordNet senses matched with their pronunciation as stored in Wiktionary.

In the future we consider using the lists compiled by Martin et al. (1981). The authors have compiled an extensive list of 54 strong and 62 weak heteronyms. They came up

[2] The code and data are available here: https://github.com/acdh-oeaw/heteronym_sl

Word	Pronunciation 1	Pronunciation 2	N° of senses
bass	bæs	beɪs	9
bow	baʊ	boʊ	14
desert	dɪˈzɛːt	ˈdɛzət	4
house	haʊs	haʊz	14
lead	lɛd	liːd	31
live	lɪv	laɪv	19
raven	ˈɹeɪvən	ˈɹævən	5
row	ɹaʊ	ɹəʊ	10
subject	ˈsʌb.dʒɛkt	səbˈdʒɛkt	15
wind	wɪnd	waɪnd	15

Table 1: Gold standard

with this classification to reflect the distinctiveness of meaning between two senses which have different pronunciations. For example, "subject" is considered an example of weak heteronym, because the differently pronounced senses denote the same concept in verb and noun form. On the other hand, "row" is considered a strong heteronym, since the meanings it conveys are completely different. According to this classification, our list has examples of 3 weak heteronyms: "live", "house", and "subject".

3.2 Sense Linking

In order to parse Wiktionary files we extract headwords, parts of speech, definitions, examples and of course the pronunciation info from the XML Wiktionary database dumps as provided by the Wikimedia Foundation. The main body of Wiktionary articles are stored in a Wikitext format, which is a semi-structured format. Each article is centred around the "Etymology" section, and words which have several meanings have several etymologies. After extracting and packing all the relevant information from the Wiktionary article, we are left with several documents — one for each of the etymologies. For simplicity, we have chosen to work with those examples that have two possible etymologies, which in our case translates to two possible pronunciations, so our task can be understood as binary classification.

For each of the words, we retrieve all the senses from WordNet. Then, for each sense we extract the synonyms with their definitions, examples, and the hypernym hierarchy. By combining this information we create a short document for each sense. Finally, for each of the pairings of WordNet senses and their two corresponding Wiktionary articles, we have provided a final label of True or False. This means that our training dataset consists of 272 examples, half of which are correctly linked.

3.3 Features

The classifiers rely on five features:

- Wiktionary POS
- WordNet POS

- S-BERT similarity score
- Laser similarity score
- TFIDF similarity score

We have decided to use the POS tags because this is an intuitive and easy idea. This feature has proven useful, but less so in comparison with the similarity metrics.

In order to get the S-BERT similarity score, we use the cosine distance of the sentence embeddings from a transformer model which is pre-trained for paraphrase identification (paraphrase-distilroberta-base-v1) and a model which is pre-trained for semantic textual similarity (stsb-roberta-base). Sentence-BERT is a modification of the pretrained BERT network which uses Siamese and triplet network structures to derive semantically meaningful sentence embeddings that can be easily compared using cosine-similarity (Reimers & Gurevych, 2019). Sentence transformers are the current state-of-the-art approach in Semantic Textual Similarity (STS) tasks, and they perform very competitively in all sentence similarity tasks. The performance of S-BERT is evaluated for common STS tasks using the cosine-similarity to compare the similarity between two sentence embeddings, which is exactly the approach we have followed also.

The LASER similarity score is obtained in the same way, with the distinction of using LASER sentence embeddings[3]. LASER stands for Language-Agnostic SEntence Representations, and it uses a single pre-trained BiLSTM encoder for 93 languages, obtaining very strong results in various scenarios without any fine-tuning, including cross-lingual textual similarity (Artetxe & Schwenk, 2018). Since we intend to expand this research to other languages, we have decided it is important to explore multilingual options as well as English language specific ones, despite the fact LASER scores on monolingual tasks are usually not as good as the ones obtained using monolingual BERT-based sentence transformers (Artetxe & Schwenk, 2018).

Finally, the TFIDF similarity score is created by following the approach laid out by Meyer & Gurevych (2011). In their work, they utilize cosine distance between bag-of-words vectors as a similarity measure. The cosine similarity calculates the cosine of the angle between a vector representation of the two senses s1 and s2:

$$COS(s1, s2) = \frac{BoW(s1) \cdot BoW(s2)}{||BoW(s1)||||BoW(s2)||}$$

Following this approach, for each word in the gold standard we simply create a corpus of short documents explaining senses from WordNet, and we create pairs of Wiktionary documents which explain the two different pronunciations. Then, we calculate the value of the cosine distance for all document combinations.

3.4 Classifiers

After feature extraction, data is split into training and testing subsets with 2:1 ratio, and we use it to train several simple classifiers from sklearn[4]:

[3] https://github.com/yannvgn/laserembeddings

[4] All the classifiers can be found here: https://scikit-learn.org/stable/supervised_learning.html

- Naive Bayes
- Decision Tree
- Random Forest

Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes' theorem with the "naive" assumption of conditional independence between every pair of features given the value of the class variable. Despite this over-simplified assumption, naive Bayes classifiers have performed quite well in many tasks, especially document classification and spam filtering. They require a small amount of training data to estimate the necessary parameters, and for this reason we have decided to try it. However, this model has not proven so good in our task, and consistently achieved scores lower than other classifiers.

Decision Trees are a non-parametric supervised learning method used for classification and regression. The model aims to predict the value of a target variable by learning simple decision rules inferred from the data features. Decision trees are simple to understand and to interpret, and they require little data preparation. As we can see in the Results section, this model has shown good results but not as good as the random forest classifier.

A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. Since the random forest classifier has proven to have the best results, we have decided to fine-tune it by trying different sets of parameters. More specifically, we explored different values for the number of trees in the forest (estimators) and the maximum number of levels in each decision tree (max depth). First, we used GridSearch from sklearn library to determine the best set of parameters from Table 2, and then we trained several classifiers with those parameters, but experimenting with different number of estimators and max depth. A graph of the classifiers' accuracy depending on the hyperparameters value is shown in Figure 1 and Figure 2.

The results so far do not show clearly which is the best parameter set. This is most likely due to the small training set which is the biggest limitation of our model. Before obtaining a larger set it is hard to get definitive results which is the best model, we can only notice some trends regarding the potential shown by some features or classifiers.

Parameter	Values
bootstrap	True, False
max features	auto, sqrt
min samples leaf	1, 2, 4
min samples split	2, 5, 10
max depth	None, 2, 4, 6, 8, 10, 15, 20, 30, 40, 50, 60, 70, 80, 90, 100
n estimators	5, 10, 15, 20, 50, 100, 150, 200, 400, 600, 800, 1000

Table 2: Different parameters tried for Random Forest

3.5 Connecting to Open English WordNet

Open English WordNet (McCrae et al., 2020) is an open-source fork of the Princeton WordNet (Miller, 1995). The wordnet is freely available on GitHub and is formatted according to the XML schemas defined at https://globalwordnet.github.io/schemas/ (McCrae et al., 2021). Currently, there are no distinctions between heteronyms in WordNet, so these would need to be introduced. As a result of adopting the XML schema, it is now possible to define two lexical entries with the same lemma that was not possible in the previous formats used by Princeton WordNet. In addition, recently support was added for indicating the pronunciation in the schema files. An example of this new modelling is as follows:

```
<LexicalEntry id="ewn-bass-n-1">
  <Lemma writtenForm="bass" partOfSpeech="n">
    <Pronunciation notation="fonipa">bæs</Pronunciation>
  </Lemma>
  <Sense id="bass%1:05:00::"
         synset="ewn-02568204-n"/>
  ...
</LexicalEntry>
<LexicalEntry id="ewn-bass-n-2">
  <Lemma writtenForm="bass" partOfSpeech="n">
    <Pronunciation notation="fonipa">beIs</Pronunciation>
  </Lemma>
  <Sense id="bass%1:06:02::"
         synset="ewn-02806515-n"/>
  ...
</LexicalEntry>
```

In this example, we see two entries `ewn-bass-n-1` with a pronunciation to rhyme with 'mass' and `ewn-bass-n-2` with a pronunciation that rhymes with 'face', the senses are assigned to one of each of the entries. Note that, each of the entries are actually organized into distinct lexicographer files, so in this case it is merely the task of identifying which of the lexicographer files corresponds to which pronunciation, e.g., `noun.food` and `noun.animal` for the first pronunciation and `noun.attribute`, `noun.communication`, `noun.person`, `noun.artifact` and `adj.all` for the second.

3.6 Representation of heteronyms in OntoLex-Lemon

The work of Declerck & Bajčetić (2021) discusses the addition of pronunciation information in wordnets, with a focus on heteronyms. Those cases are particularly relevant for wordnets, as they do carry specific senses that need to be accounted for in such lexical semantics repositories. The authors make use of the OntoLex-Lemon representation model, as it has proven to be well adapted for linking the conceptual type of resources, as exemplified by wordnets, with the full lexicographic descriptions of the lemmas, which in wordnets are only minimally represented (just the written form and the associated part-of-speech). OntoLex-Lemon introduces form variants of lexical entries as full ontological objects, which can therefore carry information on a number of grammatical properties, like gender, case, and number. Those "form" objects also include the corresponding written and phonetic representations. So that (Declerck & Bajčetić, 2021) could propose a way to represent in OntoLex-Lemon the combination of wordnet entries and lexical entries, which are themselves pointing to form variants displaying the corresponding pronunciation information. The challenge would be now to extend this approach to compound words, and we are investigating for this the use of the *decomp* module of OntoLex-Lemon.[5]

4. Results

First we will compare the results of the four classifier models, namely naive Bayes, Decision Tree, and two versions of the random forest classifier. Then, we will take a closer look at the relevance of each feature for the classification. In the end, we will see how the variance in the parameter set for training the random forest classifier affects the results.

As we have previously mentioned, we employed two different pre-trained models for obtaining the S-Bert similarity feature, namely a model pre-trained on paraphrase detection and a model trained for semantic textual similarity task. As we can see in the results below, the similarity feature extracted using the paraphrase model has proven to give better results in our case. This makes sense, as our documents are usually not aligned with each other and consist of examples and definitions glued together, sometimes incoherently. It appears that for this kind of data, paraphrase detection serves as a better benchmark task than a typical STS task. Of course, we cannot know this for certain before we obtain a larger training set to experiment with.

Classifier	Accuracy	Precision	Recall	F1-score
Naive-Bayes	0.65	0.67	0.65	0.65
Decision Tree	0.71	0.70	0.71	0.71
Random Forest - STS [6]	0.79	0.80	0.79	0.79
Random Forest - Paraphrase[7]	0.84	0.85	0.84	0.84

Table 3: Performance of different classifiers on our gold standard test

In order to compare the benefits provided by different similarity metrics, we have tried using them as a basis for very simple classifiers with a threshold value. This is quite a simple, yet effective, way to investigate the capacity of each feature in our task. Since for this purpose you do not need to train a classifier, we have used the whole gold standard as the test set for this simple one-feature threshold-based classifiers. As we can see in Figure 1, both classifiers which are based on S-Bert similarity score can reach a score on our task of up to 0.7, with the right threshold value. This shows that S-Bert similarity score has great potential as a feature, even though it is not sufficient as a classifier by itself. On the other hand, the classifiers based on LASER and TFIDF similarity scores are not quite as useful to work on their own. As another way to check the usability of our features, we have also used the feature importance function from sklearn's library. The results of this can be seen in Table 4.

[5] See https://www.w3.org/2016/05/ontolex/#decomposition-decomp for more details.

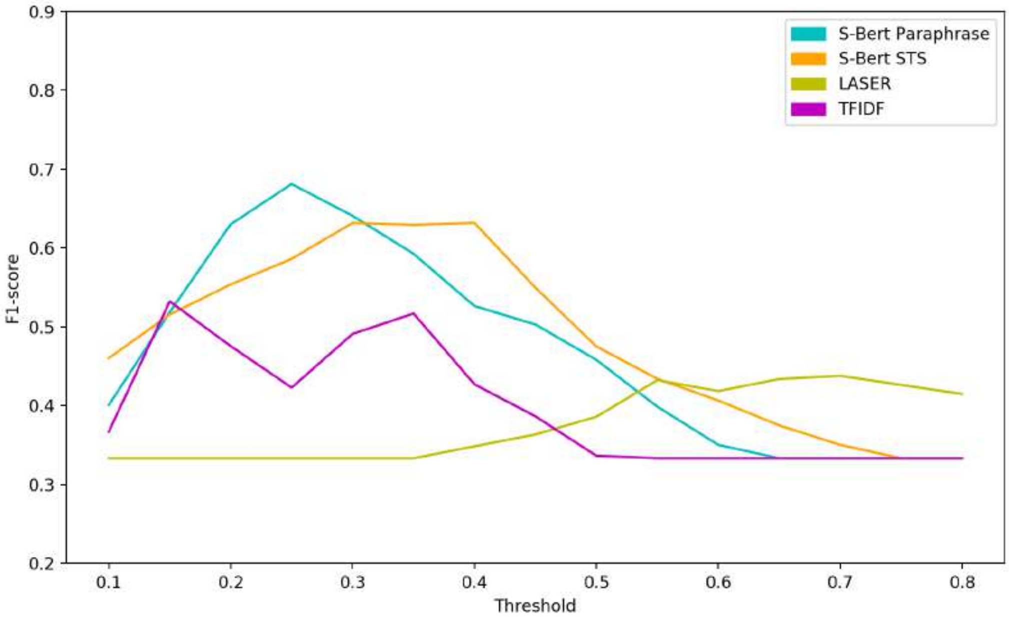


Figure 1: F1 score of different similarity metrics, depending on the threshold

Classifier	S-Bert	LASER	TFIDF	POS1	POS2
Decision Tree	0.36	0.15	0.43	0.05	0.02
Random Forest	0.39	0.23	0.28	0.06	0.04

Table 4: Relevance of different features for the classifiers

What we can see from the table is that similarity scores prove to be much more valuable predictors in comparison to the POS tags. In fact, it even looks like the POS tags can be considered irrelevant, but we have discovered that without them the results for all classifiers drop significantly (up to 10%). When it comes to similarity metrics, we see that all three of them are quite useful. Interestingly, the random forest classifier utilizes the S-Bert value most, while Decision Tree relies mostly on TFIDF. It is expected that LASER is the least useful of the three similarity metrics, due to the fact that these embeddings are multilingual, while other metrics are fine-tuned with English language in mind. Although our dataset is still quite small and the models are limited, we can assume that all three of the similarity metrics can provide valuable input to a future model.

Since we noticed that the parameters of maximum depth and the number of estimators affect the results the most, we have decided to explore them in greater length. First we use the GridSearch from sklearn to determine the best parameter set from Table 2, and then we trained several versions of the best classifier, while changing the desired two parameters. Results of this exploration can be seen in Figure 2 and Figure 3 below. We can conclude that there is no clear choice for the best value for the number of estimators nor maximum depth, but there is a distinctive trend. For maximum depth, lower values seem to perform better, which makes sense for a small dataset like ours. On the other hand, for the number of estimators very low values are not giving high performance, and neither are very high ones — the best choice are values around 200.

5. Conclusion

Since this is ongoing work, there is quite some space for future work. One of the most important things to be done next is to compile a bigger gold standard dataset, also in a multilingual setting. Another possibility to increase our dataset is to explore ways to up-sample data, or generate artificial data to increase the size of our corpus. Since the size of the data can negatively affect generalization and create difficulty in reaching the global optimum, this is an important issue when creating supervised classifiers.

A promising next step to increase the impact of our work includes handling compounds or phrasal entries in which a component is a heteronym, like for example "lead pencil". Ultimately, we hope this work will prove beneficial for handling heteronyms in text-to-speech systems as well (Henton & Naik, 2014) and (Wang et al., 2011), with the help of enriched wordnets.

6. Acknowledgements

Contributions by the German Research Center for Artificial Intelligence (DFKI GmbH) were supported in part by the H2020 project Prêt-à-LLOD with Grant Agreement number 825182. Contributions by the Austrian Centre for Digital Humanities and Cultural Heritage at the Austrian Academy of Sciences were supported in part by the H2020 project "ELEXIS" with Grant Agreement number 731015. The work described in this paper was also pursued in part in the larger context of the COST Action CA18209 — NexusLinguarum "European network for Web-centred linguistic data science". We also thank the anonymous reviewers for their insightful comments.

7. References

Ahmadi, S. & McCrae, J.P. (2021). Monolingual Word Sense Alignment as a Classification Problem. In *Proceedings of the 11th Global Wordnet Conference*. University of South Africa (UNISA): Global Wordnet Association, pp. 73–80. URL https://www.aclweb.org/anthology/2021.gwc-1.9.

Artetxe, M. & Schwenk, H. (2018). Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *CoRR*, abs/1812.10464. URL http://arxiv.org/abs/1812.10464. 1812.10464.

Bond, F. & Foster, R. (2013). Linking and Extending an Open Multilingual Wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Sofia, Bulgaria: Association for Computational Linguistics, pp. 1352–1362. URL https://www.aclweb.org/anthology/P13-1133.

Cimiano, P., McCrae, J.P. & Buitelaar, P. (2016). Lexicon Model for Ontologies: Community Report.

Declerck, T. & Bajčetić, L. (2021). Towards the Addition of Pronunciation Information to Lexical Semantic Resources. In *Proceedings of the 11th Global Wordnet Conference*. University of South Africa (UNISA): Global Wordnet Association, pp. 284–291. URL https://www.aclweb.org/anthology/2021.gwc-1.33.

Devlin, J., Chang, M., Lee, K. & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR*, abs/1810.04805. URL http://arxiv.org/abs/1810.04805. 1810.04805.

Henton, C. & Naik, D. (2014). Disambiguating heteronyms in speech synthesis. URL https://patents.google.com/patent/US9711141B2/en.

Ishida, T. (2006). Language grid: an infrastructure for intercultural collaboration.

Martin, M., Jones, G., Nelson, D. & Nelson, L. (1981). Heteronyms and polyphones: Categories of words with multiple phonemic representations. *Behavior Research Methods & Instrumentation*, 13, pp. 299–307.

McCrae, J.P., Goodman, M.W., Bond, F., Rademaker, A., Rudnicka, E. & Costa, L.M.D. (2021). The GlobalWordNet Formats: Updates for 2020. In *Proceedings of the 11th Global Wordnet Conference*. University of South Africa (UNISA): Global Wordnet Association, pp. 91–99. URL https://www.aclweb.org/anthology/2021.gwc-1.11.

McCrae, J.P., Rademaker, A., Rudnicka, E. & Bond, F. (2020). English WordNet 2020: Improving and Extending a WordNet for English using an Open-Source Methodology. In T. Declerk, I. Gonzalez-Dios & G. Rigau (eds.) *Proceedings of the LREC 2020 Workshop on Multimodal Wordnets, MMW@LREC 2020, Marseille, France, May 2020*. The European Language Resources Association (ELRA), pp. 14–19. URL https://www.aclweb.org/anthology/2020.mmw-1.3/.

Meyer, C.M. & Gurevych, I. (2011). What Psycholinguists Know About Chemistry: Aligning Wiktionary and WordNet for Increased Domain Coverage. In *Proceedings of 5th International Joint Conference on Natural Language Processing*. Chiang Mai, Thailand: Asian Federation of Natural Language Processing, pp. 883–892. URL https://www.aclweb.org/anthology/I11-1099.

Miller, G.A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, 38(11), pp. 39–41.

Reimers, N. & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *CoRR*, abs/1908.10084. URL http://arxiv.org/abs/1908.10084. 1908.10084.

Schlippe, T., Ochs, S. & Schultz, T. (2010). Wiktionary as a source for automatic pronunciation extraction. *INTERSPEECH-2010*, pp. 2290–2293.

Wang, X., Lou, X. & Li, J. (2011). Speech synthesis with fuzzy heteronym prediction using decision trees. URL https://patents.google.com/patent/US9058811B2/en.

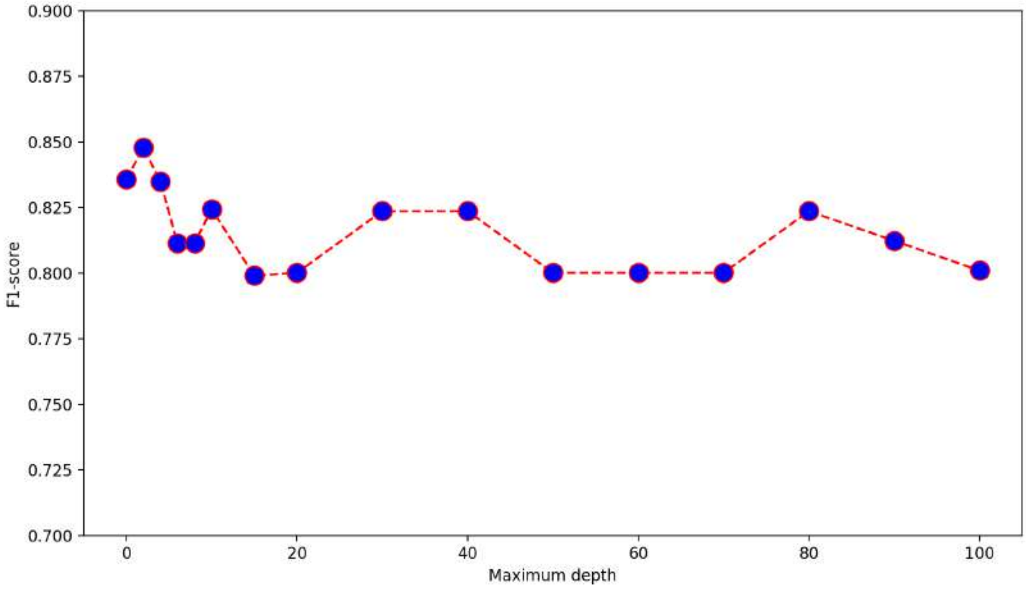


Figure 2: Random Forest Classifier F1-score depending on maximum depth

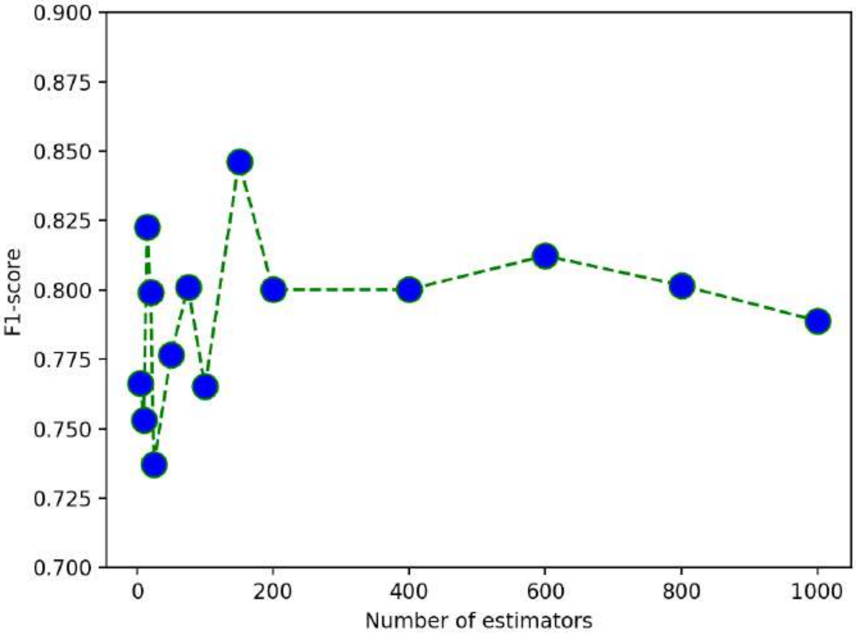


Figure 3: Random Forest Classifier accuracy depending on number of estimators

Language Monitor: tracking the use of words in contemporary Slovene

Iztok Kosem[1,2], Simon Krek[1,2], Polona Gantar[1], Špela Arhar Holdt[1], Jaka Čibej[1]

[1] Centre for Language Resources and Technologies (CJVT), University of Ljubljana
[2] Jožef Stefan Institute, Ljubljana, Slovenia
E-mail: iztok.kosem@cjvt.si, simon.krek@ijs.si, apolonija.gantar@guest.arnes.si, spela.arhar@cjvt.si, jaka.cibej@ff.uni-lj.si

Abstract

In this paper, we present Language Monitor 1.0, a new online resource for monitoring language changes in Slovene, developed at the Centre for Language Resources and Technologies at the University of Ljubljana. The resource is another part of the newly developed infrastructure for researching and describing contemporary Slovene. Language Monitor 1.0 offers four sections to observe word usage: (1) a single-word list, (2) word groups, (2) a neologism section, and (4) word comparisons. The words for a single-word list are manually validated from a list of salient word candidates, which are identified using the Simple Maths method. The paper also describes future plans, including the setup of a relational database linked with a data warehouse solution for analysis purposes, which will include various statistical information on different language phenomena relevant for researchers, lexicographers, and other users, and will provide possibilities for adding several new features to the Language Monitor.

Keywords: Language Monitor; trends; neologisms; language change; corpus

1. Introduction

One of the most challenging tasks of dictionary makers has always been ensuring that the dictionary content remains up-to-date. Modern lexicography now has all the means to address this – large corpora that can be updated on a daily basis, advanced tools for analysing the use of words over time, etc. As a result, the duration of periods between dictionary updates has decreased dramatically, from several years to months. This change has also been driven by user expectations, and the perception of dictionaries, or rather lexical resources, in modern society. The COVID-19 pandemic has exposed such a need even more – new words and word meanings have been entered into dictionaries more rapidly than ever before. It should be noted that updating the dictionary with neologisms solves only part of the problem. What about updating collocations, examples, spelling, and even definitions? It could be argued that having outdated content in a dictionary is just as problematic as lacking information on contemporary language use.

There is another element of language change that dictionary entries do not cover, namely trends in the use of existing vocabulary. Some words or their meanings, which are already established in the language, can suddenly be used much more frequently, or can be replaced by another word for a certain period. Such information can also be relevant for users, both language experts and the general public.

Another challenge brought on by monitoring language change is data modelling, as one wants to ensure that information on different language phenomena can be constantly updated, and at the same time remain compatible with databases of dictionaries and other relevant resources. Furthermore, all this information needs to be made (immediately) available to different interested parties and propagated across different resources in order to reach as many user groups as possible.

The challenges above are those faced by the Slovene lexicographic community, and probably many others, with an additional problematic factor, which is that the entire Slovene language description is in need of a significant update. This means that the language changes that need to be described may reach as far as 30 years in the past (the last general dictionary of Slovene was published in 1991[1]), and such efforts are underway. However, other solutions and methodologies have been developed to partially address this issue. These solutions include responsive dictionaries (Arhar Holdt et al., 2018), using a combination of automatic lexical data extraction and ongoing validation (e.g. Collocations Dictionary of Modern Slovene; Kosem et al., 2018[2]), and resources that focus on temporal information such as the resource presented in this paper.

In this paper, we present a new free online resource for Slovene, Language Monitor, which has been developed at the Centre for Languages and Resources at the University of Ljubljana. First, we make an overview of existing research and dictionary practices of monitoring language use. Then, we present Language Monitor 1.0, both the backend, i.e. data collection and processing procedures, and frontend, i.e. the interface. Next, we outline future plans, which include the development of a data warehouse that will be used by not only the Language Monitor, but also other resources and tools. We conclude by summarising the main points and considering potential future challenges.

2. Monitoring language use

There is a great deal of research on detecting changes in language (see e. g. Geeraerts, 2014 for an overview), with much more focus being on new words and meanings, i.e. lexical and semantic neologisms, than on changes in usage of existing meanings. Relatedly, a number of corpus-based statistical approaches and tools have been developed for neologism detection in longitudinal corpora, for example NeoCrawler

[1] There was an updated version published in 2014, but as the reviews (Ahlin et al., 2014; Krek, 2014) have pointed out, the changes introduced were not that significant.

[2] https://viri.cjvt.si/kolokacije/eng/

(Kerremans et al., 2012), NeoTrack (Janssen, 2008), ZeitGeist (Veale, 2006), Neoveille (Cartier, 2019). Similar functionality is offered by the Trends feature (Herman & Kovar, 2013) in the Sketch Engine corpus tool (Kilgarriff et al., 2004). However, the main aim of Trends is to flesh out words with significant increase or decrease in use over time.

Specifically in the area of semantic neology, a number of corpus-based techniques have been developed in the distributional semantic framework to detect semantic changes in large corpora (Sagi et al., 2011; Cook et al., 2014; Gulordava & Baroni, 2011). Such studies approach semantic neologisms from a computational perspective, while Heylen et al. (2015) present a more lexicologically oriented approach based on word space models. A similar study for Slovene was done by Fišer and Ljubešić (2016), who explored semantic shifts in Slovene tweets.

N-grams and collocations can play a pivotal role in the detection of semantic neologisms, as shown for example by projects such as AVIATOR (Renouf, 1993) and WebCorpLSE (Kehoe & Gee, 2009; Renouf, 2009). Nimb et al. (2020) used bigrams to detect new meanings of existing words in Danish for the purposes of updating the Danish dictionary. Pollak et al. (2019) conducted a similar study for Slovene, using collocations to detect new meanings in computer-mediated communication. But as Renouf (2013) points out, collocations can also help us track the life-cycle of a word, i.e. phenomena such as birth, increased use (through productivity, creativity, etc.), death, and possible revival. These aspects of collocations in Slovene have been explored in the Collocations in Slovene project (KOLOS; Kosem et al. 2020).

Translating linguistic methods into lexicographic practice, several authors have discussed the criteria of including neologisms into dictionaries (e.g. Barnhart, 1985; Metcalf, 2002; Ishikawa, 2006; O'Donovan & O'Neill, 2008; Cook, 2010; Freixa & Torner 2020). In this respect, the study by Nimb et al. (2020) is particularly valuable as it describes the decisions made and criteria used on a concrete dictionary project. What is particularly noteworthy is that Nimb et al. report (ibid. 2020: 122) that the results of their analyses lead not only to the addition of new meanings, compounds, and collocations to the dictionary, but also to the revisions of definitions and the inclusion of new usage examples.

Dictionaries use different methods and different types of data in conveying the information on language change to their users. First and foremost, announcements on newly added words and word meanings are made by dictionary publishers. The periods between these announcements have become increasingly shorter. They can now be made every few months, depending on the amount of new vocabulary that needs to be explained. During the COVID-19 pandemic, we have seen dictionaries all around the world react in an unprecedentedly rapid manner, introducing pandemic-related vocabulary within months if not weeks of the start of the pandemic.

The second approach used by dictionaries is to include the information on word usage over time directly in dictionary entries. An example of such an approach can be found in the *Digitales Wörterbuch der deutschen Sprache* (DWDS)[3] where each headword is accompanied with a line graph showing its use from 1946 (or 1600) onwards, with the frequency data coming from German corpora. This approach in principle shows the change in usage for every word (but not its individual meanings), but the information needs to compete with other more often consulted information in an entry such as definitions, collocations, etc. A different method is used by Dictionary.com where the information on trends is displayed only for words whose usage has recently increased significantly; however, this information is displayed much more prominently, on the dictionary homepage, in a manner similar to that used by stock-exchanges (Figure 1).

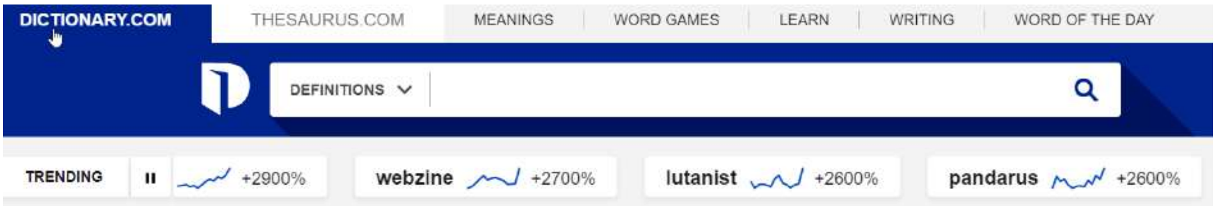


Figure 1: Trending words offered by Dictionary.com.

Some dictionaries rely on user provided information to detect language change, either indirectly or directly. An example of such practice is exhibited by *Merriam-Webster Dictionary*,[4] showing a list of the 10 most frequent words looked up by users, which is refreshed every 30 seconds.[5] While the users may not necessarily look up only words trending in frequency of use or new words (Table 1), it can be argued that many of the words from the list are probably a reaction to a current event or trending topic. As such, they not only reflect the individual's personal activities (e.g. reading), but a general topic that is relevant in a given language community at that moment.

love, infrastructure, racism, erotic, watering hole, fore, fascism, consort, hi, integrity, ambivalent, nonce, perseverance, drub, anti-sex, nexus, joke, berate, nickname, cisgender, sexi-, countenance, inclination, democracy, humility, answer, pandemic, diversity, esoteric, cognitive, autonomous, obtuse, innovation, fraud, insight, et al., pron, communism

Table 1: Words featured in the top 10 looked up by users in *Merriam-Webster Dictionary* (over a 10-minute span).

One shortcoming of the approaches mentioned so far is that they mainly promote the content already included in the dictionary. In other words, lexical or semantic

[3] https://www.dwds.de/

[4] https://www.merriam-webster.com/

[5] A similar approach is used by Oxford Dictionary at https://lexico.com, although it is not completely clear whether “Trending words (most popular in the world)” is a list of searches or corpus frequency.

neologisms that may have been detected by lexicographers but still need to be described are not included. Some dictionaries address this gap by using the crowdsourcing approach, asking users for suggestions for new words to be added to the dictionary. This approach is used by *Collins English Dictionary* in its *Word submissions* section. What is particularly commendable in the case of this particular dictionary is that the users are given publicly visible feedback on their suggestions in the form of a status note (Pending Investigation, Rejected, or Published).

As for dictionaries of Slovene, the coverage of language change has been focussed mainly on neologisms through the *Growing Dictionary of the Slovenian Language* (*Sprotni slovar slovenskega jezika*; Krvina, 2014-). Changes in the usage of existing Slovene vocabulary are much less documented, and the data has so far not been available to the general public. We decided to address this gap by developing a new resource – the Language Monitor.

3. Language Monitor 1.0

Version 1.0 of the Language Monitor (*Jezikovni sledilnik* in Slovene, or *Sledilnik* for short; https://viri.cjvt.si/sledilnik/slv/) was published in January 2021 and offers an overview of a number of salient words that have significantly impacted the language of Slovene online media in 2020 by visualising the information on temporal trends of words, i.e. the changes in their relative frequencies over a period of time. The main aim of Language Monitor in the current version is to inform users about the most prominent words in a certain period, and about new words coming into the language.

In the following subsections, we describe the data used (Section 3.1) and the process of obtaining the most salient words (Section 3.2), as well as the features of the Language Monitor 1.0 (Section 3.3).

3.1 Data

Language Monitor uses the data from the Gigafida 2.0 Reference Corpus of Written Standard Slovene (Krek et al., 2020), which covers the period between 1991 and 2018, and from the IJS NewsFeed service (Trampuš & Novak, 2012), which has been used since 2019 for daily extraction of texts from over 100 Slovene online sources, including the website of the main national television station MMC RTV Slovenija and the Slovene newspaper with the largest readership, *Delo*. The top 10 sources (by number of articles in 2020) are listed in Table 2. The output of the IJS NewsFeed service is processed through a custom pipeline that tokenises, lemmatises, morphosyntactically annotates, and segments the texts into sentences, resulting in XML files in TEI P5 format.[6]

[6] TEI P5 Guidelines - https://tei-c.org/guidelines/p5/

Our list of NewsFeed sources currently contains 102 sources. Only the sources providing at least 10 news items per year are included, but new sources or sources exceeding the minimum limit are regularly added to the list. The size of the yearly corpus from these sources was approx. 130 million tokens for 2019 and approx. 146 million tokens for 2020. Monthly subcorpora thus contain between 10 and 12 million tokens, with daily sizes ranging from 200,000 to 400,000 tokens. For reference, the yearly subcorpora from Gigafida 2.0 (1991-2018) contain an average of almost 46 million tokens, which is three times less than the yearly corpora from NewsFeed.

Source	Description	URL-domain	IJS Newsfeed articles from 2020
Slovenska tiskovna agencija (STA)	Slovenian Press Agency news portal	sta.si	101,060
MMC RTV Slovenija	National radio and television news portal	rtvslo.si	35,723
Siol.net Novice	Online news portal	siol.net	23,968
Delo	Newspaper website	delo.si	22,765
24ur.com	Commercial radio and television news portal	24ur.com	21,293
Žurnal24	Newspaper website	zurnal24.si	18,082
preberi.si	News aggregator	preberi.si	17,079
Večer	Newspaper website	vecer.com	17,054
Dnevnik	Newspaper website	dnevnik.si	15,400
Svet24	Newspaper website	novice.svet24.si	15,243

Table 2: List of sources providing most news texts in 2020.

3.2 Extraction of Salient Words

The salient words included in the Language Monitor 1.0 are obtained by comparing two corpora representing the reference period and the current period, respectively. For the most salient words of 2020, the reference corpus used was the amalgamation of Gigafida 2.0 (covering the period between 1991 and 2018) and the IJS Newsfeed output from 2019. The contemporary corpus contained the IJS Newsfeed output from 2020 (January-December).

Frequency lists of word forms[7] were extracted from both corpora using LIST (Krsnik

[7] Word forms were extracted instead of lemmas in order to prevent the merging of potential homonyms in the vein of *pot* (masculine noun, 'sweat') and *pot* (feminine noun, 'path'). Lists of word forms extracted with LIST contain lemmas and full morphosyntactic descriptions using the MTE-6 annotation schema (http://nl.ijs.si/ME/V6/msd/html/msd-sl.html), while lists of

et al. 2019), a custom-made open-source software tool for the extraction of corpus data that can be used to generate frequency lists of characters, word parts, word forms/lemmas, or word sets (n-grams). LIST supports the TEI P5 XML format and the VERT format, and outputs .TSV files.

The extracted frequency lists of word forms were then converted to frequency lists of lemmas (keeping the relevant discriminatory information such as gender for nouns and aspect for verbs). Next, the entries from both frequency lists were compared in terms of their relative frequencies using the Simple Maths formula (Kilgarriff, 2009), where fr1 is the relative frequency of a word in the reference corpus, fr2 is the relative frequency of the word in the contemporary corpus, and N is the smoothing parameter (in case the word is not found in the contemporary corpus and fr2 equals zero; the smoothing parameter was set to 1 in our case):

$$\text{sm} = (\text{f}_{\text{r2}} + \text{N}) \;/\; (\text{f}_{\text{r1}} + \text{N})$$

Table 3 shows the top 10 most salient words of 2020, along with their MTE-6 lexical features, absolute and relative frequencies, and Simple Maths scores.

Lemma	MTE-6 Lexical Features	f_a (1991-2019)	f_a (2020)	f_r (1991-2019)	f_r (2020)	Simple Maths Score
koronavirus	Som	175	214,947	0.120	1463.444	1307.997
covid	Som	0	90,054	0	613.123	614.123
pandemija	Soz	1,668	76,873	1.140	523.382	245.034
covid	Kag	0	22,870	0	155.708	156.708
karantena	Soz	2,852	48,976	1.949	333.448	113.400
epidemija	Soz	11,028	118,082	7.537	803.949	94.285
protikoronski	Pp	0	11,880	0	80.884	81.884
koronavirusen	Pp	1	10,148	0.000683	69.092	70.044
epidemiološki	Pp	1,771	21,253	1.210	144.700	65.914
Covid	Slm	0	9,419	0	64.128	65.128

Table 3: The top 10 most salient words of 2020 compared to 1991-2019.

The list of most salient words of 2020 contains neologisms (*covid-19*, *protikoronski* 'anti-corona (adjective)') as well as existing words with a significant increase in usage during 2020 (*epidemiološki* 'epidemiological', *karantena* 'quarantine', *pandemija* ('pandemic', noun), *koronavirusen* 'adjective; related to coronavirus'), *epidemija* 'epidemic'). However, the list also contains a number of problems caused by errors in automatic lemmatisation and morphosyntactic tagging. For instance, 'covid' is lemmatized as both

lemmas contain only parts-of-speech, which would merge the frequencies for *pot* (masculine) and *pot* (feminine).

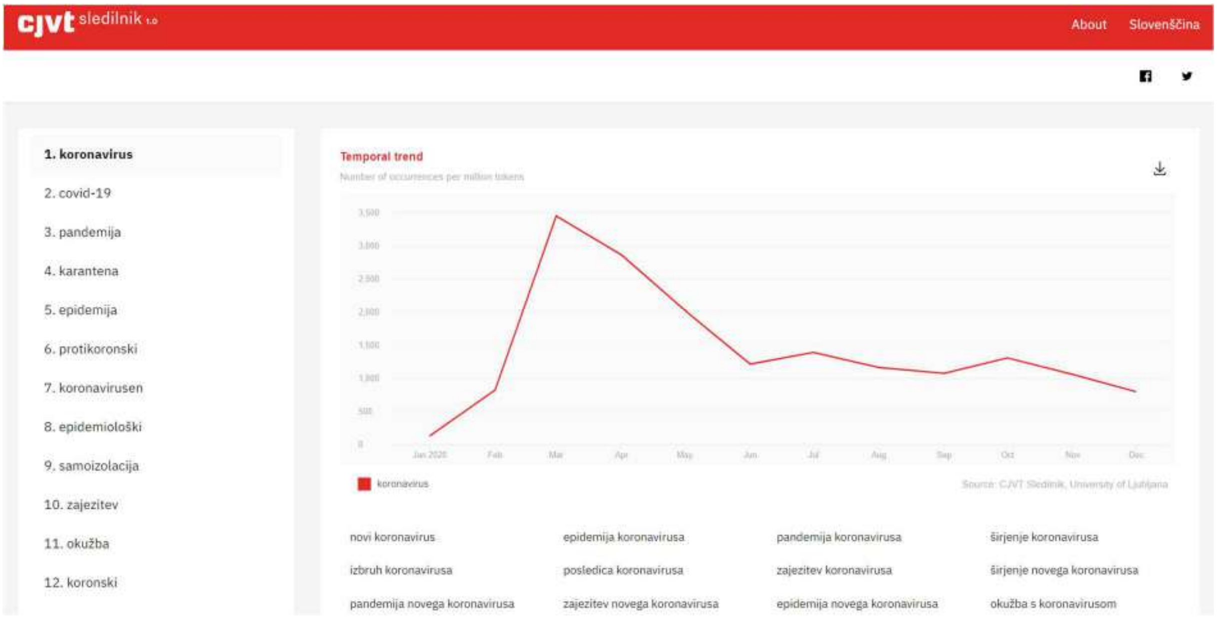
'covid' and 'Covid' and tagged as a common noun (Som), a proper noun (Slm) or even as a numeral (Kag). There is also the problem of the overlap with n-grams: 'covid' mostly often occurs as 'covid-19', which is treated as a 3-gram by our tokeniser ('covid', '-', '19'). We have amended this during manual analysis (changing *covid* to *covid-19*), as version 1.0 of the Language Monitor only focuses on single words. N-grams will be treated in future versions (as described in Section 4).

The obtained lists of salient words were manually analysed to remove noise. The relevant words were then included in the Language Monitor 1.0 database along with their frequencies.

3.3 Features

The Language Monitor 1.0 offers four sections to observe word usage: (1) a single-word list, (2) word groups, (2) a neologism section, and (4) word comparisons.

The first option (shown in Figure 2) features a list of 100 words that have been identified as the most salient in 2020 compared to the period between 1991 and 2019. The user can click on a word in the list and is provided with a line graph showing the trend of the word's relative frequency between January 2020 and December 2020. Figure 2 shows the temporal trend of the word *koronavirus* ('coronavirus'), the most salient word of 2020. The line graph shows a steep increase of usage between February and March 2020, when an epidemic was officially declared in Slovenia. After the initial surge, the usage of *koronavirus* stabilises and remains relatively unchanged in the period between June and December 2020.



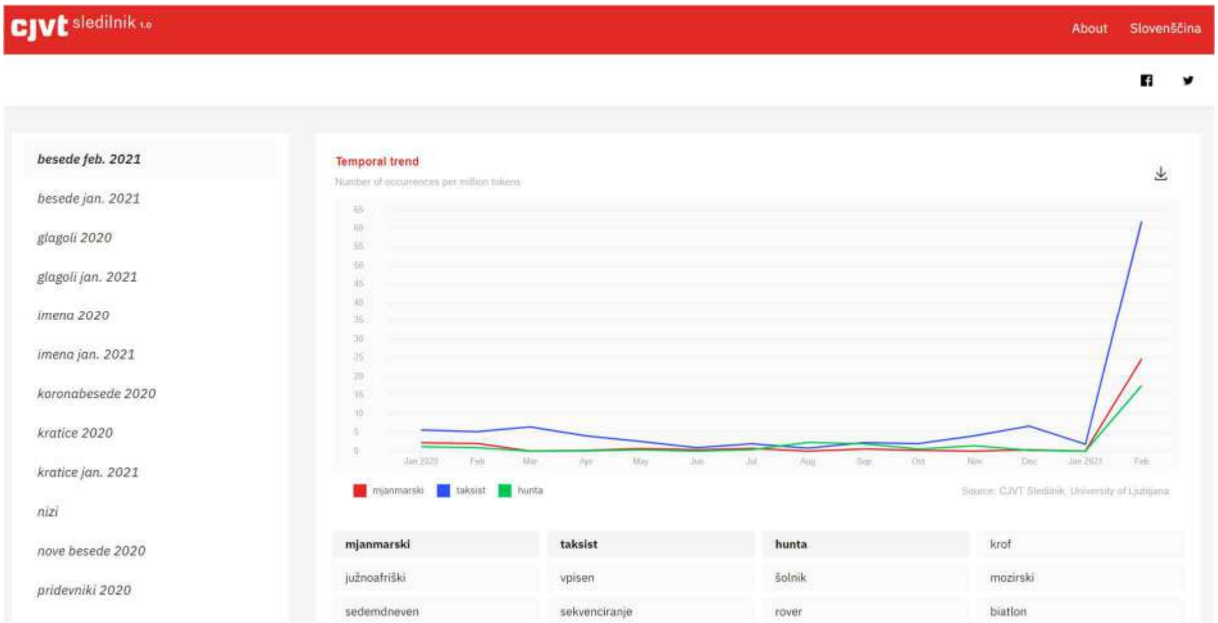
1. koronavirus
2. covid-19
3. pandemija
4. karantena
5. epidemija
6. protikoronski
7. koronavirusen
8. epidemiološki
9. samoizolacija
10. zajezitev
11. okužba
12. koronski

novi koronavirus	epidemija koronavirusa	pandemija koronavirusa	širjenje koronavirusa
izbruh koronavirusa	posledica koronavirusa	zajezitev koronavirusa	širjenje novega koronavirusa
pandemija novega koronavirusa	zajezitev novega koronavirusa	epidemija novega koronavirusa	okužba s koronavirusom

Figure 2: Line graph of the temporal trend for *koronavirus*.

Below the line graph, the most frequent n-grams featuring the word in question are listed. In this case, they contain expressions such as *novi koronavirus* (‘novel coronavirus’), *izbruh koronavirusa* (‘coronavirus outbreak’), *posledica koronavirusa* (‘consequence of coronavirus’), *širjenje novega koronavirusa* (‘spread of the novel coronavirus’), and so on.

The second section features temporal trends of word groups, i.e. groups of words that share a certain characteristic. At the end of March 2021, a total of 13 groups were available, for instance *Neologisms - February 2021* (containing salient words that first appeared in February 2021), *Words - February 2021* and *Words - January 2021* (salient words from January and February 2021, respectively), *Proper Nouns - January 2021* (prominent proper nouns from January 2021), and *Verbs - 2020* (salient verbs from 2020). Figure 3 shows the visualisation for *Words - February 2021* and features the list of available word groups on the left (the currently viewed word group is set in bold), a line graph with temporal trends of one or more salient words on the right (the first three are shown in the line graph by default; up to six can be visualised), and a clickable list of salient words below the graph. By selecting or unselecting words, the user can modify the line graph to visualise the relevant words. By clicking on the Download icon in the upper right corner of the line graph, the user can also export the line graph in .PNG format for further use.



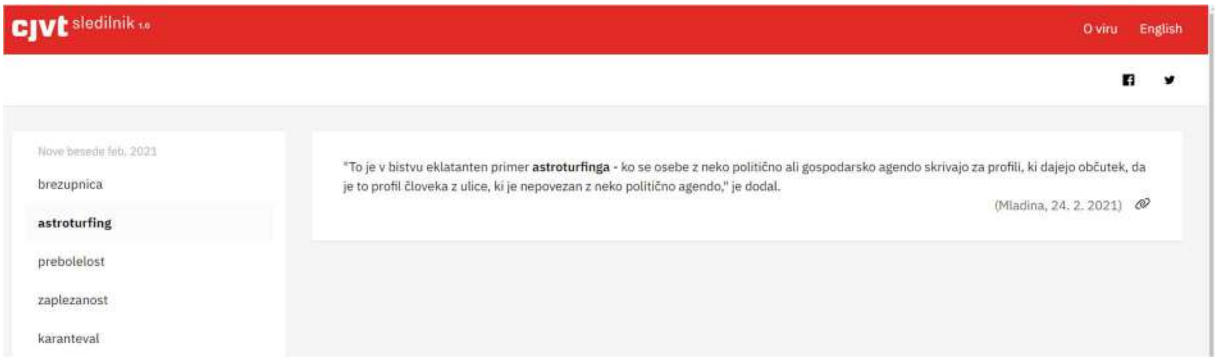
besede feb. 2021
besede jan. 2021
glagoli 2020
glagoli jan. 2021
imena 2020
imena jan. 2021
koronabesede 2020
kratice 2020
kratice jan. 2021
nizi
nove besede 2020
pridevniki 2020

mjanmarski	**taksist**	**hunta**	krof
južnoafriški	vpisen	šolnik	mozirski
sedemdneven	sekvenciranje	rover	biatlon

Figure 3: Line graph for the *Words - February 2021* word group.

The most salient words from February 2021 reflect most of the major events (both local and global) reported by Slovene media in that month, such as the coup d'etat in Myanmar (*mjanmarski* 'adjective, related to Myanmar', *hunta* 'junta'), seasonal holidays (*pusten* 'adjective, related to Mardi Gras', *krof* 'doughnut', *valentinovo* 'Valentine's Day'), the ongoing coronavirus epidemic (*južnoafriški* 'South African', *sekvenciranje* 'sequencing', *cepljen* 'vaccinated'), political turmoil in the Slovene parliament (*nezaupnica* 'vote of no confidence'), sexual harassment revelations in Slovene society and subsequent changes to Slovene legislation regarding sexual violence (*nadlegovanje* 'harassment', *redefinicija* 'redefinition'), and NASA's rover mission to Mars (*rover* 'rover').

The third option is the neologism section, a special word group section which features salient words that are found in the compared corpus but have never appeared in the reference corpus. Shown in Figure 4 is the February 2021 neologism section, which features, for example *karanteval* (a lockdown version of a Mardi Gras parade; a portmanteau of *karantena* 'quarantine' and *karneval* 'carnival') and *astroturfing* (an English loanword which experienced a surge in use after a Slovene politician accidentally revealed their use of a fake Twitter profile to attack political opponents). Each neologism also features a sentence exemplifying its use, along with a link to the original article, its source and date of publication. In version 1.0, no line graph is provided for neologisms since the word has just entered language use and no trends are yet available.



Nove besede feb. 2021

brezupnica

astroturfing

prebolelost

zaplezanost

karanteval

"To je v bistvu eklatanten primer **astroturfinga** - ko se osebe z neko politično ali gospodarsko agendo skrivajo za profili, ki dajejo občutek, da je to profil človeka z ulice, ki je nepovezan z neko politično agendo," je dodal.

(Mladina, 24. 2. 2021)

Figure 4: The neologism section (February 2021) of the Language Monitor 1.0.

The last section offers trend comparisons between words with data available in the Language Monitor 1.0. A total of 184 salient words were available for comparison by the end of March. The user can either select one of the preset comparisons (which have been prepared in advance) or generate a custom comparison by selecting up to six words from the list of available words (similar to the word group comparison, but this section allows for comparisons among all available words, not just within the relevant group). Figure 5 shows a preset comparison of the words *samoizolacija/samoosamitev* (both meaning 'self-isolation') and *izolacija/osamitev* ('isolation'). The trends show that the words *samoizolacija* (red) and *izolacija* (yellow) are both more frequent than their counterparts *samoosamitev* (blue) and *osamitev* (green).

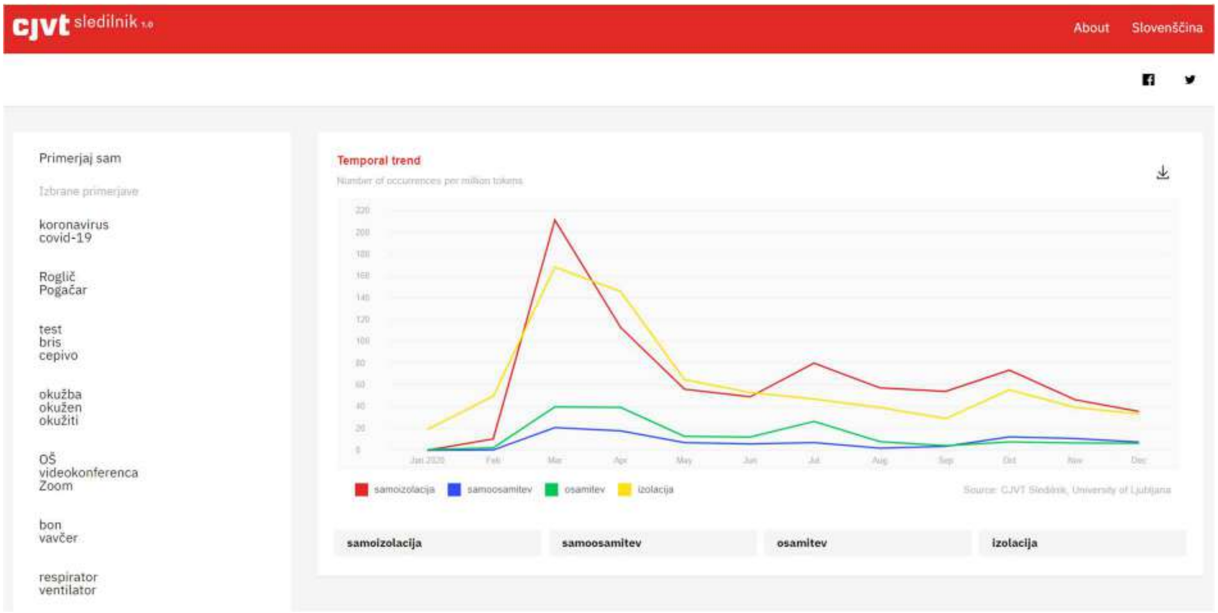


Figure 5: Trend comparisons in the Language Monitor 1.0.

4. Conclusions and future plans

The Language Monitor is a new addition to the infrastructure for contemporary Slovene, a resource that has made first strides towards consistent and constant monitoring of language change. Version 1.0 has focussed on presenting this information to the general public, using word lists in combination with different visual (line graphs) and interactive methods such as word groups and comparisons.

It was clear to us from the very beginning that the current methods of updating the Language Monitor were not sustainable nor desirable long-term, especially in view of the needs and wishes of researchers, lexicographers, and users. Considering the progress made in the area of lexical data extraction from Slovene corpora (e.g. Gantar et al., 2016) and the ongoing development of the Digital Dictionary Database for Slovenian (Klemenc et al., 2017; Kosem et al., forthcoming), which will consolidate different monolingual and bilingual lexical resources for Slovene, it is our aim to integrate the Language Monitor into this infrastructure.

Consequently, we have started preparing a pipeline that will extract various statistical information (e.g. raw frequency, number of different texts, source) on lemmas, collocations, multiword lexical units, etc., along with links to corpus examples, on a daily basis. In order to ensure data compatibility, the Gigafida 2.0 reference corpus for the years up to 2018 will need to be reprocessed with the same pipeline, using the latest versions of tools for morphosyntactic tagging, parsing and other annotation layers. This was not done for the Language Monitor 1.0, and we have already observed a number of issues caused by differences in lemmatisation and morphosyntactic tagging during our manual analyses.

All the data extracted from the text using our pipeline will be fed into a relational database, which will store various information on different language phenomena in Slovene. Importantly, the database will hold the information on data from different types of corpora from different periods. Then, using a data warehouse solution, the information in the database will be analysed using different statistical methods (including Simple Maths, various association measures for collocations, etc.) and the results made available to lexicographers working on various lexical resources. Many of these calculations are already offered by corpus tools. However, lexicographers often need to take additional calculation steps during concordance analysis in order to obtain such information, and then make decisions based on it. It is our intention to use the data warehouse solution to provide lexicographers with alerts about significant changes in the usage of lexical items over time, or about important usage patterns in general (e.g. text type dispersion).

On the other hand, the database will directly feed the resources aimed at the general public, particularly the Language Monitor, which will offer users the possibility to not only observe but also explore the usage of words and collocations over time. Specifically, the ideas for the Language Monitor currently in preparation include implementing three methodologies: automatic extraction, manual analysis by linguists/lexicographers, and user involvement (crowdsourcing). In this manner, experts and users will work together in shaping the Language Monitor, and by feeding the results back into the database, their work will be of benefit to lexicographers and researchers.

5. Acknowledgements

The authors acknowledge the financial support from the Slovenian Research Agency (research core funding No. P6-0411, *Language Resources and Technologies for Slovene*) and that the projects *Collocation as a basis for language description: semantic and temporal perspectives* (J6-8255) and *New grammar of modern standard Slovene: resources and methods* (J6-8256) were financially supported by the Slovenian Research Agency.

6. References

Ahlin, M., Lazar, B., Praznik, Z. & Snoj, J. (2014). Slovar slovenskega knjižnega jezika. Druga, dopolnjena in deloma prenovljena izdaja. Izdali Slovenska akademija znanosti in umetnosti, Znanstvenoraziskovalni inštitut Slovenske akademije znanosti in umetnosti, Inštitut za slovenski jezik Frana Ramovša. *Jezik in slovstvo*, 59 (4), pp. 121–127.

Arhar Holdt, Š., Čibej, J., Dobrovoljc, K., Gantar, A., Gorjanc, V., Klemenc, B., Kosem, I., Krek, S., Laskowski, C., Robnik Šikonja, M. (2018). Thesaurus of Modern Slovene: By the Community for the Community. In: J. Čibej, V. Gorjanc,

I. Kosem, S. Krek (eds.): Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts. ISBN 978-961-06-0097-8). Ljubljana: Znanstvena založba Filozofske fakultete. 2018, pp. 401-410. https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/118/211/3000-1

Barnhart, D.K. (1985). Prizes and pitfalls of computerized searching for new words for dictionaries. *Dictionaries* 7, pp. 253-260.

Cartier, E. (2019). Neoveille, web platform for finding and monitoring neologisms in monitor corpora. *Neologica*, 13, pp. 23–54.

Cook, P. (2010). *Exploiting Linguistic Knowledge to Infer Properties of Neologisms.* PhD Dissertation. Toronto: University of Toronto.

Cook, P., Rundell, M., Lau, J. H. & Baldwin, T. (2014). Applying a word-sense induction system to the automatic extraction of dictionary examples. In A. Abel et al. (eds.) *Proceedings of the XVI EURALEX International Congress. Bolzano, Italy: EURAC*, pp. 319–328.

Fišer, D. & Ljubešić, N. (2016). Detecting semantic shifts in Slovene Twitterese. In A. Horák, P. Rychlý & A. Rambousek (eds.) *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2016*, pp. 1–8.

Freixa, J. & Torner, S. (2020). Beyond frequency: On the dictionarisation of new words in Spanish. *Dictionaries* 41(1), pp. 131-154.

Gantar, P., Kosem, I., & Krek, S. (2016). Discovering Automated Lexicography: The Case of the Slovene Lexical Database. *International Journal of Lexicography*, 29(2), pp. 200–225.

Geeraerts, D. (2014). How words and vocabularies change. In J. Taylor (ed.) *The Oxford Handbook of the Word.*

Gulordava, K. & Baroni, M. (2011). A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. In *Proceedings of the GEMS 2011 Workshop on Geometrical Models of Natural Language Semantics*, pp. 67–71.

Herman, O. & Kovář, V. (2013). Methods for Detection of Word Usage over Time. In *Seventh Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 201*3. Brno: Tribun EU, pp. 79–85.

Heylen, K., Wielfaert, T., Speelman, D. & Geeraerts, D. (2015). Monitoring polysemy: Word space models as a tool for large-scale lexical semantic analysis. *Lingua*, 157, pp. 153-72.

Ishikawa, S. (2006). When a word enters the dictionary: A data-based analysis of neologism. In *JACET Society of English Lexicography, English Lexicography in Japan.* Bunkyo-ku: Taishukan, pp. 39-52.

Janssen, M. (2008). NeoTrack: Un analyseur de néologismes en ligne. In M.T. Cabré, O. Domènech, R. Estopà & J. Freixa (eds.) *Proceedings of CINEO 2008*, pp. 1175-1188.

Kehoe, A. & Gee, M. (2009). Weaving Web data into a diachronic corpus patchwork. In A. Renouf & A. Kehoe (eds.) Corpus Linguistics: Refinements and Reassessments. Leiden: Brill, pp. 255-279.

Kerremans, D., Stegmayr, S., & Schmid, H. J. (2011). The NeoCrawler: identifying and retrieving neologisms from the internet and monitoring ongoing change. In Allan & Robinson (eds) *Current Methods in Historical Semantics*, 73, pp. 59.

Kilgarriff, A. (2009). Simple maths for keywords. In M. Mahlberg, V. González-Díaz & C. Smith (eds.), *Proceedings of Corpus Linguistics Conference CL2009, University of Liverpool, UK, July 2009.* https://www.sketchengine.eu/wp-content/uploads/2015/04/2009-Simple-maths-for-keywords.pdf.

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, EURALEX 2004 Lorient, France July 6-10, 2004.* Lorient: Universite de Bretagne – sud, pp. 105–116.

Klemenc, B., Robnik-Šikonja, M., Fürst, L., Bohak, C. & Krek, S. (2017). Technological Design of a State-of-the-art Digital Dictionary. In V. Gorjanc, P. Gantar, I. Kosem & S. Krek (eds) *Dictionary of modern Slovene: problems and solutions. 1st ed.* Ljubljana: Ljubljana University Press, Faculty of Arts, pp. 10-22.

Kosem I., Krek, S. & Gantar, P. (forthcoming). Semantic data should no longer exist in isolation: the Digital Dictionary Database of Slovenian. *Proceedings of EURALEX 2020, Volume II.*

Kosem, I, Krek, S., Gantar, P., Arhar Holdt, Š., Čibej, J. & Laskowski, C A. (2018). Collocations dictionary of modern Slovene. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (eds.) *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts.* Ljubljana: Znanstvena založba Filozofske fakultete, pp. 989-997. https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/118/211/3000-1

Kosem, I., Krek, S., Čibej, J., Gantar, P., Arhar Holdt, Š., Logar, N., Laskowski, C. A., Klemenc, B., Ljubešić N., Dobrovoljc, K., Gorjanc, V. & Pori, E. (2020). *The Orange workflow for observing collocation clusters ColEmbed 1.0*, Slovenian language resource repository CLARIN.SI. https://www.clarin.si/repository/xmlui/handle/11356/142.

Krek, S. (2014). Prva in druga izdaja SSKJ. *Slovenščina 2.0,* 2(2), pp. 114–160. Accessed on 11 April 2021. https://doi.org/10.4312/slo2.0.2014.2.114-160.

Krek, S., Arhar Holdt, Š., Erjavec, T., Čibej, J., Repar, A., Gantar, P., Ljubešić, N., Kosem, I. & Dobrovoljc, K. (2020). Gigafida 2.0: The Reference Corpus of Written Standard Slovene. *Proceedings of the 12th Language Resources and Evaluation Conference" European Language Resources Association"*, pp. 3340-3345. https://www.aclweb.org/anthology/2020.lrec-1.409.

Krsnik, Luka; et al., (2019). *Corpus extraction tool LIST 1.2.* Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1276.

Metcalf, A. (2002). *Predicting New Words.* Boston, MA: Houghton Mifflin Company.

Nimb, S., Sørensen, N. H. & Lorentzen H. (2020). Updating the dictionary: Semantic change identification based on change in bigrams over time. *Slovenščina 2.0: Empirical, Applied and Interdisciplinary Research*, 8(2), pp. 112-138.

https://doi.org/10.4312/slo2.0.2020.2.112-138.

O'Donovan, R. & O'Neill, M. (2008). A systematic approach to the selection of neologisms for inclusion in a large monolingual dictionary. In E. Bernal & J. DeCesaris (eds.) *Proceedings of the XIII EURALEX International Congress* (Barcelona, 15-19 July 2008). Barcelona: IULA-UPF, pp. 571-579.

Pollak, S., Gantar, P. & Arhar Holdt, Š. (2019). What's New on the Internetz? Extraction and Lexical Categorisation of Collocations in Computer-Mediated Slovene, *International Journal of Lexicography*, 32 (2), pp. 184–206, https://doi.org/10.1093/ijl/ecy026.

Renouf, A. (1993), A Word in Time: first findings from dynamic corpus investigation. In J. Aarts, P. de Haan, & N. Oostdijk (eds.) *English Language Corpora: Design, Analysis and Exploitation.* Amsterdam: Rodopi, pp. 279-288.

Renouf, A. (2009). Corpus Linguistics beyond Google: the WebCorp Linguist's Search Engine. In R. Siemens & G. Shawver (eds.) *New Paths for Computing Humanists*, in Digital Studies / Le champ numérique Vol 1, No 1, the Society for Digital Humanities / Société pour l'étude des médias interactifs (SDH/SEMI).

Renouf, A. (2013). A finer definition of neology in English: The life-cycle of a word. In H. Hasselgård, J. Ebeling & S. Oksefjell Ebeling (eds.) *Corpus Perspectives on Patterns of Lexis* (Studies in Corpus Linguistics, 57), pp. 177-208.

Sagi, E., Kaufmann, S. & Clark, B. (2011). Tracing semantic change with latent semantic analysis. In K. Allan & J. A. Robinson (eds*.) Current Methods in Historical Semantics*. De Gruyter Mouton, Berlin, Germany.

Trampuš, M. & Novak, B. (2012). The Internals Of An Aggregated Web News Feed. *Proceedings of 15th Multiconference on Information Society 2012 (IS-2012).* http://ailab.ijs.si/dunja/SiKDD2012/Papers/Trampus_Newsfeed.pdf

Veale, T. (2006). Tracking the Lexical Zeitgeist with Wikipedia and WordNet. In *Proceedings of ECAI'2006, the 17th European Conference on Artificial Intelligence.*

LeXmart: A platform designed with lexicographical data in mind

Alberto Simões[1], Ana Salgado[2,3], Rute Costa[3]

[1]2Ai – School of Technology, IPCA, Barcelos, Portugal
[2] Academia das Ciências de Lisboa, Lisboa, Portugal
[3] NOVA CLUNL, Centro de Linguística da Universidade NOVA de Lisboa, Lisboa, Portugal
E-mail: asimoes@ipca.pt, anacastrosalgado@gmail.com, rute.costa@fcsh.unl.pt

Abstract

LeXmart is an open-source web platform used to support the lexicographer's work through editing, control, validation, management, and publication of lexical resources. This tool was specifically developed to facilitate the compilation of general monolingual dictionaries in which data is encoded according to the Text Encoding Initiative (TEI) schema (chapter 9). Here, we will describe the challenges of adapting LeXmart to deal with TEI Lex-0 and distinct types of lexical resources, namely *Dicionário da Língua Portuguesa* (DLP) and *Vocabulário Ortográfico da Língua Portuguesa*, lexicographic works from Academia das Ciências Lisboa, and *Dicionário Aberto*, the retro-digitised version of the Cândido de Figueiredo dictionary. This article describes the steps taken to update the LeXmart platform to deal with the TEI Lex-0 schema and describe the challenges on properly encoding these three projects while allowing the lexicographical team to work continuously. This work builds on automatic operations performed on top of the original resources. It also includes the changes made to the editor to make it capable of dealing with the encoding updates and the new types of resources.

Keywords: dictionary editing system; e-lexicography; online dictionary; TEI Lex-0

1. Introduction

Compiling a dictionary is both a challenging and time-consuming task. For centuries, data collection and compilation of lexicographic data had been done on pen and paper, making the lexicographic work as a Herculean task. Nowadays, there are various computerised tools that can support the writing of dictionaries.

Since the beginning of computer-aided dictionary editing, publishers and some academic institutions have developed their own software to create dictionaries for commercial purposes. The first-generation dedicated dictionary writing systems were developed in the previous century in order to make life easier on the entry-writing front (Rundell & Kilgarriff, 2011). On the other hand, as secrecy is always the lifeblood of a business, these systems were not shared with third parties, which nowadays has a significant impact on issues concerning interoperability between different lexicographic resources.

The irreversible transition to a digital environment in the past two decades has imposed new challenges on lexicography in terms of adopting new methods, mainly due to technological advances, the fall of many publishers, and the changes introduced in their business models (Rundell, 2010). Nevertheless, independent software continues to be developed to assist lexicographers from different institutions. LeXmart[1] is one of these tools, designed from scratch to support an existing dictionary in an era where there is no great commercial interest in a dictionary distributed in physical mode, i.e. a printed version. Our main concern is to develop a lexicographic tool that responds to heterogeneous lexicographic structures and ensures that the structural components of a lexicographic article, known for their extreme complexity, are well identified and represented in a well-defined hierarchical organisation and appropriate metalanguage.

[1] Available at http://lexmart.eu/.

In the next section, we briefly discuss the LeXmart tool. Then, in Section 3, we describe the lexicographic resources that are currently under development using LeXmart. These resources are then analysed in Section 4 in terms of their structure and encoding details, using the TEI Lex-0 specifications. This is followed by Section 5, where we describe how LeXmart is being designed to help produce valid TEI Lex-0 documents keeping, at the same time, the interface as simple as possible for the lexicographer. Finally, in Section 6, we conclude the paper by presenting some insights into the project's status and propose several promising future research areas.

2. LeXmart

LeXmart is an open-source web platform used to support the lexicographer's work. It aims to support the activities involved in the whole lifecycle of preparing a dictionary, including editing the lexicographic articles, controlling, validating, and managing the dictionary and its content Simões et al. (2019).

LeXmart was developed using a bottom-up approach to solve a specific problem: storing and allowing the editing and quality management processes of the *Dicionário da Língua Portuguesa Contemporânea* (DLPC) (ACL, 2001). Further details on this project will be elaborated in Section 3.1.

This bottom-up approach means that, instead of creating a dictionary editing system from scratch (thereby restricting how dictionaries are defined), a basic version is first built and then further refined according to users' data management needs. In this way, the LeXmart tool was shaped as per the lexical data, rather than requiring the data to fit the tool.

Despite the evident benefits of the bottom-up approach, LeXmart was clearly created as a biased tool used to deal with a single lexical resource. DLPC was encoded following version 5 of the TEI Guidelines for Electronic Text Encoding and Interchange[2] while including some adaptations to match the TEI standard to preserve the original structure of the dictionary. Although LeXmart responds to the editing needs of the DLPC, the flexibility of the scheme was very restricted as it was designed specifically for only this dictionary. This limitation severely limited the advantages of using LeXmart in other lexicographic resources, which, in a way, are characterised by high structural heterogeneity. Meanwhile, the LeXmart platform has been heavily used to edit DLP and make it robust enough to deal with an actual-sized dictionary. Therefore, the team associated with it has an interest in using the tool to edit and maintain other lexical resources, namely:

- The *Dicionário Aberto* (DA) (Simões & Farinha, 2011), a transcription of a 1913 dictionary in the Portuguese language that was encoded using a custom TEI schema.
- *The Vocabulário Ortográfico da Língua Portuguesa* (VOLP-1940) (ACL, 1940), published by *Academia das Ciências Lisboa* (ACL) in 1940, which is currently being encoded using the TEI Lex-0.

These resources have different structures and have been encoded using different schemas. We cannot maintain LeXmart with a specific bias for each resource it includes. Therefore,

[2] Available at https://tei-c.org/guidelines/.

the current task in progress is rewriting LeXmart to focus on a specific and strict schema that can adequately encode all projects currently under development, following TEI Lex-0 specifications. Nevertheless, this target requires that the current LeXmart database's lexical resources be properly transformed and encoded into TEI Lex-0. Therefore, this article focuses not only on the tool and its changes, but also on the original dictionary's encoding process and the newly added resources.

3. Lexicographic Resources

This section presents the three resources for which LeXmart is used. For each one, we share some insights into their origins and the goals of including each of those resources into LeXmart.

3.1 Dicionário da Língua Portuguesa

The *Dicionário da Língua Portuguesa* (DLP) (ACL, 2021) is a scholarly dictionary of the Portuguese language being developed by the ACL. DLP aims at being the first digital academic Portuguese dictionary. The main objective of this endeavour is to update the DLPC 2001 edition by presenting an entirely new lexicographic resource. The database will be available online for free, and currently there are no plans to publish a printed version of the dictionary. It is a monolingual dictionary that is descriptive in nature, but with normative indications, as can be expected from a dictionary prepared by an academy of sciences. It is based on a retro-digitised dictionary created by converting the DLPC, described in the previous section, that was last published in the year 2001. This retro-digitisation process was previously described by Simões et al. (2016).

The result of this retro-digitisation was the creation of a database with over 68,000 entries, each of them stored independently in an XML file. These entries did not follow the guidelines in the chapter 9 of the TEI, on dictionaries. It was designed in accordance with the metainformation that was possible to extract from a PDF file, which was the only information source. This resulted in well-formed XML files, which included non-standard XML elements and attributes. Some examples of newly added elements are the `group` tag for enclosing a set of senses with the same morphological information and the `syn` and `ant` elements for encoding the lists of synonyms and antonyms. Similarly, custom attributes were also added. One of them is the `@fem` attribute, added to the `orth` tag, that registers the feminine suffixes for the lemmas.

To guarantee interoperability, the DLP is being transformed to ensure its compliance with the TEI Lex-0 format (Salgado et al., 2019b), a streamlined version of the TEI dictionary chapter. This decision is also behind the adaptation of LeXmart to follow this specific schema. Section 4.1 elucidates this conversion process.

3.2 Dicionário Aberto

The *Dicionario Aberto* (DA) (Simões & Farinha, 2011) is a Portuguese-language dictionary obtained by the OCR and a fully manual validation of the *Nôvo Diccionário da Língua Portuguêsa*, authored by Cândido de Figueiredo in 1913. This retro-digitisation process was done in close cooperation with the Distributed Proofreaders project of the Project Gutenberg[3]. The transcription took nearly four years to complete, and in 2010

[3] Available at https://www.pgdp.net/c/.

its full version was made publicly available on the Project Gutenberg website. The DA contains 128,521 entries: almost twice the number of entries in the DLP. This significant difference is explained by the DA registering orthographic variants of the same entry, as its original dictionary was published in troubled times for Portuguese language orthography.

This transcription was performed by volunteers with no lexicographic background. Thus, they were asked to encode the dictionary following quite a simple set of rules, which are used across all transcriptions performed in the Distributed Proofreaders website: each line in the original document should be presented independently (only hyphenated words were glued to the end of the top line), and bold and italics should be encoded using a custom markup, surrounding words by one asterisk character to encode bold words and one underscore character to encode italic words.

This simple markup was then converted to a custom TEI schema. The details on this encoding are in Section 4.2, where we discuss the process of transforming this original encoding into TEI Lex-0.

For years, DA has been subject to different transformations. The most relevant was the automatic orthography update, which allowed the dictionary to be used for experiments in natural language processing tasks, such as the automatic extraction of information to create Wordnets and ontologies (Gonçalo Oliveira, 2018; Gonçalo Oliveira & Gomes, 2014).

In the future, DA will be included in another broader project that aims to encode different dictionaries currently in the public domain into a single, more comprehensive resource.

3.3 Vocabulário Ortográfico da Língua Portuguesa

The *Vocabulário Ortográfico da Língua Portuguesa* [Orthographic Vocabulary of the Portuguese Language] (VOLP-1940) is the first orthographic vocabulary published by the ACL, in 1940. The *Digital Edition of the VOLP-1940*[4] (Salgado & Costa, 2020) aims at the digitisation of all the vocabularies of the ACL. The goal is to analyse the vocabularies with computational methods to better assess the importance of this work for the evolution of the Portuguese language in the 20th century and to contribute to the current movement of creating innovative, data-driven computational methods for text digitisation, encoding, and analysis. VOLP-1940's digitisation aims to create a lexicographical resource encoded in TEI, with structured information in the Simple Knowledge Organisation System (SKOS), to guarantee its future connection to other systems and resources, particularly in the Portuguese-speaking world.

The digitisation of the VOLP-1940 resulted in a series of image files of the original PDF manuscript that were converted to plain text using a commercial character recognition program (OCR) — the *Omnipage Pro*. The text was later exported to an editing program — *Microsoft Word* — to correct typos and inconsistencies generated by OCR.

Identifying the VOLP-1940 lexicographic conventions (for example, the comma used after each lemma or the use of abbreviations listed on the initial pages of the paperwork)

[4] Further details of the project at https://clunl.fcsh.unl.pt/en/investigacao/projetos-curso/edicao-digital-do-vocabulario-ortografico-da-lingua-portuguesa-volp-1940/ and at https://www.volp-acl.pt/index.php/vocabulario-1940/projeto.

was carried out to experiment a possible automated annotation of the entire work. Using *Microsoft Word* styles, we identified the different VOLP lexicographic article components, such as grammatical information, geographic information, etc.

4. TEI Lex-0 Encoding

LeXmart is being adapted to support the TEI Lex-0 standard properly. Although it would be interesting to have the tool dealing with different encoding formats, we are only targeting TEI Lex-0 as its community is currently growing, and it is being applied in projects such as BASnum[5] and Nénufar[6].

This format's groundwork started in 2016, and it is currently led by the Digital Research Infrastructure for the Arts and Humanities (DARIAH) Lexical Resources Working Group[7]. TEI Lex-0 aims to define a clear and versatile, albeit not too permissive, annotation structure to facilitate heterogeneously encoded lexical resources' interoperability. TEI Lex-0 should be regarded as "a format that existing TEI dictionaries can be unequivocally transformed to, so that they can be queried, visualised or mined uniformly" (Tasovac et al., 2018). As this format's layout has not been finished yet, we have been actively contributing to its development by raising GitHub[8] issues.

4.1 Dicionário da Língua Portuguesa

The *Dicionário da Língua Portuguesa* (DLP) is being developed, both lexicographically and computationally, without any direct funds. This results in a slower pace of work. As such, its conversion from the custom TEI schema to TEI Lex-0 is being done progressively, using small steps that fix some specific aspect of the original encoding. Simultaneously, as the lexicographic work is being performed concurrently, the LeXmart tool also needs adaptations to support the new elements.

The designed approach is cyclical, consisting of the following steps:

1. A specific detail of the original encoding is chosen for conversion.
2. Then, its conversion to TEI Lex-0 is discussed and evaluated.[9]
3. This is followed by the complete rewrite of the dictionary files, considering that specific encoding structure.
4. While this process runs[10], the LeXmart code is edited to support this specific TEI Lex-0 encoding.

As soon as this cycle ends, the complete dictionary is validated accordingly with the TEI Lex-0 and RelaxNG schema (REgular LAnguage for XML Next Generation), so that we can account for the progress and choose what the next conversion step is.

[5] Available at https://anr.fr/Projet-ANR-18-CE38-0003.

[6] Available at http://nenufar.huma-num.fr/?article=3813.

[7] See https://www.dariah.eu/activities/working-groups/lexical-resources/.

[8] Available at https://github.com/DARIAH-ERIC/lexicalresources/projects/1.

[9] In some specific situations, the TEI Lex-0 team is contacted in order to understand and/or discuss how some information should be encoded.

[10] It can take from a few minutes to more than half an hour.

Before putting this approach into practice, the original TEI Lex-0 schema was included in another RelaxNG schema that allows the dictionary to be stored in different XML files, without repeating the whole TEI Header[11], and allows the inclusion of an extra element, named meta, that includes some metadata about the entry state. To keep the XML files as compliant as possible, this extension was done properly, using XML namespaces.

To give an idea of the adaptation process, a list of steps that were taken during the conversion is shown below:

1. To each entry, the required `@xml:id` attribute was added, using the entry filename as the base, thus guaranteeing uniqueness. At the same time, the attribute `@xml:lang` was also added.
2. The `@type` attributes for the `usg` element were normalised using the standard values for geographic and domain instead of the suggested names from the TEI schema: '*geo*' and '*dom*' (Salgado et al., 2019a).
3. As noted before, one of the adaptations during the bootstrap process was the addition of the `group` tag. For all entries which contain only one `group` element, it was removed, keeping its contents intact.
4. According with the TEI Lex-0 schema, every sense element should include the `@xml:id` attribute. These attributes were also added automatically, taking as the base the entry identifier, and adding a suffix with the sense number.
5. The `cit` elements need a `@type` attribute. This was easy to add as, at this specific stage, any occurrence of this element was a bibliography example. Thus, the attribute `@type` was added to all `cit` elements with the same value: '*example*.'
6. To encode the page part of a citation (under the `bibl` element), the original schema used the `pag` element. TEI Lex-0 suggests the usage of the `citedRange` element.
7. In the etymology, references to words in the dictionary, and references to words in other languages, were both encoded with the `mentioned` element. To be able to perform the replacement correctly we needed to use some context. Thus, the sequence

   De `<mentioned>`*word*`</mentioned>`

   was replaced by

   De `<ref type="entry">`*word*`</ref>`.
8. As every reference needs a `@type` attribute, as seen in the previous item, every `ref` element present in the dictionary was edited to include this attribute, with the entry value.
9. In the original dictionary the `ph` element was used in expressions that required placeholders (specific multiword expressions, where a specific token is a word from a class, and not a concrete word). As this element is not supported by TEI Lex-0, but the `hi` (from highlight) is valid, these were replaced.
10. Synonyms and antonyms have initially been encoded with the `syn` and `ant` elements. These were changed to a more complex structure of a reference with a specific type (synonymy or antonymy), as shown in the example below.

[11] We are dividing the dictionary into individual files, for easy concurrent editing. Nevertheless, while specified individually, the whole set of files constitutes the real document. Therefore, a TEI Header will be generated every time the full dictionary is exported in a single XML document. While in the database, that information would be redundant.

```
<xr type="synonymy"><ref type="entry">word</ref></xr>
```

11. Non-bibliographic examples were originally encoded as quotes, directly inside the `sense` element. This is not supported by the TEI Lex-0, requiring every occurrence to be replaced by the more complex structure shown below.

    ```
    <cit type="example"><quote type="example">...</quote></cit>
    ```

    Note that the `@type` attribute in the `quote` element is not required but useful for us to distinguish between bibliographic citations.
12. While DLP is being developed with the Internet as the target media, the project keeps track of entries or senses that should not be included in a paper dictionary. For this, the attribute `@digital` was originally created. To keep it with TEI Lex-0, the `@rend` attribute was chosen to encode this information. Thus, digital-only entries include the attribute `@rend="digital"`.
13. The references to words in other languages present in the etymology were encoded as mentioned elements. These were changed to citations, as shown in the next example:

    ```
    <cit type="etymon">
        <form><orth xml:lang="la">word</orth></form>
    </cit>
    ```

Even though we already converted much of the original syntax, the mentioned changes achieved 33,093 of the 70,726 entries in the dictionary as valid with regard to TEI Lex-0 (about 46.79%). There are some details needing changes that have not yet beenadequately discussed. One example is the `@fem` attribute in the `orth` element, which currently holds the suffix to generate the feminine form. One of the possibilities to encode this in TEI Lex-0 is to replace it with a full form entry. Nevertheless, for that to be done automatically we will require a morphological analyser to derive the feminine forms automatically.

4.2 Dicionário Aberto

Although the DA is also available in XML, following the dictionary chapter of TEI's general guidelines, the annotation granularity is bigger than DLP. This simplicity is derived from the lack of detailed annotation in the original document after the volunteer transcription, which only marked bold and italic words. Thus, the conversion to TEI was based only on that information, the knowledge of the dictionary's microstructure and a set of abbreviation lists (Simões & Farinha, 2011). These hints allowed a quite interesting structure to present the dictionary online with some quality but lack detailed annotations. Thus, its conversion to TEI Lex-0 is also simpler, as only the top-level structure is required.

As can be seen in Figure 1, originally each entry was encoded with only one sense. Only words with more than one grammatical class have more than one sense element. Different definitions are currently encoded in a single `def` element, where new lines are used to distinguish between senses.

While this structure is quite poor, its conversion to TEI Lex-0 is straightforward: the sense elements are removed from their current places. As for definitions (`def` element), their content is split by a new line and, for each line, a pair of `sense`/`def` elements is added. What follows is the addition of the required attributes, the identifier (`@xml:id`)

```
<entry id="drogaria">
  <form><orth>Drogaria</orth></form>
  <sense>
    <gramGrp>f.</gramGrp>
    <def>
      Porção de drogas.
      Estabelecimento, em que se vendem drogas.
    </def>
  </sense>
</entry>
```

Figure 1: Example of an entry before the TEI Lex-0 conversion.

```
<entry xml:id="drogaria" xml:lang="pt">
  <form><orth type="lemma">Drogaria</orth></form>
  <gramGrp>f.</gramGrp>
  <sense xml:id="drogaria-1"><def>Porção de drogas.</sense>
  <sense xml:id="drogaria-2"><def>Estabelecimento, em que se vendem drogas.</def></sense>
</entry>
```

Figure 2: Entry from Figure 1 after the TEI Lex-0 conversion.

and the language (`@xml:lang`). After these changes, we obtain a simple but valid TEI Lex-0 document.

While there are entries with some more annotation than in the presented example, in their transformation into a TEI Lex-0 file it is possible to keep the same basic structure.

4.3 Vocabulário Ortográfico da Língua Portuguesa

In microstructural terms, a lexicographical article from the VOLP-1940 may, as a rule, include the following elements: lemma, orthoepy, part of speech, and a gloss.

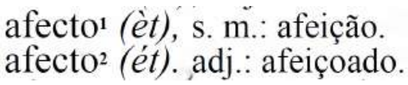
A lexicographical article in the VOLP-1940 starts with a base structure corresponding to the entry, followed by the grammatical information. Figure 3 shows the basic and regular structure of a VOLP-1940 entry to which the TEI Lex-0 annotation was applied.

```
<entry xml:id="..." xml:lang="pt" type="...">
   <form type="lemma">
      <orth>...</orth>
   </form>
   <gramGrp>
      <gram type="pos">...</gram>
      <gram type="gen">...</gram>
   </gramGrp>
</entry>
```

Figure 3: Basic and regular structure of a VOLP-1940 entry.

While the entry element encompasses all the information contained in the lexicographical article, the form element is used to note the information relating to the base, detailing its `@type` attribute as "lemma," and the orthographic form is provided in the `orth` element. It is important to note that in TEI Lex-0, the `entry` element requires the attributes

`@xml:id`, the entry identifier and `@xml:lang`, the appropriate language code. Since we are dealing with vocabulary entries, we use the form `@type="lemma"`.



afecto[1] *(èt),* s. m.: afeição.
afecto[2] *(ét).* adj.: afeiçoado.

Figure 4: Example of homonymous words on VOLP-1940.

In the particular case of homonymous words, as shown in Figure 4, “afecto”, the lemma is split. In TEI Lex-0, avoiding possible structural ambiguities, the `superEntry` element originally available in TEI (which groups a sequence of entries, such as a set of homographs) is no longer allowed, and therefore we use entry element systematically. To mark the numeric index, the element `lbl` preserves the digit of the original document while the attribute `@n` of the entry will, in turn, provide the information for the further processing of the entry by computational tools.

There is also information about words that are almost exclusively used in phrases. For example, when a particular word is only used in a particular phrase, this indication appears as an entry in what is considered the core word of that phrase — for instance, “cavalitas, el. nom. f. pl. na loc. adv. mod. às cavalitas” [riding piggyback, plural feminine noun element].

Another indication of a prescriptive nature concerns constructions that begin with the expression “Melhor que” [Better than]. The forms indicated as preferable are those that are considered to be closest to their origin or more correct for specific reasons, such as “canon” and “cânone” — “cânone, s. m. Melhor que canon” [cânone [canon], s. m. better than canon (Portuguese orthographic variant of the first form)][12]. So far, we have identified the essential and most relevant elements of the VOLP-1940’s microstructure.

5. Simplifying TEI Lex-0 Interface

TEI Lex-0 is an interesting format, as it is much less permissive than the original guidelines in the chapter 9 of the TEI, on dictionaries. To make this process more straightforward and structured, the TEI Lex-0 team is reusing some elements for different, although near, semantics. As an example, TEI allows the use of the quote element by itself, to add an authorless quote, while quotes with bibliographic information are stored inside the `cit` element. TEI Lex-0 does not allow the direct usage of the quote element and suggests the use of a `cit` element in both situations. While this makes the automatic processing of the resource easier, as element trees are shared, it creates a large overhead of XML annotations. There are other examples of such situations, namely the inclusion of synonyms or antonyms, which have already been mentioned, that require a complex reference structure, or the encoding of foreign words, that could be encoded with the mentioned element in the original TEI schema, and that requires a more complex nested entry when properly encoded using TEI Lex-0.

As an option during its development, the LeXmart editor shows entries in a format very close to its XML structure. That is interesting for experienced users, as it clearly shows

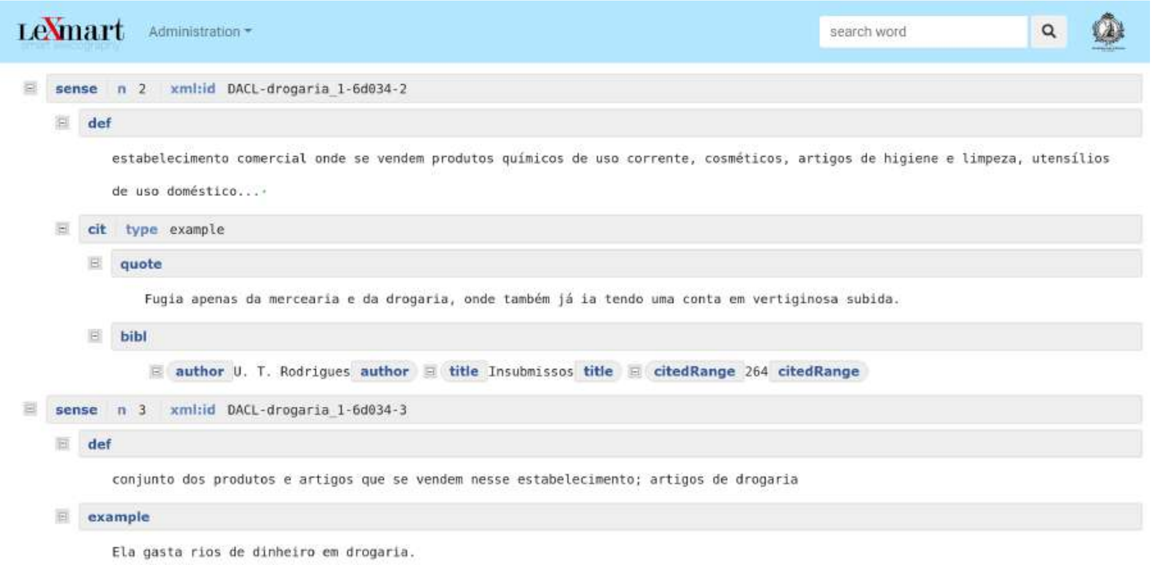
[12] However, today the non-preferential form is the most common.

the annotation details. Nevertheless, if this editor includes the full structures for some of the situations described above, entries would be challenging to edit on a web browser.

During the development we also faced some issues regarding the versatility of Xonomy[13], the JavaScript library that implements the LeXmart web editor. While Xonomy has a very interesting application programming interface (API), and allows a high level of customisation, we faced some issues during the implementation of some functionalities, as they would require a large amount of coding.

The solution for both of these problems is the XML rewrite before the editing process, removing some complex structure and hiding it under a set of custom elements, and a post-processing pipeline that transforms this custom XML back into TEI Lex-0. This process is an excellent approach to make entry editing simpler and a straightforward way to guarantee the correct usage and respective element structure for some specific constructions.

This mapping is done automatically by the eXist Database backend that supports LeXmart, running a pair of eXtensible Stylesheet Language Transformations (XSLT) that transform the document structure.



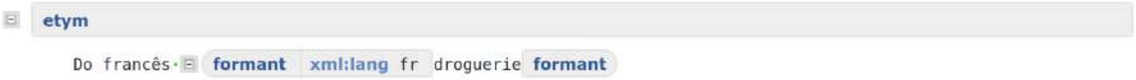
```
sense  n 2  xml:id DACL-drogaria_1-6d034-2
  def
    estabelecimento comercial onde se vendem produtos químicos de uso corrente, cosméticos, artigos de higiene e limpeza, utensílios
    de uso doméstico...
  cit  type example
    quote
      Fugia apenas da mercearia e da drogaria, onde também já ia tendo uma conta em vertiginosa subida.
    bibl
      author U. T. Rodrigues author  title Insubmissos title  citedRange 264 citedRange
sense  n 3  xml:id DACL-drogaria_1-6d034-3
  def
    conjunto dos produtos e artigos que se vendem nesse estabelecimento; artigos de drogaria
  example
    Ela gasta rios de dinheiro em drogaria.
```

Figure 5: LeXmart editor, showing two types of examples: bibliographic or not.

In Figure 5 we show two senses for the entry “drogaria” [drugstore] from DLP. Note that the first block, that corresponds to the second sense, shows a citation, of type example, that includes the quote and its bibliography information. The second block, which corresponds to the third sense, shows an `example` element. Although this element is not part of the TEI Lex-0 standard, it gets converted back and forth from the following structure:

$$\texttt{<cit type="example"><quote type="example">} \Longrightarrow \texttt{<example>}$$

[13] Available at https://github.com/michmech/xonomy.



```
etym
Do francês·⊟ formant xml:lang fr droguerie formant
```

Figure 6: LeXmart editor, showing etymological information with formant information.

Figure 6 shows a different situation for this same entry. To keep the editor as clean as possible, a `formant` element was created to hide the structure behind the inclusion of a foreign word in the etymology:

```
<etym>Do francês <cit type="etymon"><form>
  <orth xml:lang="fr">droguerie</orth></form></cit></etym>
```

These simple changes allow quicker editing for the lexicographer without jeopardising the document structure's adequacy to the TEI Lex-0 schema. In order to reduce the ambiguity, these new elements have different designations from the entries available either in TEI Lex-0 or the original TEI schema[14].

6. Conclusions

This article briefly described three different lexicographic resources, with different origins, and belonging to projects with independent goals. Nevertheless, it was shown that these resources can be encoded using the TEI Lex-0 schema, and therefore, their editing can be performed in a tool supporting this specific structure.

With this in mind, LeXmart has been modified to comply with this schema, and therefore allow their editing. To keep the tool as simple to use as possible, a set of mechanisms were developed to hide some of the XML encoding's verbosity.

For the future of LeXmart, a diverse number of features are already planned:

- The codebase of the tool requires generalisation, as much of it was developed with DLP in mind. While the code itself is easy to apply to different resources, the configuration of the system is currently hardcoded.
- LeXmart aims at allowing the lexicographer to manage labels (domain labels, geographic labels, etc.): not just to add or remove labels, but also to account for their usage. We also intend to have a taxonomy or an ontology to structure the labels. This would allow a very detailed annotation of the entries and allow interesting search scenarios for the end-user.
- With DLP going online during 2021, the system is being tested for exporting the dictionary database to a non-XML, but still document-oriented database for fast querying. Using the eXist database is quite helpful during the editing process, as the tool is aware of the XML structure, but it is relatively inefficient for simple querying. This will also allow the creation of dictionary snapshots, keeping the lexicographers' work on a non-public version of the dictionary.

[14] The designations currently in use might be changed in the future, as they were not yet a matter of discussion with all the involved parties.

- LeXmart, by itself, needs further improvements. A lot of the code is still too specific for DACL. Nevertheless, given it is available as an open-source project, we expect to have, sooner or later, new users testing the system with other languages and other kinds of resources, thus allowing for the development of new features but also the possibility of the customisation.

7. Acknowledgements

This paper was partially funded by Portuguese national funds (PIDDAC), through the FCT – Fundação para a Ciência e Tecnologia and FCT/MCTES under the scope of the projects UIDB/05549/2020 and UID/LIN/03213/2020, and the European Union's Horizon2020 research and innovation programme under grant agreement No 731015 (ELEXIS — European Lexicographic Infrastructure).

8. References

ACL (1940). *Vocabulário Ortográfico da Língua Portuguesa.* Lisboa: Academia das Ciências de Lisboa & Imprensa Nacional.

ACL (2001). *Dicionário da Língua Portuguesa Contemporânea.* Lisboa: Academia das Ciências de Lisboa & Editorial Verbo.

ACL (2021). *Dicionário da Língua Portuguesa.* Lisboa: Academia das Ciências de Lisboa.

Gonçalo Oliveira, H. & Gomes, P. (2014). ECO and Onto.PT: A flexible approach for creating a Portuguese wordnet automatically. *Language Resources and Evaluation*, 48(2), pp. 373–393.

Gonçalo Oliveira, H. (2018). A Survey on Portuguese Lexical Knowledge Bases: Contents, Comparison and Combination. *Information*, 9(2).

Rundell, M. (2010). What future for the learner's dictionary? In I.J. Kernerman & P. Bogaards (eds.) *English Learners' Dictionaries at the DSNA 2009.* Jerusalem: Kdictionaries, pp. 169–175.

Rundell, M. & Kilgarriff, A. (2011). Automating the creation of dictionaries: Where will it all end? *Studies in Corpus Linguistics*, 45, pp. 257–282.

Salgado, A. & Costa, R. (2020). O projeto 'Edição Digital dos Vocabulários da Academia das Ciências': o VOLP-1940. *Revista Da Associação Portuguesa De Linguística*, 7, pp. 275–294.

Salgado, A., Costa, R. & Tasovac, T. (2019a). Improving the consistency of usage labelling in dictionaries with TEI Lex-0. In *Lexicography ASIALEX 6.* pp. 133–156.

Salgado, A., Costa, R., Tasovac, T. & Simões, A. (2019b). TEI Lex-0 In Action: Improving the Encoding of the Dictionary of the Academia das Ciências de Lisboa. In I. Kosem, T. Zingano Kuhn, M. Correia, J.P. Ferreira, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek & C. Tiberius (eds.) *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference.* pp. 417–433.

Simões, A., Almeida, J.J. & Salgado, A. (2016). Building a Dictionary using XML Technology. In M. Mernik, J.P. Leal & H.G. Oliveira (eds.) *5th Symposium on Languages, Applications and Technologies (SLATE)*, volume 51 of *OASIcs*. Germany: Schloss Dagstuhl, pp. 14:1–14:8.

Simões, A. & Farinha, R. (2011). Dicionário Aberto: um recurso para processamento de linguagem natural. *Viceversa: revista galega de traducción*, 16, pp. 159–171.

Simões, A., Salgado, A., Costa, R. & Almeida, J.J. (2019). LeXmart: A Smart Tool for Lexicographers. In I. Kosem, T. Zingano Kuhn, M. Correia, J.P. Ferreira, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek & C. Tiberius (eds.) *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference*. pp. 453–466.

Tasovac, T., Romary, L., Banski, P., Bowers, J., de Does, J., Depuydt, K., Erjavec, T., Geyken, A., Herold, A., Hildenbrandt, V., Khemakhem, M., Petrović, S., Salgado, A. & Witt, A. (2018). TEI Lex-0: A baseline encoding for lexicographic data. Version 0.8.6. Technical report, DARIAH Working Group on Lexical Resources. URL https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html.

The ELEXIS System for Monolingual Sense Linking in Dictionaries

John P. McCrae[1], Sina Ahmadi[1], Seung-Bin Yim[2], Lenka Bajčetić[2]

[1] Data Science Institute, NUI Galway
[2] Austrian Academy of Sciences
E-mail: john@mccr.ae, sina.ahmadi@insight-centre.org,
seung-bin.yim@oeaw.ac.at, lenka.bajcetic@oeaw.ac.at

Abstract

Sense linking is the task of inferring any potential relationships between senses stored in two dictionaries. This is a challenging task and in this paper we present our system that combines Natural Language Processing (NLP) and non-textual approaches to solve this task. We formalise linking as inferring links between pairs of senses as exact equivalents, partial equivalents (broader/narrower) or a looser relation or no relation between the two senses. This formulates the problem as a five-class classification for each pair of senses between the two dictionary entries. The work is limited to the case where the dictionaries are in the same language and thus we are only matching senses whose headword matches exactly; we call this task Monolingual Word Sense Alignment (MWSA). We have built tools for this task into an existing framework called Naisc and we describe the architecture of this system as part of the ELEXIS infrastructure, which covers all parts of the lexicographic process including dictionary drafting. Next, we look at methods of linking that rely on the text of the definitions to link, firstly looking at some basic methodologies and then implementing methods that use deep learning models such as BERT. We then look at methods that can exploit non-textual information about the senses in a meaningful way. Afterwards, we describe the challenge of inferring links holistically, taking into account that the links inferred by direct comparison of the definitions may lead to logical contradictions, e.g., multiple senses being equivalent to a single target sense. Finally, we document the creation of a test set for this MWSA task that covers 17 dictionary pairs in 15 languages and some results for our systems on this benchmark. The combination of these tools provides a highly flexible implementation that can link senses between a wide variety of input dictionaries and we demonstrate how linking can be done as part of the ELEXIS toolchain.

Keywords: sense linking; lexicography; natural language processing; linked data; tools

1. Introduction

Monolingual word sense alignment is the task of finding the equivalent or related senses among two dictionary entries with the same headword from two different dictionaries. In this paper, we present our framework and tool for creating such a mapping between two dictionaries, called Naisc McCrae & Buitelaar (2018)[1]. This architecture is intended as an experimental framework into which many components can be integrated. In this paper, we give an overview of this system and examples of some of the methods that that can be integrated into this framework. For this work, we focus on only the monolingual word sense alignment task, but many of the techniques discussed here can also be used to create multilingual linking between dictionaries and also linking between other kinds of datasets.

We understand that there are three major aspects to consider when building a linking system in the framework provided by Naisc. Firstly, we have the task of textual similarity, which takes the textual content of each sense, principally the definition and estimates the similarity between them. Secondly, we have non-textual similarity, an iterative process that can be used to link dictionaries that contain links between entries, such as WordNet. These tools become especially useful in the context of linking to external encyclopaedic resources such as Wikipedia or Wikidata. Finally, we look at linking as a holistic step, where we consider the linking task as one of predicting one of four relationships between senses: equivalent, narrower, broader or partially related. This turns the task into a

[1] https://github.com/insight-centre/naisc

five-class classification task (with 'unrelated' as the fifth class), but in addition there are constraints that logically follow, and we formalise this and show how we can generate an optimal overall mapping between senses.

These elements are all being integrated into the framework and we present some preliminary results about the individual component performance as well as insight into the motivations of the architecture and the design of the system. In addition, we also summarise the development of a benchmark dataset for this task (Ahmadi et al., 2020). The rest of this paper is structured as follows. In Section 2 we present the overall architecture of the Naisc system. We then look at textual features in Section 3, non-textual features in Section 4 and constraints for linking in Section 5. Finally, we describe the development of a benchmark dataset in Section 6 and conclude in Section 7.

2. Architecture

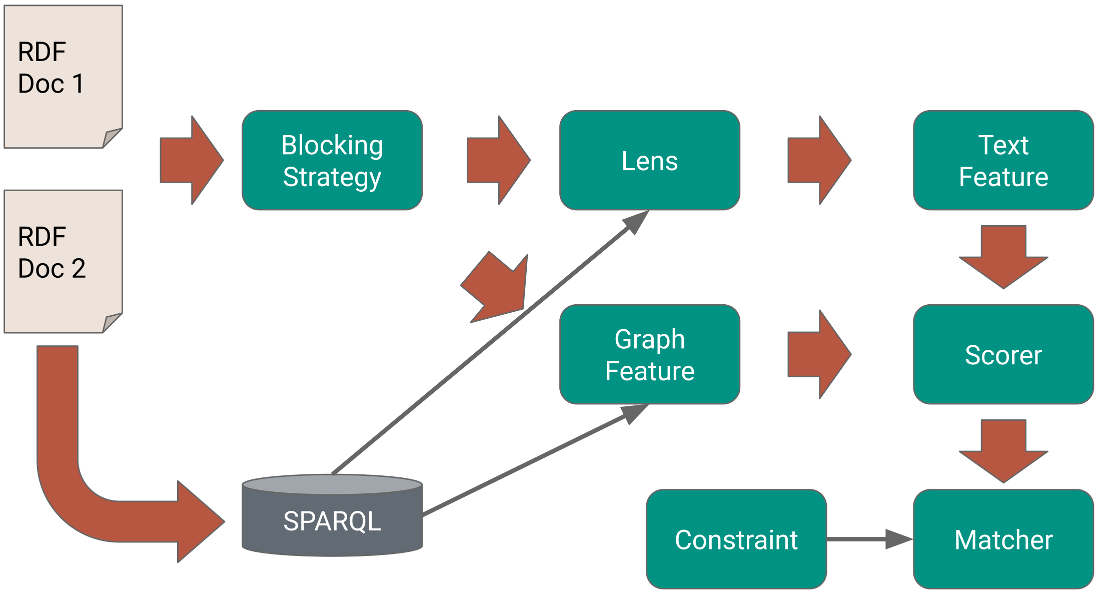


Figure 1: The architecture of the Naisc system for sense linking

The Naisc architecture is depicted in Figure 1. The architecture of Naisc was originally designed for linking any RDF datasets and this can be applied to the MWSA task by converting the dictionaries into an RDF format such as OntoLex (McCrae et al., 2017; Cimiano et al., 2016). The process of linking is broken down into a number of steps that are described as follows:

- **Blocking**: The blocking step finds the set of pairs that are required to be linked. For more general linking tasks and for the multilingual linking task this is quite challenging and error-prone. However, for the MWSA task we only link based on matching headwords so the blocking task has a single implementation that simply finds matching headwords and outputs every sense pair between these two entries. Signature: (Dataset, Dataset) ⇒ (Sense, Sense)*

- **Lens**: The lens examines the data around the sense pair to be linked and extracts text that can be compared for similarity. Clearly, the most important lens for this task extracts the senses' definitions. However, other information such as examples can also be extracted here.
  Signature: (Sense, Sense) ⇒ (Text, Text)
- **Text features**: The text features extract a set of similarity judgements about the texts extracted with the lenses and are described in more detail in the following section. Signature: (Text, Text) ⇒ $\mathbb{R}^*$
- **Graph features**: Graph (or non-textual) features do not rely on the text in the dataset but instead look at other features. They are described in more detail later in the document.
  Signature: (Sense, Sense) ⇒ $\mathbb{R}^*$
- **Scorer**: From a set of features extracted either from the text or from other graph elements, a score must be estimated for each of the sense pairs. This can be done in either a supervised or unsupervised manner and we implement standard methods for supervised classification such as SVMs and unsupervised classification using voting.
  Signature: $\mathbb{R}^* \Rightarrow [0,1]^*$ - *Output corresponds to a probability distribution over the relation classes*
- **Matcher and Constraint**: There are normally some constraints that we wish to enforce on the matching and these are applied by the matcher
  Signature: (Sense, Sense, $[0,1]^*$)* ⇒ (Sense, Sense)* - *Output is a subset of the input*

Naisc is implemented in Java and the configuration of each run can be specified by giving a JSON description of the components that can be used. For example, this is a default configuration for the MWSA task (presented using YAML syntax):

```
blocking:
  name: blocking.OntoLex
lenses:
- name: lens.Label
  property:
  - http://www.w3.org/2004/02/skos/core#definition
  id: label
textFeatures:
- name: feature.BasicString
  wordWeights: models/idf
  ngramWeights: models/ngidf
  labelChar: true
- name: feature.WordEmbeddings
  embeddingPath: models/glove.6B.100d.txt
scorers:
- name: scorer.LibSVM
  modelFile: models/default.libsvm
matcher:
  name: matcher.BeamSearch
  constraint:
    name: constraint.Taxonomic
description: The default setting for processing two OntoLex dictionaries
```

This configuration assumes that the dictionary is in the OntoLex format for blocking and processes it as such, it then extracts the definitions using the 'Label' lens and applies both some basic string text features as well as text features based on GloVe vectors (Pennington et al., 2014a). The scores for each property type are calculated using LibSVM (Chang & Lin, 2011) and finally the overall linking is calculated using the taxonomic constraints, which will be defined later in this document.

3. Text Similarity Methods

The comparison of the definitions of the lexical entries is the most obvious and effective method for establishing similarity between senses in two dictionaries and is the primary method that humans would use. As such, it makes sense to focus our efforts on developing an artificial intelligence approach for the task of estimating the similarities of definitions, which is a kind of Semantic Textual Similarity (STS) as explored in tasks at SemEval (Agirre et al., 2016). We have explored three main approaches to this, firstly using simple text features to provide a baseline for the task. Secondly, we use deep learning methods including BERT and finally we move beyond simple similarity to also predict the taxonomic type of the relationship between senses.

3.1 Basic Methods

The basic methods use frequency and surface forms of the strings to compute features; the following methods are implemented by the Naisc tool. Most of these methods can work on words or on characters.

Longest common subsequence The longest subsequence of words (characters) that match between the two strings as a ratio to the average length between the two strings.

Longest common prefix/suffix The longest subsequence of words (characters) from the start/end of each string, as a ratio to the average length.

N-gram The number of matching subsequences of words (characters) of length n between the two strings as a ratio to the average maximum number of n-grams that could match (e.g. length of string minus n plus one)

Jaccard/Dice/Containment The match between the words of the two definitions using the Jaccard and Dice coefficients. Let A and B be the set of words in each definition: Jaccard $= \frac{|A \cap B|}{|A \cup B|}$, Dice $= \frac{2|AB|}{|A|+|B|}$, Containment $= \frac{|A \cap B|}{\min(|A|,|B|)}$

Sentence Length Ratio The ratio of the length of the sentences as $\mathrm{SLR}(x, y) = 1 - \frac{min(|x|,|y|)}{max(|x|,|y|)}$

Average Word Length Ratio The ratio of the average word length in each sentence normalized to the range [0,1] as for SLR.

Negation Whether either both sentences contain negation words or both don't (1 if true, 0 if false).

Number If both sentences contain numbers do these numbers match (1 if all numbers match).

Jaro-Winkler, Levenshtein Standard string similarity functions, we use the Apache Commons Text implementations.

Monge-Elkan This is defined as follows where sim is a word similarity function (we use either Jaro-Winkler of Levenshtein) ME$(s,t) = \frac{1}{|s|}\sum_{i=1}^{|s|}\max_{j=1,\ldots t} sim(s_i, t_j)$

In addition, we implement the following approach based on using GloVe vectors (Pennington et al., 2014b), where we calculate the word embeddings for each word in the two definitions and then compare pairwise the words of each definition. These are turned into a single feature using methods described in McCrae and Buitelaar (McCrae & Buitelaar, 2018).

3.2 Beyond Similarity

Dictionaries are valuable resources which document the life of words in a language from various points of view. Senses, or definitions, are important components of dictionaries where dictionary entries, i.e. lemmata, are described in plain language. Therefore, unlike other properties such as references, cross-references, synonyms and antonyms, senses are unique in the sense that they are more descriptive but also highly contextualised. Moreover, unlike lemmata which remain identical through resources in the same language, except in spelling variations, senses can undergo tremendous changes based on the choice of the editor, lexicographer and publication period, to mention but a few factors. Therefore, the task of word sense alignment (WSA) will facilitate the integration of various resources and the creation of inter-linked language resources.

Considering the literature, various components of the WSA task have been the focus of previous research (Ahmadi & McCrae, 2021). However, few of the previous papers address WSA as a specific task on its own. As a preliminary study, our focus is on providing explainable observations for the task of WSA using manually-extracted features and analysing the performance of traditional machine learning algorithms for word sense alignment as a classification problem. Despite the increasing popularity of deep learning methods in providing state-of-the-art results in various NLP fields, we believe that evaluating the performance of feature-engineered approaches is an initial and essential step to reflect the difficulties of the task, and also the expectations from the future approaches.

We define our task of WSA and semantic induction as the detection of the semantic relationship between a pair of senses in two monolingual resources, as follows:

$$rel = sem(p, s_i, s_j)$$

where p is the part-of-speech of the lemma, s_i and s_j are senses belonging to the same lexemes in two monolingual resources and rel is a semantic relation, namely exact, broader, narrower, related and none. Our goal is to predict a semantic relation, i.e. rel given a pair of senses. Therefore, we define three classification problems based on the relation:

Binary classification which predicts if two senses can possibly be aligned together. Otherwise, none is selected as the target class.

SKOS classification which predicts a label among exact, broader, narrower and related semantic relationships.

SKOS+none classification which predicts a label given all data instances. This is similar to the previous classifier, with none as a target class.

3.2.1 Approach

Assuming that the textual representation of senses in definitions can be useful to align them, we define a few features which use the lengths of senses along with their textual and semantic similarities. In addition, we incorporate word-level semantic relationships to determine the type of relation that two senses may possibly have. Our features are defined in Table 1.

Feature Extraction

In this step, we extract sense instances from the MWSA datasets (Ahmadi et al., 2020), as $t = (p, s_i, s_j, r_{ij})$. This instance is interpreted as sense s_i has relation r_{ij} with sense s_j. Therefore, the order of appearance is important to correctly determine the relationship. It should also be noted that both senses belong to the same lemma with the part-of-speech p.

#	feature	definition	possible values
1	`POS_tag`	part of speech of the headword	a one-hot vector of {N, V, ADJ, ADV, OTHER}
2	`s_len_no_func_1/2`	number of space-separated tokens in s_1 and s_2	$\mathbb{N}$
3	`s_len_1/2`	number of space-separated tokens in s_1 and s_2 without function words	$\mathbb{N}$
4	`hypernymy`	hypernymy score between tokens	sum of weights in ConceptNet
5	`hyponymy`	hyponymy score between tokens	sum of weights in ConceptNet
6	`relatedness`	relatedness score between tokens	sum of weights in ConceptNet
7	`synonymy`	synonymy score between tokens	sum of weights in ConceptNet
8	`antonymy`	antonymy score between tokens	sum of weights in ConceptNet
9	`meronymy`	meronymy score between tokens	sum of weights in ConceptNet
10	`similarity`	similarity score between tokens	sum of weights in ConceptNet
11	`sem_sim`	semantic similarity score between senses using word embeddings	averaging word vectors and cosine similarity [0-1]
12	`sem_sim_no_func`	semantic similarity score between senses without function words	averaging word vectors and cosine similarity excluding function words [0-1]
13	`sem_bin_rel`	target class	1 for alignable, otherwise 0
14	`sem_rel_with_none`	target class	{`exact, narrower, broader, related, none`}
15	`sem_rel`	target class	{`exact, narrower, broader, related`}

Table 1: Manually extracted features for semantic classification of sense relationships

Given the class imbalance where senses with a *'none'* relationship are more frequent than the others, we carry out a data augmentation technique based on the symmetric property of the semantic relationships. By changing the order of the senses, also known as relation direction, in each data instance, a new instance can be created by semantically reversing the relationship. In other words, for each $t = (p, s_i, s_j, r_{ij})$ there is a $t' = (p, s_j, s_i, r'_{ij})$ where r'_{ij} is the inverse of r_{ij}. Thus, exact and related as symmetric properties remain the same, however, the asymmetric property of the broader and narrower relationships yields narrower and broader, respectively.

Once the senses are extracted, we create data instances using the features in Table 1. Features 2 and 3 concern the length of senses and how they are different. Intuitively speaking, this regards the wordings used to describe two concepts and their semantic relationship. In features 4 to 11, we calculate this with and without function words, words with little lexical meaning. One additional step is to query ConceptNet to retrieve semantic relations between the content words in each sense pair. For instance, the two words "gelded" and "castrated" which appear in two different senses are synonyms, and therefore the whole senses can possibly be synonyms. In order to measure the reliability of the relationships, we sum up the weights, also known as assertions, of each relationship according to ConceptNet. Finally, features 12 and 13 provide the semantic similarity of each sense pair using word embeddings. The data instances are all standardised by scaling each feature to the range of [0-1].

Feature learning and classification

A Restricted Boltzmann Machine (RBM) is a generative model representing a probability distribution given a set of observations (Fischer & Igel, 2012). An RBM is composed of two layers: a visible one where the data instances are provided according to the manually created features, and a latent one where a distribution is created by the model by retrieving dependencies within variables. In other words, the relation of the features in how the target classes are predicted is learned in the training phase. We follow the description of Hinton (2012) in implementing and using an RBM for learning further features from our data instances. Regarding the classification problem, instead of training our models using the data instances described in the previous section, we train the models using the latent features of an RBM model. These new features have binary values and can be configured and tuned depending on the performance of the models.

For this supervised classification problem, we use support vector machines (SVMs) using various hyper-parameters, as implemented in Scikit. After a preprocessing step, where the datasets are shuffled, normalized and scaled, we split them into train, test and validation sets with 80%, 10% and 10% proportions, respectively.

3.3 Experiments

Table 2 provides the evaluation results of our classification approach for MWSA. Despite the high accuracy of the baseline systems for most languages, they do not perform equally efficiently for all languages in terms of precision and recall. Although our classifiers outperform the baselines for all the relation prediction tasks and perform competitively when trained for the binary classification and also given all data instances, there is significantly lower performance when it comes to the classification of SKOS relationships. This can be explained by the lower number of instances available for these relations. Moreover, distinguishing certain types of relationships, such as related versus exact, is a challenging task even for an expert annotator. Regarding the performance of the RBM, we do not observe a similar improvement in the results of all classifiers.

One major limitation of the current approach is the usage of crafted features. We believe that as a future work further techniques can be used, particularly thanks to the current advances in word representations and neural networks. Furthermore, incorporating knowledge bases and external language resources such as corpora can be beneficial in improving the ability of the system to address sense ambiguity for polysemous entries.

Language	Metric	Baseline	Binary	All	SKOS	RBM-Binary	RBM-all	RBM-SKOS
Basque	Accuracy	78.90	78.79	58.47	49.77	70.37	54.17	28.85
	Precision	21.10	71.40	59.21	43.65	62.14	59.08	20.73
	Recall	5.00	72.78	58.45	46.01	74.93	52.55	50.87
	F-measure	8.10	**72.08**	**58.83**	**44.80**	67.94	55.62	29.46
Bulgarian	Accuracy	72.80	70.60	65.91	34.05	73.51	63.38	36.47
	Precision	25.00	68.75	64.79	31.75	77.46	34.46	36.85
	Recall	1.10	69.32	65.44	31.83	72.91	49.87	24.86
	F-measure	2.00	69.03	**65.11**	**31.79**	**75.11**	40.76	29.69
Danish	Accuracy	81.70	66.47	34.82	27.87	73.85	50.08	29.67
	Precision	3.00	74.54	23.70	36.49	60.59	60.96	30.47
	Recall	2.30	75.51	62.90	22.87	55.66	66.92	73.04
	F-measure	4.30	**75.02**	34.43	28.12	**58.02**	63.80	**43.00**
Dutch	Accuracy	93.60	82.55	59.99	24.75	83.90	51.47	36.34
	Precision	0.00	86.97	78.59	31.38	59.78	77.82	30.66
	Recall	0.00	88.24	79.22	33.10	67.33	39.65	66.03
	F-measure	0.00	**87.60**	**78.90**	32.22	63.33	52.54	41.88
English	Accuracy	75.20	89.00	81.00	49.00	80.16	65.03	48.57
	Precision	0.00	82.35	73.03	39.31	64.36	63.67	55.53
	Recall	0.00	82.87	76.41	46.63	82.13	79.35	34.51
	F-measure	0.00	**82.61**	**74.68**	**42.66**	72.17	70.65	42.57
Estonian	Accuracy	48.20	78.98	58.92	46.11	75.96	62.75	47.82
	Precision	54.50	76.06	68.83	40.81	63.53	60.67	36.63
	Recall	9.30	20.76	57.82	44.02	28.18	49.35	22.44
	F-measure	15.90	32.62	**62.85**	**42.35**	**39.05**	54.43	27.83
German	Accuracy	77.77	73.14	61.99	49.58	77.97	43.23	44.21
	Precision	0.00	77.72	64.74	41.89	80.44	66.34	40.99
	Recall	0.00	54.41	59.95	43.73	22.88	27.92	48.99
	F-measure	0.00	**64.01**	**62.25**	42.79	35.63	39.30	**44.63**
Hungarian	Accuracy	94.00	79.65	58.40	22.95	81.46	36.27	15.20
	Precision	5.30	49.96	30.14	23.41	68.50	59.80	26.58
	Recall	1.20	54.47	37.95	68.08	56.72	73.85	29.23
	F-measure	2.00	52.12	33.60	**34.85**	**62.05**	**66.09**	27.84
Irish	Accuracy	58.30	75.00	55.75	26.27	79.61	60.84	24.75
	Precision	68.00	84.42	46.58	31.84	79.03	42.52	30.25
	Recall	18.50	84.46	39.85	46.15	52.47	54.65	25.40
	F-measure	29.10	**84.44**	42.95	37.68	63.06	**47.83**	27.61
Italian	Accuracy	69.30	59.08	55.43	44.48	77.23	46.26	43.01
	Precision	0.00	52.55	42.98	28.80	75.69	46.31	40.56
	Recall	0.00	66.47	52.64	42.16	45.05	68.67	31.27
	F-measure	0.00	**58.69**	47.32	**34.22**	56.49	**55.32**	35.32
Serbian	Accuracy	59.90	80.05	32.53	27.55	82.35	41.43	32.96
	Precision	19.00	76.78	48.57	43.06	73.51	37.70	21.49
	Recall	46.40	65.73	69.40	27.10	77.46	48.45	55.53
	F-measure	26.90	70.83	**57.15**	**33.26**	**75.43**	42.40	30.99
Slovenian	Accuracy	44.20	84.29	36.13	26.13	78.93	39.57	31.63
	Precision	17.30	73.08	23.19	46.98	78.62	38.59	20.97
	Recall	58.70	83.22	45.07	28.61	41.64	28.09	33.02
	F-measure	26.80	**77.82**	30.62	**35.56**	54.45	**32.51**	25.65
Spanish	Accuracy	-	73.79	54.67	30.28	80.71	54.38	58.48
	Precision	-	79.78	55.07	33.21	79.40	42.54	39.57
	Recall	-	80.37	53.15	40.04	60.18	20.68	38.59
	F-measure	-	**80.07**	**54.10**	36.31	68.47	27.83	**39.07**
Portuguese	Accuracy	92.10	71.31	66.62	51.71	73.14	55.69	42.87
	Precision	8.30	49.29	58.23	53.52	77.72	69.41	40.45
	Recall	2.40	37.47	70.41	53.47	54.41	22.32	38.15
	F-measure	3.70	42.57	**63.74**	**53.49**	**64.01**	33.78	39.26
Russian	Accuracy	75.40	60.88	58.90	37.75	75.80	59.76	33.10
	Precision	43.80	72.92	63.83	27.28	73.38	73.77	32.71
	Recall	17.90	82.21	44.43	36.74	68.23	70.39	47.75
	F-measure	25.50	**77.29**	52.39	31.31	70.71	**72.04**	**38.82**

Table 2: Results of the classification results with and without an RBM.

3.4 Deep Learning Methods

Besides employing feature-based approaches, we additionally utilise fine-tuned pre-trained neural network language models (NNLM), Bidirectional Encoder Representations from Transformers (Devlin et al., 2019) and the Robustly optimised BERT pretraining approach (RoBERTa) (Liu et al., 2019). This is done by using the Hugging Face transformers library, which provides the API for finetuning of transformer models.

Recently, transformer-architecture-based approaches have been proven to be beneficial for improving different downstream NLP tasks. For this reason we have decided to explore how well those models are suited for the MWSA task.

BERT is designed to pre-train deep bidirectional representations from unlabelled text by jointly conditioning on both the left and right context in all layers and is trained on masked word prediction and next sentence prediction tasks. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks (Devlin et al., 2019).

The MWSA task can be ultimately regarded as sentence pair classification task and BERT can easily be fine-tuned for it, since its use of self-attention mechanism (Vaswani et al., 2017) to encode concatenated text pairs effectively includes bidirectional cross attention between two sentences. We have followed the fine-tuning approach presented in the original paper (Devlin et al., 2019).

In order to get the best results, we have experimented with different pre-trained models, such as BERT Base, BERT Large and RoBERTa for English. RoBERTa is a variation of BERT created by tweaking different aspects of pre-training, such as bigger data and batches, omitting next sentence prediction, training on longer sequences and changing the masking pattern (Liu et al., 2019)

3.4.1 Fine-tuning transformer models

A transformer-based approach was conducted for English and German. Different parameter settings have been tried out to find the best performing model for both languages. Due to the size of the pre-trained language models and limitations in computation powers, we were only able to explore hyper-parameter combinations selectively. Different pre-trained language models were used and were evaluated in the early phase of the experiments, to limit the parameter exploration space.

Preprocessing

Representation of word senses The transformers architecture requires input to be in certain structures depending on the pretrained models used. For our MWSA task, which we basically regard as sentence pair classification, transformer models require two sentences concatenated by separation token, and a preceding classification token. The Hugging Face transformers library provides tokenisers for different pre-trained models.

Labels and class weight Labels are one-hot encoded and class weights are calculated to mitigate the class imbalance problem.

Model training

Training Environment

The training was done on an NVIDIA Tesla P100 GPU hosted on Google Cloud Platform.

Hyperparameters

Our early explorations with the pre-trained models quickly showed that bigger models deliver better results. The tendency that bigger pre-trained models perform better on MWSA is in line with observations made by the original BERT paper authors by comparing BERT Base and Large for different downstream tasks (Devlin et al., 2019) or RoBERTa performing better than the original BERT on selected downstream tasks (Liu et al., 2019). For this reason, we have conducted more hyperparameter test combinations for those models (RoBERTa Large for English, and DBMDZ for German). When using bigger models, such as RoBERTa or BERT Large, smaller train-batch-size was selected due to resource limitations. The original BERT models were trained with 512 sequence lengths, but since the MWSA datasets mostly have short sentence pairs, we experimented with shorter sequence length of 128 and 256 to save memory usage and be more flexible with respect to batch size.

Parameter	value set	English	German
used model	BERT English(Large) German BERT(deepset.ai, DBMDZ cased)	RoBERTa(Large)	DBMDZ German BERT
label weights		NONE: 0.23 EXACT: 2.08 BROADER: 42.05 NARROWER:5.37 RELATED:32.69	NONE: 0.27 EXACT: 2.74 BROADER: 2.31 NARROWER:3.13 RELATED:8.32
max-seq-length	64, 128, 256, 512	256	256
train-batch-size	8, 16, 32	16	32
num-train-epochs	2,3,5,7,10,15	2	7
weight-decay	0.3, 0.5	0.3	0.3
learning-rate	1e-6, 8e-6, 9e-6, 1e-5, 3e-5, 4e-5,5e-5	9e-6	3e-5

Table 3: Language model and Hyperparameters used for fine-tuning NNLM to MWSA

Loss function

As the MWSA task is a multi-class classification task, we use categorical cross entropy as our loss function for fine-tuning the models.

Model Evaluation

For evaluation of the trained models, we use weighted the Matthews correlation coefficient (Matthews, 1975), F1-measure and balanced accuracy, to take data imbalance into account. We also monitored the three metrics during training to determine when the model starts to overfit and adjusted the hyperparameters for further tuning.

Comparison of the fine-tuned models were not only done in regard to different hyperparameter settings, but also with respect to feature-based classification models, which we took as the baseline models.

With appropriate hyperparameters, English and German classifiers based on BERT (German) and RoBERTa (English) showed convergence with respect to the categorical cross-entropy loss function. Classes were weighted according to the distribution for loss calculation. Both models selected deliver better results than feature-based models. Noteworthy is that transformer-based models were able to classify some of the “narrower” relations correctly, where feature-based models failed. The general performance of the models leaves room for improvements, and data imbalance probably plays a significant role in improving them.

Language	Model	5-class accuracy	2-class precision	2-class recall	2-class F-measure
English	Baseline	0.752	0.000	0.000	0.000
	Feature-based	0.766	0.612	0.533	0.570
	BERT Large	0.654	0.467	0.850	0.602
	RoBERTa	0.763	0.619	0.782	0.691
German	Baseline	0.777	0.000	0.000	0.000
	Feature-based	0.777	0.709	0.448	0.549
	BERT	0.798	0.738	0.608	0.667

Table 4: Evaluation of RoBERTa and BERT models on the MWSA benchmark for English and German

4. Non-textual Linking Methods

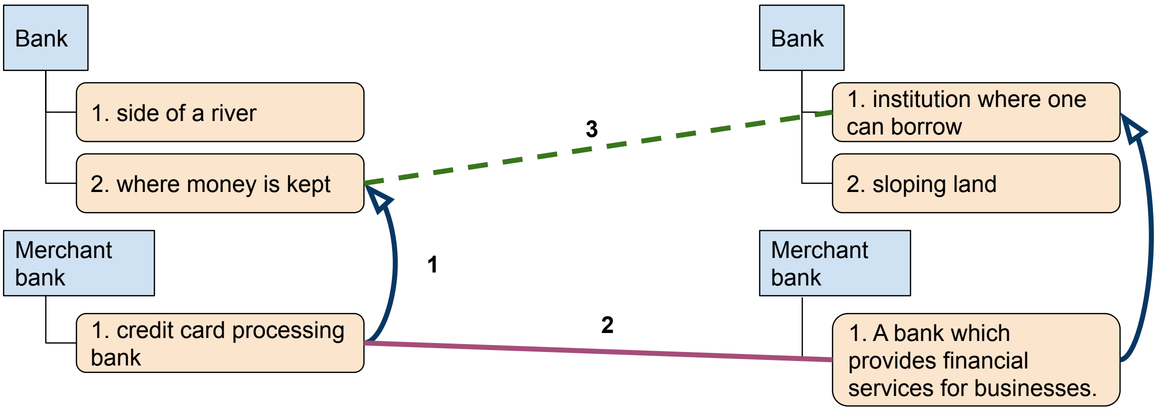

Bank
1. side of a river
2. where money is kept
Merchant bank
1. credit card processing bank
Bank
1. institution where one can borrow
2. sloping land
Merchant bank
1. A bank which provides financial services for businesses.
1
2
3
1

Figure 2: An example of the use of non-textual features for linking. Here the two senses of bank are distinguished by the hypernym links (1) and an inferred hapax legomenon link (2), so that the correct sense (3) can be selected.

In addition to using textual similarity methods, a number of non-textual methods can be used that are useful for linking dictionaries. There are two principal methods that can be used here: firstly, Naisc supports linking by means of property overlap, which creates a feature if two properties of a lexical entry are the same. These properties might be part-of-speech values or may be something more sophisticated such as register or other usage values. The second main method is graph-based similarity, which relies on there being a graph relating the senses of an entry and so is primarily used in the case of WordNet linking. Naisc implements the FastPPR method (Lofgren et al., 2014) to find graph similarity. In the case of wordnet linking, graph similarity cannot be naively applied as there are not generally links between the graphs of the two wordnets, instead we rely on the hapax legomenon links, which are links that are created when there is only one sense for the lemma in both dictionaries. These links allow us to create a graph between the two graphs, as shown in Figure 2. In another work (McCrae & Cillessen, 2021) we explored this method in the context of linking English WordNet (McCrae et al., 2019) with Wikidata, where we used the Naisc system to find equivalent senses of WordNet synsets and entities in the Wikidata database. In this paper, we found that 67,569 (55.3%) or

WordNet's synsets have a matching lemma in Wikidata, of which 16,452 (19.5%) counted as hapax legomenon links. We directly evaluated the accuracy of the hapax legomenon links and found the accuracy, when applying some simple filters, was 96.1% based on an evaluation of two annotators, who had a Cohen's kappa agreement of 81.4%. We then evaluated using the non-textual methods along with simple textual methods from the previous section and found that there was a 65-66% accuracy of the Naisc system in predicting links between WordNet and Wikidata. Divided by the prediction scores, those links predicted with a confidence of less than 60% by the system were all incorrect (0.0% accuracy), those with a 60-80% accuracy were correct 23/39 times (59.0% accuracy) and those with a greater than 80% confidence were correct 42/49 times (85.7% accuracy), indicating that the system's confidence was a good predictor of the accuracy of links.

5. Linking Constraints

Linking is a task that cannot only be achieved by looking at pairs of definitions by themselves but instead a **holistic** approach looks at all the links being generated and considers whether this leads to a good overall linking. It is clear that mapping multiple senses to the same senses or generating many more or fewer links than the number of senses is not ideal. In this section, we will look at the methods for solving the problem of sense linking holistically that are implemented in Naisc.

5.1 Bijection

The simplest constraint called **bijection** states that the senses for each dictionary entry should be marked as equivalent to at most one sense on the target side and that all senses should be linked for whichever dictionary entry has the fewest entries. This problem is known more generally as the **assignment problem** and can be formally stated for a set of source senses, $\{s_1, \ldots, s_n\}$ and target senses $\{t_1, \ldots, t_m\}$, an alignment, $A = \{a_{ij}\}$ is optimal given a score function, $s(a_{ij})$. If the following hold:

$$\begin{gathered}
\forall i \in \{1,\ldots,n\} \; \nexists j \in \{1,\ldots,m\}, j' \in \{1,\ldots,m\}, j \neq j' : a_{ij} \in A \wedge a_{ij'} \in A \\
\forall j \in \{1,\ldots,m\} \; \nexists i \in \{1,\ldots,n\}, i' \in \{1,\ldots,n\}, i \neq i' : a_{ij} \in A \wedge a_{i'j} \in A \\
\forall i \in \{1,\ldots,n\} \exists j \in \{1,\ldots,m\} a_{ij} \in A \text{ if } n \leq m \\
\forall j \in \{1,\ldots,m\} \exists i \in \{1,\ldots,n\} a_{ij} \in A \text{ if } m \leq n
\end{gathered}$$

We can weight this problem by assuming that the score is given by $\sum_{a_{ij} \in A} s(a_{ij})$ and this problem can be solved in cubic time by the Hungarian algorithm (Kuhn, 1955). To apply this we use the output probabilities from the classifiers described in the previous section and then:

$$s(a_{ij}) = \log p(a_{ij})$$

Given the high variance in the classifiers we normally further smooth this value as follows:

$$s(a_{ij}) = \log[p(a_{ij}) + \lambda]$$

Where $\lambda \simeq 0.5$. This allows the system to choose answers rejected by the classifier without an extreme penalty.

For the purpose of sense linking, the Hungarian algorithm is efficient as the problem can be divided into linking problems for each of the senses. However, for more complex cases the Hungarian algorithm can be very slow and so we have also investigated the use of approximate solvers, such as a simple greedy solver, a beam-search-based solver and the Monte-Carlo tree search algorithm (Chaslot et al., 2008).

5.2 b-Matching

WBbM, or b-matching, is one of the widely studied classical problems in combinatorial optimisation for modelling data management applications, e-commerce and resource allocation systems (Ahmed et al., 2017). WBbM is a variation of the weighted bipartite matching, also known as assignment problem. In the assignment problem, the optimal matching only contains one-to-one matching with the highest weight sum. This bijective mapping restriction is not realistic in the case of lexical resources where an entry may be linked to more than one entry. Therefore, WBbM aims at providing a more diversified matching where a node may be connected to a certain number of nodes.

Algorithm 1: Greedy WB*b*M

```
Input: $G = ((U, V), E, W)$, bounds $L$ and $B$
Output: $H = ((U, V)), E', W)$ satisfying bound constraints with a greedily-maximised score $\sum_{e \in E'} W(e)$
$E' = \emptyset$
Sort $E$ by descending W(e)
for $e$ to $E$ do
    if $H = ((U, V)), E' \cup \{e\}, W)$ does not violate $L$ and $B$ then
        $E' = E' \cup \{e\}$
return $H = ((U, V)), E', W)$
```

Algorithm 1 presents the WBbM algorithm with a greedy approach where an edge is selected under the condition that adding such an edge does not violate the lower and the upper bounds, i.e. L and B.

5.3 Taxonomic

The most typical case of sense linking consists of not only exact matches as considered in the bijective and b-matching case, but also broader, narrower and related links. As such we have investigated the use of a ‘taxonomic’ constraint that can be stated as follows:

- **Exact** links should be bijective (as defined above). Any sense that is the source or target of an exact link should not be the source or target of any other link.
- **Broader/narrower** links should be surjective/injective. This means that if a source sense is part of a broader link it may be part of other broader links, but the target sense cannot be the target of another broader link. Similarly the converse holds for narrower links.

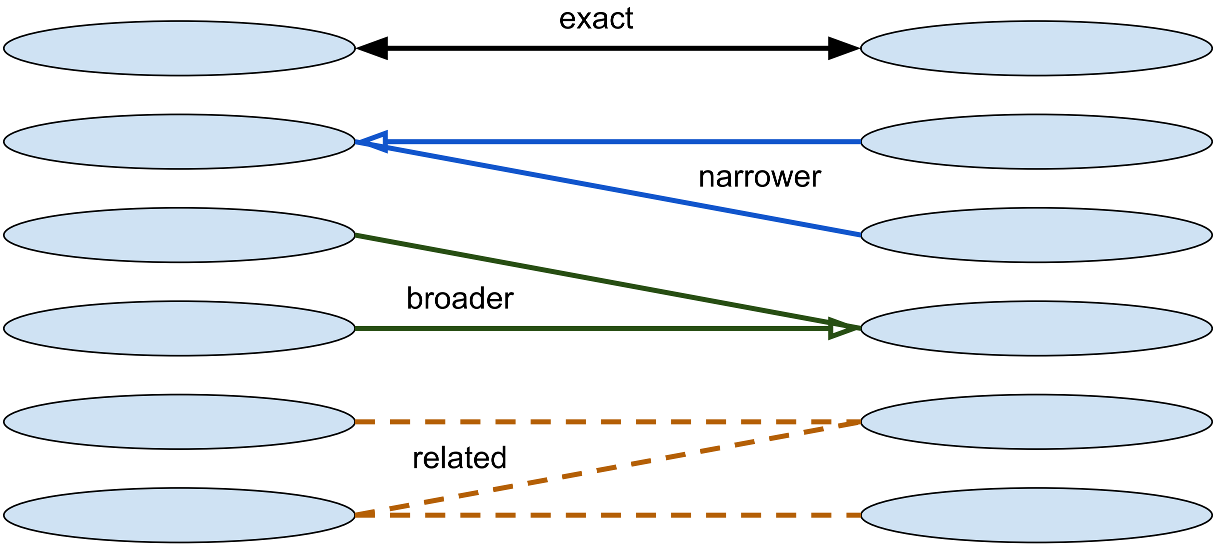


Figure 3: An example of a valid taxonomic linking according to the constraints. No further links could be added between any of the elements

- All link types are **exclusive**, that is if the source or target sense of any element is linked by one of the four relation types (exact, broader, narrower, related), then neither the source or target can be involved in a link of any other type.
- A **threshold** can be applied to ensure that only links of a certain quality are generated by the system.

An example of the links that are valid for these constraints is shown in Figure 3. With this more complex constraint, it is not clear whether there exists a polynomial-time algorithm to solve these constraints, and while, even for the small size of problems that are seen in sense linking, validating an optimal solution is not feasible, we have also observed that the greedy solver mostly returns the optimal or a near-optimal solution. As such, we simply rely on the approximate methods of linking, including the greedy solver, for this task.

6. Benchmarks and Shared Task

One major limitation regarding previous work was with respect to the nature of the data used for the WSA task. Expert-made resources, such as the Oxford English Dictionary, require much effort to create and therefore, are not as widely available as collaboratively-curated ones like Wiktionary due to copyright restrictions. On the other hand, the latter resources lack domain coverage and descriptive senses. To address this, we present a set of 17 datasets containing monolingual dictionaries in 15 languages, annotated by language experts within the ELEXIS volunteers and partners with five semantic relationships according to the simple knowledge organisation system reference (SKOS) (Miles & Bechhofer, 2009), namely, broader, narrower, related, exact and none.

The main goal of creating datasets for MWSA is to provide semantic relationships between two sets of senses for the same lemmas in two monolingual dictionaries. The actual annotation was implemented by means of dynamic spreadsheets that provide a simple but effective manner to complete the annotation. This also had the added advantage that the annotation task could be easily completed from any device. In order to collect the

Language	# Entries	# SKOS	# SKOS+`none`	# All
Basque	256	813	3661	4382
Bulgarian	1000	1976	3708	5656
Danish	587	1644	16520	18164
Dutch	161	622	20144	20766
English	684	1682	9269	10951
Estonian	684	1142	2316	3426
German	537	1211	4975	6185
Hungarian	143	949	15774	16716
Irish	680	975	2816	3763
Italian	207	592	2173	2758
Serbian	301	736	5808	6542
Slovenian	152	244	1100	1343
Spanish	351	1071	4898	5919
Portuguese	147	275	2062	2337
Russian	213	483	3376	3845

Table 5: Basic statistics of the datasets. # refers to the number

data that was required for the annotation, each of the participating institutes provided their data in some form providing the following:

- An entry identifier, that locates the entry in the resource
- A sense identifier marking the sense in the resource, for example the sense number
- The lemma of the entry
- The part-of-speech of the entry
- The sense text, including the definition

One of the challenges is that sense granularity between two dictionaries is rarely such that we would expect one-to-one mapping between the senses of an entry. In this respect, we followed a simple approach such as that in SKOS providing different kinds of linking predicates, which is described as follows:

Exact The senses are the same, for example the definitions are simply paraphrases.

Broader The sense in the first dictionary completely covers the meaning of the sense in the second dictionary and is applicable to further meanings.

Narrower The sense in the first dictionary is entirely covered by the sense of the second dictionary, which is applicable to further meanings.

Related There are cases when the senses may be equal but the definitions in both dictionaries differ in key aspects.

None There is no match for these senses.

While it is certainly not easy to decide which relationship is to be used, we found that this methodology was broadly effective, and we believe will simplify the development of machine-learning-based classifiers for sense alignment prediction. The datasets are available in JSON format and external keys such as meta_ID and external_ID enable future lexicographers to integrate the annotations in external resources. Given that some of the semantic relationships, such as narrower and broader, are not symmetric, sense_source and sense_target are important classes in determining the semantic relationship correctly.

Table 5 also provides basic statistics of the datasets such as number of entries and sense alignments. #Entries and #SKOS refer to the number of entries and senses with a relationship within SKOS. In addition, the senses within the two resources which belong to the same lemma but are not annotated with a SKOS relationship, are included with a *'none'* relationship.

Given that the datasets are publicly available, we carried out a shared task on the task of monolingual word sense alignment across dictionaries as part of the GLOBALEX 2020 – Linked Lexicography workshop at the 12th Language Resources and Evaluation Conference (LREC 2020) which took place on Tuesday, May 12, 2020 in Marseille (France).

7. Conclusions

In this paper, we have defined the monolingual word sense alignment task and a framework for solving this called Naisc. We looked at textual similarity and there are a large number of methods that are effective for estimating similarity, however the task of distinguishing between exactly equivalent senses and broader/narrower senses is still a challenging one. We then looked at non-textual linking methods that are effective for a few kinds of dictionary linking tasks, especially with large-scale knowledge graphs such as Wikidata. Finally, we examined the constraints that can be used to find the best overall linking between senses and showed how these can be solved. Further, we showed the development of a new benchmark and are working on the integration of all these tools into a single workflow that will form part of the ELEXIS dictionary infrastructure.

Acknowledgements

This work has received funding from the EU's Horizon 2020 Research and Innovation programme through the ELEXIS project under grant agreement No. 731015.

8. References

Agirre, E., Banea, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Mihalcea, R., Rigau, G. & Wiebe, J. (2016). SemEval-2016 Task 1: Semantic Textual Similarity, Monolingual and Cross-Lingual Evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. San Diego, California: Association for Computational Linguistics, pp. 497–511. URL https://www.aclweb.org/anthology/S16-1081.

Ahmadi, S. & McCrae, J.P. (2021). Monolingual Word Sense Alignment as a Classification Problem. In *Proceedings of the 11th Global Wordnet Conference*. pp. 73–80.

Ahmadi, S., McCrae, J.P., Nimb, S., Khan, F., Monachini, M., Pedersen, B.S., Declerck, T., Wissik, T., Bellandi, A., Pisani, I., Troelsgård, T., Olsen, S., Krek, S., Lipp, V., Váradi, T., Simon, L., Gyorffy, A., Tiberius, C., Schoonheim, T., Moshe, Y.B., Rudich, M., Ahmad, R.A., Lonke, D., Kovalenko, K., Langemets, M., Kallas, J., Dereza, O., Fransen, T., Cillessen, D., Lindemann, D., Alonso, M., Salgado, A., Sancho, J., Ureña-Ruiz, R., Zamorano, J.P., Simov, K., Osenova, P., Kancheva, Z., Radev, I., Stankovic, R., Perdih, A. & Gabrovsek, D. (2020). A Multilingual Evaluation Dataset for Monolingual Word Sense Alignment. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani,

H. Mazo, A. Moreno, J. Odijk & S. Piperidis (eds.) *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020.* European Language Resources Association, pp. 3232–3242. URL https://www.aclweb.org/anthology/2020.lrec-1.395/.

Ahmed, F., Dickerson, J.P. & Fuge, M. (2017). Diverse Weighted Bipartite b-Matching. In C. Sierra (ed.) *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017.* ijcai.org, pp. 35–41. URL https://doi.org/10.24963/ijcai.2017/6.

Chang, C. & Lin, C. (2011). LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3), pp. 27:1–27:27. URL https://doi.org/10.1145/1961189.1961199.

Chaslot, G., Bakkes, S., Szita, I. & Spronck, P. (2008). Monte-Carlo Tree Search: A New Framework for Game AI. In C. Darken & M. Mateas (eds.) *Proceedings of the Fourth Artificial Intelligence and Interactive Digital Entertainment Conference, October 22-24, 2008, Stanford, California, USA*. The AAAI Press. URL http://www.aaai.org/Library/AIIDE/2008/aiide08-036.php.

Cimiano, P., McCrae, J.P. & Buitelaar, P. (2016). Lexicon Model for Ontologies: Community Report. URL https://www.w3.org/2016/05/ontolex/.

Devlin, J., Chang, M.W., Lee, K. & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers).* Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. URL https://www.aclweb.org/anthology/N19-1423.

Fischer, A. & Igel, C. (2012). An Introduction to Restricted Boltzmann Machines. In L. Álvarez, M. Mejail, L.G. Déniz & J.C. Jacobo (eds.) *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications - 17th Iberoamerican Congress, CIARP 2012, Buenos Aires, Argentina, September 3-6, 2012. Proceedings*, volume 7441 of *Lecture Notes in Computer Science*. Springer, pp. 14–36. URL https://doi.org/10.1007/978-3-642-33275-3_2.

Hinton, G.E. (2012). A Practical Guide to Training Restricted Boltzmann Machines. In G. Montavon, G.B. Orr & K. Müller (eds.) *Neural Networks: Tricks of the Trade - Second Edition*, volume 7700 of *Lecture Notes in Computer Science*. Springer, pp. 599–619. URL https://doi.org/10.1007/978-3-642-35289-8_32.

Kuhn, H.W. (1955). The Hungarian Method for the Assignment Problem. In M. Jünger, T.M. Liebling, D. Naddef, G.L. Nemhauser, W.R. Pulleyblank, G. Reinelt, G. Rinaldi & L.A. Wolsey (eds.) *50 Years of Integer Programming 1958-2008 - From the Early Years to the State-of-the-Art*. Springer, pp. 29–47. URL https://doi.org/10.1007/978-3-540-68279-0_2.

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

Lofgren, P., Banerjee, S., Goel, A. & Comandur, S. (2014). FAST-PPR: scaling personalized pagerank estimation for large graphs. In S.A. Macskassy, C. Perlich, J. Leskovec, W. Wang & R. Ghani (eds.) *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*. ACM, pp. 1436–1445. URL https://doi.org/10.1145/2623330.2623745.

Matthews, B. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 405(2), pp. 442–451. URL https://www.sciencedirect.com/science/article/pii/0005279575901099.

McCrae, J.P., Bosque-Gil, J., Gracia, J., Buitelaar, P. & Cimiano, P. (2017). The OntoLex-Lemon Model: development and applications. In *Proceedings of eLex 2017*. pp. 587–597. URL https://elex.link/elex2017/wp-content/uploads/2017/09/paper36.pdf.

McCrae, J.P. & Buitelaar, P. (2018). Linking Datasets Using Semantic Textual Similarity. *Cybernetics and Information Technologies*, 18(1), pp. 109–123. URL http://www.cit.iit.bas.bg/CIT_2018/v-18-1/10_paper.pdf.

McCrae, J.P. & Cillessen, D. (2021). Towards a Linking between WordNet and Wikidata. In *Proceedings of the 11th Global Wordnet Conference*. University of South Africa (UNISA): Global Wordnet Association, pp. 252–257. URL https://www.aclweb.org/anthology/2021.gwc-1.29.

McCrae, J.P., Rademaker, A., Bond, F., Rudnicka, E. & Fellbaum, C. (2019). English WordNet 2019 – An Open-Source WordNet for English. In *Proceedings of the 10th Global Wordnet Conference*. Wroclaw, Poland: Global Wordnet Association, pp. 245–252. URL https://www.aclweb.org/anthology/2019.gwc-1.31.

Miles, A. & Bechhofer, S. (2009). SKOS Simple Knowledge Organization System Reference. W3C Recommendation, World Wide Web Consortium. URL https://www.w3.org/TR/skos-reference/.

Pennington, J., Socher, R. & Manning, C. (2014a). GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, pp. 1532–1543. URL https://www.aclweb.org/anthology/D14-1162.

Pennington, J., Socher, R. & Manning, C.D. (2014b). Glove: Global Vectors for Word Representation. In A. Moschitti, B. Pang & W. Daelemans (eds.) *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*. ACL, pp. 1532–1543. URL https://doi.org/10.3115/v1/d14-1162.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. & Polosukhin, I. (2017). Attention is All you Need. In I. Guyon, U. von Luxburg, S. Bengio, H.M. Wallach, R. Fergus, S.V.N. Vishwanathan & R. Garnett (eds.) *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. pp. 5998–6008. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Enriching a terminology for under-resourced languages using knowledge graphs

John P. McCrae[1], Atul Kr. Ojha[1], Bharathi Raja Chakravarthi[1], Ian Kelly[2], Patricia Buffini[2], Grace Tang[3], Eric Paquin[3], Manuel Locria[3]

[1]ADAPT Centre, Data Science Institute, NUI Galway, Ireland
[2] ADAPT Centre, Dublin City University, Ireland
[3] Translators without Borders
E-mail: john@mccr.ae, {atulkumar.ojha,bharathiraja.asokachakravarthi}@nuigalway.ie,
ian.anthony.kelly@gmail.com, patricia.buffini@adaptcentre.ie,
{grace,ericpaquin,manuel}@translatorswithoutborders.org

Abstract

Translated terminology for severely under-resourced languages is a vital tool for aid workers working in humanitarian crises. However there are generally no lexical resources that can be used for this purpose. Translators without Borders (TWB) is a non-profit whose goal is to help get vital information, including developing lexical resources for aid workers. In order to help with the resource construction, TWB has worked with the ADAPT Centre to develop tools to help with the development of their resources for crisis response. In particular, we have enriched these resources by linking with open lexical resources such as WordNet and Wikidata as well as the derivation of a novel extended corpus. In particular, this work has focused on the development of resources for languages useful for aid workers working with Rohingya refugees, namely, Rohingya, Chittagonian, Bengali and Burmese. These languages are all under-resourced and for Rohingya and Chittagonian there are only very limited major lexical resources available. For these languages, we have constructed some of the first corpora resources that will allow automatic construction of lexical resources. We have also used the Naisc tool for monolingual dictionary linking in order to connect the existing English parts of the lexical resources with information from WordNet and Wikidata and this has provided a wealth of extra information including images, alternative definitions, translations (in Bengali, Burmese and other languages) as well as many related terms that may guide TWB linguists and terminologists in the process of extending their resources. We have presented these results in an interface allowing the lexicographers to browse through the results extracted from the external resources and select those that they wish to include in their resource. We present results on the quality of the linking inferred by the Naisc system as well as qualitative analysis of the effectiveness of the tool in the development of the TWB glossaries.

Keywords: under-resourced languages; terminology; linking; natural language processing; knowledge graphs

1. Introduction

Terminology is a vital tool for aid workers in a wide range of crisis situations and the availability of a good-quality terminology in local languages is of vital importance. However, often these are severely under-resourced and so the development of language resources for these languages is significantly complicated. For example, after a devastating earthquake in Haiti in 2012, the natural language processing community rapidly developed tools and resources for the main language of Haiti, Haitian Creole, to help with the aid effort (Lewis, 2010). As such, the development of language resources for under-resourced languages is of critical importance and this is one of the main goals of the non-profit organisation, Translators without Borders (TWB).

The use of natural language processing technologies and existing open resources is a potentially huge benefit for the development of lexical resources for under-resourced languages, and, with this objective, we created a collaboration between the ADAPT Centre and TWB to develop tools to enrich the existing terminologies. For this collaboration, we focused on the work related to the Rohingya refugee crisis and as such the languages of relevance to this population, namely, Bengali, Burmese, Rohingya and Chittagonian. These languages vary in the availability of resources to being under-resourced languages but have significant online presence, namely Bengali and Burmese, which have large resources such as Wikipedia and support from language technologies such as Google Translate, to Rohingya and Chittagonian, which have nearly no resources or language tool support. Our strategy for expanding these resources was first to increase the corpus resources available for these languages so that we can train natural language processing tools on them. Secondly, we looked at linking them with open resources including WordNet (McCrae et al., 2020; Miller, 1995) and Wikidata so that extra information such as semantic relations, images and translations can easily be added into the glossaries. We examined some techniques for automatically finding candidates from these open resources using the Naisc (McCrae & Buitelaar, 2018) framework. We then have built this into a tool that allows terminologists to validate the data coming into the resources from external sources and thus semi-automatically extend this resource.

The rest of this paper is structured as follows, in Section 2 we lay out some related work and then we present the use case from Translators without Borders in Section 3. We then look at how we constructed the extended corpus in Section 4 and how we linked the existing glossaries with terms from open resources in Section 5. Finally, we show how we built a prototype for semi-automatic enrichment of the glossaries in Section 6 and finish with a conclusion in Section 7.

2. Related Work

As discussed in Section 1, unlike low-resourced languages, such as Bengali and Burmese, high-resourced languages, such as English and French are endowed with ample lexical and other linguistic resources such as WordNet, translated terminologies, corpora, and crowd-sourced resources such as Wikipedia or Wikidata.

Princeton WordNet (Miller, 1995) was the first WordNet which also formed the base for versions in all the other languages. Non-English languages gained focus in 1996 when EuroWordNet (Vossen, 1997) was founded to develop WordNets for several European languages giving way to a multilingual database.

When it comes to Asian WordNets the efforts started late, but significant milestones have been reached. In Asia, Indo-WordNet (Bhattacharyya, 2010) is a huge effort that was built in India to incorporate the major official Indian languages used in the Indian sub-continent, including Bengali. These languages were taken from three language families Indo-Aryan, Dravidian and Sino-Tibetan (Chakravarthi et al., 2018; Bhattacharyya, 2010). A few years ago, the University of Bangladesh (Rahit et al., 2018) also built the Bengali WordNet. Burmese WordNet[1] was developed on Open Multilingual WordNet (Bond & Paik, 2012; Bond & Foster, 2013). EuroWordNet, Indo-WordNet, Burmese and the recent Bangladeshi Bengali WordNet were built using an expand approach. However, Rohingya and Chittagonian do not have a WordNet or any lexical resources. While some effort has been made in the direct translation of WordNets into under-resourced languages (Chakravarthi et al., 2019), the results are still of poor overall quality. Similarly, some work has been done on the automatic development of terminologies for under-resourced languages (Pinnis et al., 2012; McCrae & Doyle, 2019).

Out of the various WordNets, Bengali and Burmese have large text corpora which can be scraped from Wikipedia, CURL (collecting Web Pages for Under-Resourced

[1] https://wordnet.burmese.sg/

Language) (Goldhahn et al., 2016) and An Crúbadán (Scannell, 2007). To the best of our knowledge, Rohingya and Chittagonian do not have any other existing corpora.

3. Use Case

The effectiveness of any aid program depends on delivering the correct information in the correct language. Historically, humanitarian agencies and aid workers have focused on maintaining capacity in major or "world" languages such as English, Spanish, and French. While these may constitute the "official" language of an affected country, they are often not used or well-comprehended by the affected populations. Furthermore, in humanitarian response, field workers must communicate important, sometimes life-saving information to those in need. In many cases, the critical link to ensuring affected people understand is the interpreter. However, too often, that link is broken, either because concepts do not translate well into the target language or because the interpreter does not have the tools to understand the concepts clearly.

TWB is addressing this problem by focusing on under-resourced local languages commonly used by marginalised populations in humanitarian crises. TWB's Glossaries, a critical real-time translation tool, assists front-line aid workers with an online repository of vetted, translated, simplified, and localised emergency-related terminology. It enables interpreters, cultural mediators, and any other field workers to access key concepts, terms, and phrases commonly used in crisis response. Themes include protection; housing, land, and property rights; and water, sanitation and hygiene (WASH). They were developed in collaboration with technical specialists and language partners.

TWB partnered with ADAPT to strengthen and expand TWB Glossaries, specifically the Bangladesh use case through a semantic uplift of the tools. ADAPT is a national research centre in Ireland focused on the digital media technology hosted at Trinity College Dublin and including seven other partner universities in Ireland. The main goal of the partnership was to increase the number of terms available in our glossaries and the discoverability of associated terms. We also used the collaboration to enhance the user experience and explorablity of the glossary content and the functionality e.g., keyword search, linked term review and approval, and search.

4. Corpus Building

We collected corpora from various sources for the target languages. Our target languages are from Bangladesh namely, Bangla (Bengali) (ISO 639-3 ben), Burmese (ISO 639-3 mya), Chittagonian (ISO 639-3 ctg) and Rohingya (ISO 639-3 rhg). All these languages are low-resourced languages.

Bengali is an Indo-European language spoken in Bangladesh and the West Bengal state of India and other places. Bengali is an agglutinative language and there are more than 150 different inflected forms of single verb root in Bengali. Presently, there are several dialects of Bengali that vary mainly in terms of the verb inflections and intonation. For this project, we downloaded the data for the Bangladesh version of Bengali language.

The Burmese language belongs to the Sino-Tibetan language family, it is the largest non-Chinese language from that Sino-Tibetan language family. It is the official language of the Republic Union of Myanmar and the native language of the Bamar people. The Myanmar script is an abugida system. It consists of 33 characters of standalone consonants, four dependent consonants, and tens of diacritic marks that represent vowels and tones. The orthography of Myanmar is generally syllable - based, although syllables may be merged in special writing forms. One word can be composed of multiple syllables and one syllable can be composed of multiple characters.

Chittagonian is an Indo-European language mainly spoken in the Chittagong Division in Bangladesh Country. Its sister languages include Sylheti, Rohingya, Chakma, Assamese, and Bengali. It is derived through an Eastern Middle Indo-Aryan from Old Indo-Aryan, and ultimately from Proto-Indo-European. Historically Arabic script was used for writing systems. The Bengali script is the most common script used nowadays.

Rohingya is also an Indo-European language spoken by Rohingya people of Rakhine State. The Hanifi Rohingya script is a unified script for the Rohingya language. Rohingya was first written in the 19th century with a version of the Perso-Arabic script.

We downloaded the data from CURL (Collecting Web Pages for Under-Resourced Languages) and WikiDump for Bengali and Burmese languages. For Chittagonia and Rohingya, there were no corpora available in CURL and WikiDump. However, we managed to collect the corpus for Rohingya from Rohingya Poems in Rohingyalish (Basu, 2014), Qur'an Foóila Síarah (Quran translation in Rohingya) and Rohingya Language (Mohammed & Ahmed, n.d.) books. After gathering the data, we cleaned the collected corpora following these steps:

- Removed HTML/file tags, metadata information, non-UTF/illegal characters, etc.
- Split it into one sentence per line
- Removed extra spaces and blank lines
- Removed duplicate sentences

We were able to collect 1,207,285 and 1,883 sentences for the Bengali and Burmese languages, respectively, from CURL. From WikiDump, 1,243,811 and 710,122 sentences were collected for Bengali and Burmese, respectively. From OPUS (http://opus.nlpl.eu/), we collected 681,789 sentences for Bengali and 962,654 sentences for Burmese. A total of 7,177 Rohingya sentences were extracted from books, while 5,100 Chittagonian sentences were extracted from Bible.is, Facebook and YouTube. Details of the of the corpus statistics are presented below:

Language	**Total sentences**	**Total words**
Bengali/Bangla	3,139,915	36,340,082
Burmese	1,674,659	21,568,615
Rohingya	7,177	206,089
Chittagonian	5,100	28,313

5. Linking

5.1 Objectives

The main goal of this project is to enrich the terminologies developed by TWB with the data found in resources such as WordNet and Wikidata. In order to do this, we need to establish which of the entities in these resources correspond to the terms found in these resources. This is not a trivial task as there are a large number of potential matches in general domain resources such as WordNet and Wiktionary, so it is not clear which of these resources would be a suitable match for which term. For example, the TWB glossary has highly generic terms such as ‘cut’, which is defined as “to injure a part of your body with something sharp that cuts the skin.” Similarly, WordNet has 41 verb senses for the word ‘cut’ and Wikidata has eight pages whose main label in ‘cut’. For WordNet, the most appropriate sense for ‘cut’ has the definition of “penetrate injuriously”, which is quite distinct from the definition give in the TWB terminologies. For Wikidata, none of the main definitions labelled as ‘cut’ are appropriate and the best match would actually be the page for ‘laceration’, it should also be noted that a complexity here is that as Wikidata is an encyclopaedic resource, all the concepts are nominal and so any link between these senses necessarily crosses part-of-speech boundaries. However, establishing a linking in a fully manual matter is likely very time-consuming and could be further helped by means of automatic linking tools.

In order to support automatic linking of tools, we have developed a toolkit called Naisc (McCrae & Buitelaar, 2018)[2], which acts as a toolkit for linking resources. This toolkit is designed for general purpose linking of datasets and is highly configurable, such that it can be used for a wide variety of linking tasks. In particular, we have focused a lot of work on the development of this tool for dictionary linking in the context of ELEXIS (Krek et al., 2018) infrastructure, which is developing a new infrastructure for electronic lexicography. As part of this infrastructure we envisage the development of a single large, interlinked matrix of dictionaries, which we refer to as the Dictionary Matrix. A key enabling technology for this is obviously automatic dictionary linking technology, and this is where the contribution of Naisc plays a key contribution to the ELEXIS infrastructure. As such, we have developed specialised modules for dictionary linking in Naisc, that we can also take advantage of for linking the TWB glossaries with WordNet and Wikidata.

5.2 Methodology

Naisc is a pipeline of processes which analyse two input datasets and outputs the set of links between them. This is done in a series of steps that analyse the datasets and find the best link between the elements of these datasets. The first step in this process is the **blocking step**, in which we find all potential matches between the two datasets, and as such the output of this step is superset of the final output, i.e., we can only output links that are identified at this step. As with all steps in Naisc, there are a number of different implementations that can be applied here, however in this case we restricted ourselves to only finding the elements in the target dataset (WordNet or Wikidata) for which we have a matching label. This means that we cannot find links such as ‘cut’ to ‘laceration’ described above. More exhaustive blocking strategies could be applied to find such links, however this can be computationally very expensive and lead to a large number of false positive results, so we did not attempt this here. The second step is called the **lens** step, where we analyse the input data in order to find text from each of the datasets that can be compared. In the case of this linking task, this step is fairly trivial as we only

[2] ’Naisc’ is pronounced ‘nashk’ and means ‘links’ in Irish, the software is open source and available at https://github.com/insight-centre/naisc

extract the definitions from both datasets, but it is easy to see how further information from a dictionary, such as examples or etymological information, could also be extracted and compared. The next step is then the **text feature** step, where we apply natural language processing techniques in order to estimate the similarity of the two pieces of texts. We have several methods implemented for this within the Naisc framework, but in the context of this paper we experimented with two sets of features, firstly a set of text similarity metrics based on surface characteristics, that is referred to as the 'basic string' features of Naisc. These are defined as follows:

Longest common subsequence This measures the largest number of consecutive characters in both strings.

Longest common prefix/suffix The number of characters that these two string share from the start/end of the strings.

Jaccard/Dice/Containment We measure the n-grams in each string in terms of both word n-grams and character n-grams and compare them using the standard methods of Jaccard, Dice and containment as defined below:
$\text{Jaccard} = \frac{|A \cap B|}{|A \cup B|}, \text{Dice} = \frac{2|AB|}{|A|+|B|}, \text{Containment} = \frac{|A \cap B|}{\min(|A|,|B|)}$

Sentence Length Ratio The relative length in words of the two inputs. This is symmetrised using the following formula:
$\text{SLR}(s,t) = 1 - \frac{min(|s|,|t|)}{max(|s|,|t|)}$

Average Word Length Ratio A comparison of the length of the words, symmetrized as above.

Negation A Boolean feature checking for the presence of negation keywords (such as 'not') in both or neither description.

Number Another Boolean feature comparing if all mentioned numbers match.

Jaro-Winkler, Levenshtein String similarity measures based on the edit distance between the strings as implemented by Apache Commons Text.

Monge-Elkan This metric (Monge & Elkan, 1997) uses Jaro-Winkler or Levenshtein as the base similarity function *sim* and is defined as:
$\text{ME}(s,t) = \frac{1}{|s|} \sum_{i=1}^{|s|} \max_{j=1,\dots t} sim(s_i, t_j)$

In addition, we use the Sentence-BERT model introduced by Reimers & Gurevych (2019), which produces a single vector to represent each of the definitions, we simply take the cosine of these vectors in order to estimate the similarity of the two sentences and this is used as a single feature.

The next step in the Naisc processing extracts features in parallel with the previous two steps and is referred to as the **graph feature** step. Both Wikidata and WordNet are complex graphs with many relations between the elements so we can take advantage of this to ensure that we are linking semantically similar terms. The TWB dataset did not have any links between its terms, however it did group these terms into domains and we created a graph over the TWB dataset by means of linking each term to a pseudo-node for the domain. In this way, we constructed a graph over the TWB dataset and this allows us to compare the graphs using a link prediction methodology. In particular, we take advantage of non-ambiguous nodes within the graph, that is terms which have a single sense in WordNet or in Wikidata, and use these to link the two graphs together. This creates a single graph over both the TWB data as well as the target dataset. We then applied the node proximity metric called *personalised page rank (PPR)* (Page et al., 1999) to score the likelihood of two terms being linked and in particular we used the FastPPR implementation (Lofgren et al., 2014).

The penultimate step of the algorithm starts with the prediction of the probability of a particular link by means of a **scorer**, which combines the features extracted from both the textual and graph analysis and converts them into a single score. We have explored two methods for this in the current work. Firstly an unsupervised methodology that works by means of micro-ranking the features. In particular, this feature works as follows: the values for each of the features are extracted and these are all ranked. Then we translate each feature value to its relative rank, such that, for example if a feature is the 100th highest value out of 1,000 values returned we would normalise its score to $1 - \frac{100}{1000} = 0.9$. Then we output the final score for each pair as the average of its normalised features. In addition, we used a supervised method, which is a support vector machine (Vapnik, 2000) as implemented by LibSVM (Chang & Lin, 2011).

The final step of the process, **matching**, is to find the most likely link between the terms in TWB and the target datasets. In this case, this is as simple as finding the highest scoring result for each element, however this would be substantially more complex if we also attempted to find multiple non-exact links, such as broader/narrower links. This is an active area of research, but our results in this task are not yet of a high enough quality to be reliable.

5.3 Evaluation

TWB Dataset	Target Dataset	Method	Precision	Recall	F-Measure
COVID	WordNet	Basic Unsupervised	82.30%	85.32%	83.78%
COVID	WordNet	Basic Supervised	79.65%	82.57%	81.08%
COVID	WordNet	BERT Unsupervised	87.61%	90.83%	89.19%
COVID	WordNet	BERT Supervised	87.61%	90.83%	89.19%
COVID	Wikidata	Basic Unsupervised	68.22%	84.88%	75.64%
COVID	Wikidata	Basic Supervised	68.22%	84.88%	75.64%
COVID	Wikidata	BERT Unsupervised	71.70%	88.37%	79.17%
COVID	Wikidata	BERT Supervised	71.70%	88.37%	79.17%
Bangladesh	WordNet	Basic Unsupervised	90.50%	90.53%	90.51%
Bangladesh	WordNet	Basic Supervised	81.05%	81.05%	81.05%
Bangladesh	WordNet	BERT Unsupervised	75.76%	75.79%	75.77%
Bangladesh	WordNet	BERT Supervised	75.76%	75.59%	75.67%
Bangladesh	Wikidata	Basic Unsupervised	84.79%	85.90%	85.34%
Bangladesh	Wikidata	Basic Supervised	70.01%	83.30%	76.08%
Bangladesh	Wikidata	BERT Unsupervised	76.42%	91.03%	83.09%
Bangladesh	Wikidata	BERT Supervised	76.42%	91.03%	83.09%
Average	Average	Unsupervised	81.45%	86.66%	83.82%
Average	Average	Supervised	74.73%	82.95%	78.46%
Average	Average	BERT	77.87%	86.51%	81.80%
Average	Average	BERT + Supervised	77.87%	86.46%	81.78%

Table 1: The results of the linking quality between the two datasets

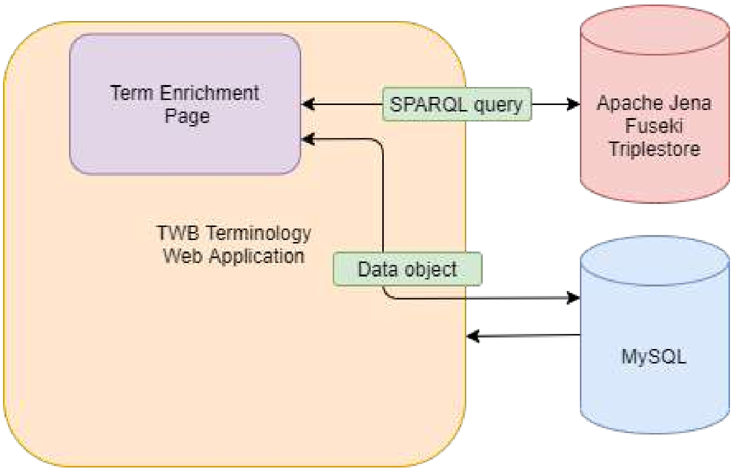

Term Enrichment Page
SPARQL query
Apache Jena Fuseki Triplestore
TWB Terminology Web Application
Data object
MySQL

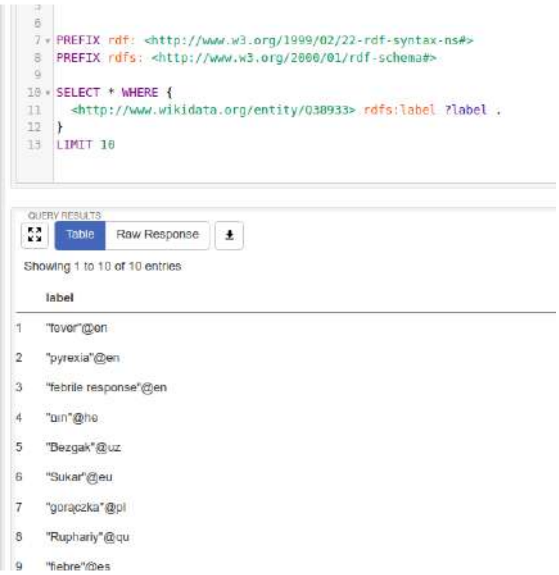
Figure 1: An overview of the enrichment system of the TWB terminology application

In order to evaluate the results of the linking we manually corrected some of the results of the Naisc linking in order to establish a partial gold standard linking. We then applied this to the four linking tasks which were based on the combination of TWB glossaries (on COVID and on Bangladesh) with the two target datasets (WordNet and Wikidata). We also tried four settings, based on whether we were using the 'basic' textual features of the BERT analysis and whether we were using the 'unsupervised' micro-ranking methodology or the 'supervised' SVM methodology; the results are presented in Table 1. Overall the results with all settings are quite strong with nearly four fifths of the links being correct automatically. Perhaps surprisingly the strongest overall system is the 'basic unsupervised' method. This is actually in line with our previous experience, where we have found that the supervised methodology does not fit well with the matching maximisation step, as it tends to predict probabilities that are close to zero or one, whereas the unsupervised method gives a good overall score to each element. Secondly, the use of Sentence-BERT while effective was not fine-tuned to the task and would have had challenges handling the short (and highly variable) nature of textual definitions. It is likely that further experiments could improve these results.

6. Terminology Enrichment

The Naisc linking output is a collection of Resource Description Framework (RDF) data in Turtle or N-Triple format files. These files were uploaded to a Jena Fuseki triplestore. The terminology enrichment (see Figure 1) was implemented into the existing Translators without Borders (TWB) terminology web application. An enrichment page is created for each term in the glossary and is constructed using the dynamic SPARQL Protocol and RDF Query Language (SPARQL) queries to the triplestore based on the ID of each term (see Figure 2).

Due to the open source nature of both Wikidata and WordNet, the results for each term may differ and as such the enrichment page needs to facilitate dynamic results. The SPARQL results are parsed and a page element is built for each returned data object. The Wikidata and WordNet results are separated and broken into sections based on result categories for visual clarity and ease of search. As an example of some of the extra information that would be available through this linking we take the example of the term 'vaccination'. The extra information is as follows:



```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT * WHERE {
  <http://www.wikidata.org/entity/Q38933> rdfs:label ?label .
}
LIMIT 10
```

Showing 1 to 10 of 10 entries

	label
1	"fever"@en
2	"pyrexia"@en
3	"febrile response"@en
4	"חום"@he
5	"Bezgak"@uz
6	"Sukar"@eu
7	"gorączka"@pl
8	"Ruphariy"@qu
9	"fiebre"@es

Figure 2: An example SPARQL database query for labels for a certain wikidata term.

- **From Corpora**
  - Examples found from the corpora developed in Section 4.
- **From WordNet**

  Alternative Definition : "taking a vaccine as a precaution against contracting a disease"

  Alternative Terms : inoculation

  Related Terms : immunization, immunisation, immunize, immunise, inoculate, vaccinate
- **From Wikidata**

  Alternative Definition : "administration of a vaccine to protect against disease"

  Alternative Terms : *(none for 'vaccination', example for 'treatment')* medical treatment, therapeutics, treating, intervention, therapy

  Related Terms : treatment, active immunotherapy, active immunity, antibody injection, vaccine, injection

  Translations : 'Impfung' *(German)*, 'vaccination' *(French)*, 'vacsaíniú' *(Irish)*, ... (about 100 languages)

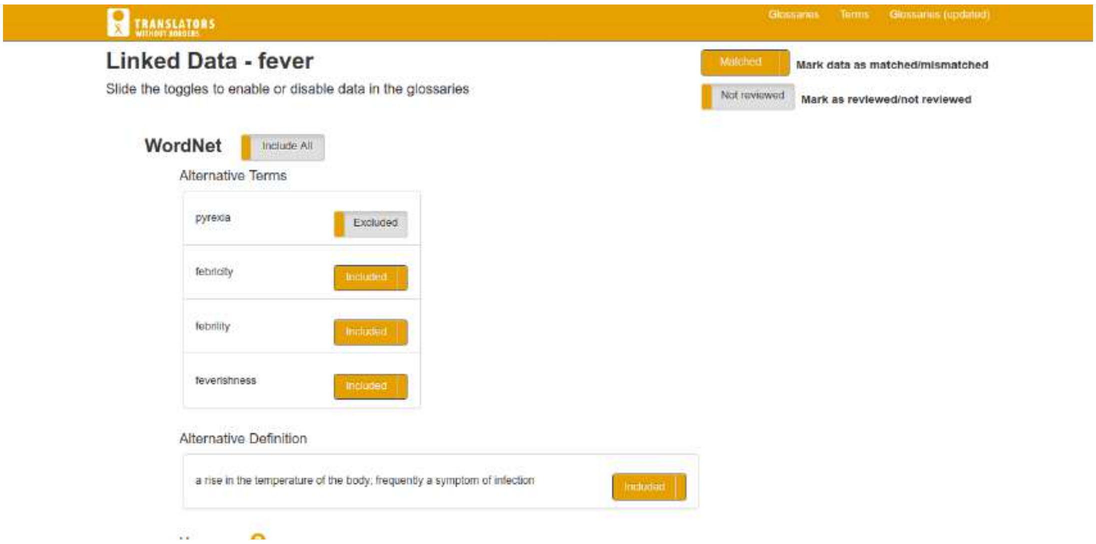
  Images : [3]

  Wikipedia Link : https://en.wikipedia.org/wiki/Vaccination (and other languages)

The processed data is then stored in a MySQL database for page load persistence and for use in the TWB glossaries. Each term can be included or excluded from the database using an accompanying slider, and includes additional sliders to allow for bulk inclusion and exclusion, indicating that a term has been reviewed, or that the data is mismatched. In the event of the linking generating incorrect term results, the matched slider allows

[3] Public domain image from https://en.wikipedia.org/wiki/File:Typhoid_inoculation2.jpg

for flagging of incorrect terms on the terminology term list page. Similarly, the reviewed slider allows for flagging that the term has been manually reviewed. A screenshot of the application is shown in Figure 3.



Linked Data - fever

Slide the toggles to enable or disable data in the glossaries

Matched — Mark data as matched/mismatched

Not reviewed — Mark as reviewed/not reviewed

WordNet — Include All

Alternative Terms

pyrexia — Excluded

febricity — Included

febrility — Included

feverishness — Included

Alternative Definition

a rise in the temperature of the body; frequently a symptom of infection — Included

Figure 3: An enrichment page for the term fever showing alternative terms as well as all sliders.

7. Conclusion

We have looked at how we can use terminological resources in order to extend a glossary of terms that are used by front-line aid workers. We examined the use case and saw how we could use open resources in order to improve the data that is available in the glossaries. We first looked at how we can compile a corpus to support these terms and found methods of finding corpus information from social media and other sources that were effective even though the languages were not well-documented. Then, we showed how we could link to Wikidata and WordNet and how to apply the Naisc framework to develop high-quality linking. We experimented with the use of machine learning and deep learning techniques here, but found that the main issues were related to finding suitable candidates in the open resources. We then developed this into a glossary tool that can be used to enrich the terminology and examined some of the extra kinds of data that can be added as the result of this analysis.

Acknowledgements

This work was supported by Science Foundation Ireland and co-funded by the European Regional Development Fund through the ADAPT Centre for Digital Content Technology [grant number 13/RC/2106].

8. References

Basu, E.M.S. (2014). *Rohingya Poems In Rohingyalish.* n.p.

Bhattacharyya, P. (2010). IndoWordNet. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner & D. Tapias (eds.) *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23*

May 2010, Valletta, Malta. European Language Resources Association. URL http://www.lrec-conf.org/proceedings/lrec2010/summaries/939.html.

Bond, F. & Foster, R. (2013). Linking and Extending an Open Multilingual Wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers*. The Association for Computer Linguistics, pp. 1352–1362. URL https://www.aclweb.org/anthology/P13-1133/.

Bond, F. & Paik, K. (2012). A Survey of WordNets and their Licenses. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*. Matsue. 64–71.

Chakravarthi, B.R., Arcan, M. & McCrae, J.P. (2018). Improving Wordnets for Under-Resourced Languages Using Machine Translation. In *Proceedings of the 9th Global Wordnet Conference*. Nanyang Technological University (NTU), Singapore: Global Wordnet Association, pp. 77–86. URL https://www.aclweb.org/anthology/2018.gwc-1.10.

Chakravarthi, B.R., Arcan, M. & McCrae, J.P. (2019). WordNet Gloss Translation for Under-resourced Languages using Multilingual Neural Machine Translation. In *Proceedings of the Second Workshop on Multilingualism at the Intersection of Knowledge Bases and Machine Translation*. Dublin, Ireland: European Association for Machine Translation, pp. 1–7. URL https://www.aclweb.org/anthology/W19-7101.

Chang, C. & Lin, C. (2011). LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3), pp. 27:1–27:27. URL https://doi.org/10.1145/1961189.1961199.

Goldhahn, D., Sumalvico, M. & Quasthoff, U. (2016). Corpus collection for under-resourced languages with more than one million speakers. *Proc. of Collabo ration and Computing for UnderResourced Languages: Towards an Alliance for Digital Language Diversity (CCURL)*, pp. 67–73.

Krek, S., McCrae, J., Kosem, I., Wissek, T., Tiberius, C., Navigli, R. & Pedersen, B.S. (2018). European Lexicographic Infrastructure (ELEXIS). In *Proceedings of the XVIII EURALEX International Congress on Lexicography in Global Contexts*. pp. 881–892. URL http://euralex.org/wp-content/themes/euralex/proceedings/Euralex%202018/118-4-2986-1-10-20180820.pdf.

Lewis, W. (2010). Haitian Creole: How to Build and Ship an MT Engine from Scratch in 4 days, 17 hours, & 30 minutes. In F. Yvon & V. Hansen (eds.) *Proceedings of the 14th Annual conference of the European Association for Machine Translation, EAMT 2010, Saint Raphaël, France, May 27-28, 2010*. European Association for Machine Translation. URL https://www.aclweb.org/anthology/2010.eamt-1.37/.

Lofgren, P., Banerjee, S., Goel, A. & Comandur, S. (2014). FAST-PPR: scaling personalized pagerank estimation for large graphs. In S.A. Macskassy, C. Perlich, J. Leskovec, W. Wang & R. Ghani (eds.) *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*. ACM, pp. 1436–1445. URL https://doi.org/10.1145/2623330.2623745.

McCrae, J.P. & Buitelaar, P. (2018). Linking Datasets Using Semantic Textual Similarity. *Cybernetics and Information Technologies*, 18(1), pp. 109–123. URL http://www.cit.iit.bas.bg/CIT_2018/v-18-1/10_paper.pdf.

McCrae, J.P. & Doyle, A. (2019). Adapting Term Recognition to an Under-Resourced Language: the Case of Irish. In *Proceedings of the Celtic Language Technology Workshop*. Dublin, Ireland: European Association for Machine Translation, pp. 48–57. URL https://www.aclweb.org/anthology/W19-6907.

McCrae, J.P., Rademaker, A., Rudnicka, E. & Bond, F. (2020). English WordNet 2020: Improving and Extending a WordNet for English using an Open-Source Methodology. In *Proceedings of the Multimodal Wordnets Workshop at LREC 2020*. pp. 14–19. URL https://lrec2020.lrec-conf.org/media/proceedings/Workshops/Books/MMW2020book.pdf#page=20.

Miller, G.A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, 38(11), pp. 39–41.

Mohammed, M. & Ahmed, R.M. (n.d.). *Rohingya Language Text Book 3*. n.p.

Monge, A.E. & Elkan, C. (1997). An Efficient Domain-Independent Algorithm for Detecting Approximately Duplicate Database Records. In *Workshop on Research Issues on Data Mining and Knowledge Discovery, DMKD 1997 in cooperation with ACM SIGMOD'97, Tucson, Arizona, USA, May 11, 1997*.

Page, L., Brin, S., Motwani, R. & Winograd, T. (1999). The PageRank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.

Pinnis, M., Ljubešic, N., Stefanescu, D., Skadina, I., Tadic, M. & Gornostay, T. (2012). Term extraction, tagging, and mapping tools for under-resourced languages. In *Proceedings of the 10th Conference on Terminology and Knowledge Engineering (TKE 2012), June*. pp. 20–21.

Rahit, K.T.H., Hasan, K.T., Al-Amin, M. & Ahmed, Z. (2018). BanglaNet: Towards a WordNet for Bengali Language. In *Proceedings of the 9th Global Wordnet Conference*. pp. 1–9.

Reimers, N. & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In K. Inui, J. Jiang, V. Ng & X. Wan (eds.) *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*. Association for Computational Linguistics, pp. 3980–3990. URL https://doi.org/10.18653/v1/D19-1410.

Scannell, K.P. (2007). The Crúbadán Project: Corpus building for under-resourced languages. In *Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop*, volume 4. pp. 5–15.

Vapnik, V.N. (2000). *The Nature of Statistical Learning Theory, Second Edition*. Statistics for Engineering and Information Science. Springer.

Vossen, P. (1997). EuroWordNet: a multilingual database for information retrieval. In *Proceedings of the DELOS workshop on Cross-language Information Retrieval, March 5-7, 1997 Zurich*. Vrije Universiteit.

From term extraction to lemma selection for an electronic LSP-dictionary in the field of mathematics

Theresa Kruse[1], **Ulrich Heid**[1]

[1]Institute for Information Science and Natural Language Processing (IwiSt), University of Hildesheim,
Universitätsplatz 1, 31141 Hildesheim, Germany
E-mail: theresa.kruse@uni-hildesheim.de, ulrich.heid@uni-hildesheim.de

Abstract

We work on term extraction for a corpus-based LSP-dictionary. Our field of study is the mathematical domain of graph theory. Our working hypothesis is that mathematics lends itself to a specific approach for term and information extraction with a lexicographical purpose. We compare different methods for term extraction: The first one combines pattern-based and statistical mean implemented by Schäfer et al. (2015), the second one has been developed especially for mathematical texts using domain-specific definition patterns based on work in the tradition of Meyer (2001). Further comparisons are made with a list of term candidates which are not part of the general language lexicon used in a version of TreeTagger trained on news text (Schmid, 1994) and with the term extraction provided by Sketch Engine (Kilgarriff et al., 2014). We use manual annotation by three expert raters and inter-rater agreement with κ-statistics to compare and evaluate the approaches. Additionally, we qualitatively analyse the extracted results. For selecting the lemmas, we work with a German corpus of lecture notes, textbooks and papers.

Keywords: LSP-dictionaries; mathematics; pattern-based extraction; automatic creation; semantic relation

1. Introduction

Our work on term extraction and lemma selection evolved as part of a project on creating an online LSP-dictionary[1] covering the domain of graph theory, a part of mathematics. Its target group are students. The dictionary is based on scientific and didactic literature from the domain: textbooks, course material and specialised publications. It is intended to cover the central terminology of graph theory as well as items from other mathematical domains which are needed to understand graph-theoretical literature. The dictionary will have an ontology as its backbone and will give equivalents in German and English, as well as definitions and semantically related terms. Most of these relations correspond to lexical semantic relations known from linguistics, such as hyperonymy, only some relations are domain-specific.

Our working hypothesis is that we can rely exclusively on (definitional) patterns to extract terms from the graph-theoretical texts and that we do not need statistical approaches, because mathematical texts contain highly standardised definitions.

In Section 2, we give a short overview of methods for term extraction. We compare different methods for term extraction to investigate our hypothesis: A rather traditional pattern-based one combined with statistics as described by Schäfer et al. (2015) and one that only relies on domain-specific patterns in the tradition of Meyer (2001). We extract a list of term candidates with these tools and add items from the corpus which are not part of the general language lexicon used in a version of TreeTagger trained on news text (Schmid, 1994).

Three expert raters decided in two rounds which of these candidates should become lemmas of the dictionary. We present the results of this rating in Section 3. As a consensus on refined guidelines for lemma selection preceded the second selection round, we also use

[1] LSP stands for Language for Special Purposes.

the resulting lemma list to evaluate the contribution of each term extraction method to the creation of the lemma list, i.e. the results of the rating are used to evaluate the different methods in Section 4. We are aware of the methodological problem that lies in the bias towards the tool output, because we may systematically miss lemma candidates not found by any of the approaches.

A further comparison is made with the term extraction provided by *Sketch Engine* (Kilgarriff et al., 2014). Section 5 brings together the results of the evaluations. We conclude in Section 6.

2. Related work on term extraction

Different approaches to term extraction appeared over the last 20 years: Cabre & Vivaldi Palatresi (2013) give an overview of the state of the art of around 2010 and distinguish linguistic, statistical and hybrid methods. An overview of current experiments based on Machine Learning (ML) can be found e.g. in Hätty (2020), while Hätty herself combines different traditional as well as ML approaches.

Cabre & Vivaldi Palatresi (2013) name three criteria for terms: unithood, termhood and specialised usage. Unithood and termhood are also common benchmarks in evaluating the results of automated term extraction (cf. Zadeh & Handschuh, 2014). *Termhood* is the extent to which a candidate is actually a term. *Unithood* is a measure of the association between different components of a multiword term candidate and is thus similar to some measures of collocational strength.

Extraction tools have to find single word terms (SWT) as well as multiword terms (MWT). Cabre & Vivaldi Palatresi (2013) indicate frequency counts, frequency comparison and pattern search as the main methods for extracting SWT and linguistically based pattern search, keyword-in-context and statistical techniques for MWT. All methods may be combined.

Frequency comparisons include contrastive approaches in which the frequency of each candidate in the specialised text is compared to a reference corpus from general language. Several measurements exist for such a comparison: e.g. frequency profiling (Rayson & Garside, 2000), the C-NC value (Bonin et al., 2010) or the modified weirdness measure (Kochetkova, 2015). *Sketch Engine* (Kilgarriff et al., 2014) also uses a frequency comparison technique.

Often, a scoring or ranking of the results follows the extraction itself. Depending on the domain, terminologies may exist as a reference for evaluation. Especially when working on variants this constitutes a useful approach (cf. Zadeh & Handschuh, 2014). Bernier-Colborne (2012) introduce a method for creating a gold standard from a corpus which may also be used for these purposes.

Recent automatic term extraction uses Machine Learning (ML). Different approaches have been developed in recent years (Rigouts Terryn et al., 2020). We give some examples in the following.

Dobrov & Loukachevitch (2011) combine frequencies from domain-specific texts and search engines with a domain-specific thesaurus. Conrado et al. (2013) combine multiple features like term frequency, part of speech and context for their ML-based extraction. Fedorenko et al. (2014) compare term extraction based on ML using different features with voting algorithms and conclude that the ML-methods outperform the others.

It has been shown that word embeddings are also helpful for term extraction. Amjadian et al. (2016) use distributed vectors based on the regression model GloVe (Pennington et al., 2014), which constitutes a step towards language independent term extraction and combines linguistic and statistical approaches. They also evaluate their method on mathematical texts, namely five English high school textbooks. They do not indicate any difficulties that would be due to the domain. Their distributed vectors work best as a filter and not directly applied to a corpus (Amjadian et al., 2018).

Wang et al. (2016) also use word embeddings with a focus on reducing the amount of labelled data. Therefore, they use co-training (Blum & Mitchell, 1998): First, only a part of the data is labelled and the most probable labels are taken into consideration. The tool works iteratively this way.

Some term extraction tools were especially developed for lexicographic purposes, such as the *Sketch Engine* term extraction or the procedure used by Heid & Weller (2010) based on dependency parsing to extract MWT. *Sketch Engine* (Kilgarriff et al., 2014; Jakubíček et al., 2014) annotates with the RFTagger (Schmid & Laws, 2008) for a pattern-based term extraction. We use this tool on our data in Section 5.4 to have another comparison. Pollak et al. (2019) present a different approach for lexicography which combines frequency methods with word embeddings.

One of the tools tested in our lemma selection experiments is the term extractor implemented by Schäfer et al. (2015) in line with the traditional hybrid approach (cf. also Roesiger et al., 2016). Schäfer et al. (2015) focus on adjectives and nouns and implement three steps: First, they select nominal candidates by part-of-speech tagging; secondly, they take the syntactic validity of noun phrases into account and thirdly, they use statistical measures. They extract the following POS-patterns based on regular expressions, where *N* is the POS tag noun, *Adj* adjective, *P* preposition, *Adv* adverb and *D* determiner:

- (Adv? Adj? Adj)? N
- (N D)? (Adv? Adj)? N P D? (Adv? Adj)? N
- (Adv? Adj)? N D (Adv? Adj)? $\text{N}_{genitive}$

For removing noise they use the *c*-value score (Frantzi & Ananiadou, 1996) and combine constituency and dependency parsing (Bohnet, 2010; Choi et al., 2015; Roesiger et al., 2016). The *c*-value is an established domain-independent (Frantzi et al., 2000) measure for ranking extracted terms based on frequency and on the usage of an item in MWTs. Schäfer et al. (2015) evaluate their tool on texts from the domain of do-it-yourself projects and get an F-score of 0.59 with a precision of 0.48 and a recall of 0.77. We present our results with this tool in Section 4.1.

3. Extracting and categorizing the lemmas

3.1 Expert raters

We work with a corpus of German lecture notes, textbooks and papers from the mathematical sub-domain of graph theory. It contains 882,910 tokens with 31,106 types.

We extract a list of 4205 lemma candidates from it and give it to three expert raters. Section 4 describes the process of selecting the terms in the list.

The classification consists of two steps: First, the raters work individually and independently. We analyse their results and make out systematic differences and disagreement concerning lexicographic and linguistic aspects. In a subsequent adjudication step, the raters discuss the (types of) phenomena which led to their divergent classifications. Finally, we ask them to agree on common guidelines for these cases.

The three expert raters come from different backgrounds in graph theory. All of them have studied mathematics, have didactic experience in mathematics and work with academic graph theory from different perspectives. In the first selection round, we simply ask them to decide for each candidate whether it should be given lemma status in the planned dictionary: „Bitte beantworten Sie für jeden Begriff in der Liste die folgende Frage: Soll es im geplanten elektronischen Wörterbuch einen Eintrag zu diesem Lemma geben?“[2]. We also ask them to propose further terms and to comment on their choices in cases of uncertainty.

All raters are familiar with the idea of the project to create an electronic dictionary for the domain of graph theory which can be used by students. One of the raters is aware of the semantic category system which is used on the lemma list at a later point in the lexicographic process.

	individual classification		after discussion	
	number of terms	percentage	number of terms	percentage
3 votes	383	9.11 %	1077	25.64 %
2 votes	783	18.62 %	376	8.94 %
1 vote	897	21.33 %	334	7.94 %
0 votes	2142	50.94 %	2417	57.48 %

Table 1: Results of the expert raters

Table 1 shows the results of the individual classification. The raters consider only about half of the extracted items as useful for the dictionary. In the later sections we investigate the reasons for the low quality of the extraction tools.

We calculate the inter-rater agreement with κ-statistics (Fleiss, 1971) and get $\kappa = 0.3484$. The agreement within the categories is $\kappa_{\text{in}} = 0.3489$ and $\kappa_{\text{out}} = 0.3479$. A pairwise comparison between the raters is provided in Table 2. The agreement in this first round is only *fair* or at most *moderate*, in terms of the terminology proposed by Landis & Koch (1977). This result confirms the observation made by Hätty (2020) that intuitive notions of termhood vary considerately between individual raters; this also seems to be the case with experts from the same domain.

We subsequently initiate the adjudication discussion mentioned above to understand the raters’ reasoning underlying their decisions, and to jointly develop refined guidelines for lemma selection.

[2] Engl.: For each term in the list, please answer the following question: Should there be an entry for this lemma in the planned electronic dictionary?

		Rater 1		
		in	out	
Rater 2	in	0.2499	0.0404	0.2903
	out	0.1960	0.5137	0.7097
		0.4459	0.5541	

$\kappa = 0.5047$

		Rater 2		
		in	out	
Rater 3	in	0.0968	0.0259	0.1227
	out	0.1936	0.6837	0.8773
		0.2904	0.7096	

$\kappa = 0.3578$

		Rater 1		
		in	out	
Rater 3	in	0.1127	0.01	0.1227
	out	0.3332	0.5441	0.8773
		0.4459	0.5541	

$\kappa = 0.2526$

Table 2: Agreement between raters

Among other case-by-case decisions, the following aspects seem to be crucial reasons for different classifications: First, the degree to which the translation between German and English is considered as difficult for the intended public of the dictionary. We work on a bilingual dictionary containing equivalents as well as onomasiological and definitional information on the terms. The annotators have a different focus on these aspects and therefore terms like *Satz von Petersen* (Engl. *Petersen's theorem*) are excluded by one rater because the students should have no difficulties in translating them. A similar reasoning holds for certain compound terms. In the discussion, the raters decide to include these terms as they belong to a given conceptual category in the final dictionary (Kruse & Heid, 2020).

A second difficulty is common mathematical terminology which is not particularly typical for the sub-domain of graph theory, such as terms referring to set theory. This issue is an instance of the more general problem of the delimitation of (sub-)domains in terminology, as addressed e.g. in the model of Roelcke (2010) of intra-subject vs. inter-subject terminology (*intrafachlicher* vs. *interfachlicher Fachwortschatz*). Some annotators include these terms because they are basic for anyone learning graph theory, and others exclude them because they are not specific of the sub-domain. Hence, we add a category for these general terms to our classification system (cf. Section 3.2).

The third main aspect that leads to differences among the raters are term variants. We already gave an overview on variants of our domain in Kruse & Giacomini (2019). Two annotators decide to only include one (primary) variant into the lemma list of the dictionary. After the discussion, they include all variants into the lemma list. Possibly, some of them will appear in the dictionary as cross-reference entries, i.e. as links to another variant.

Another issue is the handling of mathematical symbols. The raters decide to exclude them because the symbols need a verbalisation which requires some further, possibly very specific, lexicographic devices.

The raters decide to include compounds of variables and words like *2-regulär* only with the most common abbreviation, like *k-regulär*. Only few exceptions are made for terms which have a special significance in graph theory, e.g. *2-dimensional*. We cannot treat these cases like variants because on a semantic level they at most hypernyms. For example, *2-regulär* is a special case of *k-regulär* and some properties are valid for only certain values of k. Therefore, they cannot be treated on the same level.

Another discussion point are combinations of terms with words from general language like *Anzahl an...* (Engl. *number of...*). In these cases, the raters decide to only include the terminological parts as long as the added word is terminologically irrelevant. Otherwise, obviously the whole term is included, as is the case with *Kuratowskimenge* (Engl. *Kuratowski set*).

Further, the raters discuss which combinations are considered as a MWT. One example are combinations with *maximal* and *minimal*. Mostly, these combinations are not terminologically relevant, but there are specific exceptions, e.g. *maximal Matching* which is a lot more used than *minimal Matching*. Thus, these decisions are made on a case-by-case basis.

It is also very common in mathematics to have negated compounds with *nicht-* (Engl. *not-*) and *-frei* (Engl. *-free*). If one knows the other part, they are self-explanatory and therefore not included in the dictionary, but their positive counterparts will be.

		Rater 1		
		in	out	
Rater 3	in	0.1085	0.0040	0.1173
	out	0.2627	0.62	0.8827
		0.3712	0.6288	

$\kappa = 0.7145$

		Rater 1		
		in	out	
Rater 2	in	0.2133	0.041	0.2543
	out	0.1579	0.5878	0.7457
		0.3712	0.6288	

$\kappa = 0.7909$

		Rater 2		
		in	out	
Rater 3	in	0.0843	0.033	0.1173
	out	0.17	0.7127	0.8827
		0.2543	0.7457	

$\kappa = 0.7476$

Table 3: Rater results after discussion

After re-annotating the data and taking the results of the discussion into account we get $\kappa = 0.7500$. Table 1 gives the results of this second step and Table 3 the pairwise comparison between the raters. We include the candidates with at least two votes in the dictionary, and thus our final lemma list contains 1,453 lemmas.

Thus, overall, the adjudication process was also a process of refining the lexicographic lemma selection principles, and it was massively dependent on the peculiarities of the domain and on the specialised vocabulary to be dealt with, but also on decisions concerning a homogeneous lexicographic treatment of certain classes of items. Nevertheless, we have to admit that the selection remains partly random because the raters' prompt does not give clear criteria and can be individually interpreted, as the discussion has shown. It might be useful to use this criteria for another annotation with new raters to get more generalisable results.

3.2 Categorisation

We manually assign the chosen terms to the following categories: ALGORITHM, MAPPING, PART (of a graph), PERSON, PROBLEM, THEOREM, TYPE (of a graph), PROPERTY (of a graph), ACTIVITY and GENERAL. Kruse & Heid (2020) provide a detailed description of these categories, except for GENERAL which is the category mentioned above containing all the general mathematical terms which are a prerequisite to but no direct part of graph theory. In the final dictionary the category of each item defines the microstructure of its entry.

Category	Number	Percentage
Part	411	28.29 %
Property	263	18.10 %
Type	162	11.15 %
General	153	10.53 %
Theorem	146	10.05 %
Mapping	128	8.81 %
Person	89	6.13 %
Algorithm	58	3.99 %
Problem	35	2.24 %
Activity	8	0.55 %

Table 4: Distribution of lemmas over categories

Table 4 shows the distribution of the 1,453 lemmas over the ten categories. Almost a third of the lemmas belongs to the category PART (of a graph), followed by PROPERTY (of a graph) and TYPE (of a graph). These three constitute the majority of the concepts used in graph theory and in mathematics in general, as one has certain objects (PARTS and TYPES) for which PROPERTIES are defined.

4. Term extraction

In the following, we present our methods for term extraction. One has to keep in mind that our results are biased because the raters could only decide upon the extracted terms, not on an independent list. Nevertheless, they had the opportunity to add terms to the list on their own. We choose this workflow because there were no capacities for our raters to annotate the whole corpus of almost 900,000 tokens for establishing an independent gold standard.

4.1 Combination of frequencies and patterns

We extract 2,416 potential lemmas with the method by Schäfer et al. (2015). In the following, we refer to this method as the T-method. We remove candidates from the list which result from noise in the corpus data, e.g. because of formatting fragments of formulas like *IJI-IJI*. 2,229 (92.26 %) lemma candidates remain. Only then did the raters receive the list. For precision and recall we calculate with this figure.

We use the 1,453 lemmas retained in the selection process from Section 3 as a gold standard for calculating precision p, recall r and F-score F. We can do that because we asked the raters to name further terms which they would like to include into the dictionary, and they did not give any. 643 candidates in the T-list got a vote by at least two raters.

$p_T = \frac{643}{2229} = 0.2885, r_T = \frac{643}{1454} = 0.4422, F_T = 0.3492$

In their paper Schäfer et al. (2015) get $p = 0.48$, $r = 0.77$ and $F = 0.59$, which is higher than in our experiment. Nevertheless, their data is not completely comparable with ours, because our data contains lemmas which might be terminological for mathematics but not in our specific sub-domain of graph theory.

4.2 Domain-specific patterns

The second extraction method is based on the hypothesis that we do not need any frequency measurements for term extraction in mathematics because the language is highly structured. Thus, we solely use domain-specific patterns. We call this method the P-method and identify the following words as pattern indicators: *bestehen aus, bezeichnen, definieren, erklären, haben, heißen, sein, Name, nennen, sagen, schreiben, sprechen, verstehen*[3]. The P-method returns $3,071$ lemma candidates.

We carry out the same adjustments as described in Section 4.1 before we give the list to the raters. 1,797 (58.52 %) of the candidates remain after the adjustments. The raters give 506 of the remaining terms at least two votes. This percentage of potential useful lemmas is lower than what we got with the T-method. This is maybe due to the fact that we did not include any measures of frequency. We get the following results for precision, recall and F-score:

$p_P = \frac{506}{3072} = 0.1647, r_P = \frac{506}{1454} = 0.3480, F_P = 0.2236$

These values are also lower than those of the T-method. There are some possible reasons for that which we examine in Section 5.

4.3 Comparison with unknowns

The candidate list for the raters combines the terms extracted by the two methods described in the previous sections. The list is supplemented with data generated during the

[3] Engl.: *consist of*, *denote*, *define*, *explain*, *have*, *be called*, *be*, *name*, *called*, *say*, *write*, *speak of*, *understand*

correction process of the corpus. It contains words which were labeled as unknown by the TreeTagger (Schmid, 1994), trained on general language data (news text). We refer to this list as the U-list; and it contains 1478 potential terms. As the tagger operates on single word forms, only SWT appear on the list, including compounds like *(k+1)-elementig* (Engl. *(k+1)-element*) or *nicht-planar* (Engl. *non-planar*). We also calculate precision, recall and F-score to compare with the other methods.

$p_U = \frac{830}{1478} = 0.5616, r_U = \frac{830}{1454} = 0.5708, F_U = 0.5662$

These values are much higher than those obtained with the other methods because the U-list is not produced by means of data extraction, but through manual additions to the tagger lexicon. Thus, it can only be regarded as a sort of baseline with the downside that it does not contain any graph-theoretical terms that are polysemous with general language words (e.g. *Kante*, Engl. *edge* or *Ecke*, Engl. *node*).

5. Comparison of different methods

We see that the T-method produces less noise than the P-method because the T-method also includes a frequency measure whereas the other one does not. A pattern-based method works best on absolutely clean data, but formulas and abbreviations in mathematical texts lead to noise. Our corpus consists of sources with different formatting and file types, and we did not have the workforce to establish the same formatting for all texts. This has to be considered when working with mathematical texts, especially when they are combined from different sources.

The U-list has the best values, but here only SWT were included, and a lot of noise has been removed beforehand. Therefore, it can only serve as a reference. When using it for the lemma selection, it might be useful to include frequency figures and to only take lemmas with at least two mentions into consideration to improve the results of the P-method.

5.1 Comparison based on frequency

The Jaccard index J is a measure to determine how similar certain sets are (Jaccard, 1902). It is defined the following way for a number of n sets $A_1, ..., A_n$:

$$J(A_1, ..., A_n) := \frac{|A_1 \cap ... \cap A_n|}{|A_1 \cup ... \cup A_n|}$$

The Jaccard index takes values between $J = 0$ (if and only if $A_1 \cap ... \cap A_n = \emptyset$) and $J = 1$ (if and only if $A_1 = ... = A_n$). We have three sets: In P are the terms extracted by the pattern-based method, T gives the extracted terms with the tool by Schäfer et al. (2015) and U comprises the list of unknown words based on the lexicon by Schmid (1994). First, we take into account all the terms extracted:

$\lvert P\rvert = 3071$	$\lvert P \cap T\rvert = 596$	$\lvert P \cup T\rvert = 4712$	$J(P,T) = 0.1265$
$\lvert T\rvert = 2237$	$\lvert P \cap U\rvert = 240$	$\lvert P \cup U\rvert = 4308$	$J(P,U) = 0.0557$
$\lvert U\rvert = 1477$	$\lvert T \cap U\rvert = 262$	$\lvert T \cup U\rvert = 3452$	$J(T,U) = 0.0759$
	$\lvert P \cap T \cap U\rvert = 113$	$\lvert P \cup T \cup U\rvert = 5800$	$J(P,T,U) = 0.0195$

As these values are based on noisy data, it is preferable to compare only the terms which were finally chosen for the dictionary. However, still the values show no particularly high agreement between the sets:

$$\begin{array}{llll}
|P_s| = 506 & |P_s \cap T_s| = 234 & |P_s \cup T_s| = 913 & J(P_s, T_s) = 0.2563 \\
|T_s| = 641 & |P_s \cap U_s| = 178 & |P_s \cup U_s| = 1157 & J(P_s, U_s) = 0.1538 \\
|U_s| = 829 & |T_s \cap U_s| = 200 & |T_s \cup U_s| = 1270 & J(T_s, U_s) = 0.1575 \\
 & |P_s \cap T_s \cap U_s| = 89 & |P_s \cup T_s \cup U_s| = 1453 & J(P_s, T_s, U_s) = 0.0613
\end{array}$$

Another interesting set are those terms which are chosen for the final dictionary but only extracted by one of the tools. This affects 139 terms selected by the P-method, 193 terms from the T-method and 452 from the U-list.

5.2 Comparison based on categories

In Section 3.2 we divided the chosen lemma candidates into different categories. Now, we investigate how these categories are distributed among the terms depending on the extraction method. Most of the categories are evenly distributed over the different methods (cf. Table 5). The number of THEOREMS extracted by the P-method is so low because names of THEOREMS usually cannot be found with patterns as they are not part of definitions. The same applies for PERSONS. The number of ACTIVITIES extracted by the T-method is so high because it concerns nominalisations of verbs.

	Total		P-method		T-method		U-list	
PART	411	28.29 %	171	33.79 %	226	35.26 %	220	26.54 %
PROPERTY	263	18.10 %	105	20.75 %	32	4.99 %	187	22.56 %
TYPE	162	11.15 %	80	15.81 %	93	14.51 %	64	7.72 %
GENERAL	153	10.53 %	56	11.07 %	69	10.76 %	66	7.96 %
THEOREM	146	10.05 %	16	3.16 %	48	7.49 %	118	14.23 %
MAPPING	128	8.81 %	47	9.29 %	67	10.45 %	82	9.89 %
PERSON	89	6.13 %	12	2.37 %	67	10.45 %	21	2.53 %
ALGORITHM	58	3.99 %	10	1.98 %	20	3.12 %	41	4.95 %
PROBLEM	35	2.41 %	7	1.38 %	17	2.65 %	24	2.90 %
ACTIVITY	8	0.55 %	2	0.40 %	2	0.31 %	6	0.72 %
Σ	1453		506		641		829	

Table 5: Distribution over categories depending on extraction method

5.3 Error analysis

An error analysis in terms of classes of term candidates not found by the P- or the T-method is hard to realize, since almost no patterns emerge from these data. Nevertheless, some superficial remarks are possible: The P-method extracts some adjective-noun combinations, e.g. a few with the adjective *orientiert* (Engl. *oriented*). But something similar holds for the T-method, too: There are several combinations with *aufspannend*, *disjunkt* and *binär*, *hamiltonsch*, *eulersch*, *maximal*, *minimal*, *(stark)* *zusammenhängend*, *trennend*, *vollständig*[4]. All of them are combinations which appear in the texts with a certain frequency but are not part of the definitions on which the P-method mainly focuses.

The P-method and the T-method miss out systematically on MWT when they show up in a context such as *NN heißt ADJ wenn*[5], i.e. in a non-adjacent form that fills the 'slots' of definition phrases. Thus, the P-method extracts the individual words but not their combination. This issue is an instance of the well-known problem of distinguishing clearly between SWT and MWT, and between MWT and collocations of SWT. As mentioned, the U-list does not contain MWT.

The U-list contains several unique terms not found by the other methods, e.g. combinations of a number and a word, like *3-regulär*. Such items cannot be found by the P-method, because definitions will only contain their generalised form, i.e. *k-regulär*. As different values are possible for k, low frequencies of the individual instances may also prevent the T-method from extracting words. The U-list also contains many compound nouns, e.g. with the heads *Kante* (Engl. *edge*), *Graph* (Engl. *graph*), *Ecke* (Engl. *node*), which are unknown to the tagger lexicon.

In the P-list we find two further classes of noise: Combinations of only two uppercase letters like *G N* and combinations of a nominal term and a single capital letter like *Graph G*. As they are excluded from the final lemma list we remove these 550 candidates. Such items do not appear in the results of the other two methods. With this modification, we calculate the Jaccard index again:

$$\begin{array}{llll} |P| = 2018 & |P \cap T| = 597 & |P \cup T| = 3670 & J(P,T) = 0.1627 \\ & |P \cap U| = 243 & |P \cup U| = 3254 & J(P,U) = 0.0747 \\ & |P \cap T \cap U| = 116 & |P \cup T \cup U| = 4758 & J(P,T,U) = 0.0244 \end{array}$$

The precision of the P-method is now $p'_P = 0.2096$, thus much closer to $p_T = 0.2885$. Recall does not change for obvious reasons. The new F-score is $f'_P = 0.2666$. We conclude that the T-method and the P-method work almost equally well but are still outperformed, at least for SWT, by a simple list of words not being in a general language dictionary.

5.4 Comparison with *Sketch Engine*

We also extract terms from the corpus with the keyword extraction method provided by *Sketch Engine* (Kilgarriff et al., 2014), for further comparison. These terms have not been considered for the candidate list given to the raters because this extraction was done after the raters' work. Thus, the results are not completely comparable to the others.

We extract 1000 MWT and 1000 SWT with *Sketch Engine* with a minimum frequency of 1 to create conditions comparable to those of the P-method. The reference corpus for the term extraction by *Sketch Engine* is the German Web 2013 (deTenTen13) (Jakubíček et al., 2013).

[4] Engl. *spanning, disjoint, binary, Hamiltonian, Eulerian, maximum, minimum, (strongly) connected, separating, complete*

[5] Engl. *NN is called ADJ if*

Sketch Engine finds 198 terms which were not in the list given to the raters. 139 of them are SWT and 59 MWT. One of the raters annotates these 198 items with the criteria which resulted from the discussion. 65.94 % of the SWT and 35.59 % of the MWT are considered as useful for the dictionary. However, some of the selected SWT already appear in our candidate list as a part of MWT because the different tools use different criteria to distinguish between SWT and MWT.

We also calculate the Jaccard index between the results of the three different tools introduced in Section 4 and the list provided by *Sketch Engine*. S stands for the *Sketch Engine* in the calculations given below. We use the original P-list without the above-mentioned modifications.

$$\begin{array}{lll}
|P \cap S| = 177 & |P \cup S| = 4895 & J(P,S) = 0.0362 \\
|T \cap S| = 108 & |T \cup S| = 4140 & J(T,S) = 0.0261 \\
|U \cap S| = 76 & |U \cup S| = 3402 & J(U,S) = 0.0223 \\
|P \cap T \cap U \cap S| = 0 & |P \cup T \cup U \cup S| = 7535 & J(P,T,U,S) = 0
\end{array}$$

Now, we only take those terms into consideration which were selected for the final lemma list:

$$\begin{array}{lll}
|P_s \cap S_s| = 62 & |P_s \cup S_s| = 545 & J(P_s,S_s) = 0.1138 \\
|T_s \cap S_s| = 22 & |T_s \cup S_s| = 724 & J(T_s,S_s) = 0.0304 \\
|U_s \cap S_s| = 59 & |U_s \cup S_s| = 873 & J(U_s,S_s) = 0.0676 \\
|P_s \cap T_s \cap U_s \cap S_s| = 0 & |P_s \cup T_s \cup U_s \cup S_s| = 1453 & J(P_s,T_s,U_s,S_s) = 0
\end{array}$$

The results show that the terms extracted by *Sketch Engine* are closest to those extracted by the P-method, but the Jaccard index is still under 0.1 and only slightly above 0.1 for the selected terms. All the values here are below those calculated above.

6. Conclusion and future work

The described methods led to the definition of the final lemma list for creating the electronic dictionary on graph theory. Which information is given in the microstructure of a particular lemma is defined by its category. For example, the entry of a lemma from the category PERSON provides information on THEOREMS named after this PERSON, whereas a lemma from the category TYPES gives the information which PROPERTIES this TYPE of graphs has or can have. The information needed to provide such items will also be extracted by means of patterns and interactive corpus exploration.

The objective of our study was to compare the output of different term extractors, to understand to which degree such output can be used as a lemma list of the dictionary, and which amount of post-processing is needed to end up with an adequate lemma list. We note that a combination of different techniques may still be needed to cover the domain adequately. And two lessons are, if not learned form the exercise, at least recapitulated: deciding on termhood is also hard for experts, as long as no very strict guidelines are given; and lexicographic lemma selection also depends on the lexicographer's intuition about the dictionary's target group as well as on their strategy to ensure a homogeneous treatment of lexical items with respect to lemma selection.

With a view to further automating the lemma selection process, one could suggest a comparison of the term extraction with existing lemma lists for the domain; but such list do not really exist for graph theory terminology in German. One approach could be to use the titles of articles in Wikipedia which belong to the category *Graphentheorie*[6], but this list only comprises 100 items and thus is on a totally different scale than the amounts in our work. Furthermore, such a comparison does not take the available corpus into account; its results would thus only be significant to a very limited extent.

Not only the methods to identify lemma and item candidates, but also the evaluation methods are adaptable to other mathematical fields. This way it becomes easier to create electronic LSP-dictionaries for mathematical domains.

For selecting the lemmas, we worked with a German corpus. As we also have a comparable corpus of English texts, we will experiment with a similar (semi-)automatic approach for English.

To answer the question how the methods can be improved to also extract the terms which were only given in the U-list requires further research. In the end, we can see that a combination of different term extraction tools might work best because their pairwise Jaccard index is really low. We will take these results into consideration when working with the English data. Nevertheless, an extra difficulty is that we are only interested in the terminology of a sub-domain, not of a whole domain. Thus, some issues remain although we have chosen our corpus data according to this prerequisite.

7. References

Amjadian, E., Inkpen, D., Paribakht, T. & Faez, F. (2016). Local-Global Vectors to Improve Unigram Terminology Extraction. In *Proceedings of the 5th International Workshop on Computational Terminology (Computerm2016)*. Osaka, Japan: The COLING 2016 Organizing Committee, pp. 2–11. URL https://www.aclweb.org/anthology/W16-4702.

Amjadian, E., Inkpen, D., Paribakht, T.S. & Faez, F. (2018). Distributed specificity for automatic terminology extraction. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 24(1), pp. 23–40. URL https://www.jbe-platform.com/content/journals/10.1075/term.00012.amj.

Bernier-Colborne, G. (2012). Defining a gold standard for the evaluation of term extractors. In *Proceedings of the Terminology and Knowledge Representation Workshop, LREC 2012,*. pp. 15–18.

Blum, A. & Mitchell, T. (1998). Combining Labeled and Unlabeled Data with Co-Training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, COLT' 98. New York, NY, USA: Association for Computing Machinery, pp. 92–100. URL https://doi.org/10.1145/279943.279962.

Bohnet, B. (2010). Top Accuracy and Fast Dependency Parsing is not a Contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*. Beijing, China: Coling 2010 Organizing Committee, pp. 89–97. URL https://www.aclweb.org/anthology/C10-1011.

Bonin, F., Dell'Orletta, F., Montemagni, S. & Venturi, G. (2010). A Contrastive Approach to Multi-word Extraction from Domain-specific Corpora. In N.C.C. Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner & D. Tapias (eds.) *Proceedings of the Seventh International Conference on Language Resources and*

[6] https://de.wikipedia.org/wiki/Kategorie:Graphentheorie

Evaluation (LREC'10). Valletta, Malta: European Language Resources Association (ELRA), pp. 3222–3229.

Cabre, M.T. & Vivaldi Palatresi, J. (2013). 110. Acquisition of terminological data from text: Approaches. In R.H. Gouws, U. Heid, W. Schweickard & H.E. Wiegand (eds.) *Supplementary Volume Dictionaries. An International Encyclopedia of Lexicography.* De Gruyter Mouton, pp. 1486–1497. URL https://doi.org/10.1515/9783110238136.1486.

Choi, J.D., Tetreault, J. & Stent, A. (2015). It Depends: Dependency Parser Comparison Using A Web-based Evaluation Tool. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Beijing, China: Association for Computational Linguistics, pp. 387–396. URL https://www.aclweb.org/anthology/P15-1038.

Conrado, M., Pardo, T. & Rezende, S. (2013). A Machine Learning Approach to Automatic Term Extraction using a Rich Feature Set. In *Proceedings of the 2013 NAACL HLT Student Research Workshop*. Atlanta, Georgia: Association for Computational Linguistics, pp. 16–23. URL https://www.aclweb.org/anthology/N13-2003.

Dobrov, B. & Loukachevitch, N. (2011). Multiple Evidence for Term Extraction in Broad Domains. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*. Hissar, Bulgaria: Association for Computational Linguistics, pp. 710–715. URL https://www.aclweb.org/anthology/R11-1103.

Fedorenko, D., Astrakhantsev, N. & Turdakov, D. (2014). Automatic Recognition of Domain-Specific Terms: an Experimental Evaluation. *Proceedings of the Institute for System Programming of RAS*, 26(4), pp. 55–72. URL https://doi.org/10.15514/ispras-2014-26(4)-5.

Fleiss, J.L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), pp. 378–382. URL http://search.ebscohost.com/login.aspx?direct=true&db=pdh&AN=1972-05083-001&lang=de&site=ehost-live.

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. URL https://doi.org/10.1007/s007999900023.

Frantzi, K.T. & Ananiadou, S. (1996). Extracting Nested Collocations. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*. pp. 41–46. URL https://www.aclweb.org/anthology/C96-1009.

Hätty, A. (2020). *Automatic term extraction for conventional and extended term definitions across domains*. Ph.D. thesis, Institut für Maschinelle Sprachverarbeitung der Universität Stuttgart. URL http://elib.uni-stuttgart.de/handle/11682/11136.

Heid, U. & Weller, M. (2010). Corpus-derived data on German multiword expressions for lexicography. In A. Dykstra & T. Schoonheim (eds.) *Proceedings of the 14th EURALEX International Congress*. Leeuwarden/Ljouwert, The Netherlands: Fryske Akademy, pp. 331–340.

Jaccard, P. (1902). Lois de distribution florale dans la zone alpine. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 28(144), pp. 69–130.

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P. & Suchomel, V. (2014). Finding Terms in Corpora for Many Languages with the Sketch Engine. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Gothenburg, Sweden: Association for Computational Linguistics, pp. 53–56. URL https://www.aclweb.org/anthology/E14-2014.

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P. & Suchomel, V. (2013). The TenTen Corpus Family. In *7th International Corpus Linguistics Conference CL 2013*. Lancaster, pp. 125–127. URL http://ucrel.lancs.ac.uk/cl2013/.

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

Kochetkova, N.A. (2015). A method for extracting technical terms using the modified weirdness measure. *Automatic Documentation and Mathematical Linguistics*, 49(3), pp. 89–95. URL https://doi.org/10.3103/s0005105515030036.

Kruse, T. & Giacomini, L. (2019). Planning a domain-specific electronic dictionary for the mathematical field of graph theory: definitional patterns and term variation. In I. Kosem, T.Z. Kuhn, M. Correia, J.P. Ferreira, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek & C. Tiberius (eds.) *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference*. Brno: Lexical Computing CZ, s.r.o., pp. 676–693.

Kruse, T. & Heid, U. (2020). Lemma Selection and Microstructure: Definitions and Semantic Relations of a Domain-Specific e-Dictionary of the Mathematical Domain of Graph Theory. In Z. Gavriilidou, M. Mitsiaki & A. Fliatouras (eds.) *Euralex Proceedings*, volume 1. pp. 227–233.

Landis, J.R. & Koch, G.G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1), pp. 159–174. URL http://www.jstor.org/stable/2529310.

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*, volume 2 of *Natural Language Processing*. Amsterdam/Philadelphia: John Benjamins, pp. 279–302.

Pennington, J., Socher, R. & Manning, C. (2014). GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, pp. 1532–1543. URL https://www.aclweb.org/anthology/D14-1162.

Pollak, S., Repar, A., Martinc, M. & Podpečan, V. (2019). Karst Exploration: Extracting Terms and Definitions from Karst Domain Corpus. In I. Kosem, T.Z. Kuhn, M. Correia, J.P. Ferreira, M. Jansen, J.K. Isabel Pereira, M. Jakubíček, S. Krek & C. Tiberius (eds.) *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference*. Sintra: Lexical Computing CZ s.r.o, pp. 934–956.

Rayson, P. & Garside, R. (2000). Comparing Corpora using Frequency Profiling. In *The Workshop on Comparing Corpora*. Hong Kong, China: Association for Computational Linguistics, pp. 1–6. URL https://www.aclweb.org/anthology/W00-0901.

Rigouts Terryn, A., Hoste, V., Drouin, P. & Lefever, E. (2020). TermEval 2020: Shared Task on Automatic Term Extraction Using the Annotated Corpora for Term Extraction Research (ACTER) Dataset. In *Proceedings of the 6th International Workshop on Computational Terminology*. Marseille, France: European Language Resources Association, pp. 85–94. URL https://www.aclweb.org/anthology/2020.computerm-1.12.

Roelcke, T. (2010). *Fachsprachen*. Berlin: Erich Schmidt, 3rd, newly revised edition edition.

Roesiger, I., Bettinger, J., Schäfer, J., Dorna, M. & Heid, U. (2016). Acquisition of semantic relations between terms: how far can we get with standard NLP tools? In *Proceedings of the 5th International Workshop on Computational Terminology*

(Computerm2016). Osaka, Japan: The COLING 2016 Organizing Committee, pp. 41–51. URL https://www.aclweb.org/anthology/W16-4706.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*. Manchester.

Schmid, H. & Laws, F. (2008). Estimation of Conditional Probabilities With Decision Trees and an Application to Fine-Grained POS Tagging. In *Proceedings of the 22nd International Conference on Computational Linguistics*. Manchester, pp. 777–784.

Schäfer, J., Rösiger, I., Heid, U. & Dorna, M. (2015). Evaluating noise reduction strategies for terminology extraction. In *Proceedings of the conference Terminology and Artificial Intelligence 2015 (Granada, Spain)*. Granada, Spain: Universidad de Granada, pp. 123–131.

Wang, R., Liu, W. & McDonald, C. (2016). Featureless Domain-Specific Term Extraction with Minimal Labelled Data. In *Proceedings of the Australasian Language Technology Association Workshop 2016*. Melbourne, Australia, pp. 103–112. URL https://www.aclweb.org/anthology/U16-1011.

Zadeh, B.Q. & Handschuh, S. (2014). The ACL RD-TEC: A Dataset for Benchmarking Terminology Extraction and Classification in Computational Linguistics. In *Proceedings of the 4th International Workshop on Computational Terminology (Computerm)*. Dublin, Ireland: Association for Computational Linguistics and Dublin City University, pp. 52–63. URL https://www.aclweb.org/anthology/W14-4807.

GIPFA: Generating IPA Pronunciation from Audio

Xavier Marjou

Lannion, Brittany, France
E-mail: xavier.marjou@gmail.com

Abstract

Transcribing spoken audio samples into the International Phonetic Alphabet (IPA) has long been reserved for experts. In this study, we examine the use of an Artificial Neural Network (ANN) model to automatically extract the IPA phonemic pronunciation of a word based on its audio pronunciation, hence its name Generating IPA Pronunciation From Audio (GIPFA). Based on the French Wikimedia dictionary, we trained our model which then correctly predicted 75% of the IPA pronunciations tested. Interestingly, by studying inference errors, the model made it possible to highlight possible errors in the dataset as well as to identify the closest phonemes in French.

Keywords: audio; transcription; phonemes; Artificial Neural Network; sataset

1. Introduction

Some dictionaries such as Wiktionary offer a choice of both listening to words spoken by real users and reading phonemic pronunciations in the form of the International Phonetic Alphabet (IPA).

However, in the case of the French Wiktionary, the phonemic IPA transcripts are subject to a small percentage of errors. Several reasons can explain these errors. First, Wiktionary contributors may not be IPA experts; second, even IPA experts sometimes may make careless mistakes; third, the audio may be inconsistent because it is generally recorded independently without taking IPA pronunciation into account, which can lead to important discrepancies; fourth, some sounds such as /o/ and /ɔ/ may be very close to each other and can depend on the speaker.

This article examines whether such errors could be avoided by using a Natural Language Processing (NLP) tool to automatically extract phonemic IPA pronunciation from audio pronunciation.

For this purpose, we made use of Automatic Speech Recognition (ASR), which has already been the subject of in-depth studies. In particular, many recent implementation approaches have successfully used a deep Artificial Neural Network (ANN), such as in Han et al. (2020) and Das et al. (2019), hence our choice to design a new ANN called Generating IPA Pronunciation From Audio (GIPFA). In order to train and test it, we also assembled a new experimental dataset based on 80400 samples from the French Wiktionary.

Despite a dataset containing an unknown percentage of erroneous data samples, our GIPFA model succeeded in providing reasonable accuracy. Although it failed to replace IPA experts, it nevertheless proved to be particularly useful in identifying the biggest errors in the dataset.

2. Methodology

In order to predict the IPA pronunciation of a word, two main steps were necessary: identifying a relevant dataset and designing an ANN model capable of inferring an IPA pronunciation from an audio pronunciation.

2.1 Dataset

Word	Audio filename	IPA pronunciation
bonjour	LL-Q150 (fra)-LoquaxFR-bonjour.wav	bɔ̃ʒuʁ

Table 1: Dataset

Our dataset came from a Wikimedia dump[1] containing all pages and articles of the French Wiktionary. In this dump, each page generally contains three essential features: one *word* along with n main *IPA pronunciations* and m examples of *audio pronunciations* recorded by several speakers.

- A word is a text string containing Unicode characters. The *word* terminology has to be taken in the broad sense as a Wiktionary word contains common names, proper names words, abbreviations, numbers, and even sayings. Although our ANN did not use it, we kept the word in our dataset for debugging purposes, in order to have the possibility to again find the Wiktionary page containing the pronunciations.
- An audio pronunciation refers to an audio file generally recorded in a Waveform Audio File (WAV) format containing the pronounced word. Wiktionary pages can contain one or more audio pronunciations for the same word. When an audio file is generated with LinguaLibre (LL)[2] software, it benefits from three useful features: the audio file is under the Creative Commons sharing license[3]; the file can be fetched from Wikimedia Commons[4] based on its audio filename; the audio filename also contains a label representing a user name which can be used to identify audio files generated by users.
- An IPA pronunciation is a text string containing IPA symbols. For learning purposes, each audio pronunciation of a word should ideally be associated with a single IPA pronunciation transcribing this precise audio content; a ranking of the most common pronunciations might also be calculated and indicated in the page describing the word. However, most words have a single IPA pronunciation (i.e. $n = 1$) even when multiple audio pronunciations are available. Although some words have multiple IPA pronunciations (e.g. *coût*), a Wiktionary page rarely indicates which of these pronunciations corresponds to an audio file.

For our purposes, we restricted our dataset to samples containing:

- words in the French Wiktionary[5];
- French words, given that each Wiktionary describes words of several languages;
- words with a single IPA pronunciation, given that multiple IPA per audio sample introduce ambiguities;

[1] https://dumps.wikimedia.org/frwiktionary/20200501/

[2] https://lingualibre.org

[3] https://creativecommons.org/licenses/by-sa/4.0/

[4] https://commons.wikimedia.org/

[5] https://fr.wiktionary.org/

- IPA pronunciation containing symbols making part of the 37 traditional French phonemes (i.e. ’i’, ’e’, ’ɛ’, ’a’, ’ɑ’, ’ɔ’, ’o’, ’u’, ’y’, ’ø’, ’œ’, ’ə’, ’ɛ̃’, ’ɑ̃’, ’ɔ̃’, ’œ̃’, ’j’, ’w’, ’ɥ’, ’p’, ’k’, ’t’, ’b’, ’d’, ’g’, ’f’, ’s’, ’ʃ’, ’v’, ’z’, ’ʒ’, ’l’, ’ʁ’, ’m’, ’n’, ’ɲ’, ’ŋ’);
- IPA pronunciation containing less than 20 phonemes, in order to keep our ANN model reasonable in size regarding our resources;
- audio files recorded with LL, in order to easily fetch audio files.

We also discarded 9 symbols that appear as optional in the IPA pronunciation of the French Wiktionary (’⁀’, ’.’, ’ ’, ’‿’, ’’’ and ’ː’, ’(’, ’)’, ’-’).

The resulting dataset contained 80200 samples from 102 different speakers. As depicted in Table 1, each sample contained three features: a *word*, an *audio filename* and an *IPA pronunciation*.

In addition, we also preprocessed the WAV files to have a fixed length of 2 seconds, and then converted them into a Mel-Frequency Cepstral Coefficients (MFCC) format so that they could serve as direct inputs into our model. Although processing audio files under a WAV format would be possible as in Sainath et al. (2015), it requires significant RAM memory, hence our choice to transpose them into an MFCC format, as usually performed in many studies, such as in Alcaraz Meseguer (2009) and Nahid et al. (2017).

2.2 Experiments

2.2.1 Model architecture

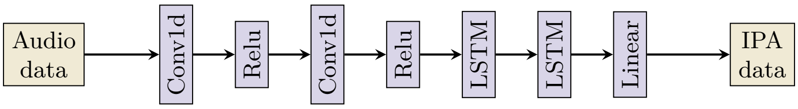


Figure 1: The GIPFA ANN model used for transcribing audio samples into IPA samples.

We modelled our GIPFA ANN as depicted in Figure 1. It contains typical components found in many ANN models used for ASR. However, given that we only had to translate a single word per sample, we did not use any Transformer component (Vaswani et al., 2017). Each audio input sample (MFCC data) first traversed a stack of two 1D convolution layer (Conv1D) layers to extract the shape of the MFCC data; followed by two Long Short Term Memory (LSTM) filters (Hochreiter & Schmidhuber, 1997) to extract temporal sequences; and finally followed by a linear layer in order to allow a Connectionist Temporal Classification (CTC) loss calculation (Graves, 2012). We did not allow the succession of two identical phonemes because this is rare in French words. In addition, we used an AdamW optimiser (Loshchilov & Hutter, 2017) with a learning rate of 1×10^{-4}.

2.2.2 Hyperparameters

We used Ray Tune (Moritz et al., 2018) for fine-tuning our hyperparameters with respect to accuracy results. This led us to identify a set of best values among a larger set of experimented values as summarised in Table 2. The resulting model contained 9,609,558 trainable parameters. Slight variations in the best values did not lead to significant improvement. Although it is believed that a wider network may lead to better results (Nakkiran et al., 2019), we limited our model to these $10M$ parameters due to our limited computing resources.

Hyperparameter	Tested values	Best value
mfcc_coefficients	40	40
conv1d_activ	none, relu	relu
conv1d_layers	0, 1, 2, 3	2
conv1d_units	32, 64, 128	128
conv1d_bn	False, True	True
lstm_layers	0, 1, 2	2
lstm_units	128, 256, 512	512
lstm_dropout	0.1, 0.25, 0.5	0.5
lstm_bidir	False, True	True
lstm_bn	False, True	True
optimizer	Adam, AdamW	AdamW
lr	1e-3, 1e-4	1e-4

Table 2: GIPFA hyperparameters values

2.2.3 Training

For the training step, we used 79,326 samples distributed over 3,966 batches of 20 samples (3,927 training batches and 39 evaluation batches). During a preprocessing step, all audio samples were standardised with the mean (-11.48) and standard deviation (80.30) pre-observed with regard to the dataset.

Before each run, the data samples were randomly shuffled. Each training run took approximately 10 epochs of 3 minutes each on a single GPU (GeForce RTX 2080, 8 GB).

2.2.4 Test

For the testing step, we used 1,000 unseen samples to evaluate the performances of the GIPFA ANN.

2.2.5 Accuracy

Since solving the translation problem requires correct inference of the entire IPA pronunciation, we simply set for each tested sample an accuracy of 1 when our model predicted an IPA pronunciation equal to the tested target IPA pronunciation, or 0 otherwise. After each training run, we then calculated the average accuracy across all samples (i.e. an average accuracy between 0.0 and 1.0).

We performed 11 runs (with one training step and one test step for each) to allow reasonable confidence in the average accuracy results. We finally computed the mean accuracy and the associated standard deviation (std) for the 11 tests.

Since the dataset had not been studied further, there was unfortunately no baseline reference to challenge our results.

2.2.6 Further details on errors

To our knowledge, no study has examined the exactness and coherence of the audio files and IPA pronunciations of the French Wiktionary, meaning that the dataset may contain errors, making it difficult to assess whether a prediction error comes from the dataset or from the ANN.

In order to obtain more in-depth information on errors, we therefore also calculated three other metrics related to the 80000 samples in the dataset:

- At the word level
    - *Edit distance error*: the Levenshtein distance (Levenshtein, 1965) between the predicted IPA pronunciation and the target IPA pronunciation, in order to estimate how far the prediction was from the target.
- At the phoneme level
    - *Average phoneme accuracy*: the percentage of correct translations for each phoneme;
    - *Error pair percentage*: Since each of the 37 target phonemes can be incorrectly translated as one of the other 36 phonemes, the results can contain up to 37 * 36 categories of error pairs. To assess the representativeness of each pair, we calculated its number of occurrences divided by the number of phonemic errors.

The code is available on Github [6].

3. Results

In this section, we describe two different results: first, the accuracy of the model, then a more detailed observation of errors at phoneme level and at word level.

3.1 Accuracy

Table 3 presents the accuracy results which were consistent across the 11 runs; our GIPFA ANN model successfully predicted around 75 IPA pronunciations out of 100 audio samples.

Correctly inferred pronunciations had a mean length of 7.51, whereas incorrectly inferred pronunciations had a mean length of 8.65, thus indicating a slightly higher probability of error as the length of the IPA pronunciation increased.

[6] Code available at https://github.com/marxav/gipfa

Training samples	Tested samples	Pronunciation accuracy (mean)	Pronunciation accuracy (std)
79326	1000	0.75	0.02

Table 3: Pronunciation accuracy

3.2 Insight into the errors

Performing inferences on 80,000 samples of the dataset enabled a better understanding of the reasons for the errors.

3.2.1 Phoneme accuracy

Table 4 reports the translation accuracy of each phoneme. One phoneme (/ɑ/) had poor accuracy (less than 50%), five phonemes (/o/, /ŋ/, /œ̃/, /ɲ/ and /oe/) had moderate accuracy (between 65% and 89%), while the remaining thirty-one phonemes had high accuracy (over 90%).

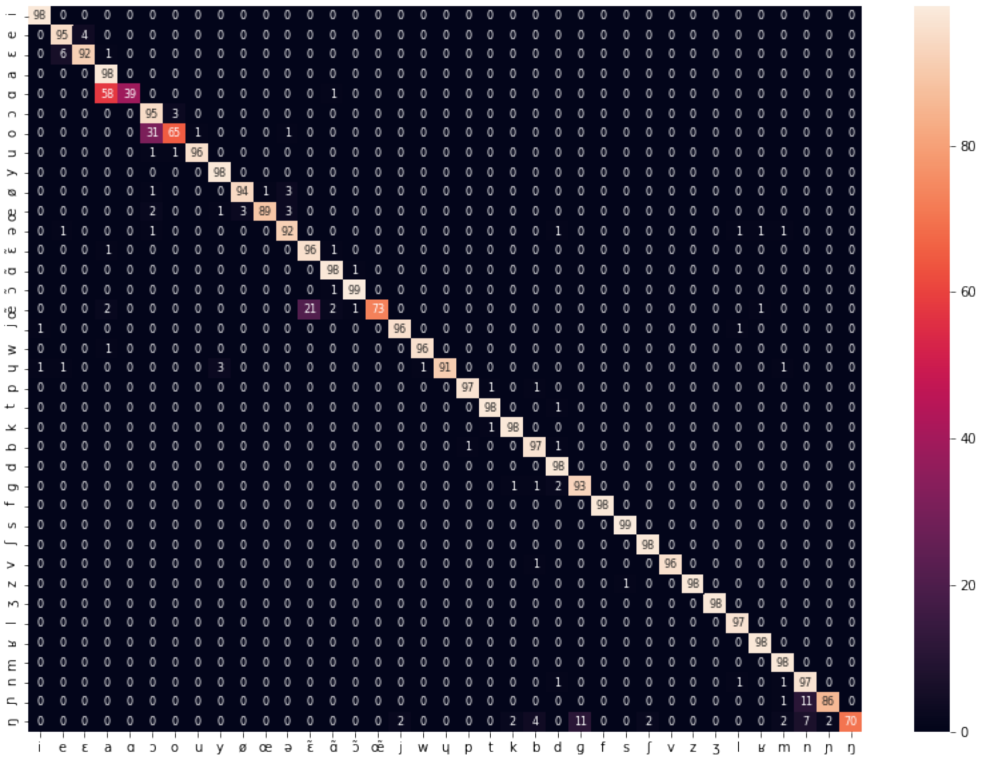


Figure 2: Confusion Matrix

To better observe the details, we also detailed these phoneme translation errors in a confusion matrix, as shown in Figure 2. Each row in the matrix represented a target phoneme while each column represented the distribution of the predicted phonemes. For instance, it turned out that the target phoneme /ɛ/ was predicted to be /e/ 6% of the

Target phoneme	Correct translation	Incorrect translation	Average accuracy
ɑ	392	605	0.39
o	4,615	2,485	0.65
ŋ	40	17	0.70
œ̃	241	89	0.73
ɲ	697	110	0.86
œ	2,459	301	0.89
ɥ	1,185	113	0.91
ɛ	15,859	1,472	0.92
ə	7,918	732	0.92
ɡ	5,911	427	0.93
ø	2,587	169	0.94
ɔ	18,655	1,074	0.95
e	30,018	1,608	0.95
w	4,357	159	0.96
v	7,469	282	0.96
u	6,712	250	0.96
ɛ̃	4,527	192	0.96
j	12,567	547	0.96
b	12,753	434	0.97
n	13,165	472	0.97
p	14,845	464	0.97
l	23,181	684	0.97
ɑ̃	13,704	226	0.98
f	9,632	225	0.98
y	8,235	183	0.98
z	7,730	146	0.98
i	34,772	664	0.98
d	15,975	323	0.98
k	23,159	503	0.98
ʃ	4,407	92	0.98
a	44,575	707	0.98
m	17,334	313	0.98
ʁ	47,221	799	0.98
ʒ	5,552	137	0.98
t	29,691	713	0.98
ɔ̃	9,258	129	0.99
s	30,018	400	0.99

Table 4: Average accuracy of each phoneme

time, /ɛ/ 92% of the time, and /a/ 1%. Notable outliers were four large numbers outside the diagonal: 58% of /ɑ/ seemed to be poorly predicted as an /a/; 31% of /o/ as /ɔ/; 21% of /œ̃/ as /ɛ̃/; and 11% of /ŋ/ as /ɡ/; It turned out that, like humans, the ANN had difficulties in differentiating close elementary sounds.

3.2.2 Error pair percentage

Table 5 represents the proportion of the error associated with each phoneme pair compared to the total errors of all pairs of phonemes. Interestingly, only three pairs of phonemes generated 31% of all errors: (/o/, /ɔ/) (15% of all errors), (/e/, /ɛ/) (12% of all errors), and (/a/, /ɑ/) (4% of all errors).

Target phoneme	Predicted phoneme	Percentage of all errors
o	ɔ	12.03%
e	ɛ	6.51%
ɛ	e	5.46%
ɑ	a	3.16%
ɔ	o	3.07%
t	d	1.25%
ɛ	a	1.04%
a	ɑ	0.83%

Table 5: Most encountered error pairs

3.2.3 Word-level distance error

Computed samples	Levenshtein distance mean, std
80000	0.31, 0.66

Table 6: Levenshtein distance

Table 6 reports a small mean Levenshtein distance and gives assurance that there is strong consistency between the audio content and the IPA pronunciation for the samples in the dataset studied.

However, Table 7 focuses on the most extreme outliers by reporting the 10 samples with the highest Levenshtein distance. Upon investigation, it was found that all of these 10 samples contained either an error in the audio sample (e.g. bad word pronunciation or no word spoken at all) or an error in the target IPA pronunciation, which meant that all these errors were in the dataset itself. These results therefore suggest that data samples whose pronunciations have a high Levenshtein distance probably contain an error.

Additional work would be required to identify the best threshold distance to identify possible errors in the dataset.

Word	IPA Target	IPA Prediction	Levenshtein distance
1337	/lit/	/mitasɑ̃tʁɑ̃mzɔt/	13
agent innervant	/aʒɑ̃inɛʁvɑ̃/	/go/	11
brut de décoffrage	/bʁytdədekɔfʁaʒ/	/sbɔʁdedtɔʁ/	10
Michel	/miʃɛl/	/stɛ̃dəsɑ̃mʃɛl/	10
phalange proximale	/falɑ̃ʒpʁɔksimal/	/falɑ̃ʒ/	9
analyse calorimétrique	/analɔgʃimik/	/analiskalɔʁimetik/	9
àtha	/atɔ̃nœ̃blavi/	/ata/	9
Wikitionnaire	/gazaefɛdəsfɛʁ/	/gɔʒifisølɛʁ/	9
arrondir par défaut	/aʁɔ̃diʁpaʁdefo/	/aʁɑ̃diʁ/	8
Luxembourg	/lyksɑ̃buʁ/	/yseʁzɔnb/	8

Table 7: Top-10 pronunciations with the highest Levenshtein distance

4. Discussion and Conclusion

Previous work has documented the effectiveness of the ANN model for ASR. However most studies have focused on the direct translation of audio samples into words.

In this study, we focused instead on the translation of audio samples into phonemes. We first proposed an ANN predicting with 75% accuracy the French pronunciations of the French Wiktionary.

Since to our knowledge no existing work has been done on this specific task and dataset, there was no basis for comparison or assurance as to the accuracy and consistency of the data.

We have shown that the translations of certain phonemes were more problematic since some phonemes are close elementary sounds (/o/ and /ɔ/; /ɛ/ and /e/; /ɑ/ and /a/) and thus difficult to distinguish. Future work may consider carefully checking the audio samples and IPA pronunciations containing these close phonemes, which would in turn enhance the efficiency of the ANN. In addition, future work could also involve synthesised audio examples and use them as additional samples to reinforce training data.

However, we have also shown that the Levenshtein distance between our GIPFA prediction and the target (as it exists in the dataset and therefore in the Wiktionary) can highlight the most suspect samples in the dataset. Such results therefore suggest that our GIPFA ANN would be a valuable tool to help verify the consistency of Wiktionary regarding pronunciation.

Therefore, integrating it into a tool like LL should be useful in order to suggest an IPA transcription. It could even be used to suggest an IPA transcription associated with each recorded audio sample, since having one IPA transcription per audio file should further improve the performances of the ANN.

Finally, we believe this method should be applicable to other languages provided that a sufficient number of training samples are available.

Acknowledgements

We thank all Wiktionary and LinguaLibre contributors for their contributions, without which there would be no wonderful free dictionary nor a free dataset.

5. References

Alcaraz Meseguer, N. (2009). *Speech analysis for automatic speech recognition*. Master's thesis, Institutt for elektronikk og telekommunikasjon.

Das, A., Li, J., Ye, G., Zhao, R. & Gong, Y. (2019). Advancing Acoustic-to-Word CTC Model with Attention and Mixed-Units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(12), pp. 1880–1892.

Graves, A. (2012). Connectionist temporal classification. In *Supervised Sequence Labelling with Recurrent Neural Networks*. Springer, pp. 61–93.

Han, W., Zhang, Z., Zhang, Y., Yu, J., Chiu, C.C., Qin, J., Gulati, A., Pang, R. & Wu, Y. (2020). ContextNet: Improving Convolutional Neural Networks for Automatic Speech Recognition with Global Context. *arXiv preprint arXiv:2005.03191*.

Hochreiter, S. & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), pp. 1735–1780.

Levenshtein, V.I. (1965). Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. *Reports of the USSR Academy of Sciences*, 163(4), pp. 845–848.

Loshchilov, I. & Hutter, F. (2017). Decoupled Weight Decay Regularization. 1711.05101.

Moritz, P., Nishihara, R., Wang, S., Tumanov, A., Liaw, R., Liang, E., Elibol, M., Yang, Z., Paul, W., Jordan, M.I. et al. (2018). Ray: A distributed framework for emerging {AI} applications. In *13th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 18)*, pp. 561–577.

Nahid, M.M.H., Purkaystha, B. & Islam, M.S. (2017). Bengali speech recognition: A double layered LSTM-RNN approach. In *2017 20th International Conference of Computer and Information Technology (ICCIT)*. IEEE, pp. 1–6.

Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B. & Sutskever, I. (2019). Deep double descent: Where bigger models and more data hurt. *arXiv preprint arXiv:1912.02292*.

Sainath, T., Weiss, R.J., Wilson, K., Senior, A.W. & Vinyals, O. (2015). Learning the Speech Front-end with Raw Waveform CLDNNs. In *Interspeech*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008.

A workflow for historical dictionary digitisation: Larramendi's Trilingual Dictionary

David Lindemann, Mikel Alonso

UPV/EHU University of the Basque Country, Vitoria-Gasteiz, Spain
E-mail: david.lindemann@ehu.eus, mikelalon@gmail.com

Abstract

In this paper, we present a workflow for historical dictionary digitisation, with a 1745 Spanish-Basque-Latin dictionary as the use case. We start with scanned facsimile images, and get to represent attestations of modern standard Basque lexemes as Linked Data, in the form they appear in the dictionary. We are also able to produce an index of the dictionary, i. e. a Basque-Spanish version, and to map extracted Spanish and Basque lexical items to reference dictionary lemma list entries. The workflow is entirely based on freely available software. OCR and information extraction are performed using Machine Learning algorithms; data exhibits and the transcription curation environment are provided using Wikisource and Wikidata. Our evaluation of a first iteration of the workflow suggests its capability to deal with early modern printed dictionary text, and to reduce manual effort in the different stages significantly.

Keywords: Historical Lexicography; Digitisation; OCR; information extraction; Linked Data

1. Introduction

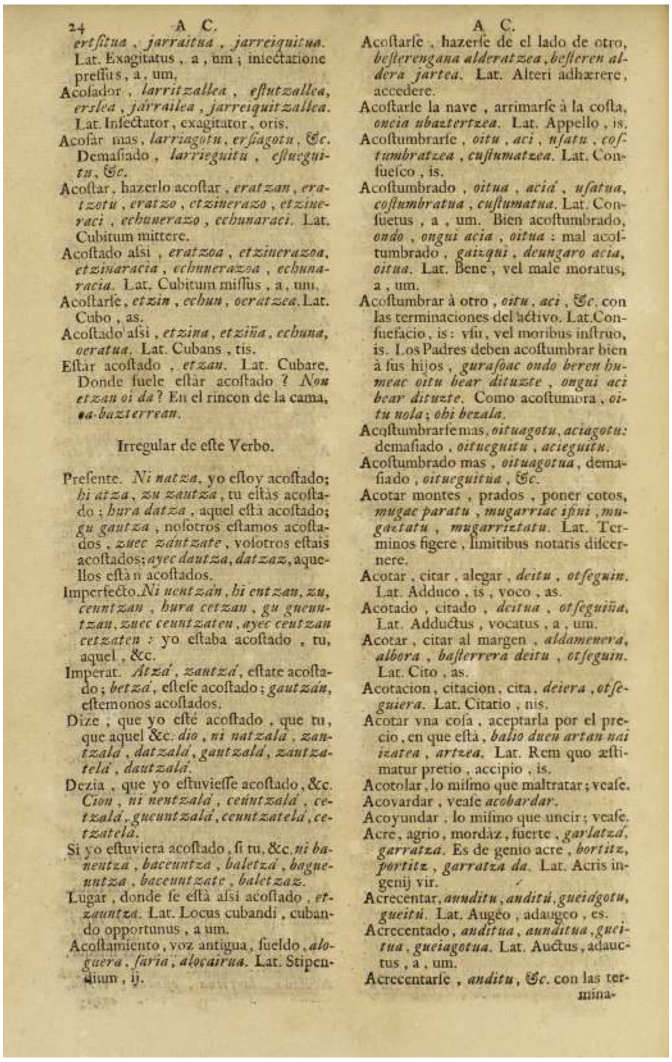
Manuel de Larramendi's Spanish-Basque-Latin Trilingual Dictionary in two volumes Larramendi (1745), henceforth LAR, for more than a century and a half has been the outstanding reference resource for Basque, and can be regarded the classic lexicographic work that brought a significant shift in the periodisation of Basque Lexicography (Urgell, 2002); it represents the beginning of modern Basque Lexicography. Nevertheless, this important classic is still available only as print dictionary, the digitisation of which has not overcome the stage of scanned images. The dictionary has been subject to in-depth philological and lexicographical research (Urgell, 1998a,b), which had to resort to manually compiled sets of examples, and thus was not able to include full-fledged quantitative methods that would take into consideration the content as a whole. For example, we do not have anything else than approximate estimations regarding the overall amount of headwords and distinct lemmata, and regarding the relation to the headword list of the 1725-1739 Spanish-Latin Diccionario de Autoridades (Real Academia Española, 2013), henceforth DA, the outstanding lexicographic work for Spanish at that time, which Larramendi used as primary reference for his dictionary.

In this project report, we reach out to propose and evaluate a workflow for digitisation, using the cited early modern print dictionary as showcase. Starting point is a collection of scanned images of both volumes of LAR dictionary, produced and provided by Koldo Mitxelena public library.[1] Following the digitisation stages outlined in Lindemann & San Vicente (2020), we apply a semi-automatic toolchain, and measure its rendering. This includes Optical Character Recognition (OCR), information extraction, and a first proposal for modeling attestations according to the Resource Description Framework (RDF), having in mind its integration in Wikidata.

Our main goal is the evaluation of the tested workflow, which includes an assessment of the precision reached by the employed tools, in order to make predictions concerning

[1] The item's first volume is available at https://www.kmliburutegia.eus/Record/26577, the second at http://www.kmliburutegia.eus/Record/203133.

manual validation and editing effort regarded necessary for a complete and accurate digitisation. We want to point out that the dictionary on hand is doubtlessly one of the harder nuts to crack, due to the early modern typefont, and lexicographic features. One working hypothesis therefore is the following: If we are able to get acceptable results for this dictionary with a predictable and limited manual workload, printed lexical resources published later than 1745 should require less effort to get digitised.



24 A C.

ertſitua, *jarraitua*, *jarreiquitua*. Lat. Exagitatus, a, um; iniectatione preſſus, a, um.

Acoſador, *larritzallea*, *eſtutzallea*, *erslea*, *jarrailea*, *jarreiquitzallea*. Lat. Infectator, exagitator, oris.

Acoſar mas, *larriagotu*, *erſiagotu*, *&c.* Demaſiado, *larrieguitu*, *eſtueguitu*, *&c.*

Acoſtar, hazerlo acoſtar, *eratzan*, *eratzotu*, *eratzo*, *etzinerazo*, *etzineraci*, *echunerazo*, *echunaraci*. Lat. Cubitum mittere.

Acoſtado aſsi, *eratzoa*, *etzinerazoa*, *etzinaracia*, *echunerazoa*, *echunaracia*. Lat. Cubitum miſſus, a, um.

Acoſtarſe, *etzin*, *echun*, *oeratzea*. Lat. Cubo, as.

Acoſtado aſsi, *etzina*, *etziña*, *echuna*, *oeratua*. Lat. Cubans, tis.

Eſtàr acoſtado, *etzan*. Lat. Cubare. Donde ſuele eſtàr acoſtado? *Non etzan oi da*? En el rincon de la cama, *oa-bazterrean*.

Irregular de eſte Verbo.

Preſente. *Ni natza*, yo eſtoy acoſtado; *hi atza*, *zu zautza*, tu eſtàs acoſtado; *hura datza*, aquel eſtà acoſtado; *gu gautza*, noſotros eſtamos acoſtados, *zuec zautzate*, voſotros eſtais acoſtados; *ayec dautza*, *datzaz*, aquellos eſtàn acoſtados.

Imperfecto. *Ni nentzán*, *hi entzan*, *zu*, *ceuntzan*, *hura cetzan*, *gu gueuntzan*, *zuec ceuntzaten*, *ayec ceutzan cetzaten*: yo eſtaba acoſtado, tu, aquel, &c.

Imperat. *Atzá*, *zautzá*, eſtate acoſtado; *betzá*, eſteſe acoſtado; *gautzán*, eſtemonos acoſtados.

Dize, que yo eſté acoſtado, que tu, que aquel &c. *dio*, *ni natzalá*, *zautzalá*, *datzalá*, *gautzalá*, *zautzatelá*, *dautzalá*.

Dezia, que yo eſtuvieſſe acoſtado, &c. *Cion*, *ni nentzalá*, *ceúntzalá*, *cetzalá*, *gucuntzalá*, *ceuntzatelá*, *cetzatelá*.

Si yo eſtuviera acoſtado, ſi tu, &c. *ni banentza*, *baceuntza*, *baletzá*, *bagueuntza*, *baceuntzate*, *baletzaz*.

Lugar, donde ſe eſtà aſsi acoſtado, *etzauntza*. Lat. Locus cubandi, cubando opportunus, a um.

Acoſtamiento, voz antigua, ſueldo, *aloguera*, *ſaria*, *alocairua*. Lat. Stipendium, ij.

A C.

Acoſtarſe, hazerſe de el lado de otro, *beſterengana alderatzea*, *beſteren aldera jartea*. Lat. Alteri adhærere, accedere.

Acoſtarſe la nave, arrimarſe à la coſta, *oncia ubaztertzea*. Lat. Appello, is.

Acoſtumbrarſe, *oitu*, *aci*, *uſatu*, *coſtumbratzea*, *cuſtumatzea*. Lat. Conſueſco, is.

Acoſtumbrado, *oitua*, *aciá*, *uſatua*, *coſtumbratua*, *cuſtumatua*. Lat. Conſuetus, a, um. Bien acoſtumbrado, *ondo*, *ongui acia*, *oitua*: mal acoſtumbrado, *gaizqui*, *deungaro acia*, *oitua*. Lat. Bene, vel male moratus, a, um.

Acoſtumbrar à otro, *oitu*, *aci*, *&c.* con las terminaciones del activo. Lat. Conſuefacio, is: vſu, vel moribus inſtruo, is. Los Padres deben acoſtumbrar bien à ſus hijos, *guraſoac ondo beren humeac oitu bear dituzte*, *ongui aci bear dituzte*. Como acoſtumbra, *oitu nola*; *obi bezala*.

Acoſtumbrarſe mas, *oituagotu*, *aciagotu*: demaſiado, *oitueguitu*, *acieguitu*.

Acoſtumbrado mas, *oituagotua*, demaſiado, *oitueguitua*, *&c.*

Acotar montes, prados, poner cotos, *mugac paratu*, *mugarriac iſini*, *mugaztatu*, *mugarriztatu*. Lat. Terminos figere, limitibus notatis diſcernere.

Acotar, citar, alegar, *deitu*, *otſeguin*. Lat. Adduco, is, voco, as.

Acotado, citado, *deitua*, *otſeguiña*. Lat. Adductus, vocatus, a, um.

Acotar, citar al margen, *aldamenera*, *albora*, *baſterrera deitu*, *otſeguin*. Lat. Cito, as.

Acotacion, citacion, cita, *deiera*, *otſeguiera*. Lat. Citatio, nis.

Acotar vna coſa, aceptarla por el precio, en que eſtà, *balio duen artan nai izatea*, *artzea*. Lat. Rem quo æſtimatur pretio, accipio, is.

Acotolar, lo miſmo que maltratar; veaſe.

Acovardar, veaſe *acobardar*.

Acoyundar, lo miſmo que uncir; veaſe.

Acre, agrio, mordàz, fuerte, *garlatzá*, *garratza*. Es de genio acre, *bortitz*, *portitz*, *garratza da*. Lat. Acris ingenij vir.

Acrecentar, *aunditu*, *anditú*, *gueiágotu*, *gueitú*. Lat. Augeo, adaugeo, es.

Acrecentado, *anditua*, *aunditua*, *gueitua*, *gueiagotua*. Lat. Auctus, adauctus, a, um.

Acrecentarſe, *anditu*, *&c.* con las termina-

Figure 1: LAR, vol. 1, page 24, scanned image

LAR presents several severe deviations from an up-to-date standard in print Lexicography. First, the early modern typefont, and the scanned images made from stained and half-transparent paper are to be mentioned as strong handicaps for OCR, which is the reason for the poor quality of LAR digital text versions available today. Second, the lexicographic structure is not consistently mirrored in structural markup and layout. That is true on macrostructural level (i. e. the segmentation of the dictionary text into entries), and concerning the lexicographical microstructure, in other words, the inner organisation of entries. This makes it evident that a rule-based segmentation of the dictionary text into labelled lexicographic components like "entry", "headword" and "translation equivalent", i.e. to "extract" the information to a format that can be interpreted by machines employing fixed rules, would not lead to satisfying results. Therefore, it becomes interesting to look at applications that use neural networks for the these tasks, since algorithms based on such technology are able to predict a result also in cases where a strict rule would fail. Applications for OCR and dictionary segmentation that use such technologies are available today, and we are witnessing their consolidation in the very recent past and present.

In the following, we present our experiments, for which we have employed tools that are freely available for research purposes, so that they are fully reproducible by anybody interested in this use case, or in similar endeavours.

2. Optical Character Recognition

An OCR tool converts images of characters to digital characters, i.e. it associates pixel patterns on an image with letters. The result, a digital text (txt), unlike the pure image (a collection of pixels), enables editing, searching, and computational processing of the textual content. State-of-the-art OCR tools rely on Machine Learning (ML) algorithms, that are trained on a manually transcribed subset of the work, and predict the mappings between letters as pixel patterns and as digital characters on that basis. The advent of ML in OCR technology has made the processing of early modern printing (and even hand-written text)[2] feasible: While in modern or even digital print characters can be mapped to uniform pixel patterns, in early modern printing, the patterns for the same letter may differ from each other in a significant way. In addition, pixel patterns may be disturbed by irregularities or stains on the paper, or ink from the reverse side of the pertaining page shining through. Similar to the flexibility in human reasoning, the ML algorithm tries to associate each pixel pattern it identifies to the most probable candidate letter, which means it can resolve doubts. The shortcomings of OCR tools developed for standard (modern) print become clear if we look at the text versions of LAR offered at the moment.[3] These can be roughly classified as follows: (a) characters that do not belong to a modern standard typeset, (b) characters that match to different pixel patterns, including the impact of stains or colour changes on the paper, and (c), errors due to wrong layout identification, i. e. errors in column and line segmentation.

Kraken[4] is a freely available OCR tool that relies on ML. It requires scanned images with a minimum resolution of 300 DPI, although authors of related work argue that even lower resolutions may serve. Kraken has shown that it outperforms leading proprietary OCR solutions designed for printed and manuscript documents, for example, digitising classical Arabic-script text ((Romanov et al., 2017). In addition, the fact that Kraken produces output following ALTO XML standard (see section 3) has been a reason for choosing this tool. We have favoured Kraken over Transkribus,[5] a web-based tool with similar features, because of its ability to be flexible towards font styles, i. e. that it is able to learn not only the character but also to discriminate font styles such as italics (see section 2.2 below).[6]

2.1 Pre-processing

As input, the Kraken tool needs scanned images, which are preprocessed following the guidelines,[7] i. e., the images are converted to black-and-white binary, and, if needed, their angle is corrected, so that lines appear horizontally, misalignments due to paper curvations

[2] For example, the Transkribus software uses ML for processing hand-written text. It also offers a graphical user interface for the creation of training sets, and the manual correction of the output, see https://transkribus.eu/Transkribus/. For a use case, see (Lindemann et al., 2018).

[3] Spanish National Library BNE (http://bdh-rd.bne.es/viewer.vm?id=0000015622), Google Books (https://play.google.com/books/reader?id=whdf0XXf6gwC), and Bavarian State Library BSB (https://opacplus.bsb-muenchen.de/title/BV035479582) offer image and txt versions of LAR.

[4] See http://kraken.re. This tool is built upon OCRropus (https://github.com/ocropus/ocropy) and features a user interface for ML training set creation. Kraken has been developed in the framework of the eScripta project at Université Paris Sciences et Lettres (cf. https://escripta.hypotheses.org/tag/kraken).

[5] See note 6.

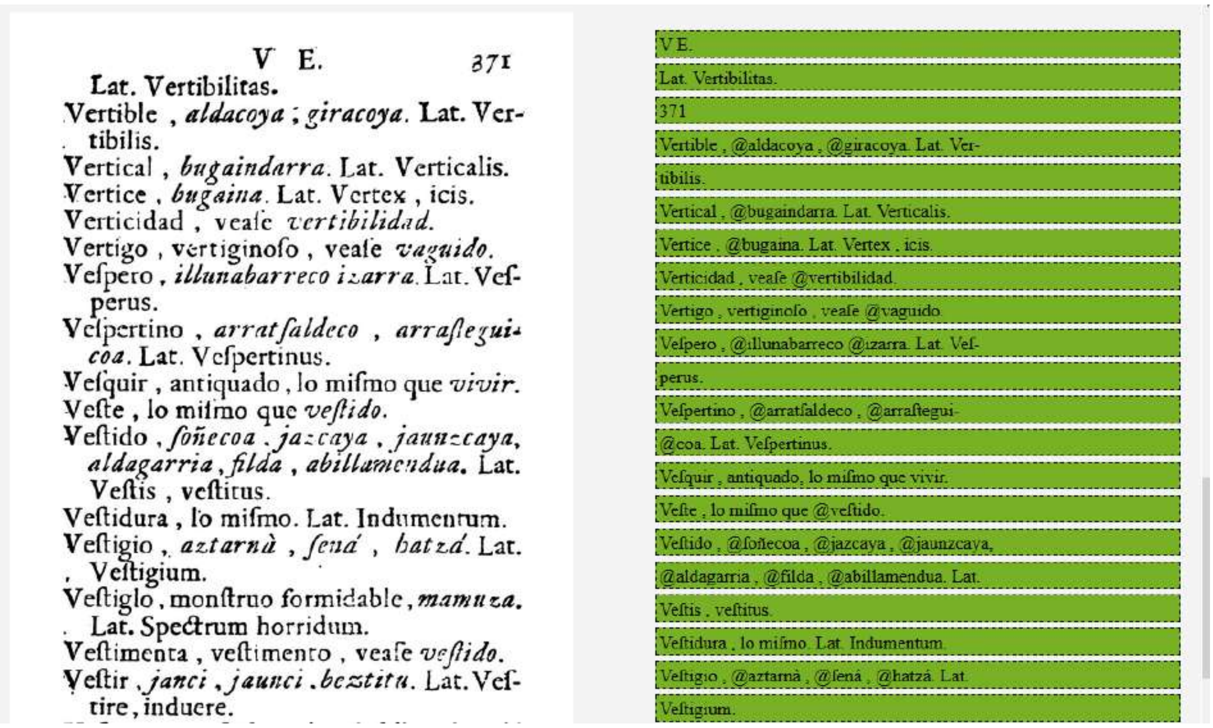
[6] This feature is not needed in hand-written text recognition, the task Transkribus was developed for.

[7] See http://kraken.re/training.html.

are eliminated, and stains are reduced. To this end, the ScanTailor application[8] has been used. We also have separated each LAR page into two, one for each of the two columns, in order to ease layout recognition to Kraken. Kraken's layout recognition module is then triggered, so that the files used in the transcription process are created.

2.2 Transcription

Kraken needs a training set, consisting of a certain amount of correctly transcribed lines. Before being given evidence from the training set, it is completely agnostic. In the guidelines, a set of 800 lines is recommended for training. It is clear that all characters that appear throughout the text to digitise have to appear in the training set. After transcribing one single two-column page of about 60 lines per column from scratch, we trained a Kraken model, and, from then on, corrected the OCR output page by page instead of transcribing from scratch. Any new corrected page was then introduced in the training set, in order to get constantly improved results which would require less corrections on the remaining pages. Despite very encouraging precision rates, that from the very first dictionary page on were clearly above the precision found in the available LAR txt versions, we realised that certain (infrequent) characters were not recognised. In the subsequent training sets, we added transcriptions of pages that contained the missing characters, mainly upper case letters that would appear massively in their corresponding alphabet sections. As the precision rates presented in Table 1 below suggest, overall precision has not significantly grown, but the infrequent characters formerly "unknown" to Kraken had now been properly identified.



V E. 371
Lat. Vertibilitas.
Vertible , *aldacoya ; giracoya*. Lat. Ver-
tibilis.
Vertical , *bugaindarra*. Lat. Verticalis.
Vertice , *bugaina*. Lat. Vertex , icis.
Verticidad , veaſe *vertibilidad*.
Vertigo , vertiginoſo , veaſe *vaguido*.
Veſpero , *illunabarreco izarra*. Lat. Veſ-
perus.
Veſpertino , *arratſaldeco* , *arraſtegui-
coa*. Lat. Veſpertinus.
Veſquir , antiquado , lo miſmo que *vivir*.
Veſte , lo miſmo que *veſtido*.
Veſtido , *ſoñecoa . jazcaya* , *jaunzcaya*,
aldagarria , *filda* , *abillamendua*. Lat.
Veſtis , veſtitus.
Veſtidura , lo miſmo. Lat. Indumentum.
Veſtigio , *aztarnà* , *ſená* , *hatzá*. Lat.
Veſtigium.
Veſtiglo , monſtruo formidable , *mamuza*.
Lat. Spectrum horridum.
Veſtimenta , veſtimento , veaſe *veſtido*.
Veſtir , *janci* , *jaunci* . *beztitu*. Lat. Veſ-
tire , induere.

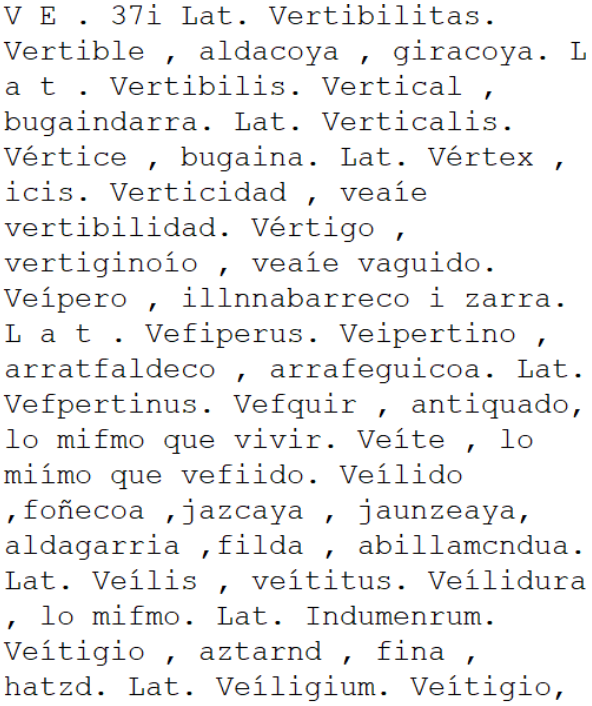
V E.
Lat. Vertibilitas.
371
Vertible , @aldacoya , @giracoya. Lat. Ver-
tibilis.
Vertical , @bugaindarra. Lat. Verticalis.
Vertice , @bugaina. Lat. Vertex , icis.
Verticidad , veaſe @vertibilidad.
Vertigo , vertiginoſo , veaſe @vaguido.
Veſpero , @illunabarreco @izarra. Lat. Veſ-
perus.
Veſpertino , @arratſaldeco , @arraſtegui-
@coa. Lat. Veſpertinus.
Veſquir , antiquado, lo miſmo que vivir.
Veſte , lo miſmo que @veſtido.
Veſtido , @ſoñecoa , @jazcaya , @jaunzcaya,
@aldagarria , @filda , @abillamendua. Lat.
Veſtis , veſtitus.
Veſtidura , lo miſmo. Lat. Indumentum.
Veſtigio , @aztarnà , @ſená , @hatzá. Lat.
Veſtigium.

Figure 2: Kraken OCR output, displayed by the transcription module

[8] Available at http://scantailor.org. ScanTailor is free software.

Transcriptions are performed inside a set of html files rendered by a web browser (see Figure 2). To each text box, which usually is a single text line on the scanned image recognised by the Kraken layout recognition module, a text field is provided. For creating the first training set, these fields are empty, and have to be filled with the text read by the user from the corresponding line. The modified page is then saved for inclusion in the training set. After a first OCR iteration, new html files are produced for a custom set of dictionary pages, and the text fields now contain the text recognised by Kraken based on the first model (derived from the first training set). From now on, the text in these fields is not typed in from scratch, but manually corrected. Corrected entire pages can be added to the training set, so that, in the next iteration, they are also considered for building the upgraded text recognition model, and so on, until the desired precision threshold is reached. Transcriptions must always reflect what is represented in typed letters in the original, without amendments or omissions, following the guidelines for Ground Truth Transcription.[9]

LAR contains two font styles, regular and italics. Kraken transcriptions are plain text without any markup, but the algorithm can deal with this using the following method: In the transcription, any word in italics is preceded by a sign not present in the whole resource, for which we chose an '@'. Kraken will learn that words written in italics, in the transcription should be preceded by this sign. As we could verify, this has worked out almost perfectly.



```
V E . 37i Lat. Vertibilitas.
Vertible , aldacoya , giracoya. L
a t . Vertibilis. Vertical ,
bugaindarra. Lat. Verticalis.
Vértice , bugaina. Lat. Vértex ,
icis. Verticidad , veaíe
vertibilidad. Vértigo ,
vertiginoío , veaíe vaguido.
Veípero , illnnabarreco i zarra.
L a t . Vefiperus. Veipertino ,
arratfaldeco , arrafeguicoa. Lat.
Vefpertinus. Vefquir , antiquado,
lo mifmo que vivir. Veíte , lo
miímo que vefiido. Veílido
,foñecoa ,jazcaya , jaunzeaya,
aldagarria ,filda , abillamcndua.
Lat. Veílis , veítitus. Veílidura
, lo mifmo. Lat. Indumenrum.
Veítigio , aztarnd , fina ,
hatzd. Lat. Veíligium. Veítigio,
```

Figure 3: BNE txt version of LAR

After transcribing 50 columns of about 60 lines each (i. e., 25 pages, cf. Figure 4), we assumed that precision would not significantly increase. In Table 1, we list the precision rates reached after each OCR iteration, with the amount of columns present in the training set. LAR, volume I and II, contains 1676 columns (two per page). This means that after manually transcribing less than 3% of the content we have gained a precision of nearly 98.5% in a txt version that covers the whole dictionary. This clearly outperforms the txt versions available before (cf. BNE version in Figure 3). Precision rates are calculated by Kraken, which uses a 10% share of the given training data as the evaluation set. Nevertheless, there is a drawback to take into account: Kraken's layout recognition module has worked out with a high precision, but still a considerable amount of lines have not been recognised. Either a line is not recognised at all, or lines are wrongly joined, so that a recognised text box range includes two real lines instead of one. Since there is no straightforward way to correct these mistakes manually, we had to leave this question for the (near) future, when our participation in a workshop related to Kraken will be possible.

[9] See https://ocr-d.de/en/gt-guidelines/.

Pages	2	3	4	5	6	7	8	9	10	11	12	13
Precision	0.9614	0.9417	0.9673	0.9648	0.965	0.9767	0.9764	0.9793	0.9808	0.9775	0.9845	0.9816
Pages	14	15	16	17	18	19	20	21	22	23	24	25
Precision	0.9773	0.9847	0.986	0.9829	0.9782	0.9794	0.9792	0.9869	0.9846	0.9839	0.9802	0.9812

Table 1: OCR precision rates.

At this stage, we cannot measure the impact of mistaken line recognitions, but our revisions of the OCR output during the transcription process and beyond suggest that it is worth digging deeper at this point, towards including a layout recognition validation step in the workflow. Anyhow, if the final resource, i.e. a digitised version of LAR has to obey quality criteria as for an edited publication, despite a very high OCR precision it seems necessary to manually validate all content, and this should include the correction of any errors due to mistakes in layout recognition. By tracking that effort we will get precise information concerning the precision of automatic layout recognition.

2.3 Result export

Kraken exports OCR results in different formats, and among them, txt, and ALTO XML,[10] an OCR result representation standard used e.g. by the Library of Congress, produced by some proprietary OCR engines, and supported as input format by the Elexifier toolchain (see section 3). While plain text contains just the recognised characters of each text box, separated by line breaks, ALTO XML also preserves the exact position of each text box on the page. This means, for example, that line indents are represented, which is an essential layout feature used for entry structure representation in LAR (headwords appear with a different indent than subsequent entry lines), and therefore is information worth considering. This information is also needed for publishing a digital version of the source document that includes active links between bits of text on the image and in the transcription.

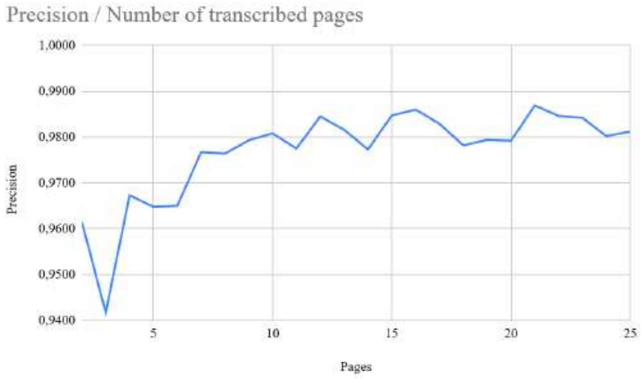


Figure 4: OCR precision rates

2.4 Wikisource

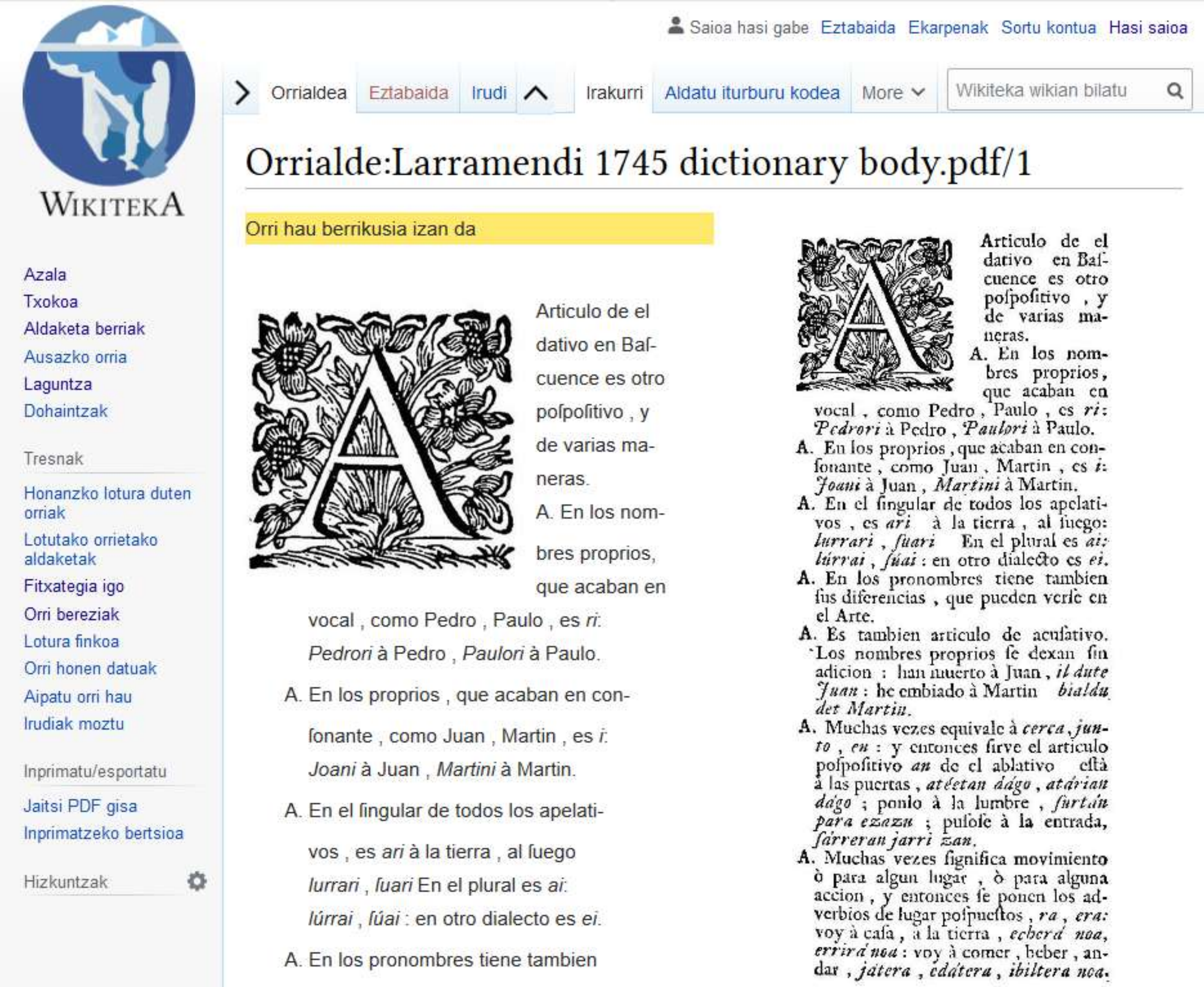
The Basque branch of Wikimedia's Wikisource platform, Wikiteka[11] (see Figure 5)[12] can be used for exhibiting and collaboratively editing OCR results: Anyone can view scanned images, along with their transcriptions, and edit the latter, in order to correct errors.

[10] See https://altoxml.github.io/.

[11] Accessible at https://eu.wikisource.org/wiki/Azala.

[12] See online at https://eu.wikisource.org/wiki/Orrialde:Larramendi_1745_dictionary_body.pdf/1.

The choice of that platform for collaborative OCR correction is due to Wikimedia Basque Country funding this small project; but a generally applying reason for that choice would be the fact that there exists an active community around Wikiteka, which has completely validated transcriptions of literary works of considerable size.[13] With this goal in mind, we have transformed the OCR result from ALTO XML format to Wikitext format.[14] Unfortunately, Wikitext format does not allow including text box position data, but



Orrialde:Larramendi 1745 dictionary body.pdf/1

Orri hau berrikusia izan da

Articulo de el dativo en Baſcuence es otro poſpoſitivo , y de varias maneras.

A. En los nombres proprios, que acaban en vocal , como Pedro , Paulo , es *ri*: *Pedrori* à Pedro , *Paulori* à Paulo.

A. En los proprios , que acaban en conſonante , como Juan , Martin , es *i*: *Joani* à Juan , *Martini* à Martin.

A. En el ſingular de todos los apelativos , es *ari* à la tierra , al ſuego *lurrari* , *ſuari* En el plural es *ai*: *lúrrai* , *ſúai* : en otro dialecto es *ei*.

A. En los pronombres tiene tambien

Articulo de el dativo en Baſcuence es otro poſpoſitivo , y de varias maneras.

A. En los nombres proprios, que acaban en vocal , como Pedro , Paulo , es *ri*: *Pedrori* à Pedro , *Paulori* à Paulo.

A. En los proprios , que acaban en conſonante , como Juan , Martin , es *i*: *Joani* à Juan , *Martini* à Martin.

A. En el ſingular de todos los apelativos , es *ari* à la tierra , al ſuego: *lurrari* , *ſuari* En el plural es *ai*: *lúrrai* , *ſúai* : en otro dialecto es *ei*.

A. En los pronombres tiene tambien ſus diferencias , que pueden verſe en el Arte.

A. Es tambien articulo de acuſativo. Los nombres proprios ſe dexan ſin adicion : han muerto à Juan , *il dute Juan* : he embiado à Martin *bialdu det Martin*.

A. Muchas vezes equivale à *cerca* , *junto* , *en* : y entonces ſirve el articulo poſpoſitivo *an* de el ablativo eſtà à las puertas , *atéetan dágo* , *atárian dágo* ; ponlo à la lumbre , *ſurtán para ezazu* ; puſoſe à la entrada, *ſárreran jarri zan*.

A. Muchas vezes ſignifica movimiento ò para algun lugar , ò para alguna accion , y entonces ſe ponen los adverbios de lugar poſpueſtos , *ra* , *era*: voy à caſa , à la tierra , *echerá noa*, *errirá noa* : voy à comer , beber , andar , *játera* , *edátera* , *ibiltera noa*.

Figure 5: LAR sample on Basque Wikisource platform

nevertheless we are able to represent line indents in Wikitext, which is the layout feature used in LAR for marking up headwords (negative indent), in opposition to consequent entry lines (normal indent). Using the text box position data present in ALTO, we have defined a filter that isolates text lines with negative indent, and from these, those lines which start with a capital letter that belongs to the pertaining alphabet section. From that subset of lines we chose the first part, i.e. until the first whitespace or punctuation.[15] These headword candidates have been enriched with a Wikitext markup that allows navigation

[13] See e.g. https://eu.wikisource.org/wiki/Gero for a Basque literature classic, or https://eu.wikisource.org/w/index.php?title=Berezi:OrrialdeGuztiak for a list of transcriptions.

[14] See documentation at https://en.wikipedia.org/wiki/Help:Wikitext.

[15] See code at https://github.com/dlindem/LBLR/blob/master/Larramendi/wikisource/wssarreraanchor.py.

inside the dictionary text (see Figure 6 below.) We have uploaded the plain text enriched in the described way to Wikiteka, together with the corresponding scanned images.[16] The task of correcting any errors, aiming to increase the transcription precision to 100%, is thus delegated to the community of Wikiteka users, which is open to anybody. General guidelines for transcription are given on the platform.[17] To that we add here some points to have in mind in this particular case, and as explanation of the sample shown in Figure 6:[18]

- Centred text (like the "A B." running title in Figure 6) will be preceded by five colons (":::::").
- Negative indent lines will be preceded by one colon (":").
- Other lines will be preceded by two colons ("::").
- Words in italics will be enclosed in pairs of single quotes (i.e. two "’", before and after the word).
- Line breaks and end-of-line hyphenations will be kept as in the scanned original.
- If the anchor markup element is not properly set, like in the second line of the example page in Figure 6, that shall be corrected. In this case, where the OCR tool has missed to identify the first capital letter ‘A’, the corrected line will start "{{sarrera|abandono}}Abandono".
- The anchor markup element that encloses headword candidates contains a single word. In the case of homograph headword candidates, that anchor includes a disambiguation number. If the anchor, instead of a single word, should enclose a multiword unit, the anchor shall be manually adapted, like for the entry with headword "abaratado demasiadamente" where "merquetueguia" and "merquequi ifinia" are listed as Basque equivalents: the anchor’s scope will be widened to two words, so that "{{sarrera|abaratado}}" (Figure 6) will be corrected to "{{sarrera|abaratado demasiadamente}}",[19] while leaving the following text as it is.

3. Information Extraction

Our method for isolating Spanish headword candidates described in the preceding section is entirely rule-based; it takes into account the text line position data present in ALTO format, and the correspondence of the first capital letter in that line to the pertaining alphabet section. We have defined as the headword candidate what precedes a whitespace or punctuation sign. Another method for defining headword candidates is to manually annotate headwords in a sample, and train a ML tool for predicting headword candidates in the whole dictionary text. Such a method can provide results that may be complementary to the rule-based approach.

Very recently, the ELEXIS project[20] launched Elexifier,[21] a toolchain supported by graphical user interfaces for information extraction from dictionaries. Dictionary content

[16] Accessible at https://eu.wikisource.org/wiki/Hiztegi_Hirukoitza. The scanned images are a processed version (see section 2.1) of the image collection distributed by Koldo Mitxelena public library.

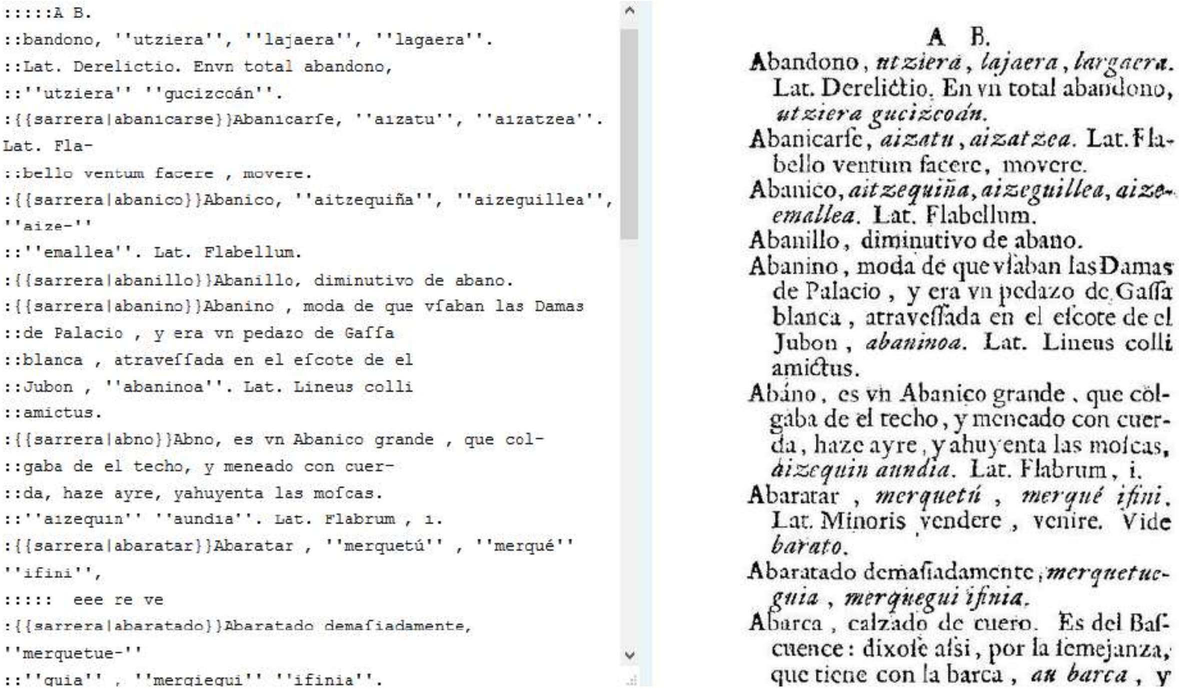
[17] For an English version, see https://en.wikisource.org/wiki/Help:Page_status.

[18] See online at https://eu.wikisource.org/w/index.php?title=Orrialde:Larramendi_1745_dictionary_body.pdf/4&action=edit.

[19] Note: "sarrera" is the Basque equivalent for "entry".

[20] See project homepage at http://elex.is.

[21] See https://elexifier.elex.is/. UPV/EHU has an observer status in the ELEXIS project, and among other activities, it is early adopter of the Elexifier toolchain, being this project a first use case. Other



```
::::A B.
::bandono, ''utziera'', ''lajaera'', ''lagaera''.
::Lat. Derelictio. Envn total abandono,
::''utziera'' ''gucizcoán''.
:{{sarrera|abanicarse}}Abanicarſe, ''aizatu'', ''aizatzea''.
Lat. Fla-
::bello ventum facere , movere.
:{{sarrera|abanico}}Abanico, ''aitzequiña'', ''aizeguillea'',
''aize-''
::''emallea''. Lat. Flabellum.
:{{sarrera|abanillo}}Abanillo, diminutivo de abano.
:{{sarrera|abanino}}Abanino , moda de que vſaban las Damas
::de Palacio , y era vn pedazo de Gaſſa
::blanca , atraveſſada en el eſcote de el
::Jubon , ''abaninoa''. Lat. Lineus colli
::amictus.
:{{sarrera|abno}}Abno, es vn Abanico grande , que col-
::gaba de el techo, y meneado con cuer-
::da, haze ayre, yahuyenta las moſcas.
::''aizequin'' ''aundia''. Lat. Flabrum , i.
:{{sarrera|abaratar}}Abaratar , ''merquetú'' , ''merqué''
''ifini'',
::::  eee re ve
:{{sarrera|abaratado}}Abaratado demaſiadamente,
''merquetue-''
::''guia'' , ''mergiegui'' ''ifinia''.
```

A B.

Abandono, *utziera*, *lajaera*, *largaera*. Lat. Derelictio. En vn total abandono, *utziera gucizcoán*.

Abanicarſe, *aizatu*, *aizatzea*. Lat. Flabello ventum facere, movere.

Abanico, *aitzequiña*, *aizeguillea*, *aizeemallea*. Lat. Flabellum.

Abanillo, diminutivo de abano.

Abanino, moda de que vſaban las Damas de Palacio, y era vn pedazo de Gaſſa blanca, atraveſſada en el eſcote de el Jubon, *abaninoa*. Lat. Lineus colli amictus.

Abáno, es vn Abanico grande, que colgaba de el techo, y meneado con cuerda, haze ayre, y ahuyenta las moſcas, *aizequin aundia*. Lat. Flabrum, i.

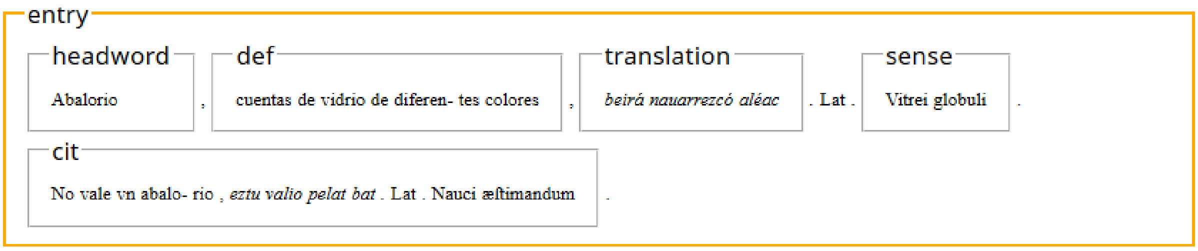
Abaratar, *merquetú*, *merqué iſini*. Lat. Minoris vendere, venire. Vide *barato*.

Abaratado demaſiadamente, *merquetueguia*, *merquegui iſinia*.

Abarca, calzado de cuero. Es del Baſcuence: dixoſe aſsi, por la ſemejanza, que tiene con la barca, *au barca*, y

Figure 6: Wikitext "source code" editable view on Wikiteka platform

available as text or rich text (in PDF format) or ALTO XML is parsed into an XML structured format that represents the structure of the dictionary, i.e. the division between entries, and inside the entry, the division into lexicographic items such as headwords, definitions, and translation equivalents. For this task, a ML application is trained by providing manually annotated training material.



Abalorio , cuentas de vidrio de diferen- tes colores , *beirá nauarrezcó aléac* . Lat . Vitrei globuli .

No vale vn abalo- rio , *eztu valio pelat bat* . Lat . Nauci æſtimandum .

Figure 7: Elexifier annotation module, graphical interface

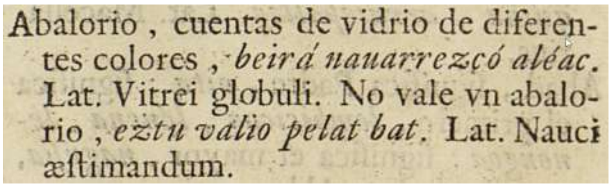
The Elexifier toolchain is currently in beta stage, and still subject to some feature restrictions. In particular, a limited tagset for the representation of microstructural items is available as for the current version: Entry, and as child elements of Entry: Headword, Translation, Sense, Part of Speech, Definition, and Example. The XML element tags that correspond to these lexicographical items are defined according to TEI-Lex0.[22]

use cases are planned. Hence, we were interested to test Elexifier in the workflow presented here. Another tool with similar features (that supports a more complete TEI tagset, but lacks a graphical interface), is GROBID-dictionaries, see https://github.com/MedKhem/grobid-dictionaries.

[22] In the framework of the Text Encoding Initiative (TEI), and DARIAH-EU working group "Lexical Resources", co-funded by the ELEXIS project, a tagset for representation of dictionary content has been developed and proposed as standard, in order to ensure interoperability of lexical datasets, see https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html.

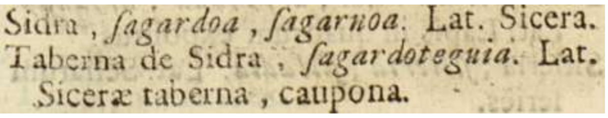
We have annotated a sample of LAR assigning tags to entries ("entry"), headwords ("headword"), definitions ("def"), Basque translations ("translation"), examples and notes ("cit"), and Latin translations (due to the limitations in the available tagset, "sense").[23] This can be observed in the screenshot image from the Elexifier annotation module reproduced as Figure 7, together with an image of the original entry (Figure 8).

Following the recommendations given in Elexifier documentation, the annotation has been carried out for twenty columns (ten dictionary pages), and then used as training set for the Elexifier segmentation ("information extraction") module, which structures the content of the whole dictionary according to what it has been given as training set.



Abalorio , cuentas de vidrio de diferentes colores , *beirá nauarrezçó aléac.* Lat. Vitrei globuli. No vale vn abalorio , *eztu valio pelat bat.* Lat. Nauci æſtimandum.

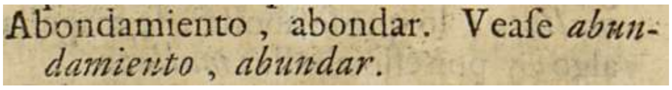
Figure 8: LAR entry example, scanned image



Sidra , *ſagardoa* , *ſagarnoa.* Lat. Sicera. Taberna de Sidra , *ſagardoteguia.* Lat. Siceræ taberna , caupona.

Figure 9: LAR entry example, scanned image

A first evaluation of the information extraction results suggests that Spanish headwords and Basque translation equivalents have been recognised by the software with high precision. Latin equivalents, the third category we have looked at, have been recognised with much lower precision. Headwords seem to be recognised seamlessly, which should be due to the fact that headwords are positioned in the entry layout in a first negatively indented line, and followed by a comma. This has been the case in all annotated entries, and thus is a very straightforward criterion for the ML algorithm. On the other hand, also items that do not describe headwords are placed in a negatively indented line, and subsequently, have been identified as headwords. This is the case for the items listed in the example shown in Figure 1 above, between "acostar" and "acostamiento", where non-canonical inflected and combined forms of the preceding headword (such as "estar acostado"), and even items representing grammatical information that serve for introducing a list of inflected forms appear in that position. Figure 9 also contains an example for a sub-entry that appears just as headwords appear, but in this case, not only breaking alphabetical order but totally out of the scope of the current alphabet section. Latin equivalents, as can be observed in Fig. 8, will appear for headwords, but also as translation of usage examples; here we have contradictory evidence that makes the algorithm unable to predict the correct annotation for Latin items in many cases.



Abondamiento , abondar. Veaſe *abundamiento* , *abundar.*

Figure 10: LAR entry example, scanned image

Basque equivalents are not that clearly identifiable, since their layout feature (italics font style) is also present in examples (Basque translation of the idiomatic usage example, see Fig. 8), and also in Spanish to Spanish

[23] According to TEI-Lex0, "sense" is not at all defined as adequate for annotating translation equivalents; it groups word senses within entries. Due to the lack of an appropriate tag in the current version of Elexifier, we have nevertheless chosen this workaround.

cross-references, as e. g. in "Acovardar" in Fig. 1 above,[24] and in the example shown in Fig. 10. In fact, "abundamiento" and "abundar", which in that example entry, correctly identified, would represent Spanish headwords where a cross-reference leads to, have been identified by the software as Basque equivalents. This should be solved by annotating cross-references (that in the dictionary text are preceded or followed by the structure markers "vease" or "lo mismo que") with a special tag, not available in the present version of Elexifier,[25] so that the segmentation algorithm gains evidence for identifying words preceded by such structure marker as cross-reference, regardless of their font style.[26]

Compiling the training set for Elexifier, in cases of multiword items as headwords, we have annotated it accordingly as multiword headword (i.e., for example, "abaratado demasiadamente", with "merquetueguia" as equivalent). Using that evidence, Elexifier has identified multiword headwords in 1,925 cases in the whole dictionary text. If we compare the results of both methods (rule-based and ML-based) regarding the whole headword list, we gain the figures shown in Table 2:

	LAR, rule-based	**LAR, ML-based**
Spanish Headwords	36,451	29,932
- of which appear in DA	33,015	25,057
- of which are multiword items	0	1,925

Table 2: Items identified as headword

The ruleset for spelling normalisation and comparison will be explained in the following.

4. Merging historical lemma lists

In order to compare the Spanish lemma list extracted from LAR to DA lemma list,[27] we have performed a normalisation of lemma-signs found in both resources. This step is necessary for defining pairs of matching lemmata that from resource to resource show different written representations. For the purpose of achieving mappings such as those represented in Table 3, we processed all lemma-signs of both dictionaries with the unidecode Python module,[28] which removes all diacritics and replaces non-canonical (non-ASCII) characters with the most similar canonical one. We excepted the "ñ" letter, canonical in Spanish, from that replacement, preventing it from being converted to "n".

[24] As can be observed, this layout feature is not strictly applied: in the follwing entry "acoyundar", the cross-referenced headword is not printed in italics.

[25] As explained above, we have used all available tags, so that, for cross-references, in this first experimental iteration, we had no remaining option.

[26] As soon as Elexifier offers a full-fledged tagset, we will repeat the process. An interim solution for identifying such cross-reference items in the output of Basque translation equivalents, we can check for the presence of the items in the DA headword list, which for the example solves the problem, since "abundamiento" and "abundar" both are listed there as headwords. The Basque item, on the other hand, should be checked for if it is a homograph translation of a Spanish headword (the headword of the same entry), such as LAR Basque equivalent "saca" for Spanish "saca".

[27] The DA lemma list is available at http://web.frl.es/DA_Preliminares/DA_lemario.pdf. This list contains headwords only. Other parts of the digitised DA are accessible only through a graphical user interface that allows one-by-one queries by lemma (available at http://web.frl.es/DA.html). The unabridged content is not publicly accessible in any other format than on paper.

[28] Available at https://pypi.org/project/Unidecode/.

Also, we converted all upper case characters to lower case, and double s to single s (historical "ſ" having been converted to "s" by the unidecode module). As the examples listed in table 3 show, "ss" (which in 18th century Spanish was still frequent) and diacritics are not used in the same way, and also their use inside LAR and DA is not concise. We will evaluate the described normalisation process in detail, having in mind related work about historical Spanish, which uses an approach based on Levenshtein distance thresholds, which is more flexible, but prone to yield false-positive mappings (Porta et al., 2013). We then wrote normalised lemma-signs from LAR and DA, their original written representations, and, for LAR, also Basque equivalents, as elements into the same XML tree, so that we were able to produce the datasets[29] listed in Table 4.[30]

LAR	**DA**	**matching normalised lemma sign**
Obsession	obsessión	obsesion
Hueſſo	huesso	hueso
Occiſſion	occisión	occision
Atràs	atras	atras

Table 3: Lemma-sign normalisation mappings

#	**List**	**Rule or ML**	**Rule**	**ML**	**Rule and ML**
1	LAR: all lemmata	32,700	30,045	27,125	24,470
2	Union of LAR and DA: all lemmata	46,843			
3	Lemmata appearing in LAR, but not in DA	4,875	2,431	4,875	2,431
4	Lemmata appearing in DA, but not in LAR	14,143	14,354	19,718	19,929
5	LAR and DA: intersection	27,825	27,614	22,25	22,039
6	All items extracted as LAR Basque equivalent candidates	60,193	58,235	38,300	36,342
7	LAR equivalents that also appear in SAR, WD, or OEH	15,152	14,886	11,551	11,285
8	LAR equivalents that also appear in SAR with "1745" datation	3,134	3,088	2,508	2,461
9	LAR equivalents that also appear in SAR with "1745" datation, and in WD	1,478	1,456	1,201	1,179
10	LAR equivalents with attestation in Wikidata (2021-01)	1,416	1,396	1,151	1,131

Table 4: Produced datasets

Besides that, we produced an index of LAR, that is, a version where Basque lexical items point to their Spanish equivalents, the original lemmata. In Table 4 (6-10), we show the amounts of Basque items extracted using both methods. Based on rules, we got all items printed in italics, that is, as explained above, not only Basque items, but all content printed in italics. We compare these amounts with those obtained from the Elexifier tool, for which we had manually annotated a sample, as explained in section 3. We have developed a set of rules for linking historical spellings of Basque lexical items to standard

[29] These datasets are available at http://lexbib.org/larramendi.

[30] For this task, we have used the TLex Dictionary Writing System, see https://tshwanedje.com/tshwanelex/.

spelling, similar to the approach used for matching Spanish LAR and DA headwords, but that contains a total of 36 regular expressions, to be executed in a certain order.[31] The ruleset is discussed in detail in Alonso Arrospide (2021). We then mapped LAR Basque equivalents in their written representation as modified by the ruleset to the lemma lists of SAR (Sarasola, 1996), OEH (Mitxelena & Sarasola, 1988) dictionaries, and Wikidata Basque lexemes (WD), with the results listed in Table 4. These datasets now are available for further research that can also include quantitative methods, although, for this version of the datasets we must stress the fact that transcription precision is below 100%.[32] Having a closer look at the data, for example, in list (4), 457 items can be found that describe superlative inflected adjective forms (e. g., "alegrissimo", "aliviadissimo"), which apparently are referenced systematically as lemmata in DA, but not in LAR. This suggests that groups of lemmata present in DA but missing in LAR can be, at least in part, identified in groups. This list obviously also contains those LAR headwords that have not been properly converted to text in the OCR process, or that constitute an orthographical variant that has not been handled in the normalisation process. List (3), in turn, also contains headwords that due to OCR errors or failed normalisation have not been mapped to their counterpart in DA, but in addition to these, it contains those headwords that have been added by Larramendi, without having had reference in DA (e.g. "derecho natural", in Basque, "sortaraudea", "sorneurtartea").

5. Enriching Wikidata

5.1 Wikidata Lexemes

In section 4, we have shown how we have performed a merging of historical lemma lists, a process that also can be seen as a linking of lexical resources, at the lemma sign level (i.e., without regard to part of speech or word sense disambiguation). We have taken two resources into consideration, LAR and DA. In order to link a lexical dataset to more and different resources, the workflow proposed for Elexifier resorts to the already mentioned TEI-Lex0 XML annotation scheme, which has been developed for that purpose (Bański et al., 2017).

Another way to link lexical data, which can be characterised as an upcoming trend regarding Linked Open Data,[33] is to make use of Wikidata lexemes. Wikidata represents entities such as concepts, lexemes, and properties that describe relations between the former, according to the Resource Description Framework (RDF). RDF uses semantic triples consisting of subject, predicate, and object, for the representation of statements, which can be visualised through the Wikidata graphical interface[34] or retrieved through a query interface using SPARQL.[35]

If we look at how a lexeme in Wikidata is linked to the concept it denotes on one hand, and to translation equivalents on the other, we find that while in dictionaries statements about

[31] See ruleset at https://github.com/dlindem/LBLR/blob/master/Larramendi/erkaketak_eus_elexifier/rules.csv.

[32] See all results at http://lexbib.org/larramendi, including detailed merged subsets of all discussed dictionaries, and access to Wikidata attestations.

[33] For this concept and a short overview on the topic, see e.g. https://en.wikipedia.org/wiki/Linked_data#Linked_open_data.

[34] See http://wikidata.org.

[35] See https://en.wikipedia.org/wiki/SPARQL.

lexemes are encoded as lexicographic items, so that a human user can discriminate them by structural design features, here we are in front of statements coded in machine-readable semantic triples. For example, the English noun "magic"[36] is furnished with statements about its attestation (with an OED online entry ID as URI for the reference), with word senses and translations that belong to a certain sense, and with a link to the (ontological) concept denoted by a one of the senses, which is shortly defined as "type of beliefs and practices involving supernatural acts", member of class "occult" and part of "Magic and Religion", which is further described in a Wikipedia article entitled "Magic (supernatural)". Another Wikipedia article, "magic", describes another Wikidata entity, member of the classes "circus skill" and "performing arts", and that is linked to a different sense of the same lexeme "magic".

Translation equivalence is expressed in two ways in and around Wikidata. On the one side, Wikidata items that correspond to word senses,[37] i.e. not lexical but ontological items, are multilingually labelled. On the other side, translation equivalence can be expressed between lexemes, and between senses of lexemes, using a set of properties and classes related to lexical data defined in Wikidata itself,[38] but also using the linked data vocabularies developed by the OntoLex-Lexica Community group inside W3C,[39] which is a collection of RDF models that is used also in Wikidata. In the following, we describe how to link the historical lexical data on hand to lexical data contained in Wikidata.

5.2 Linking attestations

As we have mentioned, Wikidata contains not only ontological concepts (entity URI starting with a "Q"), but also lexemes (URI starting with an "L"). Senses of Lexemes can be linked to the concepts they denote using Wikidata P5137 property ("item for this sense").

The Elhuyar Foundation, a major dictionary publisher in the Basque Country,[40] has recently shared the Basque lemmata contained in their Basque-Spanish bilingual dictionaries on Wikidata. In any case, we shall not propose creating new Wikidata lexemes for the (historical) Basque lexical forms extracted from LAR, but rather link them to existing lexemes, as attestation. This is not trivial, since we have to deal with historical spelling, as discussed in section 3, and with the part of speech, a property Wikidata lexemes are furnished with by default.

For a first iteration, we have chosen those lexical items identified as Basque by both the rule-based and the ML-based approach, that, at the same time, could be mapped to items present in Wikidata, and to items present in SAR, and that are marked in that dictionary with the attestation datum "1745". In other words, we chose those 1,179 items which are double-checked to appear in LAR by SAR dictionary. To be sure to avoid mistaken part-of-speech mappings, from those we chose the 1,131 items which do not appear with an ambiguous part of speech on Wikidata. Wikidata lexemes data model does not foresee

[36] See https://www.wikidata.org/wiki/Lexeme:L3.

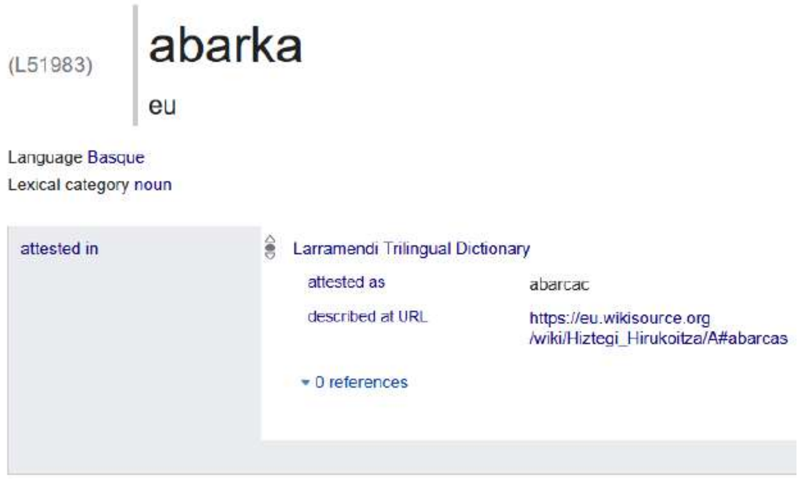
[37] For this link, property http://www.wikidata.org/entity/P5137 is used ("item for this sense").

[38] See https://www.wikidata.org/wiki/Wikidata:Lexicographical_data/Documentation.

[39] See https://www.w3.org/community/ontolex/, and related publications (McCrae et al., 2017; Bosque-Gil et al., 2017).

[40] Elhuyar dictionary portal is accessible at https://hiztegiak.elhuyar.eus/.

more than one part of speech assigned to a lexeme, and Basque lexemes are represented according to that, so that lexemes with a different part of speech that share the same written representation (which certainly is not infrequent in Basque)[41] are represented as distinct lexemes. Since LAR does not contain part of speech data, and, on the other hand, lemma and equivalents in LAR often do not share the same part of speech, such disambiguation at the homograph level could not be carried out in this first iteration; most probably, manual work will be required here.[42]



(L51983) abarka

eu

Language Basque

Lexical category noun

attested in — Larramendi Trilingual Dictionary

attested as — abarcac

described at URL — https://eu.wikisource.org/wiki/Hiztegi_Hirukoitza/A#abarcas

Figure 11: Wikidata lexeme attestation

We have used Wikidata property P5323, "attested in",[43] for the attestation statement, together with P7855, "attested as",[44] and P973, "available at URL"[45] as qualifiers to that statement, that is, the claim that a lexeme is attested in LAR is further described, providing the attested written representation, and the reference to the corresponding dictionary entry, which is a hyperlink pointing to a headword anchor in the dictionary text on the Wikiteka platform (see section 2.4).[46] Since that text is aligned at page level with the facsimile image version, full reference to the attestation source is guaranteed.

6. Outlook

It was the purpose of our small study to run through the whole digitisation process for a historical dictionary, starting from scanned images, with this being one of the harder tasks to solve for texts of this age. In this paper, we have tried to make our workflow transparent. We have pointed out achievements and drawbacks encountered at each stage. Although the OCR process did not yield 100% precision, we have sent the output to the next stage in the pipeline, i.e. information extraction, which has also not brought error free results. Nevertheless, we believe that we have showed what automatic tools can do for us, and that the datasets we have been able to create with a very reduced manual validation effort already have something to offer to further research. Since we have used open software tools provided by the research community, and Wikimedia-related communities, this workflow is easily reproducible. For the near future, we propose to manually validate the ALTO XML content we have produced, using the Wikiteka platform, which allows this to be a

[41] Also in English this is not at all infrequent (cf. items like 'sound', with three part of speech values assigned in dictionaries).

[42] Another option would be a lexical data model that foresees a part-of-speech assignation at a level lower than the lemma sign, i. e., either between lemma and sense, or inside the sense.

[43] See https://www.wikidata.org/wiki/Property:P5323.

[44] See https://www.wikidata.org/wiki/Property:P7855.

[45] See https://www.wikidata.org/wiki/Property:P973.

[46] See the statements shown in Fig. 11 online at http://www.wikidata.org/entity/L51983.

community-driven effort. Based on the tracked working time spent on transcription, we estimate an average of 15 minutes for correcting a dictionary column's transcription, that is, around 425 working hours for producing a ground truth transcription of the whole dictionary.

We then propose to take actions for improving precision in information extraction. That is, to annotate a larger training set for the Elexifier tool, and to make use of a more complex tag set. That would also mean annotating microstructural items other than translation equivalents, such as examples and cross-references.

We finally want to further develop the proposed model for integration in Wikidata. We are currently discussing the possibility to use an own instance of Wikibase,[47] i.e. the software solution that drives Wikidata, as a separate ecosystem for the development of linked (Basque) lexical datasets. Such a parallel resource would serve as infrastructure for collaborative research on converting plain dictionary text into structured datasets, its integration with other kinds of lexical resources, its representation as Linguistic Linked Data, and ultimately, regarding sufficiently validated lexical data, its transfer to the main Wikidata platform.

7. Acknowledgements

In this paper, we present methodology and findings from an eight months project started in January 2020, funded by Wikimedia Basque Country and UEU Udako Euskal Unibertsitatea. Our research has also received funding from the Basque Government Eusko Jaurlaritza (IT 665-13).

8. References

Alonso Arrospide, M. (2021). *Larramendiren Hiztegi Hirukoitzaren digitalizazioa.* Master's thesis, UPV/EHU University of the Basque Country, Vitoria-Gasteiz. URL http://doi.org/10.13140/RG.2.2.27926.68169.

Bański, P., Bowers, J. & Erjavec, T. (2017). TEI-Lex0 Guidelines for the Encoding of Dictionary Information on Written and Spoken Forms. In I. Kosem, C. Tiberius, M. Jakubíček, J. Kallas, S. Krek & V. Baisa (eds.) *Electronic lexicography in the 21st century: Lexicography from scratch. Proceedings of eLex 2017.* Brno: Lexical Computing CZ s.r.o., pp. 485–494. URL https://elex.link/elex2017/wp-content/uploads/2017/09/paper29.pdf.

Bosque-Gil, J., Gracia, J. & Montiel-Ponsoda, E. (2017). Towards a Module for Lexicography in OntoLex. In *Proceedings of the 1st Workshop on the OntoLex Model (OntoLex-2017).* Galway/Gaillimh, pp. 74–84. URL http://ceur-ws.org/Vol-1899/OntoLex_2017_paper_5.pdf.

Larramendi, M. (1745). *Diccionario trilingüe castellano, bascuence y latin dedicado a la M.N. y M.L. provincia de Guipuzcoa.* San Sebastián: Bartholomé Riesgo y Montero.

Lindemann, D., Khemakhem, M. & Romary, L. (2018). Retro-Digitizing and Automatically Structuring a Large Bibliography Collection. In *European Association for Digital Humanities (EADH) Conference.* Galway/Gaillimh, Ireland: National University of Ireland. URL https://hal.archives-ouvertes.fr/hal-01941534/.

[47] See https://wikiba.se/.

Lindemann, D. & San Vicente, I. (2020). Baliabide lexikoen sarea: Baldintza filologiko eta tekniko zenbait. In *Hitzak sarean: Pello Salabururi esker onez*. Bilbo: UPV/EHU Argitalpen Zerbitzua, pp. 79–96. URL http://www.ehu.eus/ehg/salaburu/liburua/HitzakSarean06.pdf.

McCrae, J., Bosque-Gil, J., Gracia, J., Buitelaar, P. & Cimiano, P. (2017). The OntoLex-Lemon Model: Development and Applications. In I. Kosem, C. Tiberius, M. Jakubíček, J. Kallas, S. Krek & V. Baisa (eds.) *Electronic lexicography in the 21st century: Lexicography from scratch. Proceedings of eLex 2017*. Leiden: Lexical Computing, pp. 587–597. URL https://elex.link/elex2017/wp-content/uploads/2017/09/paper36.pdf.

Mitxelena, K. & Sarasola, I. (1988). *Diccionario general vasco - Orotariko euskal hiztegia*. Euskaltzaindia; Editorial Desclée de Brouwer.

Porta, J., Sancho, J.L. & Gómez, J. (2013). Edit transducers for spelling variation in Old Spanish. In *Proceedings of the workshop on computational historical linguistics at NODALIDA 2013; May 22-24; 2013; Oslo; Norway. NEALT Proceedings Series 18*. Linköping University Electronic Press, pp. 70–79.

Real Academia Española (ed.) (2013). *Diccionario de autoridades: 1726 - 1739*. Boadilla del Monte (Madrid): JdeJ Editores, ed. facs. con motivo del iii centenario edition.

Romanov, M., Miller, M.T., Savant, S.B. & Kiessling, B. (2017). Important New Developments in Arabographic Optical Character Recognition (OCR). *arXiv:1703.09550 [cs]*. URL http://arxiv.org/abs/1703.09550.

Sarasola, I. (1996). *Euskal Hiztegia*. Donostia: Kutxa Gizarte-eta Kultur Fundazioa.

Urgell, B. (1998a). "Hiztegi Hirukoitza" eta "Diccionario de Autoridades" erkatuaz (I): oinarrizko ezaugarri zenbait. *Anuario del Seminario de Filología Vasca "Julio de Urquijo"*, 32(1), pp. 109–163. URL https://www.ehu.eus/ojs/index.php/ASJU/article/view/8709.

Urgell, B. (1998b). "Hiztegi Hirukoitza" eta "Diccionario de Autoridades" erkatuaz (II): sarreraren edukia. *Anuario del Seminario de Filología Vasca "Julio de Urquijo"*, 32(2), pp. 365–414. URL https://www.ehu.eus/ojs/index.php/ASJU/article/view/8717.

Urgell, B. (2002). *Euskal Lexikografia. Irakaskuntza proiektua*. UPV/EHU. URL https://www.academia.edu/3481533/Euskal_Lexikografia._Irakaskuntza_proiektua.

A Use Case of Automatically Generated Lexicographic Datasets and Their Manual Curation

Dorielle Lonke[1], Raya Abu Ahmad[1], Volodymyr Dzhuranyuk[1], Maayan Or Ner[1], Ilan Kernerman[1]

[1] K Dictionaries, Tel Aviv
E-mail: dorielle@kdictionaries.com, raya@kdictionaries.com, vova@kdictionaries.com, maayan@kdictionaries.com, ilan@kdictionaries.com

Abstract

This paper provides an overview of a multi-layer project combining machine and manual processes in linking multilingual lexicographic resources and leading to the generation of over 200 new language pairs and the update of over 50 existing ones. In the first phase, we create multilingual glossaries by reversing entries from the Password English multilingual dataset of K Dictionaries, reformulating the L1 translations into headwords, aligning them to the original English entries that become their translations, and adding the other language translations of those English entries. The reversal is supplemented by rule-based algorithms to reduce noise; merge, duplicate and separate entries; and check duplicate senses for similar or identical definitions and examples of usage. This is followed by manual detection and amendment of erroneous grammatical categories and faulty meanings, and editing the translation links. The next phase concerns cross-linking each semi-automatically generated multilingual glossary from the first phase with another full lexicographic resource of that L1 from the Global Multilingual Data Series, including its own bilingual versions whenever available. We present the main tasks involved in this project, featuring the automated operations combined with post-editing, the outcomes, our conclusions and further plans.

Keywords: auto-generated data; automatic post-editing; semi-automated processes; manual curation; resource cross-linking

1. Introduction

The creation of up-to-date lexical resources is increasingly facilitated and enhanced by the myriad of methodologies and technologies available for natural language understanding, generation and processing. Traditional requirements and techniques associated with manual compilation of dictionary entries are, on the one hand, empowered by a wide array of automated processes while, on the other hand, supplemented by emerging challenges that stem from these very same processes and others that open new capacities and options for merging different resources with each other.

This paper describes a pipeline of resource convergence and production facets that combine automated processes with manual curation. We begin with crosslingual datasets created by reversing the Password English multilingual dictionary into L1-

English word-to-sense glossaries – by reformulating the L1 translations into headwords, linking them to the original English headwords that become their translations, and adding the other language translations of those English entries – and merge each new L1 resource with another resource of that L1 – some of which are monolingual, bilingual or multilingual – in creating numerous new L1 pairs. The merging process can be outlined as follows:

(1) Use the Password English multilingual dictionary resource (R1).

(2) Reverse R1 – transforming the translations into headwords and the English headwords into their translations – thus producing an L1 to English dataset (R2).

(3) Add the other language translations from R1 onto R2 – using the new English translations as pivots – thus generating an L1 multilingual dataset (R3).

(4) Use another resource of each L1, which may be monolingual, bilingual or multilingual, from the Global Multilingual Data Series of K Dictionaries (R4).

(5) Merge R3 and R4, thus generating a new L1 multilingual resource (R5).

(6) Divide R5 into bilingual sets, thus producing a series of language pairs (R6).

The entire project comprises 19 source languages and 15 target languages (of which 10 are also source languages), so the total number of R6 is 275 language pairs (10x14 + 9x15), involving 25 different languages altogether. Approximately one fifth of these (a little over 50 pairs) were already available in R4, so their corresponding R6 pairs have been updated in the process, whereas all the other language pairs are new. The source and target languages are listed in Table 1.

The pipeline relies on various behind-the-scenes automatic software operations of diverse complexities, with manual editing taking place particularly in curating R2 by means of the specially designated K Index Editorial Tool (KIET), which is used by the editors to review and revise the L1 headword candidates, validate their auto-attributed parts of speech (POS), link to the English equivalents and determine their sense hierarchy, thus detecting and amending erroneous grammatical categories and faulty meanings. The automated processes include rule-based algorithms that reduce noise and merge duplicate entries and senses and check for similar or identical definitions and examples. The rules that serve in this process are devised in accordance with the structure of each target language, taking into consideration semantic variances between English senses and their corresponding translations. Missing POS categories are further provided by matching parallel headwords from a different resource, and more information is introduced from R1, which is later expanded onto matching non-identical but similar POS categories and annotating the glossary to distinguish single lemmas and multiword expressions (MWEs) based on automatic detection. The editor’s manual intervention is minimised by integrating simple rules deduced from repeated evidence of the same error, avoiding redundancies and repetitive amendments of erroneous patterns. Some of the challenges in the post-editing tasks include the detection of such repetitive rules and validating the resulting algorithm, a process which is still mostly done through manual revision and proofing.

Source Languages	**Target Languages**
Arabic	
Chinese Simplified	Chinese Simplified
Czech	Czech
Danish	
Dutch	
	English
	French
	German
Greek	
Hebrew	
Hindi	
Italian	Italian
Japanese	Japanese
Korean	Korean
Norwegian	
Polish	Polish
Portuguese Brazil	Portuguese Brazil
Portuguese Portugal	Portuguese Portugal
Russian	Russian
	Spanish
Swedish	
Thai	
Turkish	
	Ukrainian
	Vietnamese

Table 1: The source and target languages

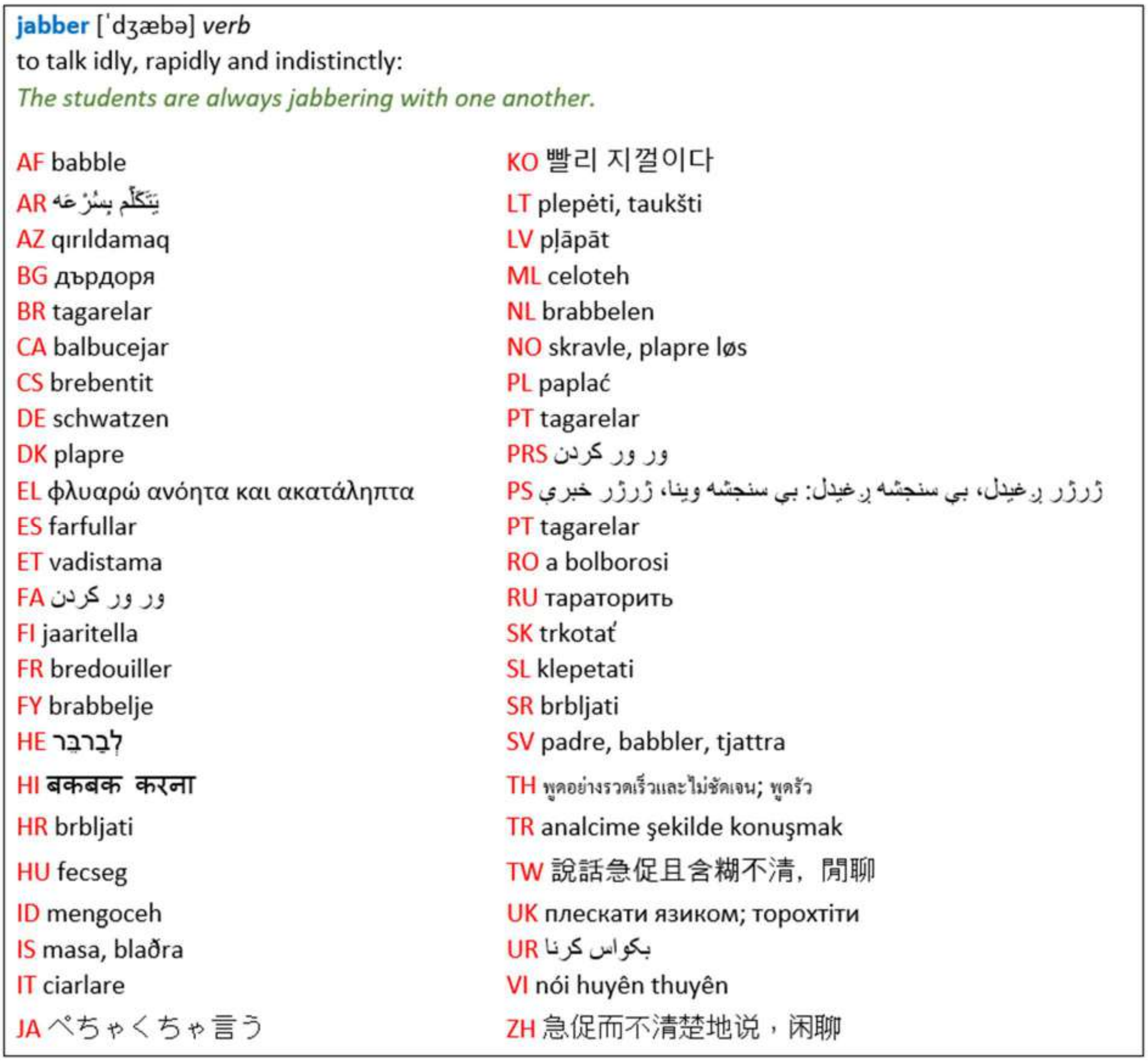
Section 2 of this paper presents R1, the automatic reversal process and KIET. The actual post-editing of R2 is described in Section 3, along with corresponding automated tasks to produce R3 and combine data components from R4, and the final convergence of R5 is described in Section 4. Section 5 summarises the outcomes of the project and forecasts next steps.

2. The K Index Editorial Tool, its Background and By-products

This section describes the automatic reversal of the English multilingual dictionary (R1), the generation of bilingual (R2) and then multilingual glossaries (R3), and post-editing R2 with the K Index Editorial Tool (KIET).

2.1 The Password English Multilingual Dictionary Resource

The Password English multilingual dictionary (R1) consists of English entries with translation equivalents in nearly fifty languages. The headwords are supplemented with phonetic transcription (IPA) and alternative scripts, POS, grammatical number and sub-categorisation. Each sense of the entry includes a definition and example(s) of usage, and MWEs appear as sub-entries. The translations offer a brief equivalent of each sense and MWE. Figure 1 presents a sample monosemous entry.



jabber [ˈdʒæbə] *verb*
to talk idly, rapidly and indistinctly:
The students are always jabbering with one another.

AF babble
AR يَتَكَلَّم بِسُرْعَة
AZ qırıldamaq
BG дърдоря
BR tagarelar
CA balbucejar
CS brebentit
DE schwatzen
DK plapre
EL φλυαρώ ανόητα και ακατάληπτα
ES farfullar
ET vadistama
FA ور ور کردن
FI jaaritella
FR bredouiller
FY brabbelje
HE לְבַרְבֵּר
HI बकबक करना
HR brbljati
HU fecseg
ID mengoceh
IS masa, blaðra
IT ciarlare
JA ぺちゃくちゃ言う

KO 빨리 지껄이다
LT plepėti, taukšti
LV pļāpāt
ML celoteh
NL brabbelen
NO skravle, plapre løs
PL paplać
PT tagarelar
PRS ور ور کردن
PS ژرژر ږغیدل، بي سنجشه ږغیدل: بي سنجشه وينا، ژرژر خبري
PT tagarelar
RO a bolborosi
RU тараторить
SK trkotať
SL klepetati
SR brbljati
SV padre, babbler, tjattra
TH พูดอย่างรวดเร็วและไม่ชัดเจน; พูดรัว
TR analcime şekilde konuşmak
TW 說話急促且含糊不清，閒聊
UK плескати язиком; торохтіти
UR بکواس کرنا
VI nói huyên thuyên
ZH 急促而不清楚地说，闲聊

Figure 1: The entry *jabber* in the Password English multilingual resource

2.2 The Reversal Process

The L1-English data are compiled in the process of reversing R1, followed by post-editing R2 as regards the new headwords and POS categories, their links to the English translations and reordering the corresponding senses, including additions or omissions for the auto-generated raw dataset. The L1 entry is created by deriving all identical translations of English entries in R1. The translations are grouped by their POS category and presented to the editor with the original English headword and definition. The editor then determines a new sense order, relying on the English definition as a basic sense indication. This process occurs within the KIET editorial interface. The compilation program follows the algorithm below:

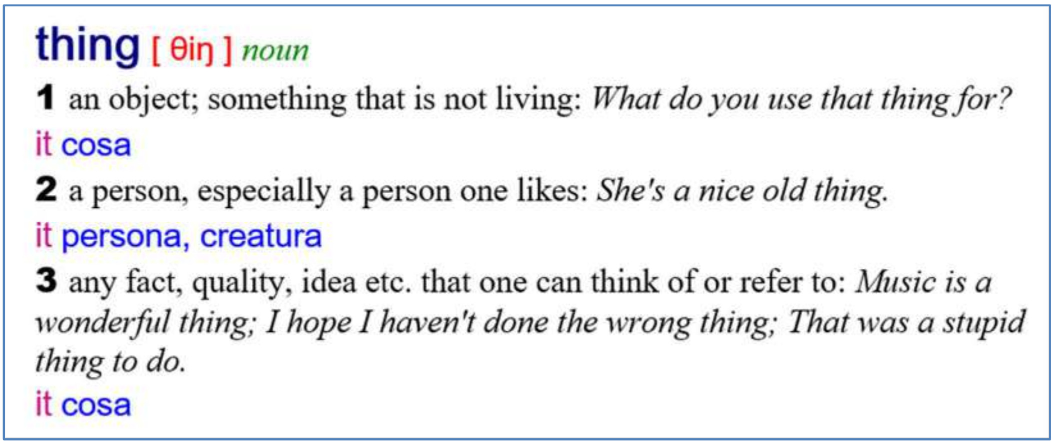
(1) The program runs through all the R1 entries and their corresponding senses. For each sense, it retrieves the translation to L1.

(2) The program creates a new entry in L1 with the same POS as the English headword from which it originated.

(3) If the translation text includes parentheses, commas or semicolons, the text within is divided into separate headwords.

(4) Each L1 headword will include all senses from which it was extracted, including their English definition. This is displayed in the editorial interface, in which the editor can now reorder or remove senses as may be appropriate.

Figure 2 shows an example of the generation of Italian entries from the English entry *thing* in R1. The translation of the second sense as 'a person, especially a person one likes' to Italian is '*persona*, *creatura*'. These translations were thus divided into two separate headwords, *persona* and *creatura*, in R2.



thing [θiŋ] *noun*

1 an object; something that is not living: *What do you use that thing for?*

it cosa

2 a person, especially a person one likes: *She's a nice old thing.*

it persona, creatura

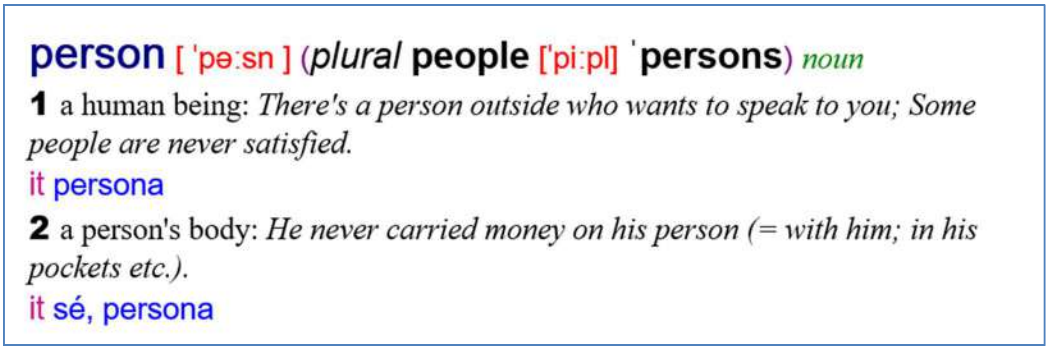
3 any fact, quality, idea etc. that one can think of or refer to: *Music is a wonderful thing; I hope I haven't done the wrong thing; That was a stupid thing to do.*

it cosa

Figure 2: The English entry *thing* with translations to Italian in R1

The English entries *person* and *soul* also contain '*persona'* as a translation of one of their senses, as shown, respectively, in Figures 3 and 4.

As a result, the entry *persona* in the Italian R2 includes all the occurrences of this word as a translation to Italian in R1. All its senses thus comprise these original English meanings, as shown in Figure 5.



person ['pə:sn] (*plural* **people** ['pi:pl] '**persons**) *noun*

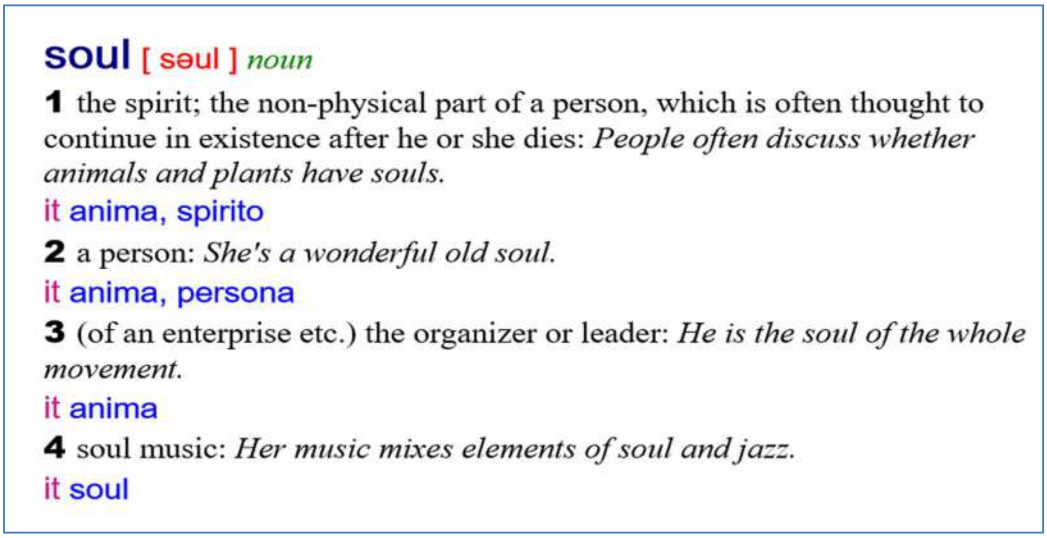
1 a human being: *There's a person outside who wants to speak to you; Some people are never satisfied.*

it persona

2 a person's body: *He never carried money on his person (= with him; in his pockets etc.).*

it sé, persona

Figure 3: The English entry *person* with translations to Italian in R1



soul [səul] *noun*

1 the spirit; the non-physical part of a person, which is often thought to continue in existence after he or she dies: *People often discuss whether animals and plants have souls.*

it anima, spirito

2 a person: *She's a wonderful old soul.*

it anima, persona

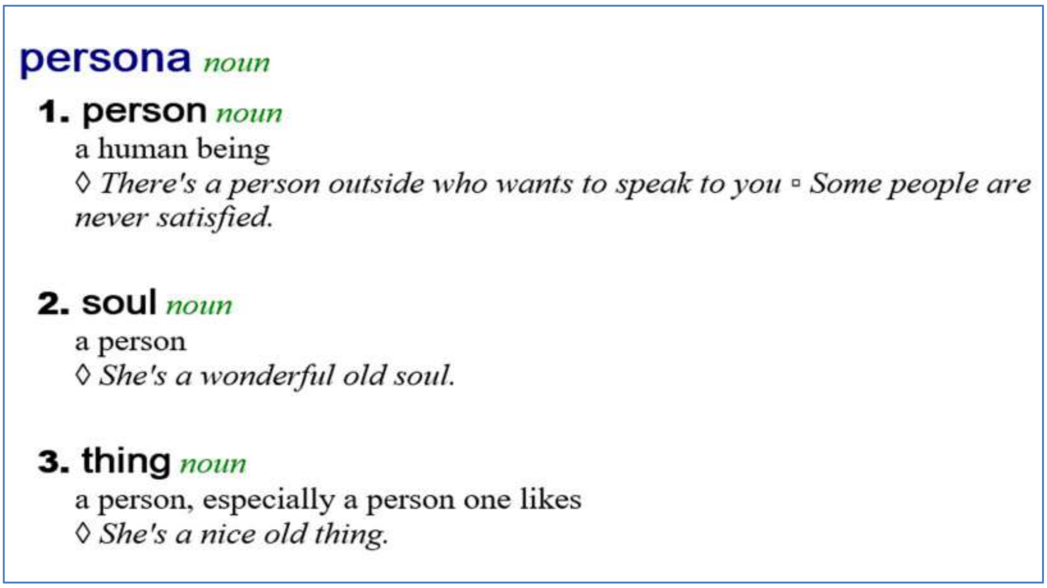
3 (of an enterprise etc.) the organizer or leader: *He is the soul of the whole movement.*

it anima

4 soul music: *Her music mixes elements of soul and jazz.*

it soul

Figure 4: The English entry *soul* with translations to Italian in R1



persona *noun*

1. person *noun*

a human being

◊ *There's a person outside who wants to speak to you* ▫ *Some people are never satisfied.*

2. soul *noun*

a person

◊ *She's a wonderful old soul.*

3. thing *noun*

a person, especially a person one likes

◊ *She's a nice old thing.*

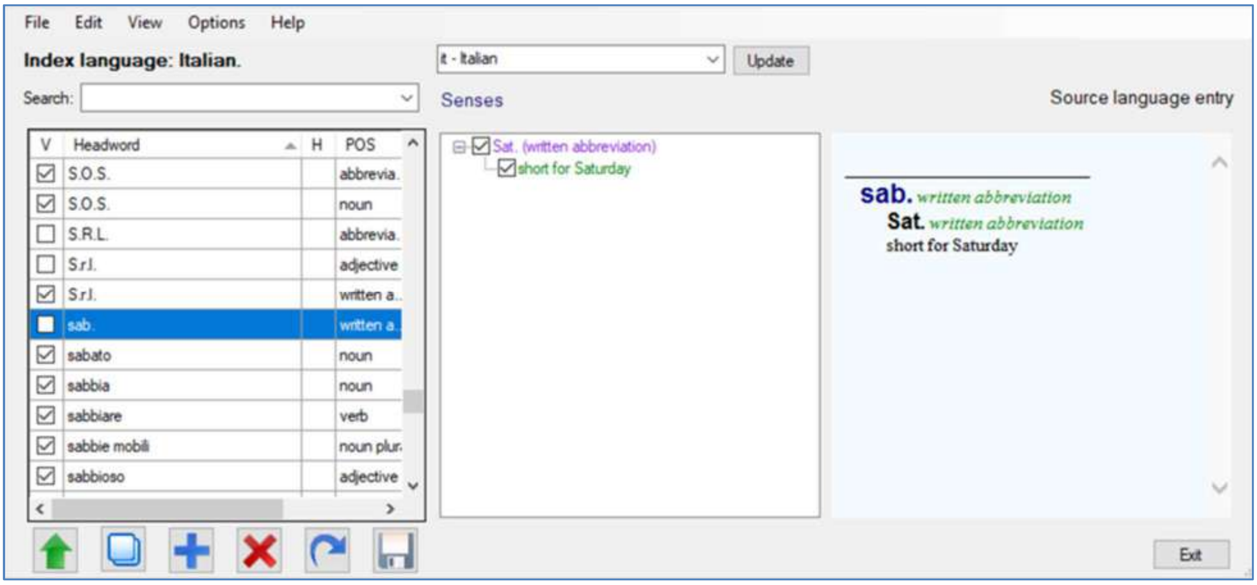
Figure 5: the Italian entry *persona* in R2 with the sense division
based on the English entries in R1

2.3 K Index Editorial Tool

This post-editing process is done with KIET, which is a bundle consisting of two programs – the admin tool and the editing tool. The admin tool has a graphic user interface (GUI) that enables the project manager to control the backend processes by which data is generated. In these processes, databases on which the editors perform the initial revision are generated from R1, and at a later step, XML files are created from the edited datasets (R2 and R3).

The current version of KIET is based on a revision of the original version developed in 2014 (cf. Egorova, 2015; Kernerman, 2015). The current generation of R2 data was prefaced with a thorough review of the 2014 version, which resulted in several improvement points. The first point of action was adding an admin interface, as the initial KIET version did not include one. The review process raised the need for a GUI on which project managers could control the process of the initial creation of R2 datasets. With the admin tool, project managers can add more languages to the datasets and create new ones by simply entering the required languages into the admin tool, without depending on a software developer to handle the creation. Second, new design features were added to the new (2020) version. It was decided to improve the design and performance in terms of user experience, a point that was previously ranked lower in priority. As more and more languages were added to the R2 project, it became evident that the user experience of the editors was crucial for smooth operation. Through productive cooperation between the software and the content teams, the KIET UX\UI was improved incrementally, with the content team providing input on whether an added feature was intuitive and easy to understand. Third, new features related to the linguistic aspect of the compilation were added in an evolving process that occurred concurrently with the R2 post-editing (described in detail in Section 3) and were added incrementally to the KIET. For example, the POS value list was updated to correspond to ongoing work on the R2 data, and new *GrammaticalNumber* and *Subcategory* fields were added to reflect newfound grammatical information. Further, automatic checks were introduced to reduce duplications (which were also handled in the post-editing stage), as well as a feature alerting the editor about missing information such as POS category. These additions were born from a trial-and-error process pertaining to the revision of the first R2 files in the 2020 project, which contained substantially more duplicates and missing categories than the consecutive versions.

Figures 6 and 7 display screenshots of the original and new KIET main interface, respectively.



Index language: Italian.

it - Italian

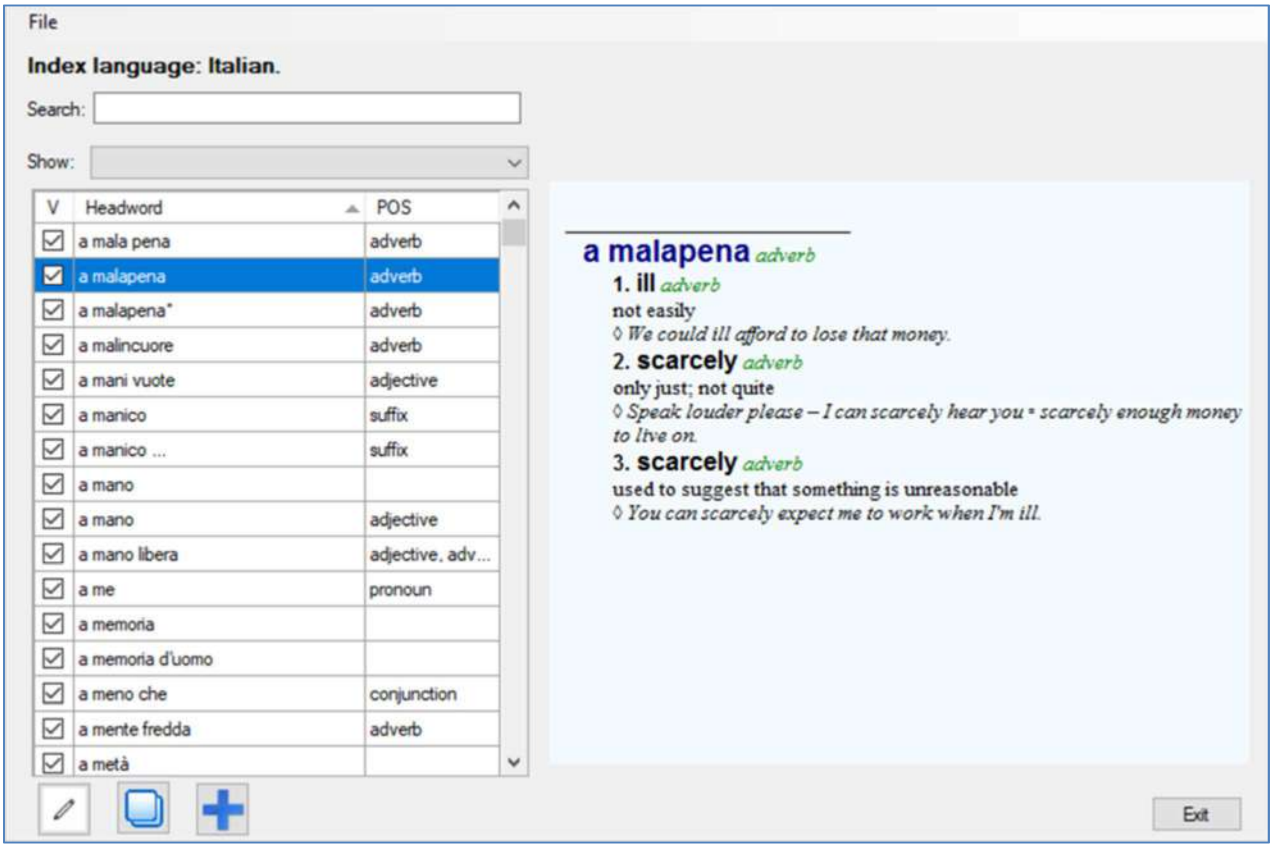
Search:

V	Headword	H	POS
☑	S.O.S.		abbrevia.
☑	S.O.S.		noun
□	S.R.L.		abbrevia.
□	S.r.l.		adjective
☑	S.r.l.		written a..
□	sab.		written a..
☑	sabato		noun
☑	sabbia		noun
☑	sabbiare		verb
☑	sabbie mobili		noun plur.
☑	sabbioso		adjective

☑ Sat. (written abbreviation)
☑ short for Saturday

sab. *written abbreviation*
Sat. *written abbreviation*
short for Saturday

Figure 6: Screenshot of the main interface of the original KIET



Index language: Italian.

Search:

Show:

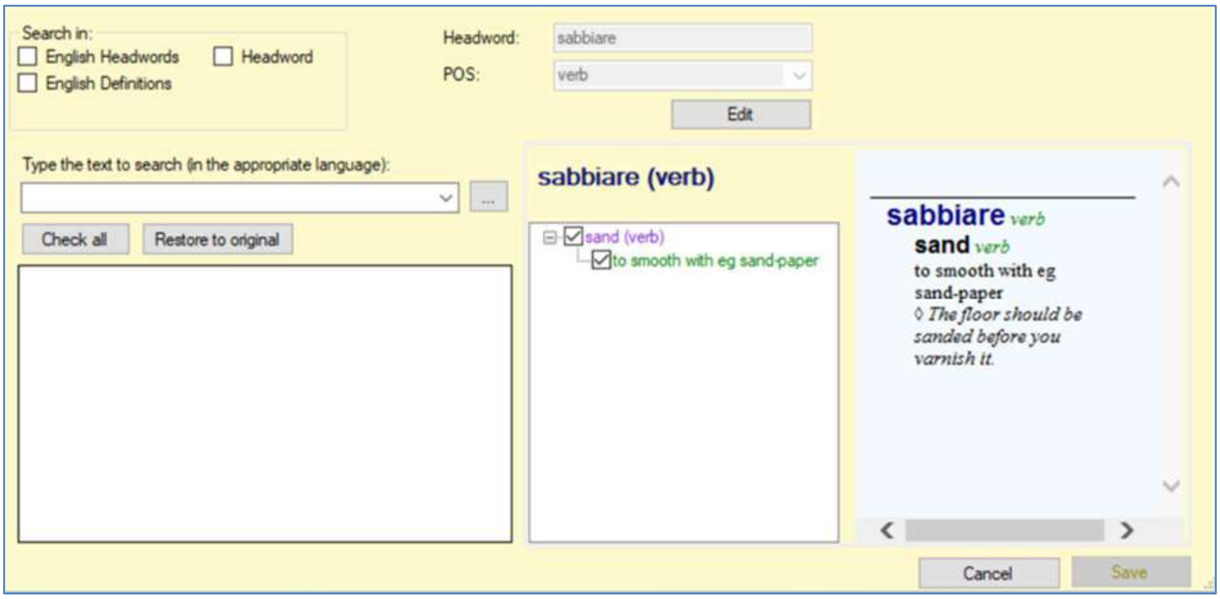
V	Headword	POS
☑	a mala pena	adverb
☑	a malapena	adverb
☑	a malapena*	adverb
☑	a malincuore	adverb
☑	a mani vuote	adjective
☑	a manico	suffix
☑	a manico ...	suffix
☑	a mano	
☑	a mano	adjective
☑	a mano libera	adjective, adv...
☑	a me	pronoun
☑	a memoria	
☑	a memoria d'uomo	
☑	a meno che	conjunction
☑	a mente fredda	adverb
☑	a metà	

a malapena *adverb*
1. ill *adverb*
not easily
◊ *We could ill afford to lose that money.*
2. scarcely *adverb*
only just; not quite
◊ *Speak louder please – I can scarcely hear you • scarcely enough money to live on.*
3. scarcely *adverb*
used to suggest that something is unreasonable
◊ *You can scarcely expect me to work when I'm ill.*

Figure 7: Screenshot of the main interface of the new KIET

2.4 KIET Editorial Interface

The editorial interface of KIET is where the R2 entries are reviewed and revised, enabling the editor to create, remove or duplicate headwords and manage the sense relations and order. Figures 8 and 9 show screenshots of the editorial interface from the initial version (2014) and the current version (2020), respectively.



Headword: sabbiare

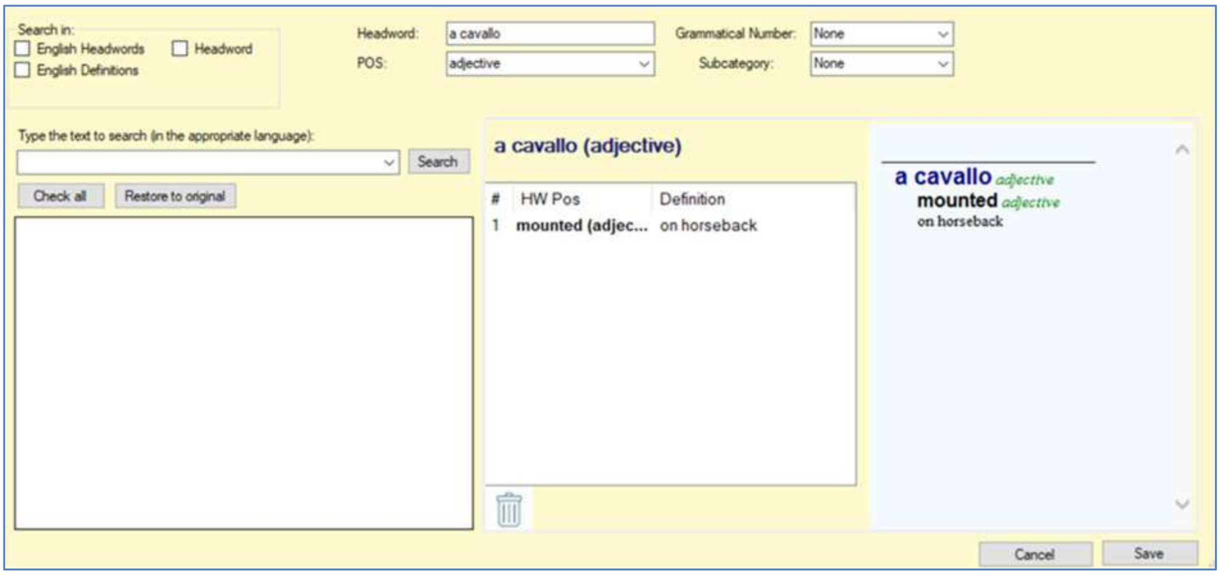
POS: verb

sabbiare (verb)

☑ sand (verb)
☑ to smooth with eg sand-paper

sabbiare *verb*
sand *verb*
to smooth with eg sand-paper
◊ *The floor should be sanded before you varnish it.*

Figure 8: Screenshot of the editorial interface of the original KIET



Headword: a cavallo

POS: adjective

Grammatical Number: None

Subcategory: None

a cavallo (adjective)

#	HW Pos	Definition
1	mounted (adjec...	on horseback

a cavallo *adjective*
mounted *adjective*
on horseback

Figure 9: Screenshot of the editorial interface of the new KIET

The main changes in the two versions include a feature that disables the appearance of duplicate entries, that is, entries consisting of the same headword and POS. The first versions of R2 were generated prior to these enhancements and contained many duplications that were handled in the post-editing stage. Post-editing also produced insights with respect to the implementation of new features, such as the rearrangement of sense order and removal of irrelevant senses. The previous version had design errors that caused the preservation of senses that did not correlate to the senses in the target language, or whose prevalence in that language was much lower than in English. Simultaneously, a newly added menu displays the valid entries as well as those either removed or edited.

The search functionality was enhanced and made more flexible. First, unaccented search was enabled in both (main and editorial) interfaces, removing diacritics and disregarding case. While the first KIET version only allowed searching for a particular entry by the exact headword text, the new version lets the editor search for specific senses of a word by entering either the original English headword or keywords from the definition.

To ensure that a certain structure is maintained, post-editing is only allowed at the entry level. That is, only information pertaining to the spelling of the headword or its grammatical information can be changed. The editor cannot edit the existing sense definitions or add new senses, which is arguably the main shortcoming of R2. The reason for this lies in the R2 structure: while the entry information is generated from a combination of the information pertaining to the original English entry and the equivalent translations in R1, the sense division is based on the English information only. Generally, the R2 senses consist of the English headword and definition, and include the English POS (as further sense indication) and examples of usage. Obliged to remain agnostic to R1, only the sense division and order can be modified in KIET.

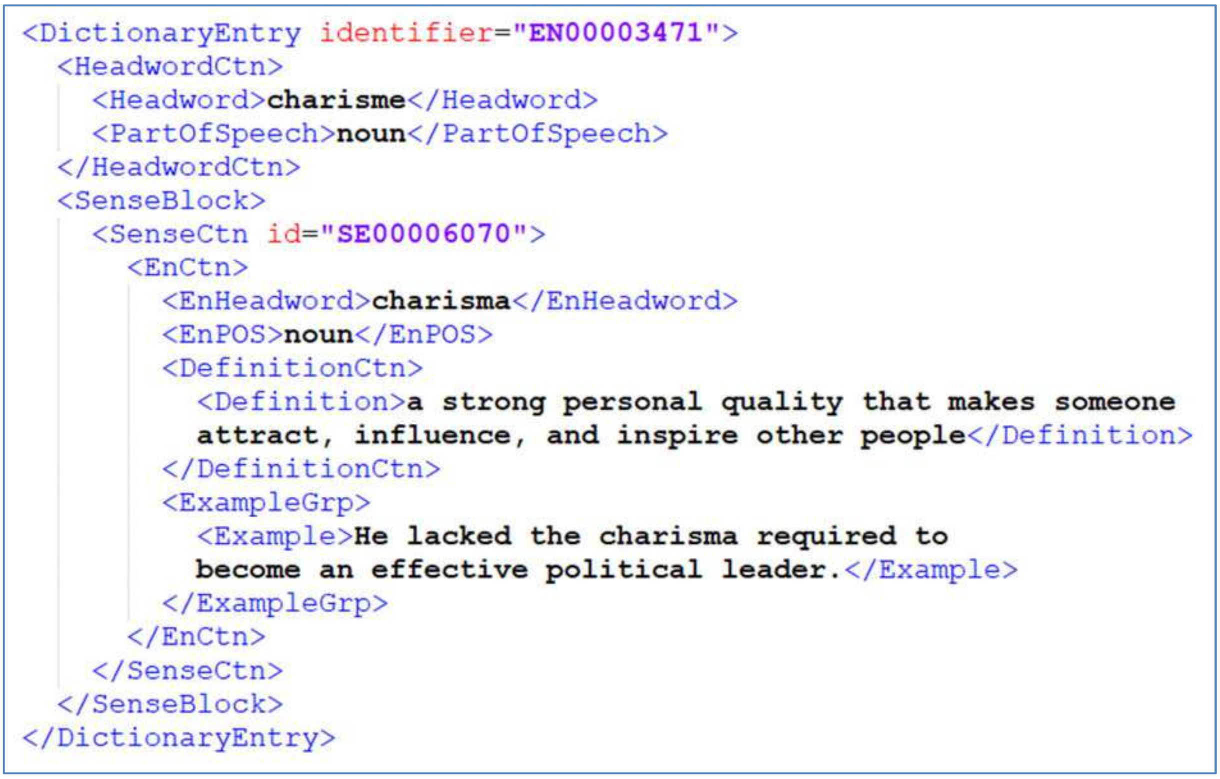
3. Post-Editing with Corresponding Automated Tasks

The KIET described in Section 2 was used for the manual editing of the raw (automatically generated) L1-English glossary R2. This post-editing process combined further automated tasks, and the main ones are described in this section. Once the R2 editing was complete, the R1 translations in other languages were added automatically in creating the English pivot-based multilingual glossary R3.

3.1 The Reversed Glossary XML Structure

The multilingual glossary (R3) data is comprised of simple XML documents with a straightforward XML schema. The initial structure consisted of a *DictionaryEntry* element containing two main components. First, the *HeadwordCtn* includes information on the lemma or phrase; initially, it comprised only the headword and POS category, but it was expanded to include more grammatical details such as number or gender, as well as inflected forms. These changes are described below and are part of the post-editing process, which combines automated methods with manual revision and editing.

The second main component of the entry is the *SenseBlock*, including a division into different meanings (represented by separate *SenseCtns*), their definitions and examples of usage in English and translation equivalents. The sense division is manifested by the original English information: the headword and POS are wrapped in their own component nested inside the sense, to allow retracing to the original sense in R1. The definition functions as the main sense indicator. Figure 10 presents a sample monosemous entry in the French R2, demonstrating the headword and sense structure.



```
<DictionaryEntry identifier="EN00003471">
  <HeadwordCtn>
    <Headword>charisme</Headword>
    <PartOfSpeech>noun</PartOfSpeech>
  </HeadwordCtn>
  <SenseBlock>
    <SenseCtn id="SE00006070">
      <EnCtn>
        <EnHeadword>charisma</EnHeadword>
        <EnPOS>noun</EnPOS>
        <DefinitionCtn>
          <Definition>a strong personal quality that makes someone
          attract, influence, and inspire other people</Definition>
        </DefinitionCtn>
        <ExampleGrp>
          <Example>He lacked the charisma required to
          become an effective political leader.</Example>
        </ExampleGrp>
      </EnCtn>
    </SenseCtn>
  </SenseBlock>
</DictionaryEntry>
```

Figure 10: XML data of the French monosemous entry *charisme* in R2

3.2 Headwords and Part of Speech Categories

Following the initial automated generation of the R2 sets, it was necessary to introduce editorial amendments reflecting a refinement of the headword forms and the grammatical categories to fit the newfound source languages. The post-editing phase started with revising the headword text and adjusting the POS categories. These modifications were performed manually by the editor of each language and were facilitated by automated processes, including revising the headword text to reflect a more common variant in that language; fixing typos; stripping characters such as slashes, commas, parentheses or brackets; and handling gender inflection. Such cases were either eliminated, inserted into a corresponding tag or divided into independent entries. The primary aim of this initial revision was to verify that all the headword text was cleaned and normalised in order to become fit for automated processing and machine readability.

Alongside the headword revision, the POS category was modified as well. When given the opportunity to redesign R2 from scratch, the leading heuristic was to simplify the dataset as much as possible, placing the relevant information in designated tags and adhering to a closed list of POS values. As part of the post-editing process, entries missing a POS element were singled out and fixed; existing categories were normalised and stripped from additional information to adhere to a predefined schema of particular POS values; and any additional information that was relevant to the grammar of the word was retained and transferred to corresponding elements, namely the *GrammaticalNumber* and *Subcategory* tags, to reduce noise and facilitate searching the data for relevant information. The POS irregularities can be attributed to two main causes:

(a) The original output did not include a POS category, or the existing category generated by KIET was removed by the editor in the initial editing phase and was not replaced with another value accidentally.

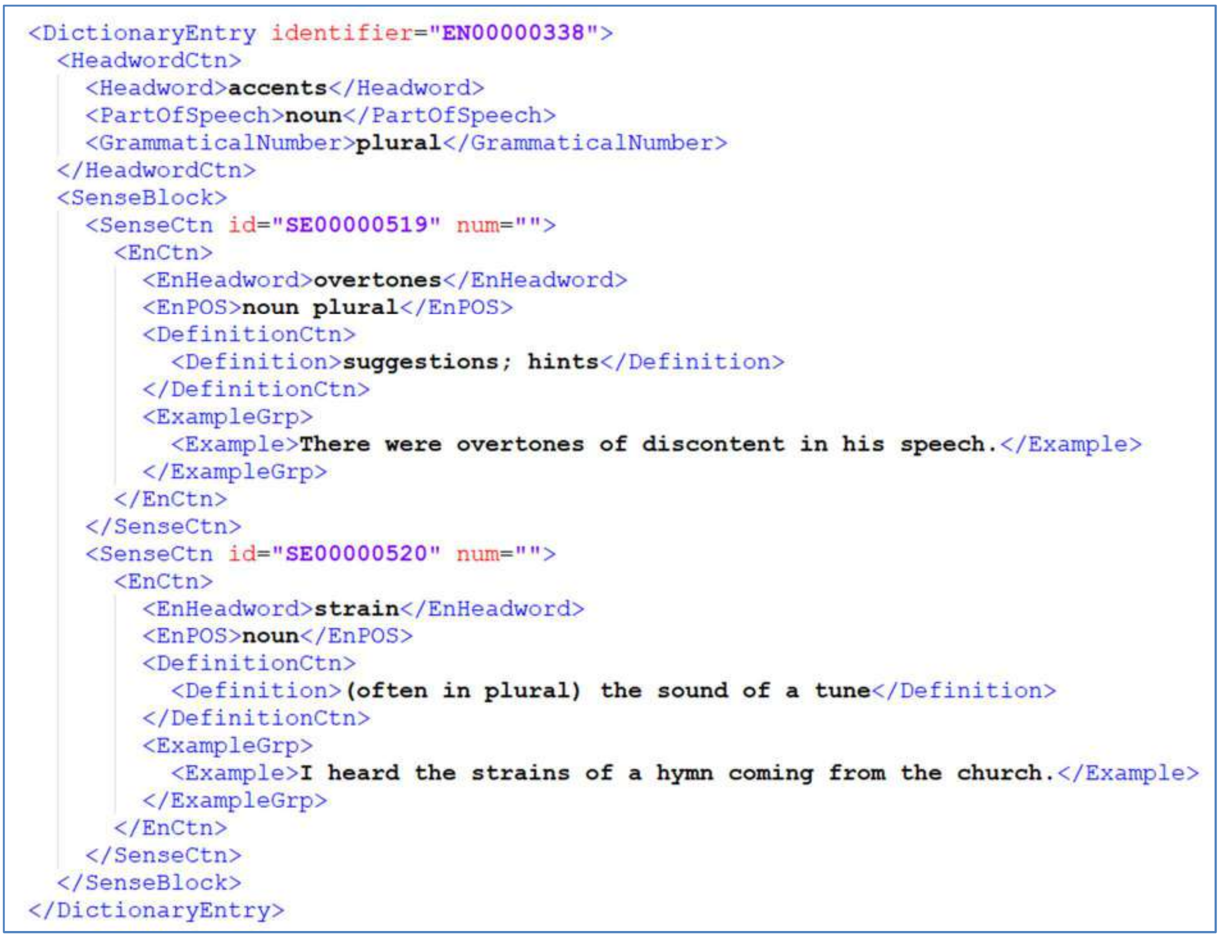
In these cases, an automated process matched the headword text with a corresponding entry in the Global Multilingual Data Series (R4) and inserted the corresponding POS category into the R2 dataset. Since the POS category does not pertain to a particular meaning, it was not necessary to perform any sense alignment prior to the matching. If there were multiple entries with different categories in R4, the information was transferred to an editor to determine the correct category.

(b) The original POS category, which was generated from the English POS category in R1, included additional information, such as grammatical number or subcategory.

In these cases, an automated process located all instances of a POS tag including additional information and separated the POS category from the grammatical information, placing the new information in a corresponding tag.

Figure 11 is a demonstration of an R2 French entry containing the newfound *GrammaticalNumber* tag whose information on plurality is evident from the original English part of speech (*EnPOS*).

As the automated process for generating R2 included attributing the POS of the original English entry in R1 to the new L1 headword in R2, the editors also received a list of headwords whose POS had to be determined or validated. In some cases, no equivalent was available in any parallel resource, so the editors supplemented the information based on their own linguistic knowledge. In other cases, multiple equivalents were found in R4 and were all given to the editor, thus facilitating the decision. In addition, a list of uncertain POS categories was curated, consisting of headwords with POS values that did not belong to a predefined closed list of values – including narrower categories such as ‘proper noun’ instead of ‘noun’ and unconventional or abbreviated text such as ‘adj’, standing for ‘adjective’ – and the editor was asked to select an appropriate POS category from a list of values. As a final step, the editors were asked to review all headwords tagged as ‘plural’ or ‘abbreviation’ (for each element respectively) and to verify whether this tagging was correct. This demonstrates how automatic retrieval of information, albeit not precise or exact, can help the manual work and speed the post-editing process.



```
<DictionaryEntry identifier="EN00000338">
  <HeadwordCtn>
    <Headword>accents</Headword>
    <PartOfSpeech>noun</PartOfSpeech>
    <GrammaticalNumber>plural</GrammaticalNumber>
  </HeadwordCtn>
  <SenseBlock>
    <SenseCtn id="SE00000519" num="">
      <EnCtn>
        <EnHeadword>overtones</EnHeadword>
        <EnPOS>noun plural</EnPOS>
        <DefinitionCtn>
          <Definition>suggestions; hints</Definition>
        </DefinitionCtn>
        <ExampleGrp>
          <Example>There were overtones of discontent in his speech.</Example>
        </ExampleGrp>
      </EnCtn>
    </SenseCtn>
    <SenseCtn id="SE00000520" num="">
      <EnCtn>
        <EnHeadword>strain</EnHeadword>
        <EnPOS>noun</EnPOS>
        <DefinitionCtn>
          <Definition>(often in plural) the sound of a tune</Definition>
        </DefinitionCtn>
        <ExampleGrp>
          <Example>I heard the strains of a hymn coming from the church.</Example>
        </ExampleGrp>
      </EnCtn>
    </SenseCtn>
  </SenseBlock>
</DictionaryEntry>
```

Figure 11: XML data of the French polysemous entry *accents* in R2

3.3 Eliminating Duplicate Entries and Senses

As mentioned in Section 3.2, as part of the automated POS attribution process, missing categories were supplemented from the Global Series, and variants of existing categories were cleaned and normalised. This process in turn resulted in another data issue, which was also handled and solved automatically as part of the post-editing pipeline. Amending the headword text and POS categories resulted in many cases in which the same headword text and POS appeared for two separate *DictionaryEntry* elements in the data, that is, two separate entries that originally included the same headword text, but different POS categories were now duplicate cases of the same entry. However, just removing one of the entries would not suffice, since the senses were in most cases different for each entry. The purpose of the automated task was to eliminate duplicate entries in the data while retaining all information from the sense level. This was divided into two steps. The first step, handling the duplicate entries, was designed according to the following algorithm and combined an automated process with extra human validation:

(1) For each entry, check if there is another entry that shares the same headword text and POS category.

(2) If one entry includes additional grammatical information (such as number or subcategory), the revision is delegated to the editor to manually verify that the entrees are indeed separate entries and make the proper modifications to distinguish them.

(3) If there is no additional information, take all senses from the second occurring entry and append them to the *SenseBlock* of the first occurring entry, then remove the second entry from the dataset.

This process is general enough to catch many cases, but at the same time remove the risk of accidentally concatenating two entries that are not in fact identical; involving the editors in the automated post-editing process allowed the flexibility and speed of an entirely automated pipeline while still retaining the benefits of humanly curated data that is checked and validated after every step. The second step, which included the revision of duplicate senses following the grouping together of senses from two separate entries, was done separately, so as to break down the deletion process into smaller, manageable steps that could be verified upon execution, thus reducing the error margin to a minimum.

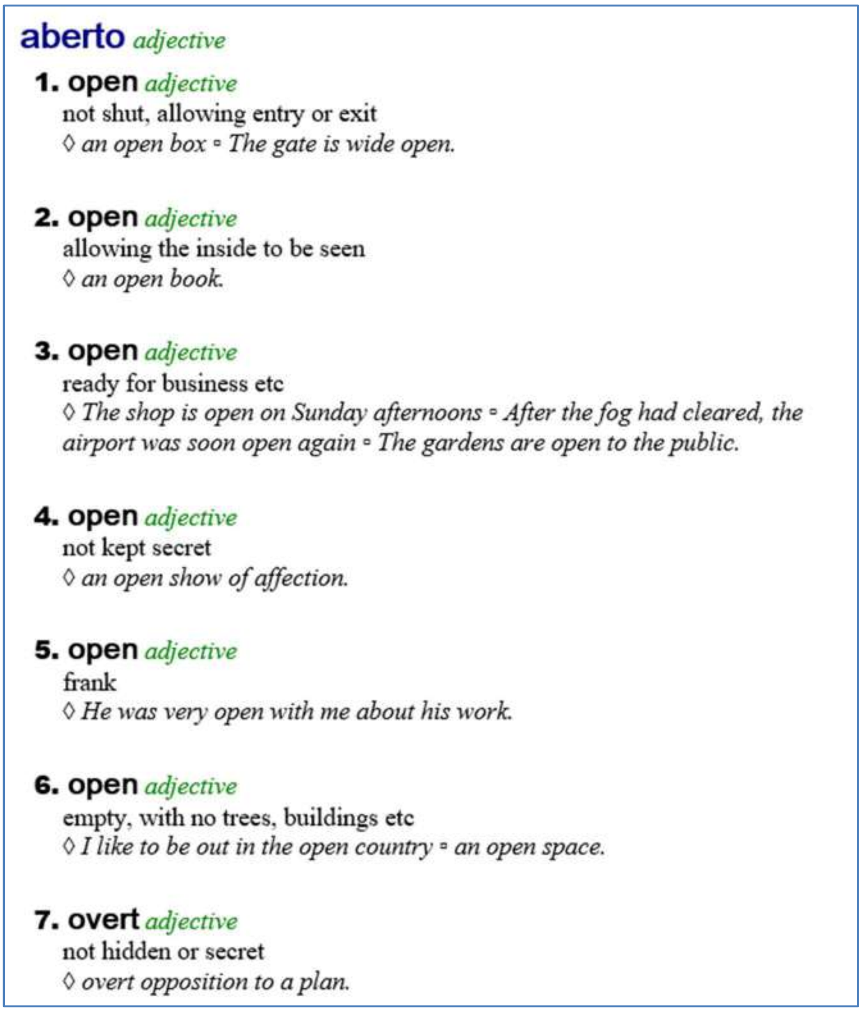
To preface the sense elimination step, it is important to reiterate the compilation process of the R2 dataset: as presented in Section 2, this data is constructed by retrieving translations from English entries in R1. Translations from different entries are grouped together by POS categories, and the editor is requested to rank the sense order by importance or prevalence, relying on the English definition as an indication for the sense (since no additional information is given for the entry in L1). Then, the L1 entry is created for that R2, including all senses belonging to the corresponding English entries sharing the POS category. Upon revision of the resulting R2, and after amending headword text and POS categories as previously described, we generate separate entries encompassing the same lemma, for which multiple and different senses belong. When concatenating together the amended entries, it is now necessary to check that no duplications occur within the collection of the different senses. This phase is slightly more complicated than the previous one of eliminating duplicate entries, as it must take into account the meaning variations and carefully consider whether two senses reflect the same meaning. This process, like the previous one, combined an automated process with manual post-editing. Relying on the four types of information that currently exist within a *SenseCtn* for an individual sense, which is the English headword, the English POS, the English definition text and examples of usage, an algorithm was constructed according to the following guidelines:

(1) Comparing each following sense to the first one as an anchor, an automated process checked whether the sense pair included the identical English headword and POS information. If so, and there was no additional information, the second sense was removed from the dataset.

(2) If additional information existed, the process then compared the definition text: if the definition text was identical, then the process merged the two senses by deleting the second sense and taking any examples it contained and appending them to the *ExampleCtn* of the first sense; if no examples existed, no action was required.

(3) If the definition text was not identical, the senses were transferred to the editor for manual editing.

The editor then had to determine whether the two definitions encompassed the same meaning, or if they were distant enough to count as separate senses. Figure 12 presents a sample of a merged entry in which the original English headword and POS information are identical, but the definitions reflect separate meanings:



aberto *adjective*

1. open *adjective*
not shut, allowing entry or exit
◊ *an open box* ▫ *The gate is wide open.*

2. open *adjective*
allowing the inside to be seen
◊ *an open book.*

3. open *adjective*
ready for business etc
◊ *The shop is open on Sunday afternoons* ▫ *After the fog had cleared, the airport was soon open again* ▫ *The gardens are open to the public.*

4. open *adjective*
not kept secret
◊ *an open show of affection.*

5. open *adjective*
frank
◊ *He was very open with me about his work.*

6. open *adjective*
empty, with no trees, buildings etc
◊ *I like to be out in the open country* ▫ *an open space.*

7. overt *adjective*
not hidden or secret
◊ *overt opposition to a plan.*

Figure 12: Italian polysemous entry *aberto* in R2

Here, the manual check was able to determine that these are all separate meanings of the English word *open*, thus leaving the initial sense division as is and retaining all relevant example phrases and sentences. Some senses containing three or more examples are the result of an automated process comparing two senses which had the same definition text and grouping together their separate examples to one sense, demonstrating uniform usage for a singular meaning.

The numbers of problematic entries varied between languages. Some, such as French or Italian, initially included a small number of suspicious duplicates, and others, such as Chinese, had much higher numbers of duplicate or erroneous headwords to be examined and modified, ranging between 100 and 5,200 entries per language. The automated process managed to reduce manual work by more than half, resulting in a significantly lower number of cases for editorial review and revision. In the case of Chinese, the initial process of eliminating duplicates covered as many as 5,000 cases, leaving approximately 200 entries only for manual post-editing and curation. This process could be further automated by relying on additional tools and resources that enable the definitions to be compared, checked for their closeness, or rated for their similarity by a particular metric (Kaltenböck and Kernerman, 2017). The current process relied on straightforward string comparison and applied human judgement to determine sense division, due to time constraints and the uncertainty of such similarity tests. However, it would be interesting to incorporate such tests in more elaborate automatic post-editing pipelines.

3.4 Further Revision and Evaluation

Nearly every step explained above required the editor to verify and validate the automatic outcome, as well as to point out additional problems with the data that might need further (automatic) tackling. The design of the pipeline itself allowed for the minimal amount of material to be manually reviewed, by taking care of tasks that can be handled entirely automatically first and delivering anomalous tasks to editors second. A list of unconventional duplicate entries and senses was also reviewed manually, bearing in mind to amend any automatically integrated information that was incorrect, while keeping all relevant information by concatenating it from the duplicate entries, thus creating one full final entry. Similarly, a list of headwords with slashes, brackets and other abnormal characters was reviewed, stating the correct text to be amended and whether another entry was to be added. For example, the original Swedish headword '[allt]sedan' was separated to two new headwords 'allt sedan' and 'sedan'.

The process of identifying and separating variants from headwords containing slashes revealed a sub-category of cases in which the text after the slash was not an individual word but rather a suffix for the feminine form of the headword for languages with gender inflection. These were identified by a dash preceding the suffix, indicating the need to replace the masculine suffix of the original word. For example, the French R2 included the headwords with text 'acteur/-trice', 'alarmant/-ante' and 'champion/-onne', which surfaced when searching for headwords with peculiar characters such as slashes. These cases were handled almost entirely automatically, by devising a rule for generating the full feminine form based on the root and the masculine form, verifying the results automatically, and then manually checking them to obtain even more security. The process is described below:

(1) Generating the feminine form was carried out according to the following rule, based on French grammar:

a. If the suffix begins with a vowel V, the root form is taken as all characters up to the same vowel in the ultimate position of the word, and the suffix, i.e., the text after /-, is then appended to the root, e.g., champion/-onne → champi + onne → championne; alarmant/-ante → alarm + ante → alarmante.

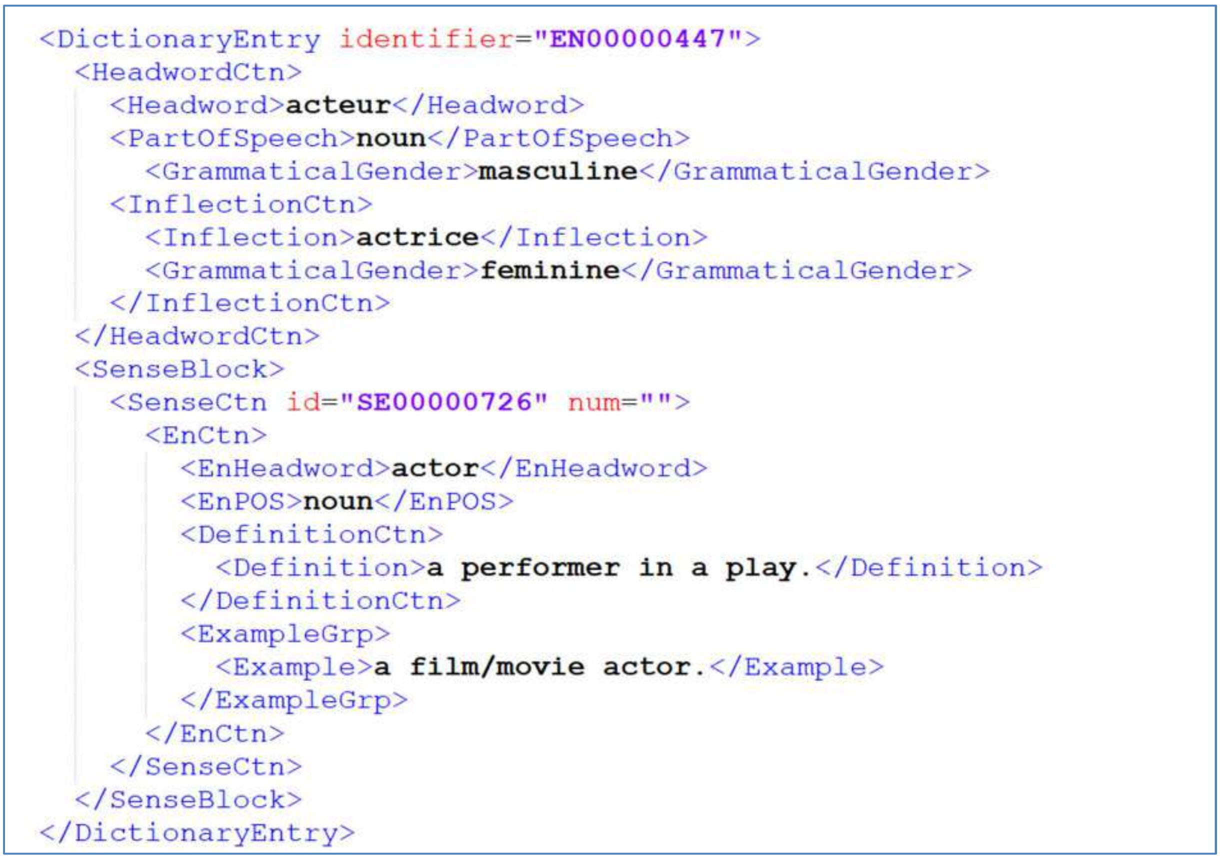
b. If the suffix begins with a consonant C, the root form is taken as all characters up to the same consonant in the ultimate position of the word, and the suffix, i.e., the text after /-, is then appended to the root, e.g., acteur/-trice → ac + trice → actrice. It should be noted that the V/C distinction is based on the existing orthography and not on French morphological rules.

The resulting forms ('acteur' and 'actrice', 'champion' and 'championne', 'alarmant' and 'alarmante') were then looked up in existing French resources or morphological lists and marked as safe if said forms existed in any such resource. If not, they underwent an automatic translation process, relying on machine translation tools to translate both forms back to English and check whether they match. A match indicates that the automatic generation succeeded in high likelihood. For example, 'champion' and 'championne' both translate to the English 'champion' and were thus marked as a success. The pair 'acteur' and 'actrice', in turn, were located in R4 and marked as a success too.

(2) Following suit, the editor reviewed the automatically generated forms and their success mark and amended the results if necessary.

The benefits of having an existing suggestion for a form as well as a metric to evaluate the success for the automatic generation is twofold: it saves time by eliminating the need to manually enter a value, and it greatly reduces the chances for typos or spelling mistakes. However, relying solely on written characters and their placement relative to each other to devise an automatic rule carries its own risks. The inclusion of manual editorial work in this case also proved to be of high importance: the editor was able to amend errors caused by the algorithm, as well as identify cases that were not marked as a success and identify whether or not they encompass a gender inflection, or a typo.

(3) The reviewed masculine and feminine forms were then incorporated in the data by keeping the masculine form in the headword and introducing an *InflectionCtn* component in which the feminine form was inserted. Grammatical information pertaining to gender was also added to *GrammaticalGender* tags. Figure 13 presents an example of the instantiation of this modelling for the entry ‘acteur’.



```xml
<DictionaryEntry identifier="EN00000447">
  <HeadwordCtn>
    <Headword>acteur</Headword>
    <PartOfSpeech>noun</PartOfSpeech>
      <GrammaticalGender>masculine</GrammaticalGender>
    <InflectionCtn>
      <Inflection>actrice</Inflection>
      <GrammaticalGender>feminine</GrammaticalGender>
    </InflectionCtn>
  </HeadwordCtn>
  <SenseBlock>
    <SenseCtn id="SE00000726" num="">
      <EnCtn>
        <EnHeadword>actor</EnHeadword>
        <EnPOS>noun</EnPOS>
        <DefinitionCtn>
          <Definition>a performer in a play.</Definition>
        </DefinitionCtn>
        <ExampleGrp>
          <Example>a film/movie actor.</Example>
        </ExampleGrp>
      </EnCtn>
    </SenseCtn>
  </SenseBlock>
</DictionaryEntry>
```

Figure 13: XML data of the French polysemous entry *acteur* in R2

Naturally, this process of further revising the headword texts for any R2 dataset may result in newfound duplicate entries. The previously described process of identifying duplicate entries, concatenating them and eliminating their duplicate senses was performed incrementally after each revision of the headword text and could be performed again and again until the revision was finalised.

To find possible misspellings among the resulting headwords, a spell-checking pipeline was defined and implemented for each language. First, all textual data was checked automatically using existing or custom spell-checkers, and then the results were reviewed by the editor, who corrected true misspellings. At the end of the process, the amended text was merged back to the dataset. Obviously, spell-checking in a multilingual environment is a rather challenging task. For some languages, existing tools or simple pipelines yield satisfying results, with a small number of false positives and high recall, that is, most of the misspellings were detected by the system. However, for other languages, mostly morphologically-rich or low-resource ones, the task requires more tuning and specific implementation. A high number of false positives is counterproductive, as it generates additional editorial work, which is expensive and impractical. Possible solutions may involve morphological analysis as a pre-processing step, mining additional "known words" vocabularies from corpora and utilisation of other available resources.

4. The Full Resource Conversion and the Final Outcomes

In the second phase of the project, the R3 resources were merged with the Global Multilingual Data Series (R4), consisting of a collection of extensive lexicographic cores for different languages. Each language core includes a wide lexical base featuring rich semantic and grammatical information arranged in well-structured datasets, within the framework of a single comprehensive macrostructure and all adhering to the same entry microstructure, with most of these language cores having bilingual and multilingual versions in varied numbers.

The main entry components of the R4 sets include phonetic transcription (in IPA) and alternative scripts, POS, irregular forms, grammatical subcategorisation, gender and number, as well as sense division based on frequency with definitions, examples of usage, related MWEs and other attributes such as synonyms, antonyms and subject domain.

To converge R3 with R4, it was necessary to develop a meticulous algorithm, first to match the headwords in each resource and then to link senses correctly for polysemous entries in either or both resources.

MWEs and nested entries were also taken into consideration so as to expand the database of entries for which the merging is performed and raise the chances of a match. The matching algorithm then searched for the headwords within the expanded collection and matched them with corresponding entries from R3. The algorithm was constructed as follows:

(1) A dataset was created for the R4 entries, including POS categories, synonyms and inflections.

(2) The matching program ran through this dataset, and for each headword or inflection, and their corresponding POS, it checked whether the pair exists in R3 as a headword and POS pair, disregarding the POS component for MWEs.

(3) If the headword and POS pair was identified within R3, it was added to a set of all matching R3 entries.

The result is a set of matching pairs – R4 entries and their corresponding entries from R3 that were found as headwords or MWEs or as an inflection of an entry in R4. The following stage, which consisted of a sense alignment of sorts, was comprised of two steps. The first relied on translations to perform the initial sense linking. The second relied on synonyms, if existing, to further expand the possibility of matching the R4 sense with a corresponding R3 sense. The sense matching algorithm is as follows:

(1) The algorithm loops through all of the senses of the matching entry, focusing on the available translations of the senses.

(2) The algorithm then loops through all the senses of the R3 entry; for each sense, if any of the translations of the R4 sense matches any of its translations, the sense is registered as a matching sense pair, and the number of matching translations is counted.

(3) If any of the R4 senses have synonyms, they are searched for within the R3 resource. If an R3 entry identical to the synonym is located, then the program runs through all its senses, comparing it to the sense of the R4 entry in which the synonym was originally found; the same process of translation comparison is performed for the matching synonym entry.

After reviewing all the R3 senses that were singled out as possible matches, the most fitting one is selected. The parameter in this case is the highest number of matching translations. The guiding principle in the process of sense linking was that each R3 sense can match no more than one R4 sense for the same entry. The percentage threshold for the matching varied for each language, mainly due to a discrepancy in the number of target languages for each source language in R4. At first, each language output included only the sense that passed a certain matching percentage threshold. Later on, it was decided to also include entries that constitute exact matches at the headword and POS level (i.e., not found as inflections), even if none of the senses passed the initial threshold.

Prior to the matching phase, there were a few issues that were taken into consideration. Similarly to the initial creation of R2, text containing slashes, commas or semicolons that separated two or more values was handled to find matches for each value separately. Further, definite articles and prepositions were cleaned from the text. Any additional information that usually accompanies the main headword and found inside parentheses was removed. Diacritics, stress and case or capital letters (uppercase vs lowercase) were disregarded. Conversion tables provided for each L1 facilitated the normalisation and mapping process.

5. Conclusions and Future Work

The endeavor of converging and transforming existing lexicographic datasets into a brand-new resource requires substantial effort. The initial manual editing is tedious yet necessary; this process encompasses the initial shift from English as the main source language to a new language that is now at the front. What was previously a target language, embodying the lexicographic resolutions of translating that which cannot always be directly translated, is now at the forefront. The following stage of post-editing enabled a combination of automated and manual processes to facilitate much of the manual labor. This also embodied a learning curve wherein insights were extracted from the work on each dataset and improved for the reiteration of the next step. In that sense, the incremental workflow, whereby each step enabled evaluation and later revision of previous steps, allowed for a flexible pipeline and immediate repairing of errors.

Some improvements of KIET could be derived from the automated post-editing pipeline. The admin tool could be enhanced with more automated features and functionalities, thus eliminating the need to perform these tasks in the next post-editing phase. For example, when post-editing revealed many duplicate entries that existed in datasets generated by previous versions of KIET, a feature to alert about possible duplications was added to both the editorial and administrative interfaces. Other processes such as the normalisation of POS and grammatical information could be added as a preliminary phase inside KIET, voiding the need for an extra step in the following automated pipeline. Generally speaking, the post-editing pipeline could be reduced to manual editing accompanied by particular automation as required, and anything that could be described as a general rule could be added to the KIET backend.

The process of merging datasets also relied heavily on automated checks, which could be further improved by expanding the arsenal of tools that are used for such revision. The R4 resource was merged with the newly generated R3 resources in a matching process consisting of a direct string-based comparison with minimal clean-up. Indeed, MWEs were included as well, and a closed list of inflections and synonyms were added to expand the pool of words in which the search for matches was conducted. The downside to this is that variants, either spelling variants or other morphologically inflected forms, could be missed even though the POS is identical, and the meaning is similar, which could result in fewer matches and a lower recall. However, many senses that may have been overlooked due to a small discrepancy in the headword form, or other small variations, might be detected with further adjustments. For example, this process could be improved by utilizing word embeddings that can provide an approximation of similarity between variants or differently spelled words. Similarly, the merge pipeline could be enhanced by employing sentence encoders to measure the similarity of two differently phrased definitions at the sense level.

The main benefit of the creation of new datasets and merging them with existing ones is the prospect of creating a larger, more extensive dataset, combining the strengths of different resources. The Global Series could be enhanced as well, not only by using the newly created R3 resource, but also taking external multilingual resources and applying the same pipeline, thus adding more components and enriching the data. In terms of evaluation, future work may include an exact documentation of numbers of matching instances for duplicate entries and mismatches that require manual review. A case could be made for the calculation of precision values for each language, as this information could be included in further identification of language-specific issues, but since this project did not implement learning algorithms and its focus was the preparation of data for production and not the training of a model, we did not explicitly document these numbers, and the current information provided in this paper is based on a retrospective examination of logs.

6. References

Egorova, K. (2015). Editing an automatically-generated index with K Index Editing 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. https://elex.link/elex2015/proceedings/paper-17

Kaltenböck, M. & Kernerman, I. (2017). Introducing LDL4HELTA: Linked data lexicography for high-end language technology application. *Kernerman Dictionary News* (25), 2–3. https://kdictionaries.com/kdn/kdn25_2017.pdf.

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. https://elex.link/elex2015/proceedings/paper-24

Codification Within Reach: Three Clickable Layers of Information Surrounding the New Slovenian Normative Guide

Helena Dobrovoljc[1,2], Urška Vranjek Ošlak[1]

[1] ZRC SAZU, Fran Ramovš Institute of the Slovenian Language, Novi trg 2, SI-1000 Ljubljana, Slovenia;

[2] University of Nova Gorica, School of Humanities, Vipavska 13, SI-5000 Nova Gorica, Slovenia

E-mail: helena.dobrovoljc@zrc-sazu.si, urska.vranjek@zrc-sazu.si

Abstract

This paper presents how language technology tools enable the integration of different types of normative data into a single language manual. The new Slovenian Normative Guide, the central normative manual consisting of normative rules and an orthographic dictionary, is based on language problems reported by language users. The normative guide consists of normative rules, and the orthographic dictionary supplements them with additional examples. The normative guide contains not only a systematic set of basic writing rules at the vowel-letter level (orthography or spelling), but also other consensual norms of the standard language. In order to effectively meet the needs of today's users of Slovenian, it was necessary to create a new concept for the orthographic dictionary so that it could effectively accompany the normative guide. In revising the normative rules, data collected on the Language Counselling Service platform were used. The normative guide is surrounded by three digitally interconnected layers of normative information; these three resources help the user navigate through the new normative view of the Slovenian language and provide arguments and explanations for the decisions made in the revision process.

Keywords: Slovenian; normative guide; orthographic dictionary; corpora research

1. Introduction

First we must point out some Slovenian peculiarities: Normative guides provide information about the acceptability of language elements for standard language use. In Slovenian, the standard language is an agreed supra-regional idiom that has been used in the written language since the middle of the 19th century. The so-called normative manuals (i.e. grammars, unabridged monolingual descriptive dictionaries of the standard Slovenian language and normative (orthographic) guides) are updated every few decades to harmonise the standard idiom with the natural language.

While grammars and descriptive dictionaries are universal concepts in linguistics, the term “normative guide”, which is the English equivalent of the Slovenian term “pravopis”, requires additional explanation. It is a manual consisting of normative rules accompanied by an orthographic dictionary. A normative guide includes not only a systematic set of basic writing rules at the vowel-letter level (orthography or spelling), but also other consensual norms of the standard language that determine the use of lower and upper case letters, writing and syntactic use of names from other languages, writing together or apart, the status and use of loan words and proper names, punctuation, and the like (Dobrovoljc, 2016).

Until the publication of the central unabridged monolingual descriptive dictionary *Slovar slovenskega knjižnega jezika* (hereafter SSKJ), it was understandable that even simpler spelling dictionaries had to contain very concise semantic and stylistic information. After the publication of the SSKJ, the orthographic dictionary needed a new concept. Unfortunately, this did not happen with the publication of the normative guide *Slovenski pravopis* 2001 (hereafter SP 2001). The codification of the Slovenian language went in two directions; discrepancies occurred not only between the normative manual as a whole and the descriptive dictionary, but also between the orthographic rules and the orthographic dictionary.

This paper presents the new *Slovenian Normative Guide* and its strategy for overcoming such inconsistencies with the help of a digital environment.

It was only after the publication of SP 2001 that it became clear that two partially overlapping dictionaries for the Slovenian language (the SSKJ and the orthographic dictionary) were not needed. A new concept for the orthographic dictionary had to be created so that it could effectively accompany the normative rules. The most typical language facts are listed in the general descriptive dictionary of the standard Slovenian language; however, the orthographic dictionary needs to include (1) the material expansion or enrichment of the normative rules (i.e. rules for the use of capital letters, borrowing, punctuation, writing together or apart) and (2) language elements that cause difficulties (i.e. atypical phrases that are difficult to use – problem-oriented approach). The starting point of this approach is the recognition that the way a word is written also depends on its meaning, which cannot be determined or represented in a dictionary without context (Moon, 2014).

2. The New Concept: Orthographic Codification of the Slovenian Language

The new approach to the elaboration of normative rules for the Slovenian language is problem-based; the problematic areas of language are identified with the help of an online language counselling service, which builds a database of Slovenian user language problems. The analysis of language use is carried out with the help of digital tools: text corpora and word sketches provide more advanced means of language processing, which allow the description of a standard language to be a more accurate representation of actual language use.

In designing the new normative guide, particular consideration will be given to the needs of language users and new linguistic facts. Organisationally, the lexical part will be produced simultaneously with the new semantic dictionary, so that its specialisation can be unambiguously normative (orthographic) and problem-oriented.

2.1 Three Layers of Information

Revision of normative rules with the help of data collected from the Language Counselling Service platform forms digitally interconnected layers of normative information.

The main source of linguistic dilemmas addressed in the new normative guide is the **Language Counselling Service** (*Jezikovna svetovalnica*) online platform, which is widely used by Slovenian language users to seek advice on linguistic choices and ambiguities; researchers use it to identify language description gaps (Dobrovoljc et al., 2020).

The platform (available at https://svetovalnica.zrc-sazu.si) has been active since 2012 and is based at the ZRC SAZU's Fran Ramovš Institute of the Slovenian Language. The different types of questions posted by the users of the language counselling site represent a rich and reliable source of difficulties with standard language usage which need to be taken into account in the process of revising the existing normative rules (Dobrovoljc et al., 2020).

User questions are answered by the staff of the Fran Ramovš Institute of the Slovenian Language, and each answer is approved by at least three members of the otherwise eight-member editorial board. Each answer is tagged with labels from three different levels of information:

- Language plane code (morphology, syntax, spelling, phonetics, etc.);
- Sub-area code (e.g. declension in morphology, verb-object agreement in syntax, etc.);
- Keyword (individual difficult cases and examples, e.g. COVID-19 related words).

The Language Counselling Service automatically creates a provisional online language guide with clickable codes and keywords. The platform is linked to other language resources available online (the *Fran* platform).

Language counselling points out gaps in language description and in language manuals; the creators of the new Slovenian Normative Guide (called *Pravopis 8.0*) use it as a source of language problems. *Pravopis 8.0* is an online normative manual and as such is part of the *Fran* Slovenian language portal (Fran Ramovš Institute for the Slovene Language ZRC SAZU, n.d.). The normative information surrounding it consists of three interconnected layers:

1. The **normative rules** are the theoretical part of the normative guide; they are available on the *Fran* platform under the name *Pravopis 8.0*. Each illustrative example in the rules is linked to the *ePravopis* orthographic dictionary.

2. The **Orthographic Dictionary** (*ePravopis*) is in its essence a normative dictionary; it is a growing dictionary the main purpose of which is to offer (additional) examples of the rules presented in the normative guide, a typical function of orthographic dictionaries (Verovnik, 2004). Dictionary entries contain information about spelling, pronunciation, text usage, morphological behaviour and word-formation possibilities of the included words.

3. Dictionary entries form problem-oriented groups linked to a publication called **Orthographic Categories** (*Pravopisne kategorije*), a collection of comments on how certain normative and orthographic difficulties are solved in the new orthographic dictionary, and a record of how the new orthographic dictionary differs from the current codification. Each dictionary entry is linked to its corresponding category, which contains a description of the linguistic problem and a list of all the entries included.

2.1.1 International Perspectives

As ZRC SAZU's Fran Ramovš Institute for the Slovenian Language is the only research institution dealing with the orthography of the Slovenian language, the new Normative Guide, together with the new Orthographic Dictionary and Orthographic Categories, is the central language resource of its kind for the Slovenian language. However, the idea of an online platform through which linguists can obtain data on language difficulties from a wide range of language users is not new. In 2011, a collection of language problems in the standard Slovenian language was formed, as part of the project "Sporazumevanje v slovenskem jeziku" (Communication in Slovene, available at http://eng.slovenscina.eu). In this project, existing online language counselling resources were used to create a manual of style. The project was based on best practices in European linguistic projects where language portals have been successfully designed using the public engagement method. There are several similar contemporary resources in other European countries, which are described below.

The "Grammis" portal of the Leibniz Institute for the German Language in Mannheim (available at https://grammis.ids-mannheim.de) was published in 1997. On this portal, users can find out about difficulties of the German language in the form of questions and answers as presented in the current grammar of the German language. Today, the portal has developed into a comprehensive hypermedia network information system, which is currently being expanded in terms of content and functionality.

The online Czech language handbook, created within the project "Internetová jazyková příručka" (Internet Language Reference Book, available at https://prirucka.ujc.cas.cz) by the Czech Language Institute of the Czech Academy of Sciences and the Faculty of Informatics of the Masaryk University (Pala & Šmerk, 2011), is currently being expanded. The content of the reference book is based on the questions and problems posed to the linguists of the Czech language counselling centre “Jazyková poradna”. The Czech linguists still offer language advice over the phone.

A similar, partially interconnected system exists for the Croatian language (Vranjek Ošlak & Černivec, 2021). The Croatian language counselling service uses questions from speakers of Croatian as a framework for their printed language manuals. The answers are also published online (https://jezicni-savjetnik.hr).

The Estonian language advisory service is organised in a similar way. The Estonian “Keelenõuanded” (Language Council, available at https://keeleabi.eki.ee) of the Institute of Estonian Language answers language questions by phone, e-mail and mail. Estonian linguists answer questions about grammatical and orthographic difficulties in Estonian; however, they do not answer questions related to language policy or teaching. Their answers are published on a website. The same applies to the language advisory service of the Institute for the Languages of Finland (to be found at https://www.kotus.fi/en/services/telephone_counselling).

The Slovak language advisory service called “Jazyková poradňa JÚĽŠ SAV” operates online; the platform is a joint project of the Sme.sk portal and the Ľudovít Štúr Institute of Linguistics of the Slovak Academy of Sciences (available at https://www.juls.savba.sk/poradna.html). During the coronavirus pandemic, language advice played an important role in helping lay people, lecturers and teachers with their language problems.

The Language Counselling Service for the Slovene language, presented below, is at its core an advisory service; however, it is also a starting point for active public involvement in the process of updating (normative) language manuals. The main goal of the committee responsible for revising the normative rules and of the associated researchers is to produce up-to-date online manuals based on real language difficulties as pointed out by language users.

2.2 Language Counselling as a Source of Language Difficulties

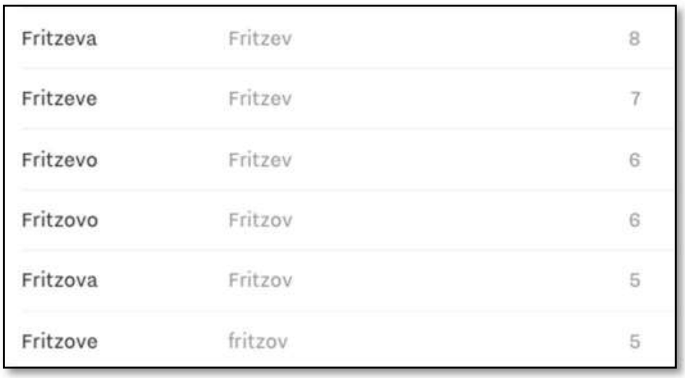
In the following, we exemplify how the questions posed in the Language Counselling Service contribute to the enrichment of the new Slovenian Normative Guide, namely the normative rules. We selected two user questions that were the basis for the corresponding corpus research; they investigate the morphological behaviour of loan words:

a) Do proper names ending in *-tz* show vowel alternation (*o* to *e*) in morphological inflexion and possessive adjective formation? Case No.1: *Fritz* – instrumental case: *Fritzem*/*Fritzom*, possessive adjective: *Fritzev*/*Fritzov*.

b) What is the gender of loan words with atypical endings? Case No.2: *karitas* – feminine or masculine.

2.2.1 Case No.1: Morphological and word-formational variability of forms

A morphosyntactically tagged text corpus allows us to extract data on morphological duplicates. For example, if we want to know where Slovenian language users are hesitant about the syntactic use of the German name *Fritz*, the corpus provides a list of duplicate forms (Figure 1) from which we could deduce: (a) that duplicate forms *Fritzom*/*Fritzem* in the instrumental case are frequent in the corpus and b) that users often question the vowel alternation in the possessive adjective ending in -*ov*/-*ev* (*Fritzov*/*Fritzev*).



Fritzeva	Fritzev	8
Fritzeve	Fritzev	7
Fritzevo	Fritzev	6
Fritzovo	Fritzov	6
Fritzova	Fritzov	5
Fritzove	fritzov	5

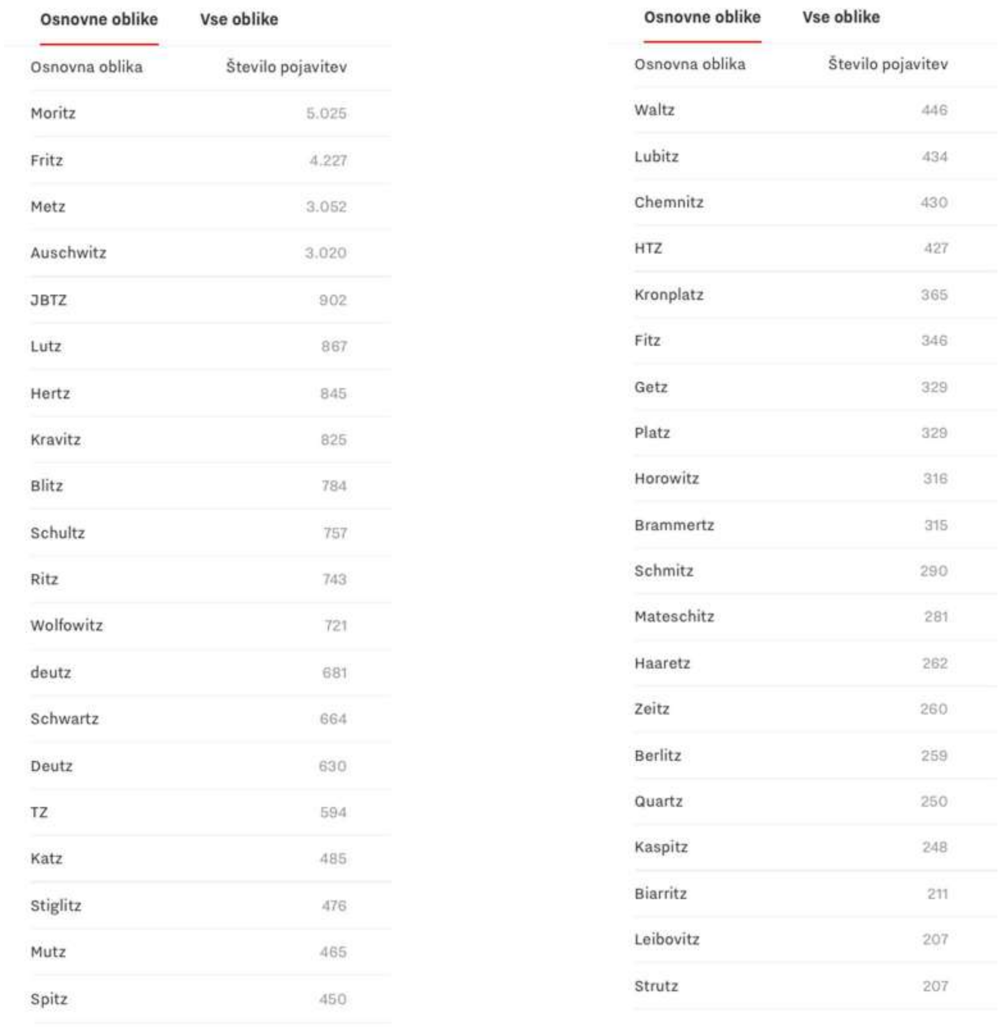
Figure 1: A list of duplicate forms ending in -*ov*/-*ev* (*Gigafida 2.0* corpus)

The norm thus established shows that vowel alternation in Slovenian depends not only on pronunciation but also on the notation. If the combination of letters ⟨t⟩ and ⟨z⟩ is understood as a digraph and pronounced as [c], this triggers the realisation of vowel alternation *o* to *e* (*Fritzev*). However, if the combination is perceived only as a sequence of letters ⟨t⟩ and ⟨z⟩ and not as a digraph, the vowel alternation does not occur and the endings in written language follow the notation (*Fritzov*).

In drafting a normative rule to represent these findings in the normative guide, it is necessary to systematically examine all possible cases that could be included in the normative guide and orthographic dictionary. For this purpose, a glossary is created (Figure 2).

A review of the corpus material shows that vowel alternation *o* > *e* is merely a possibility and that such a phenomenon is systematically observed in all names ending in the final spoken [c] sound when it is written with other letters or letter combinations, e.g. Hungarian names ending in -*cz* (*Göncz*), German names ending in -*z* (*Leibniz*), or when the stem of a proper name contains the spoken [c] sound, e.g. Italian names ending in -*zza* or -*zzo* (*Tomizza*, *Campazzo*).[1]

[1] This particular rule's status is currently at the proposal level and is not yet published in the new normative guide (*Pravopis 8.0*).



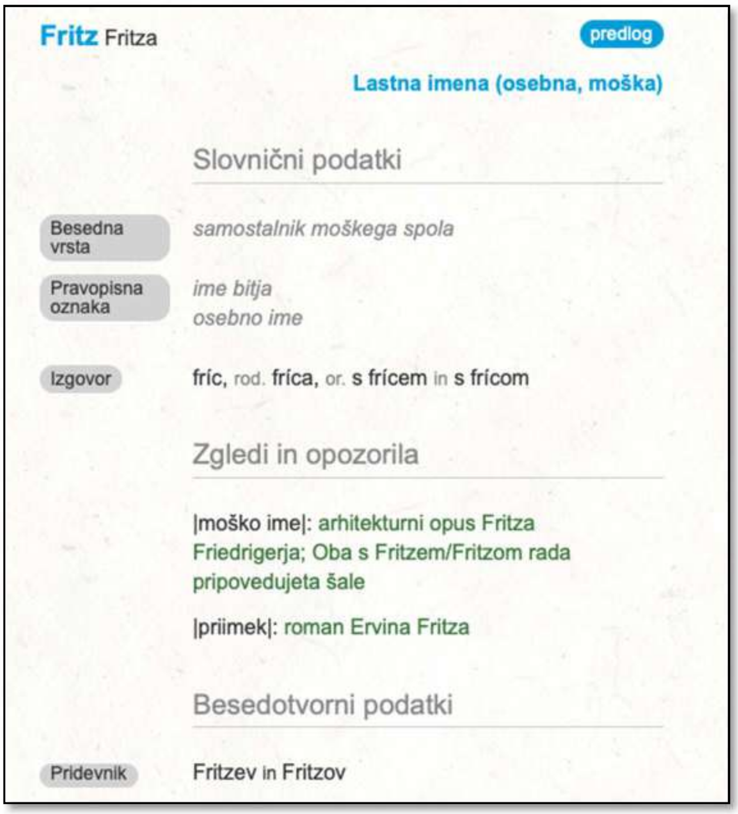
Osnovna oblika	Število pojavitev
Moritz	5.025
Fritz	4.227
Metz	3.052
Auschwitz	3.020
JBTZ	902
Lutz	867
Hertz	845
Kravitz	825
Blitz	784
Schultz	757
Ritz	743
Wolfowitz	721
deutz	681
Schwartz	664
Deutz	630
TZ	594
Katz	485
Stiglitz	476
Mutz	465
Spitz	450

Osnovna oblika	Število pojavitev
Waltz	446
Lubitz	434
Chemnitz	430
HTZ	427
Kronplatz	365
Fitz	346
Getz	329
Platz	329
Horowitz	316
Brammertz	315
Schmitz	290
Mateschitz	281
Haaretz	262
Zeitz	260
Berlitz	259
Quartz	250
Kaspitz	248
Biarritz	211
Leibovitz	207
Strutz	207

Figure 2: Case No.1 glossary example: similar borrowed names ending in -*tz* (*Gigafida 2.0* corpus)

The normative rule must therefore include general instructions for changing the pronunciation and ending attribution in cases where a borrowed name is pronounced with the final sound [c].

The normative rules thus prepared are deductively coherent with the orthographic dictionary (*ePravopis*) on the *Fran* language portal. The realisation of this ending attribution phenomenon is expected without exceptions for all names in the dictionary referring to this rule, following the presented example – *Fritz* (Figure 3). Moreover, the background of the decisions made in the process of normative rule formation is explained in the corresponding orthographic category.



Fritz Fritza

predlog

Lastna imena (osebna, moška)

Slovnični podatki

Besedna vrsta	*samostalnik moškega spola*
Pravopisna oznaka	*ime bitja* *osebno ime*
Izgovor	fríc, rod. fríca, or. s frícem in s frícom

Zgledi in opozorila

|moško ime|: arhitekturni opus Fritza Friedrigerja; Oba s Fritzem/Fritzom rada pripovedujeta šale

|priimek|: roman Ervina Fritza

Besedotvorni podatki

Pridevnik	Fritzev in Fritzov

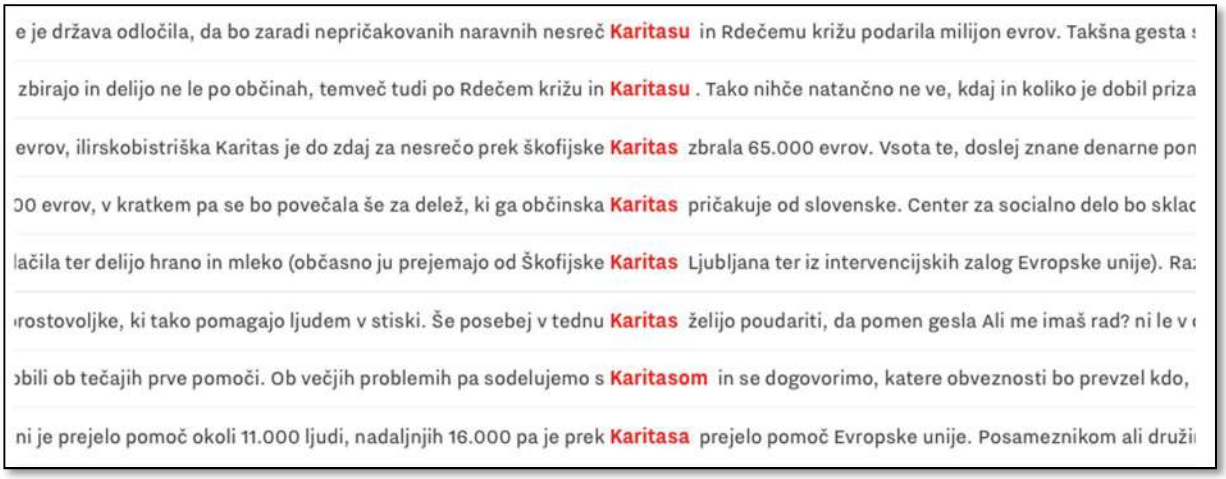
Figure 3. Orthographic dictionary entry: ending attribution phenomenon in the case of *Fritz* (see bottom line: possessive adjectives *Fritzev* and *Fritzov*)

2.2.2 Case No.2: Morphological variability of syntactic categories

Although corpus queries are in principle frequency-oriented on sets of the most common lexemes, and researchers are primarily concerned with establishing the lowest boundary of relevant hits (Holz, 2003), deviations are also important for finding out what is difficult for users of a particular language. These difficulties are not only defined by the right-or-wrong dichotomy; we are increasingly aware of standard linguistic diversity and thus language choice (Larsen-Freeman, 2000). Borrowing in Slovenian happens morphologically; we borrow both proper nouns and common nouns by adapting their grammatical categories in terms of ending to the existing system. However, often words from related languages (e.g. the Serbian and Croatian names *Užice* and *Brela*) or classical languages (*karitas* 'charity') indicate different possibilities, as the user experience with these nouns may be different.

The Language Counselling Service received this user question: "I have a question regarding the name of the organisation *Slovenska karitas* or in short *Karitas*. According to their website and the language use, the gender of this word is predominately feminine. The normative guide, however, determines it is masculine. That is the first problem. The second problem is capitalisation: *Škofijska karitas Koper* or *Škofijska Karitas Koper* or *škofijska Karitas Koper*. What is the correct capitalisation of the name of this organisation?"

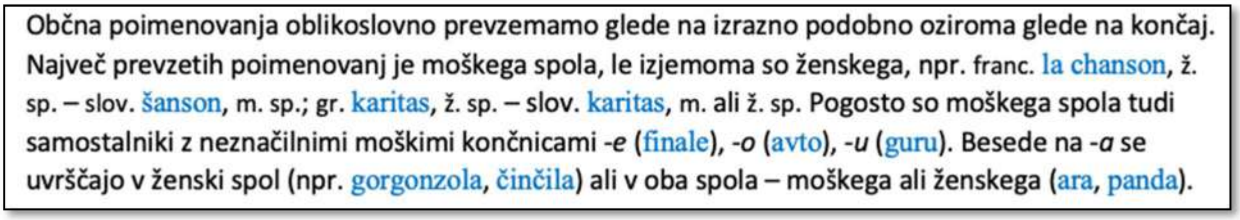
In the case of the word *karitas*, standard language manuals recommend keeping the gender of the word as it is in the donor language (partly because of the important role of the connoisseurs of classical languages, who generally keep the gender of words: SSKJ classifies this noun as feminine and indeclinable). Lay users, however, follow the system of the Slovenian language and decline this noun like its parallel e.g. *ananas* ('pineapple', masculine and declinable). The current normative guide *Slovenski pravopis* 2001 characterises both the common noun *karitas* and the proper name *Karitas* as masculine. Despite the relevant normative rule, the use of the common noun *karitas* indicates an increase in the frequency of the feminine gender (Figure 4), which is due to the related proper name *Slovenska karitas* (as the preceding adjective suggests, the noun *karitas* is feminine and indeclinable).



```
e je država odločila, da bo zaradi nepričakovanih naravnih nesreč Karitasu in Rdečemu križu podarila milijon evrov. Takšna gesta
zbirajo in delijo ne le po občinah, temveč tudi po Rdečem križu in Karitasu . Tako nihče natančno ne ve, kdaj in koliko je dobil priza
evrov, ilirskobistriška Karitas je do zdaj za nesrečo prek škofijske Karitas zbrala 65.000 evrov. Vsota te, doslej znane denarne pon
00 evrov, v kratkem pa se bo povečala še za delež, ki ga občinska Karitas pričakuje od slovenske. Center za socialno delo bo skla
lačila ter delijo hrano in mleko (občasno ju prejemajo od Škofijske Karitas Ljubljana ter iz intervencijskih zalog Evropske unije). Ra
rostovoljke, ki tako pomagajo ljudem v stiski. Še posebej v tednu Karitas želijo poudariti, da pomen gesla Ali me imaš rad? ni le v
bili ob tečajih prve pomoči. Ob večjih problemih pa sodelujemo s Karitasom in se dogovorimo, katere obveznosti bo prevzel kdo,
ni je prejelo pomoč okoli 11.000 ljudi, nadaljnjih 16.000 pa je prek Karitasa prejelo pomoč Evropske unije. Posameznikom ali druži
```

Figure 4: Declination variability: *karitas* (*Gigafida 2.0*)

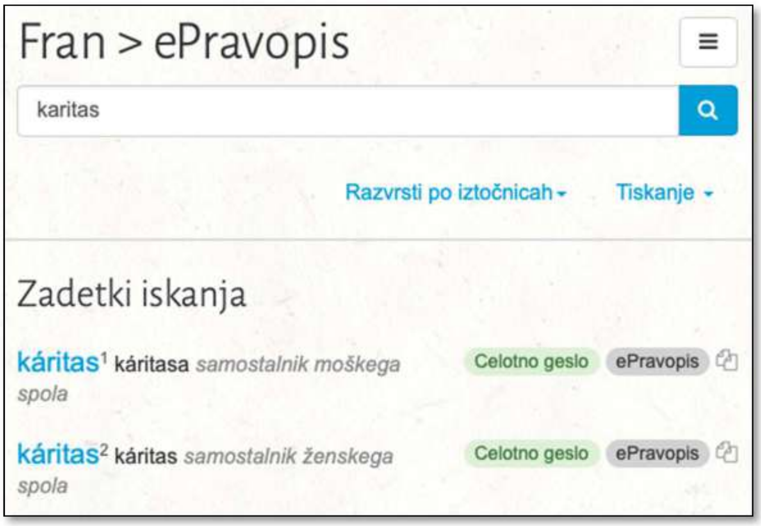
In normative rule-making, then, the material dictates the rule formulation, which must express the following linguistic fact (Figure 5): In short, most borrowed names in Slovenian are masculine, and in rare exceptions feminine. The noun *karitas* (originally feminine) is masculine or feminine in Slovenian.



Občna poimenovanja oblikoslovno prevzemamo glede na izrazno podobno oziroma glede na končaj. Največ prevzetih poimenovanj je moškega spola, le izjemoma so ženskega, npr. franc. la chanson, ž. sp. – slov. šanson, m. sp.; gr. karitas, ž. sp. – slov. karitas, m. ali ž. sp. Pogosto so moškega spola tudi samostalniki z neznačilnimi moškimi končnicami *-e* (finale), *-o* (avto), *-u* (guru). Besede na *-a* se uvrščajo v ženski spol (npr. gorgonzola, činčila) ali v oba spola – moškega ali ženskega (ara, panda).

Figure 5: The new normative guide (*Pravopis 8.0*): loan words rule formulation

In Figure 5, the examples in blue contain links to dictionary entries (Figure 6) included in the orthographic dictionary (*ePravopis*). In this way, the interconnectedness of the new Normative Guide (normative rules) and the associated orthographic dictionary is ensured.



Zadetki iskanja

káritas^{1} káritasa *samostalnik moškega spola*

káritas^{2} káritas *samostalnik ženskega spola*

Figure 6: The new orthographic dictionary (*ePravopis*): *karitas* as headword (masculine and feminine homonyms)

3. Interconnectedness: How it Works

In the following chapter, we show how the above-mentioned three-layered interconnectedness works in the case of Slovenian temporal names, namely names of days, months, seasons, historical events and points in time such as holiday names.

3.1 Temporal Names and Capitalisation

In Slovenian, temporal names are not considered proper noun categories, as for example in English, where month names (*July*, *May*), names of days (*Monday*, *Sunday*), names of historical events (*the French Revolution*) and of points in time (*All Saints' Day*) are considered as proper names and written with a capital initial letter (Langendonck, 2007). In Slovenian, all temporal names are written without a capital initial letter regardless of syntactic position, which can be typically proper (*mesec maj* ‘the month named May’) or typically common-noun (*vsi trije božiči so minili mirno* ‘all three Christmases passed peacefully’ in the sense of 'all three Christmas holidays': Christmas Eve, New Year’s Eve and the evening before Epiphany). The current fashion of writing holiday names without capital initial letters does not follow their syntactic role, but derives from a traditional agreement. In the first half of the 20th century, capitalisation was often rejected because it was characteristic of Germanic languages. Due to the nation's Austro-Hungarian past, Slovenian language speakers were constantly in a position of bilingualism, as German was the dominant language (Štih et al., 2008). Linguists therefore rejected capitalisation as something foreign and not in accordance with the history and structure of the Slovenian language (Dobrovoljc, 2004).

Certain groups in the Slovenian language community occasionally petition to change this normative rule, hoping to achieve its alternation into writing holidays with a capital initial letter. The reasons for these episodic tendencies are (1) emotional (capital initial letters are associated with respect, especially regarding religious holidays); (2) influenced by foreign practices entering the Slovenian language, through e.g. greeting cards; or (3) are the result of certain beliefs that the lowercase initial letter reflects 45 years of enforced reduction of religious practices.

The committee responsible for revising the normative rules (*Pravopisna komisija pri SAZU in ZRC SAZU*) considered the current social situation and the requirements of various social groups when revising the chapter on upper and lower case letters in accordance with the methodology outlined above.

vesel božič	161
Vesel božič	88
vesel Božič	59
Vesel Božič	43
VESEL BOŽIČ	14
veseli božič	2
veseli Božič	1
veselemu Božiču	1
veselejši božič	1
veselega božiča	1
veselega Božiča	1
vesele božič	1
vesele Božič	1
vesel BOŽIČ	1
Veselega božiča	1
Vesel BOŽIČ	1

Table 1: Corpus query: *vesel božič* in the *slWac* corpus

In both versions of the central reference corpus for the Slovenian language *Gigafida*, proofread texts predominate, making it impossible for the corpus to reflect intuitive writing practices. In previous research (Dobrovoljc & Vranjek Ošlak, 2018), it was argued that contemporary linguistic research must also be conducted on non-standard language corpora (e.g. *Janes* or *slWac*), as they often yield different results.[2] The study of linguistic material was therefore also focused on various other corpora (Table 1).[3] Since in Slovenian the word *božič* is homonymous with the surname *Božič*, which is relatively common, we looked for a characteristic greeting, namely *vesel božič* 'Merry Christmas'.

[2] One of the reviews pointed out that "users in some user-generated contents write nonstandardly on purpose or by decision. This influences the genre itself." This is true, of course, but it does not preclude the possibility of using non-standard language corpora as a means of comparison and for testing indicators of language change in particular. The predominant source of research is still reference and representative language corpora such as *Gigafida.*

[3] All corpora used are available at: https://www.clarin.si/noske/.

A comparison of the corpus queries in Table 2 shows that the use of capital letters has not increased over the years (comparing the *Gigafida* 1.1 and 2.0 corpora), but capitalisation is much more noticeable in the *slWac* corpus (Slovenian websites) and in the *Janes* corpus (blog texts, online chat rooms, tweets, etc.); another obvious feature is also non-standard notations (e.g. colloquial words written according to their pronunciation). We also noticed that language users avoid having to choose upper or lower case in certain cases, namely by using only capital letters.

	Gigafida 1.1	**Janes**	**slWac**	**Gigafida 2.0**
vesel božič	71.28%	63.33%	67.64%	83.40%
vesel Božič	18.44%	39.20%	28.12%	13.82%
vesel BOŽIČ	10.17%	6.32%	4.24%	2.81%
errors		1.49%		

Table 2: Corpus query: *vesel božič* in four different corpora

In order to place the material research described above in a linguistic and social context, we carried out two further analyses. We were interested in the relationship between single-word and multi-word holiday names, so we also examined corpora material with regard to the use of the holiday name *velika noč* 'Easter' (Table 3).

	Gigafida 1.1	**Janes**	**slWac**	**Gigafida 2.0**
velika noč	54.22%	40.39%	47.69%	63.86%
Velika noč	38.84%	50.00%	47.31%	34.87%
Velika Noč	6.42%	5.87%	4.04%	1.14%
VELIKA NOČ	0.42%	1.78%	0.96%	0.13%

Table 3: Corpus query: *velika noč* in four different corpora

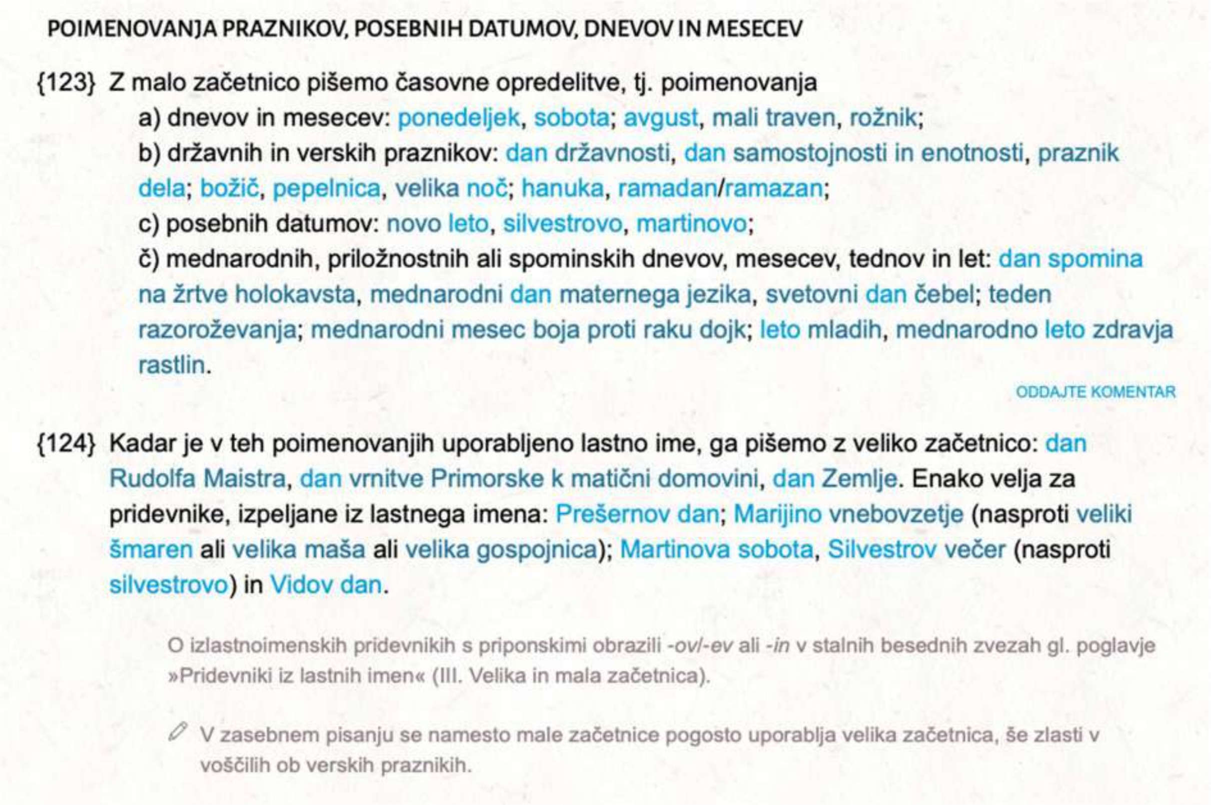
	Gigafida 1.1	**Janes**	**slWac**	**Gigafida 2.0**
dan državnosti	91.52%	67.13%	74.05%	91.52%
Dan državnosti	7.15%	29.98%	22.94%	7.15%
dan Državnosti	0.07%	0.44%	0.64%	0.07%
Dan Državnosti	0.02%	0.61%	0.20%	0.02%
DAN DRŽAVNOSTI	1.24%	1.75%	2.17%	1.24%
errors		0.09%		

Table 4: Corpus query: *dan državnosti* in four different corpora

Based on the wording in the normative rules, we checked whether the increase in the use of the capital initial was related to religious content; therefore, we conducted a parallel study of the holiday name *dan državnosti* 'Statehood Day' (Table 4).[4]

Comparison of the results shows that capitalisation is significantly more common for religious holiday names than for others, but the predominant notation manner is still lowercase, which has been preferred for over a hundred years. The influence of multi-word holiday names is negligible.

The formulation of the normative rule is thus threefold (Figure 7). (1) Holiday names, regardless of the type of holiday (religious, national, European, etc.), continue to be written with a lowercase initial letter, as normative tradition and also prevailing usage dictate – {123}. (2) A separate admonition refers to holiday names containing proper nouns; they are capitalised, e.g. *dan Zemlje* 'Earth Day', *dan svetega Patrika* 'St. Patrick's Day' – {124}. (3) The so-called *stylistic instruction* (marked with a pencil symbol) introduces the possibility of writing holiday names in private correspondence with a capital initial, as a sign of respect, especially for religious holidays.



POIMENOVANJA PRAZNIKOV, POSEBNIH DATUMOV, DNEVOV IN MESECEV

{123} Z malo začetnico pišemo časovne opredelitve, tj. poimenovanja
a) dnevov in mesecev: ponedeljek, sobota; avgust, mali traven, rožnik;
b) državnih in verskih praznikov: dan državnosti, dan samostojnosti in enotnosti, praznik dela; božič, pepelnica, velika noč; hanuka, ramadan/ramazan;
c) posebnih datumov: novo leto, silvestrovo, martinovo;
č) mednarodnih, priložnostnih ali spominskih dnevov, mesecev, tednov in let: dan spomina na žrtve holokavsta, mednarodni dan maternega jezika, svetovni dan čebel; teden razoroževanja; mednarodni mesec boja proti raku dojk; leto mladih, mednarodno leto zdravja rastlin.

{124} Kadar je v teh poimenovanjih uporabljeno lastno ime, ga pišemo z veliko začetnico: dan Rudolfa Maistra, dan vrnitve Primorske k matični domovini, dan Zemlje. Enako velja za pridevnike, izpeljane iz lastnega imena: Prešernov dan; Marijino vnebovzetje (nasproti veliki šmaren ali velika maša ali velika gospojnica); Martinova sobota, Silvestrov večer (nasproti silvestrovo) in Vidov dan.

O izlastnoimenskih pridevnikih s priponskimi obrazili *-ov/-ev* ali *-in* v stalnih besednih zvezah gl. poglavje »Pridevniki iz lastnih imen« (III. Velika in mala začetnica).

✎ V zasebnem pisanju se namesto male začetnice pogosto uporablja velika začetnica, še zlasti v voščilih ob verskih praznikih.

Figure 7: The new normative guide (*Pravopis 8.0*): holiday names rule formulation

Since the normative rules are illustratively and factually limited to only the most typical examples, which are of course linked to dictionary entries in the orthographic dictionary, an additional explanatory section called Orthographic Categories (*Pravopisne kategorije ePravopisa*) was conceived, which is also available online. This

[4] Capital initial letter occurences include those at the beginning of sentences.

section explains for each normative rule whether and how it has been changed; it also describes the possible language difficulties. For holiday names, Orthographic Categories lists all holiday names included in the orthographic dictionary, focusing not only on the initial case, but also on (a) newer holiday names in Slovenian that have not yet been included in a dictionary (e.g. *ašura* 'Ashura', *noruz* 'Novruz'), and on (b) synonymous holiday names (*judovska velika noč* 'the Jewish Easter' and *pasha* 'Passover'). Holiday identifiers were supplemented with dates where possible, otherwise with other relevant information. A link directs the language user to the relevant article of the old and new normative rules.

4. Conclusion

The new Slovenian Normative Guide draws from the knowledge of the generational, cognitive and educational diversity of language users. The creators of the new approach tried to write **the same linguistic information in different language codes and connect them**. In creating the normative manuals, an interconnected online system was designed, combining a language counselling platform, normative rules, an orthographic dictionary, and a description of normative solutions. All these levels are clickable and interlinked.

Through language-related questions asked on the Language Counselling Service platform, linguists encounter language difficulties. Through corpus material research, the working group learns about the problem and finds similar use cases. It articulates these in normative rules that concisely inform about the problem and its exceptions, and provide hints that are often stylistic in nature. All illustrative examples in the rules are included in the dictionary and linked to the corresponding dictionary entries. Normative solutions are presented in Orthographic Categories; from there, the user is referred back to the normative guide, completing the circle of interrelated linguistic information.

The same linguistic information is represented in different ways, and these different representations are interconnected so that the user can choose the most appropriate one. The choice depends on the user's prior knowledge of the language and practical experience with language use. The interconnectedness of the normative guide, orthographic dictionary, Language Counselling Service and Orthographic Categories is made possible by the digitally designed databases in the background. Language manuals designed in this way can reach a larger number of language users and be more user-friendly by making the normative information more accessible.

5. Acknowledgements

This contribution was prepared within the framework of the P6-0038 programme financed by the Slovenian Research Agency (ARRS).

6. References

Dobrovoljc, H. (2004). *Pravopisje na Slovenskem.* Ljubljana: Založba ZRC, ZRC SAZU.

Dobrovoljc, H. (2016). Povezljivost pravopisnih pravil in slovarja: sanje pravopiscev 20. stoletja. In Erjavec, T. & Fišer, D. (eds.) *Proceedings of the conference on Language Technologies & Digital Humanities.* Ljubljana, Slovenia. Available at: http://www.sdjt.si/wp/wp-content/uploads/2016/09/JTDH-2016DobrovoljcPovezljivost-pravopisnih-pravil-in-slovarja.pdf.

Dobrovoljc, H., et al. (2020). *Kje pa vas jezik žuli?.* Ljubljana: Založba ZRC, ZRC SAZU.

Dobrovoljc, H., Vranjek Ošlak, U. (2018). Zakaj ne z eno poizvedbo hkrati po različnih korpusih? (Troje korpusnih preverb pod primerjalnim drobnogledom). In Erjavec, T. & Fišer, D. (eds.) *Proceedings of the conference on Language Technologies & Digital Humanities.* Ljubljana, Slovenia. Available at: http://www.sdjt.si/wp/wp-content/uploads/2018/09/JTDH-2018_Dobrovoljc-H_Zakaj-ne-z-eno-poizvedbo-hkrati-po-razlicnih-korpusih.pdf.

Fran. Accessed at: https://fran.si/. (20 January 2021)

Grammis. Accessed at: https://grammis.ids-mannheim.de. (8 June 2021)

Holz, N. (2003). Besedilni korpus Nova beseda in geslovnik za Slovar novejšega besedja. *Jezikoslovni zapiski* 9, pp. 89–94.

Institute for the Languages of Finland. Accessed at: https://www.kotus.fi/en/services/telephone_counselling. (8 June 2021)

Internetová jazyková příručka. Accessed at: https://prirucka.ujc.cas.cz. (8 June 2021)

Jazyková poradňa JÚĽŠ SAV. Accessed at: https://www.juls.savba.sk/poradna.html. (8 June 2021)

Jezični savjetnik. Accessed at: https://jezicni-savjetnik.hr. (8 June 2021)

Jezikovna svetovalnica. Accessed at: https://svetovalnica.zrc-sazu.si/. (20 January 2021)

Keelenõuanded. Accessed at: https://keeleabi.eki.ee. (8 June 2021)

Langendonck, W. (2007). *Theory and Typology of Proper Names.* De Gruyter Mouton.

Larsen-Freeman, D. (2009). *The Grammar of Choice.* Teaching and testing grammar. The handbook of language teaching. Blackwell.

Moon, R. (2014). Meanings, Ideologies, and Learners' Dictionaries. In Abel, A., Vettori, C. & Ralli, N. (eds.) *Proceedings at the XVI EURALEX International Congress: The User in Focus. Part 3.* Bolzano. Available at: http://euralex2014.eurac.edu/en/callforpapers/Documents/EURALEX_Part_3.pdf.

NoSketchEngine. Clarin. Accessed at: https://www.clarin.si/noske/. (2 April 2021)

Pala, K. & Šmerk, P. (2011). Internetová jazyková příručka. *Zpravodaj ÚVT MU*, XXI(3), pp. 14–17. Available at: http://webserver.ics.muni.cz/bulletin/articles/666.html.

Pravopis 8.0. Accessed at: https://www.fran.si/pravopis8. (20 January 2021)

Pravopisne kategorije. Accessed at: https://fran.si/spt-kategorije. (20 January 2021)

Sporazumevanje v slovenskem jeziku. Accessed at: http://eng.slovenscina.eu. (8 June 2021)

Štih, P., et al. (2008). *Slovenska zgodovina: družba – politika – kultura.* Ljubljana: Inštitut za novejšo zgodovino. Available at: http://www.sistory.si/publikacije/pdf/zgodovina/Slovenska-zgodovina-SLO.pdf.

Toporišič, J., et al. (eds.) (1990). *Slovenski pravopis 1 – Pravila.* Ljubljana: SAZU, Državna založba Slovenije.

Toporišič, J., et al. (eds.) (2001). *Slovenski pravopis.* Ljubljana: SAZU, ZRC SAZU, Založba ZRC. Available at: https://fran.si/134/slovenski-pravopis.

Verovnik, T. (2004). Norma knjižne slovenščine med kodifikacijo in jezikovno rabo v obdobju 1950–2001. *Družboslovne razprave* 20, pp. 241–258.

Vranjek Ošlak, U. & Černivec, M. Slovenski in hrvaški pravopis – podobnosti in razhajanja na prehodu v digitalno dobo. In Nikolovski, G. (ed.) *Izzivi slavistike v 21. stoletju: 4. mednarodna znanstvena konferenca Slavistični znanstveni premisleki, zbornik povzetkov.* Maribor: Univerza v Mariboru, Univerzitetna založba. Available at: https://press.um.si/index.php/ump/catalog/view/562/713/1411-1.

An Online Tool Developed for Post-Editing the New Skolt Sami Dictionary

Mika Hämäläinen[1], Khalid Alnajjar[1], Jack Rueter[1], Miika Lehtinen[2] and Niko Partanen[1]

[1]University of Helsinki, Unioninkatu 40, 00100 Helsinki, Finland
[2] University of Oulu, Pentti Kaiteran katu 1, 90570 Oulu, Finland
E-mail: firstname.lastname@helsinki.fi, firstname.lastname@oulu.fi

Abstract

In this paper, we present our free and open-source online dictionary editing system that has been developed for editing the new edition of the Finnish-Skolt Sami dictionary. We describe how the system can be used in post-editing a dictionary and how NLP methods have been incorporated as a part of the workflow. In practice, this means the use of FSTs (finite-state transducers) to enhance connections between lexemes and to generate inflection paradigms automatically. We also discuss our work in the wider context of lexicography of endangered languages. Our solutions are based on the open-source work conducted in the Giella infrastructure, which means that our system can be easily extended to other endangered languages as well. We have collaborated closely with Skolt Sami community lexicographers in order to build the system for their needs. As a result of this collaboration, the latest Finnish-Skolt Sami dictionary was edited and published using our system.

Keywords: Skolt Sami, online dictionary, NLP

1. Introduction

In this paper, we present an online system developed in close collaboration with linguists and native speakers during the Skolt Sami dictionary project (see Alnajjar et al. 2020). We recognise that when developing lexical resources for endangered languages we must take into account various user groups and their needs, and the resource that is created is often in a very important position for the entire language community. Large dictionaries in endangered languages often play an important role in the future language development and efforts at normalisation. This means that these projects entail lots of responsibility. Establishing a common ground with knowledgeable native speakers and pencil and paper linguists with regard to online editing can present quite a challenge. Native speakers, on the one hand, need to be given an understandable and intuitive system for interacting with the growing dictionary database. Experienced linguists, on the other, may at times require an outstretched hand of enlightenment, one that introduces them to direct work in a database without interceding paper prints for contemplation of all entries with a pencil and eraser. The developers, of course, must also be prepared to design print-out and download possibilities just in case the users have difficulties managing the computer-readable data. In these instances the exported versions should also be used primarily to read and use the dictionary, and the changes should be done in the actual database, if possible.

Only this way we can ensure that the lexical resources that are being created will definitely benefit different user groups, and take into account the multiple purposes these materials can be used for. We also acknowledge that there is a need for specialised lexicographic solutions in different situations, and that the work presented here on Skolt Sami is just one of the many possibilities. At the same time there are many important lessons to be learned from our Skolt Sami work, and these can be generalised in different scenarios.

The work with Skolt Sami was started using a tabular data format. Spreadsheet editing programs are readily available and many linguists as well as native speakers are familiar with them, so it is obvious many endangered language lexicons appear in such formats. For this reason, our system has also been designed so that these can be processed. Converting different tabular files is often not trivial, and this was the case in this project too, with the original lexical resources for Skolt Sami presenting several challenges. The most prominent consisted of a malformed flat CSV file containing several character encoding issues. We built our online system so that it fixes such issues while importing the flat structure into its internal graph based representation. Similar issues are common for different materials on endangered languages, so our solutions generalise very well to this wider context. We use graphs as the internal structure for their advantages over trees (see Mechura 2016). Despite the popularity of spreadsheets, this structure is poorly suited to lexicographic work. There are relations between entries, hierarchical entries, and additional content such as example sentences that can serve as examples for multiple different headwords, in which case repeating them again and again is not desirable. Lexicographic data is by nature relatively complicated to model, but as we will describe, the approach to import tabular formats into our online tool seems to provide a very good starting point for the creation of such a more complex structure, partly through automatic conversions and deductions.

Even though Skolt Sami is severely endangered with its 300 native speakers (Moseley, 2010), thanks to previous projects on its digital revitalisation, the language has morphological analysers (Rueter & Hämäläinen, 2020) that our system can use when importing data. Our system will automatically add relation information such as derivations and compounds to lexemes with the help of the morphological tools. If the system were to be used for a language that does not have a morphological analyser, these relations would need to be created either manually or by using different heuristics. In any case, the resulting dictionary would not be as interlinked.

Our system has been in continuous use by linguists and trained native speakers, who have been editing the lexicographic material into a publishable form. We have introduced constant improvements to the system based on the feedback from our actual users. Some of the requested functionalities have been automatic morphological inflections for full inflectional paradigms for each entry with a feedback facility, the ability to have an overseer view where a super user can see the edit history of each entry and finalise/approve it, and the ability of showing lexicographic information from other sources, such as the Sami TermWiki[1].

The final product, a printed edition of the dictionary (Lehtinen et al., 2021) was recently published, and it was greatly facilitated by the fact that our system can output the desired lexicographic content in a LaTeX format for easy PDF conversion. Other output formats could be easily added, if needed by the community or researchers.

Currently, we are extending the use of our system to other endangered languages documented in the Giella infrastructure (Moshagen et al., 2014). Like Skolt Sami, these languages have morphological tools as well, which makes work with them analogous to what we have already developed for Skolt Sami.

2. Related Work

Developing dictionaries is essentially connected to language documentation and revitalisation activities in the contemporary world. With entirely undocumented languages

[1] https://satni.uit.no/termwiki/

the lexicon is built from scratch as part of the corpus building and elicitation process, whereas in many cases there are existing dictionaries and lexical resources that can be used. Common approaches are to extend existing resources, or to publish them again in a digital format. There is also extensive global variation in what kind of resources exist and what kind of challenges are connected to making them usable for the communities. We will describe some of the most relevant work next.

Especially with the work on endangered languages of North America there are many examples where unfamiliarity with the orthographic conventions of the language is an issue in language learning. Additionally, many orthographic norms are not entirely fixed, if they exist at all, which is a challenge for lexicographic work. It is also a problem for a new use of the lexical infrastructure, as the user cannot be expected to know how to find a specific entry in the dictionary. Both spell relax and morphological awareness are methods that have been used in Tsimsianic and Salish dictionaries, with the aid of language technology that has been developed for these languages (Littell et al., 2017).

Another example comes from work done with St. Lawrence Island Yupik, where the language materials have been made openly available for the community online. Different writing systems that have previously been used for this language have been taken into account as different input methods, also here with the aid of morphological modelling (Hunt et al., 2019). As similar situations with various writing systems is very common for endangered languages around the world, and there are various ways to handle this issue. Situations are also different, since in some contexts different writing systems are actively in use, whereas at times they represent different historical periods of orthography development. One approach that has been designed for some endangered languages of Russia is to develop separate transliteration conventions between different writing systems, to the extent that is possible (Bradley & Skribnik, 2021).

One challenge we also identify is that the concept of a low-resource language is often used in a very inexact manner, as discussed further by Hämäläinen (2021). Any language besides English can in some situation be called a low-resource language, which makes the category difficult to use, and the concept less practical. Still, there are important differences between languages and the existing resources for them. This governs the starting point for further work, which makes it important to be able to contextualise up to some degree. Building new lexical resources is an entirely different undertaking when other bi- or multilingual lexicons already exist, even though they would differ in various ways from a new resource currently planned. In a study by Nasution et al. (2018) existing bilingual dictionaries in individual languages were used to create new resources for different language pairs. Even in this case, some of the languages were significantly smaller than the majority languages, which were also included in the original dataset.

Our method relies heavily on an existing morphological analyser. Such tools are not available for all languages, but the number of languages with at least some degree of coverage is not small, even if we look into individual infrastructures such as GiellaLT[2], or a Python package that can access these and other analysers, described by Hämäläinen (2019). At the same integrating the development work of a morphological analyser into the whole language documentation work and dictionary creation is not unprecedented either. Pirinen 2019 has reported in detail his parallel work on Karelian treebanks,

[2] https://github.com/giellalt/

dictionaries and computational grammar. As similar approach where a morphological analyser supports language documentation work is reported in Gerstenberger et al. (2017), although this did not include a more specific discussion about lexicographic work, which is still connected to the creation of an analyser on at least the lemma level. Wilbur (2017) developed a workflow for Pite Sami where lexicographic data is stored in a database and connected to the morphological analysis, which provides a strong parallel to our work.

Lexonomy (Měchura et al., 2017) is a good all-purpose online tool for dictionary editing. However, it is not sufficient for our needs. The main reason is that our aim is to have the system built in such a fashion that it can be directly used with the existing tools for Uralic languages (XML dictionary conventions, FST morphology and so on) (see Pirinen & Tyers 2021). We also need to provide an interface for users who are not familiar with the technology, and even the mere fact of having the XML structure visible in an advanced view might startle them.

We must also emphasise that often the endangered languages with limited resources do not have a native speaker base who could participate in the lexicographical work. This also calls for very customised and specialised solutions in each situation. We see, however, that there are some general characteristics and demands upon which the specialised versions can be constructed, instead of designing everything from scratch.

3. Our Online Editor

In this section, we describe our online dictionary editor. It is fully open source[3] and based on technologies such as Django[4] and the MariaDB database[5]. One of the key design goals of the editor has been building it on top of Giella's (Moshagen et al., 2014) reusable components, this means that the system can input and output Giella formatted XML dictionaries and use the NLP tools provided in the infrastructure.

3.1 User Interface

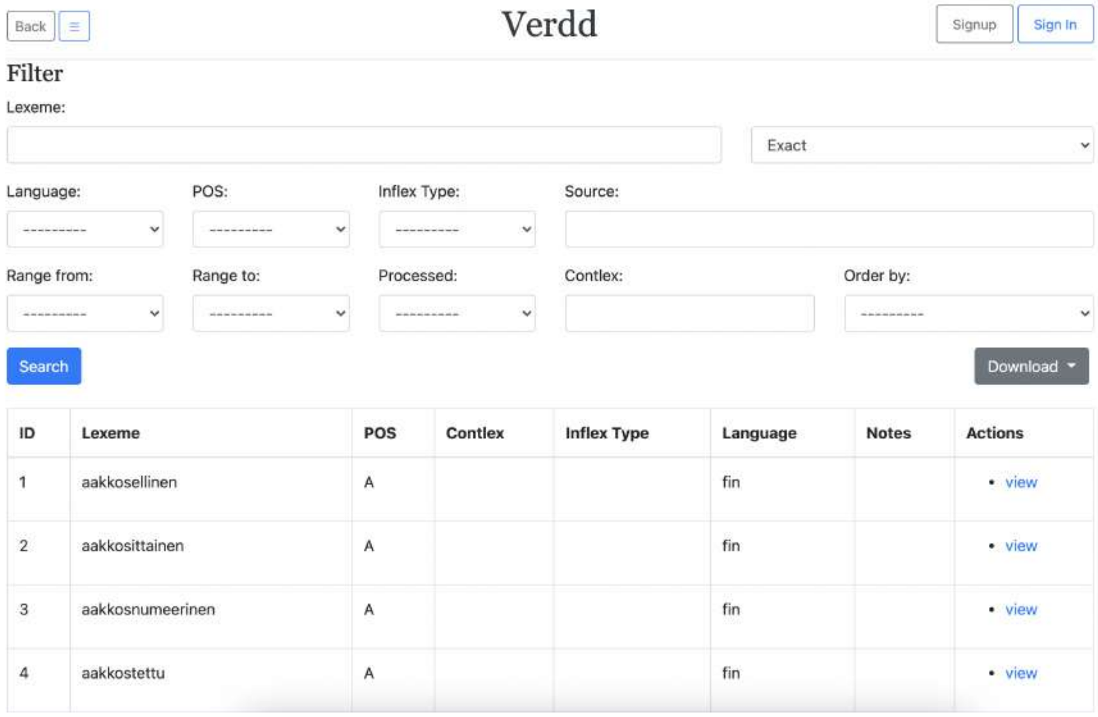
Our online system is bundled with numerous features and commands to facilitate searching, editing and producing dictionaries. These features include, but are not limited to, importing and exporting dictionaries from Giella's XMLs and CSVs, merging and cleaning lexemes, searching and approving entries in the dictionary, and generating a printable dictionary in LaTeX. In this section, we show a glimpse of the user-interface.

Figure 1 displays the homepage of the system where users can perform simple and complex search queries to find lexemes and interesting patterns. Simple filtering involves matching lexemes that either contain, start or end with, or have an exact match with the input query, whereas complex filtering can be conducted with the help of regular expressions (e.g., matching lexemes following a given pattern such as starting with "v" and ending with "ed"). Further filtering, for instance based on the part-of-speech, language, the source of the lexeme and/or whether it has been checked by an expert in the language, can be applied to retrieve relevant lexemes promptly.

[3] A GitHub link will be provided in the camera ready version

[4] https://www.djangoproject.com/

[5] https://mariadb.org/



ID	Lexeme	POS	Contlex	Inflex Type	Language	Notes	Actions
1	aakkosellinen	A			fin		• view
2	aakkosittainen	A			fin		• view
3	aakkosnumeerinen	A			fin		• view
4	aakkostettu	A			fin		• view

Figure 1: The user-interface for searching for lexemes in Ve′rdd.

When a user navigates to a given lexeme, all the information regarding the lexeme along with all relations to and from it are returned. An example of what is supplied to the user when visiting a lexeme is given in Figure 2. In this example, the lexeme is “ve′rdd”. In addition to the core information of the lexeme (e.g., its language, POS and notes), our online system utilises FSTs (Finite-State Transducers) dedicated to the language to produce mini- and full- paradigms of the language. The user has the ability to override any automatically generated paradigms or even introduce new ones, which would serve as a feedback interface for improving the state of the FST. At the end of the lexeme page, all of its relations, e.g., derivations and translations, are shown, along with any examples and metadata which might be present for each relation.

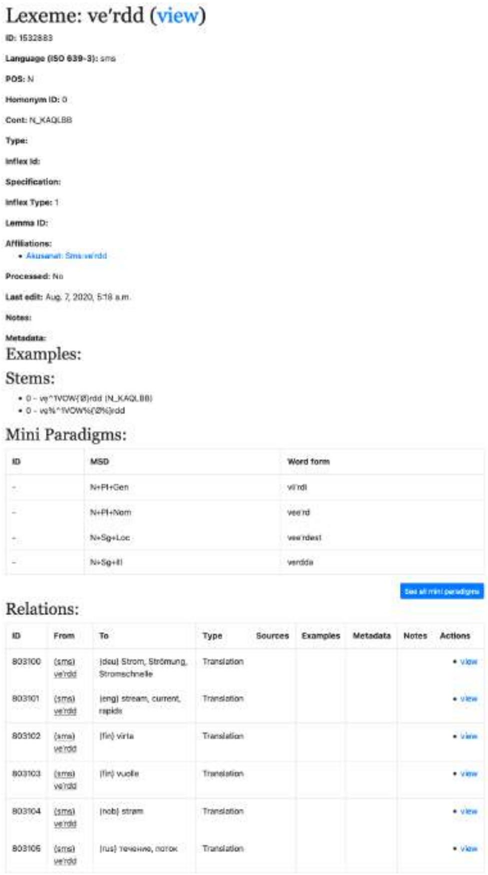
3.2 The data structure

In our system, the basic unit is a lexeme. A lexeme is just a word consisting of its lemma, part-of-speech and other metadata. If there are two words, the lemmas of which are homonyms, they will be two separate lexemes in the system with distinctive homonym IDs. *Sokk* is an example of such a case. It can be a word for *a family* or *a sock*, but it is inflected differently depending on which one of the homonyms is in question.

Lexemes are linked to each other with relations. These can be virtually anything, but in practice we have translation, derivation, compound and etymological relations. Relations can be uni- or bidirectional.

3.3 Importing and Exporting Data

Since the very beginning of the Skolt Sami dictionary project, it was evident that the system needed to support multiple different input formats. On the one hand, the original



Lexeme: ve′rdd (view)

ID: 1532883

Language (ISO 639-3): sms

POS: N

Homonym ID: 0

Cont: N_KAQLBB

Type:

Inflex Id:

Specification:

Inflex Type: 1

Lemma ID:

Affiliations:

- Akusanat: Sms:ve′rdd

Processed: No

Last edit: Aug. 7, 2020, 5:18 a.m.

Notes:

Metadata:

Examples:

Stems:

- 0 - ve^1VOW{Ø}rdd (N_KAQLBB)
- 0 - ve%^1VOW%{Ø%}rdd

Mini Paradigms:

ID	MSD	Word form
-	N+Pl+Gen	vi′rdi
-	N+Pl+Nom	vee′rd
-	N+Sg+Loc	vee′rdest
-	N+Sg+Ill	verdda

Relations:

ID	From	To	Type	Sources	Examples	Metadata	Notes	Actions
803100	(sms) ve′rdd	(deu) Strom, Strömung, Stromschnelle	Translation					• view
803101	(sms) ve′rdd	(eng) stream, current, rapids	Translation					• view
803102	(sms) ve′rdd	(fin) virta	Translation					• view
803103	(sms) ve′rdd	(fin) vuolle	Translation					• view
803104	(sms) ve′rdd	(nob) strøm	Translation					• view
803105	(sms) ve′rdd	(rus) течение, поток	Translation					• view

Figure 2: Information displayed to the user when accessing a lexeme, “ve′rdd” in this case.

material of the first Finnish-Skolt Sami dictionary (Sammallahti & Mosnikoff, 1991), which was stored in a CSV format, needed to be imported, on the other hand, we needed to import the latest advances in the Giella XML-based Skolt Sami dictionary[6].

The first issue was the inconsistent characters that were used; the recent XML dictionaries only consisted of correct characters without an extended vocabulary, while the older CSV material had many different wrong encodings. For example, Skolt Sami uses the modifier letter prime character in its orthography in a word such as *ve′rdd* (stream), however words containing this character were often written with a single quote *ve’rdd* or as an accent *ve´rdd*. For this reason, we implemented a feature in our system that takes in a list of accepted characters in the language one is importing and shows an error if an unaccepted character is being imported. The system also takes in a conversion map that it uses to resolve erroneous characters automatically.

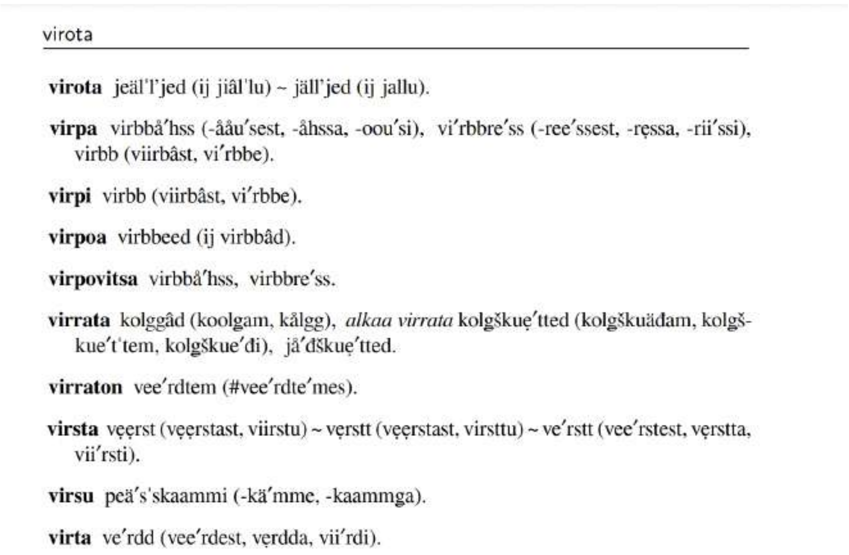
When the data was imported, we needed to support several output formats. First and foremost, Giella XML. This format is needed because several tools such as spell checkers and online language learning tools use dictionaries in this format. This means that this output format makes it possible for us to upload changes made in our system to the Giella infrastructure to benefit the higher level tools of the infrastructure.

Other output formats needed were CSV format as some lexicographers found it easier to work on that format as well, and most importantly LaTeX for producing the final

[6] https://gtsvn.uit.no/langtech/trunk/words/dicts/sms2X/

printable dictionary. The LaTeX output is generated with Django's template language[7], this means that customising the output dictionary does not require modifications to the program logic of our system, merely edits in the template file.

The interface allows downloading a printable dictionary edition in LaTeX format. Figure 3 shows part of a page of the printable dictionary that is automatically generated by our system. Our LaTeX template takes care of all the essential printed-dictionary formatting requirements, such as dividing the dictionary into alphabetised chapters, adding page headers containing guiding words and allowing single- or double- column dictionaries. The PDF output of the printable dictionary is searchable using any PDF reader, which permits distributing two versions of the dictionary: 1) an electronic version that is properly built and indexed, and 2) a physical dictionary.



virota

virota jeäl'l'jed (ij jiâl'lu) ~ jäll'jed (ij jallu).

virpa virbbå′hss (-ååu′sest, -âhssa, -oou′si), vi′rbbre′ss (-ree′ssest, -rẹssa, -rii′ssi), virbb (viirbâst, vi′rbbe).

virpi virbb (viirbâst, vi′rbbe).

virpoa virbbeed (ij virbbâd).

virpovitsa virbbå′hss, virbbre′ss.

virrata kolggâd (koolgam, kålgg), *alkaa virrata* kolgškuẹ′tted (kolgškuäđam, kolgškue′t'tem, kolgškue′đi), jå′dškuẹ′tted.

virraton vee′rdtem (#vee′rdte′mes).

virsta vẹẹrst (vẹẹrstast, viirstu) ~ vẹrstt (vẹẹrstast, virsttu) ~ ve′rstt (vee′rstest, vẹrstta, vii′rsti).

virsu peä′s'skaammi (-kä′mme, -kaammga).

virta ve′rdd (vee′rdest, vẹrdda, vii′rdi).

Figure 3: A snapshot of a page in the automatically produced printable dictionary.

3.4 Integration with NLP Tools

Our system uses FSTs (Finite-State Transducers) based on a tool called HFST by Lindén et al. (2013). These are useful as they produce morphological readings for word forms and they can be used to generate inflectional forms based on a lemma and morphological tags. We use these FSTs for two purposes: inflection and relations.

When a new word is input into the system, the first thing the system does is that it consults the FST and sees if this word is a derivational form of another word or if the new word is a compound formed of existing words. The system will then suggest to the person editing the dictionary that derivational and compound relations can be added automatically. All the editor needs to do is to either confirm or reject the automatically produced relations.

An important part of a dictionary of any morphologically rich language is the presence of certain inflectional forms in the lexicographic entry as, based on them, the user can

[7] https://docs.djangoproject.com/en/3.1/ref/templates/language/

know the full inflectional paradigm (see Hulden & Silfverberg 2021). We generate these inflectional forms automatically in our system for all input words. These can be inspected under the miniparadigm field. The dictionary editor can override these automatically generated inflectional forms by editing them. This also serves as feedback for the people editing the FSTs so that they can correct any mistakes in their output.

We are currently integrating our latest graph and deep learning based methods (Hämäläinen et al., 2021) into our system. We have been able to automatically predict new translations for the Giella dictionaries based on XML dictionaries in other languages and Wikitionaries in large languages. In short, for a lemma that has translations into at least two other languages, our method can predict more suitable synonymous translations for the two languages and translation candidates with the same meaning in other languages, with the idea that the more languages our system covers, the more nuanced its understanding of polysemy becomes.

4. Discussion and Future Directions

In the future, the dictionary editing platform has to be tested with different languages and editorial teams. This is necessary so that we understand what kinds of workflows serve different communities best, and which of the current design choices can be improved upon. At the same time, more work is needed with different dictionary search and visualisation platforms, which can be catered also to the needs of specific user groups. One of the strengths of the current implementation is that we have a large amount of lexical data from different languages in the same infrastructure, and the new work is not disconnected from earlier efforts, but instead builds upon it. However, from the user perspective it is probably necessary to differentiate language and target group specific exports and views.

In building new systems, one has to always remember the importance of the longevity of the data. We recently got an important reminder of this as the servers of our service provider caught fire[8]. We take regular backups of the data of our system both as SQL dumps and as Giella XMLs. Backing the data up in the Giella XML format comes with the additional benefit of it being convertible into the ISO standardised TEI format (Rueter & Hämäläinen, 2019), which ensures that the lexicographic data remains readable even in the distant future.

We will also consider which is the best option for digital preservation of this work in some larger and more persistent infrastructure. Zenodo[9] is one obvious option to store versioned exports as well, and the exports that relate to individual published dictionaries should be stored also digitally with particular care, so that it is always possible to go back into individual versions. This is needed, for example, to quantify the changes between different dictionary editions. These questions are also strongly related to the dictionary editing workflows of the individual teams, although we believe that periodic publications and later improved editions is a model that remains relevant for many dictionary creators. With the online platforms, naturally, the question of release based updates and continuous updates also becomes important, and may vary from situation to situation.

The system has already been used by other researchers (Koponen & Kuokkala, 2021) to study Skolt Sami word derivation. This shows that the data stored in our dictionary

[8] https://web.archive.org/web/20210310232354/ https://www.ovh.com/world/news/press/cpl1787.fire-our-strasbourg-si

[9] https://zenodo.org/

system is also accessible for other researchers and it can be a useful resource in linguistic research. However, this has not been taken into account as a possible use case when developing the system. In the future, it would be important to conduct user studies in order to better understand the needs of linguistics researchers to better support their use cases.

The most important feature that needs to be further improved and adapted is the dictionary editing workflow that the user interacts with. It is especially important that this is done in close collaboration with the system users. This calls for identification of different usage patterns that the users have, including the documentation of various steps in the usage.

Adding a new lexical entry, editing relations, adding example sentences and searching for related entries are all tasks the dictionary editor will do continuously, and the interface should allow focused work where there are minimal interruptions and pauses caused by the underlying system. Ideally the information about usage bottlenecks would be collected by observing and tracking the real user actions in the interface, with their permission, and having continuous discussions about their experiences. However, it is particularly important to be able to distinguish the true obstacles in the editing platform, and issues that are related to insufficient training and documentation: a complex expert system will inevitably have some learning curve. From this point of view it is also important to distinguish the issues novice and expert users have, and to understand the process through which the novices become fully competent expert users.

Most of the dictionaries contain a large number of example sentences for each entry and meaning groups. Some of these are created by the dictionary editors, and some originate from various sources. The sources are in all cases important to indicate. When possible, the example sentences used in the dictionaries should be linked into different corpora and related datasets, both for accountability and the possibility to further provide access into them. This also makes it clear which materials, created by who, are actually used in the dictionaries, which makes citation of all sources used easier and benefits the visibility of previously done work in our scientific community. At the same time linked data also becomes more difficult to maintain when we cannot guarantee that all linked sources remain as accessible as our system.

Another area where similar connections could be created is multimedia. There are numerous spoken language corpora, some of which are openly licensed, and using their materials in connection with the dictionary resources would be an excellent addition, since our system doesn't currently have pronunciation information. It could be possible to add this information also in IPA or other transcription system, but in this day and age actual multimedia references seem very realistic and even expected.

5. Conclusions

In this paper, we have presented our open-source dictionary editing system that was developed for post-editing the new printed Finnish-Skolt Sami dictionary. We have described the system and how it interacts with the existing open-source language technology infrastructure called Giella. By releasing our source code openly on GitHub, we hope that other people can make use of our system to meet their dictionary editing needs.

We have developed the system taking into account the latest NLP tools available for Skolt Sami. This has made the dictionary editing process easier as automatically introduced information such as inflectional forms, derivations and compounds would have taken a great deal of time to annotate manually. The fact that our system makes it possible for the dictionary editors to fix errors in the automatically generated inflectional forms also benefits the development of the NLP tools used. Finally, we aimed at building a system that not only serves in producing a paper dictionary, but forces the editors to edit the lexicographic entries in such a way that they remain structured and parseable by computational means. This meant that the final dictionary was also easy to be made available online[10] in a searchable fashion.

6. References

Alnajjar, K., Hämäläinen, M., Rueter, J. & Partanen, N. (2020). Ve'rdd. Narrowing the Gap between Paper Dictionaries, Low-Resource NLP and Community Involvement. In *Proceedings of the 28th International Conference on Computational Linguistics: System Demonstrations*. Barcelona, Spain (Online): International Committee on Computational Linguistics (ICCL), pp. 1–6. URL https://www.aclweb.org/anthology/2020.coling-demos.1.

Beesley, K.R. & Karttunen, L. (2003). *Finite-State Morphology*. Stanford, CA: CSLI Publications, pp. 451–454.

Bradley, J. & Skribnik, E. (2021). The many writing systems of Mansi: challenges in transcription and transliteration. In M. Hämäläinen, N. Partanen & K. Alnajjar (eds.) *Multilingual Facilitation*. Rootroo Ltd.

Gerstenberger, C., Partanen, N. & Rießler, M. (2017). Instant annotations in ELAN corpora of spoken and written Komi, an endangered language of the Barents Sea region. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pp. 57–66.

Hämäläinen, M. (2021). Endangered Languages are not Low-Resourced! In M. Hämäläinen, N. Partanen & K. Alnajjar (eds.) *Multilingual Facilitation*. RootRoo Ltd, pp. 1–11.

Hulden, M. & Silfverberg, M. (2021). The Principal Parts of Finnish Nominals. In M. Hämäläinen, N. Partanen & K. Alnajjar (eds.) *Multilingual Facilitation*. Rootroo Ltd.

Hunt, B., Chen, E., Schreiner, S.L. & Schwartz, L. (2019). Community lexical access for an endangered polysynthetic language: An electronic dictionary for St. Lawrence Island Yupik. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 122–126. URL https://www.aclweb.org/anthology/N19-4021.

Hämäläinen, M. (2019). UralicNLP: An NLP Library for Uralic Languages. *Journal of Open Source Software*, 4(37), p. 1345.

Hämäläinen, M., Partanen, N., Rueter, J. & Alnajjar, K. (2021). Neural Morphology Dataset and Models for Multiple Languages, from the Large to the Endangered. In *Proceedings of the the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa 2021)*.

[10] https://saan.oahpa.no/fin/sms/

Hämäläinen, M. & Rueter, J. (2018). Advances in Synchronized XML-MediaWiki Dictionary Development in the Context of Endangered Uralic Languages. In *Proceedings of the Eighteenth EURALEX International Congress*, pp. 967–978.

Koponen, E. & Kuokkala, J. (2021). Kantasaamen *-(e̮)hče̮-frekventatiivijohtimen edustuksesta nykyisissä saamelaiskielissä. In M. Hämäläinen, N. Partanen & K. Alnajjar (eds.) *Multilingual Facilitation*. Rootroo Ltd.

Lehtinen, M., Koponen, E., Fofonoff, M., Lehtola, R. & Rueter, J. (eds.) (2021). *Suomi–koltansaame-sanakirja Lääʹdd-sääʹm-sääʹnnǩeeʹrjj*. Saamelaiskäräjät.

Lindén, K., Axelson, E., Drobac, S., Hardwick, S., Kuokkala, J., Niemi, J., Pirinen, T.A. & Silfverberg, M. (2013). HFST a system for creating NLP tools. In *International Workshop on Systems and Frameworks for Computational Morphology*. Springer, pp. 53–71.

Littell, P., Pine, A. & Davis, H. (2017). Waldayu and Waldayu Mobile: Modern digital dictionary interfaces for endangered languages. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*. Honolulu: Association for Computational Linguistics, pp. 141–150. URL https://www.aclweb.org/anthology/W17-0119.

Mechura, M. (2016). Data Structures in Lexicography: from Trees to Graphs. *RASLAN 2016 Recent Advances in Slavonic Natural Language Processing*, p. 97.

Měchura, M.B. et al. (2017). Introducing Lexonomy: an open-source dictionary writing and publishing system. In *Electronic Lexicography in the 21st Century: Lexicography from Scratch. Proceedings of the eLex 2017 conference*. pp. 19–21.

Moseley, C. (ed.) (2010). *Atlas of the World’s Languages in Danger*. UNESCO Publishing, 3rd edition. Online version: http://www.unesco.org/languages-atlas/.

Moshagen, S., Rueter, J., Pirinen, T., Trosterud, T. & Tyers, F.M. (2014). Open-Source Infrastructures for Collaborative Work on Under-Resourced Languages. In *The LREC 2014 Workshop “CCURL 2014 - Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era”*. pp. 71–77.

Nasution, A.H., Murakami, Y. & Ishida, T. (2018). Designing a Collaborative Process to Create Bilingual Dictionaries of Indonesian Ethnic Languages. In *Proceedings of the 11th Language Resources and Evaluation Conference*. Miyazaki, Japan: European Language Resource Association. URL https://www.aclweb.org/anthology/L18-1536.

Pirinen, T. & Tyers, F. (2021). Building language technology infrastructures to support a collaborative approach to language resource building. In M. Hämäläinen, N. Partanen & K. Alnajjar (eds.) *Multilingual Facilitation*. Rootroo Ltd.

Pirinen, T.A. (2019). Building minority dependency treebanks, dictionaries and computational grammars at the same time—an experiment in Karelian treebanking. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*. pp. 132–136.

Rueter, J. & Hämäläinen, M. (2017). Synchronized Mediawiki Based Analyzer Dictionary Development. In *Proceedings of the Third Workshop on Computational Linguistics for Uralic Languages*. pp. 1–7.

Rueter, J. & Hämäläinen, M. (2019). On XML-MediaWiki Resources, Endangered Languages and TEI Compatibility, Multilingual Dictionaries For Endangered Languages. In M. Gürlek, A. Çiçekler & Y. Taşdemir (eds.) *AsiaLex 2019*. Turkey: Asos Publisher.

Rueter, J. & Hämäläinen, M. (2020). FST Morphology for the Endangered Skolt Sami Language. In *Proceedings of the 1st Joint SLTU and CCURL Workshop (SLTU-CCURL 2020)*. European Language Resources Association (ELRA).

Sammallahti, P. & Mosnikoff, J. (1991). *Suomi–koltansaame sanakirja.* GIRJEGIISA.

Wilbur, J. (2017). The Pite Saami lexicographic backbone From a FileMaker Pro database to published digital results. In Электронная письменность народов российской федерации: опыт, проблемы и перспективы, pp. 299–309.

eLex 2021

ORGANIZERS

LEXICAL COMPUTING

cjvt

/instituut voor de Nederlandse taal/

Univerza *v Ljubljani*

Institute of the Estonian Language

elexis european lexicographic infrastructure

eLex

elex.link/elex2021