



Electronic lexicography in the 21st century (eLex 2023)

Proceedings of the eLex 2023 conference

edited by

Marek Medved'
Michal Měchura
Carole Tiberius
Iztok Kosem
Jelena Kallas
Miloš Jakubíček
Simon Krek

27-29 June 2023

elex.link/elex2023



Edited by

Marek Medved'
Michal Měchura
Carole Tiberius
Iztok Kosem
Jelena Kallas
Miloš Jakubiček
Simon Krek

Published by

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

License

Creative Commons Attribution
ShareAlike 4.0 International License

ISSN: 2533-5626

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



ORGANIZERS

‘LEXICAL,
COMPUTING’

Univerza v Ljubljani

cjvt Centre for
Language Resources
and Technologies



; EESTI
KEELE
INSTITUUT

/instituut
voor de
Nederlandse
taal/

SPONSORS



Organizing Committee

Miloš Jakubiček
Jelena Kallas
Iztok Kosem
Simon Krek
Carole Tiberius

Ondřej Matuška
Tereza Olšanová
Michal Cukr
Vlasta OhlídaloVá

Scientific committee

Aleksandra Marković
Aleš Horák
Alexander Geyken
Amália Mendes
Andrea Abel
Arvi Tavast
Anila Çepani
Annette Klosa-Kückelhaus
Carole Tiberius
Edward Finegan
Egon Stemle
Emma Sköldbberg
Henrik Lorentzen
Hindrik Sijens
Ilan Kemerman
Iztok Kosem
Jelena Kallas
Kris Heylen
Kristian Blensenius
Kristina Koppel

Kristina Strkalj Despot
Lars Trap-Jensen
Laurent Romary
Lionel Nicolas
Lothar Lemnitzer
Lut Colman
Maarten Janssen
Margit Langemets
María José Domínguez Vázquez
Michal Kren
Michal Měchura
Miloš Jakubiček
Mojca Kompara Lukančič
Nicolai Hartvig Sørensen
Patrick Drouin
Paul Cook
Pilar León Araúz
Polona Gantar
Radovan Garabik
Ranka Stanković

Rizki Gayatri
Robert Lew
Said Zohairy
Sara Moze
Simon Krek
Stella Markantonatou
Sussi Olsen
Svetla Koeva
Špela Arhar Holdt
Tamás Varadi
Tanara Zingano Kuhn
Thierry Fontenelle
Vincent Ooi
Vít Suchomel
Vojtěch Kovář
Voula Giouli
Yongwei Gao
Yukio Tono
Zoe Gavriilidou

Contents

Evaluation of the Cross-lingual Embedding Models from the Lexicographic Perspective	1
<i>Michaela Denisová, Pavel Rychlý</i>	
Definition Extraction for Slovene: Patterns, Transformer Classifiers and ChatGPT	19
<i>Hanh Thi Hong Tran, Vid Podpečan, Mateja Jemec Tomazin, Senja Pollak</i>	
Ceci n'est pas un dictionnaire. Adding and Extending Lexicographical Data of Medieval Romance Languages to and through a Multilingual Lexico-Ontological Project	39
<i>Sabine Tittel</i>	
Towards a lexical database of Dutch taboo language	53
<i>Gerhard B van Huyssteen, Carole Tiberius</i>	
The Czechoslovak Word of the Week. Re-joining Czechs and Slovaks together in an example of invisible lexicography work	75
<i>Peter Malčovský, Michal Škrabal, Vladimír Benko, Jan Koček</i>	
The Central Word Register of the Danish language	91
<i>Thomas Widmann</i>	
Invisible meaning relations for representing near equivalents	104
<i>Arvi Tavast, Kristina Koppel, Margit Langemets, Silver Vapper, Madis Jürviste</i>	
Military Feminine Personal Nouns: A Corpus-Based Update to the Web Dictionary of Ukrainian Feminine Personal Nouns	118
<i>Olena Sychak</i>	
Improving second language reading through visual attention cues to corpus-based patterns	141
<i>Kate Challis, Tom Drusa</i>	
An Unsupervised Approach to Characterize the Adjectival Microstructure in a Hungarian Monolingual Explanatory Dictionary	160
<i>Enikő Héja, Noémi Ligeti-Nagy, László Simon, Veronika Lipp</i>	

How Lexicographers Evaluate User Contributions in The Thesaurus of Modern Slovene in Comparison to Dictionary Users	178
<i>Magdalena Gapsa, Špela Arhar Holdt</i>	
Word-sense Induction on Buddhist Sanskrit Corpus	201
<i>Matej Martinc, Andraž Pelicon, Senja Pollak, Ligeia Lugli</i>	
Word sense induction for (French) verb valency discovery	216
<i>Naïma Hassert, François Lareau</i>	
Towards a Comprehensive Dictionary of Middle Persian	230
<i>Francisco Mondaca, Kianoosh Rezania, Slavomír Čéplö, Claes Neuefeind</i>	
The Kosh Suite: APIs for Lexical Data	246
<i>Francisco Mondaca, Philip Schildkamp, Felix Rau, Luke Günther</i>	
Humanitarian reports on ReliefWeb as a domain-specific corpus	258
<i>Loryn Isaacs</i>	
A Federated Search and Retrieval Platform for Lexical Resources in Text+ and CLARIN	280
<i>Thomas Eckart, Axel Herold, Erik Körner, Frank Wiegand</i>	
From Structured Textual Data to Semantic Linked-data for Georgian Verbal Knowledge	293
<i>Archil Elizbarashvili, Mireille Ducassé, Manana Khachidze, Magda Tsintsadze</i>	
A Search Engine for the Large Electronic Dictionary of the Ukrainian Language (VESUM)	308
<i>Tamila Krashtan</i>	
The Use of Lexicographic Resources in Croatian Primary and Secondary Education	322
<i>Ana Ostroški Anić, Daria Lazić, Maja Matijević, Martina Pavić</i>	
Exploring the Capabilities of ChatGPT for Lexicographical Purposes: A Comparison with Oxford Advanced Learner’s Dictionary within the Microstructural Framework	345
<i>Chayanon Phoodai, Richárd Rikk</i>	

Thesaurus of Modern Slovene 2.0	376
<i>Špela Arhar Holdt, Polona Gantar, Iztok Kosem, Eva Pori, Marko Robnik-Šikonja, Simon Krek</i>	
Trawling the corpus for the overlooked lemmas	392
<i>Nathalie Hau Sørensen, Nicolai Hartvig Sørensen, Kirsten Lundholm Appel, Sanni Nimb</i>	
Tēzaurs.lv – the experience of building a multifunctional lexical resource	410
<i>Mikus Grasmanis, Pēteris Paikens, Lauma Pretkalniņa, Laura Rituma, Laine Strankale, Artūrs Znotiņš, Normunds Grūzītis</i>	
Novel Slovenian COVID-19 vocabulary from the perspective of naming possibilities and word formation	429
<i>Senja Pollak, Ines Voršič, Boris Kern, Matej Ulčar</i>	
Automating derivational morphology for Slovenian	449
<i>Tomaž Erjavec, Marko Pranjić, Andraž Pelicon, Boris Kern, Irena Stramljič Breznik, Senja Pollak</i>	
Lexicography for mathematics	466
<i>Theresa Kruse, Ulrich Heid, Boris Girnat</i>	
From experiments to an application: the first prototype of an adjective detector for Estonian	476
<i>Geda Paulsen, Ahti Lohk, Maria Tuulik, Ene Vainik</i>	
Collocations Dictionary of Modern Slovene 2.0	501
<i>Iztok Kosem, Špela Arhar Holdt, Polona Gantar, Simon Krek</i>	
The end of lexicography? Can ChatGPT outperform current tools for post-editing lexicography?	518
<i>Miloš Jakubiček, Michael Rundell</i>	
From a dictionary towards the Hungarian Constructicon	534
<i>Bálint Sass</i>	
Probing visualizations of neural word embeddings for lexicographic use	545
<i>Ágoston Tóth, Esra Abdelzaher</i>	
The <i>SERBOVERB</i> Language Resource and Its Multifunctionality	567
<i>Saša Marjanović</i>	

Operationalising and Representing Conceptual Variation for a Corpus-driven Encyclopaedia	587
<i>Santiago Chambó, Pilar León-Araúz</i>	
Rapid Ukrainian-English Dictionary Creation Using Post-Edited Corpus Data	613
<i>Marek Blahuš, Michal Cukr, Ondřej Herman, Miloš Jakubiček, Vojtěch Kovář, Jan Kraus, Marek Medved, Vlasta Ohlídalová, Vít Suchomel</i>	
Adding Information to Multiword Terms in Wiktionary	638
<i>Thierry Declerck, Lenka Bajčetić, Gilles Sérasset</i>	
Development of Evidence-Based Grammars for Terminology Extraction in OneClick Terms.....	650
<i>Marek Blahuš, Michal Cukr, Miloš Jakubiček, Vojtěch Kovář, Vít Suchomel</i>	
Workshop on Lexicography and CEFR: Linking Lexicographic Resources and Language Proficiency Levels ..	663
Building a CEFR-Labeled Core Vocabulary and Developing a Lexical Resource for Slovenian as a Second and Foreign Language.....	663
<i>Matej Klemen, Špela Arhar Holdt, Senja Pollak, Iztok Kosem, Eva Pori, Polona Gantar, Mihaela Knez</i>	
Author Index	681

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

Michaela Denisová¹, Pavel Rychlý²

¹Natural Language Processing Centre
Faculty of Informatics, Masaryk University
Brno, Czech Republic
E-mail: ¹449884@mail.muni.cz, ²pary@fi.muni.cz

Abstract

Cross-lingual embedding models (CMs) enable us to transfer lexical knowledge across languages. Therefore, they represent a useful approach for retrieving translation equivalents in lexicography. However, these models have been mainly oriented towards the natural language processing (NLP) field, lacking proper evaluation with error evaluation datasets that were compiled automatically. This causes discrepancies between models hindering the correct interpretation of the results. In this paper, we aim to address these issues and make these models more accessible for lexicography by evaluating them from a lexicographic point of view. We evaluate three benchmark CMs on three diverse language pairs: close, distant, and different script languages. Additionally, we propose key parameters that the evaluation dataset should include to meet lexicographic needs, have reproducible results, accurately reflect the performance, and set appropriate parameters during training. Our code and evaluation datasets are publicly available.¹

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

1. Introduction

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

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

In the natural language processing field, finding translation equivalents candidates is referred to as bilingual lexicon induction (BLI) task. In the BLI task, the target language words are induced from shared space through the nearest neighbour search for a source

¹ https://github.com/x-mia/Evaluation_of_CWE

language word. Afterwards, they are run against a gold-standard dictionary to measure the quality of the model (Ruder et al., 2019).

The BLI task is a popular way among researchers to evaluate their models (Artetxe et al., 2016; Conneau et al., 2017; Joulin et al., 2018; Glavaš & Vulić, 2020; Parizi & Cook, 2021; Tian et al., 2022: etc.). However, the evaluation is often inconsistent and differs from paper to paper, using various metrics and gold-standard dictionaries from multiple sources (Ren et al., 2020; Karan et al., 2020; Woller et al., 2021; Severini et al., 2022: etc.). This impedes our ability to correctly interpret the results and make models comparable to each other.

Moreover, many currently used gold-standard dictionaries are generated automatically (Conneau et al., 2017; Glavaš et al., 2019; Vulić et al., 2019: etc.). Therefore, they are prone to contain mistakes. For example, the most widely used gold-standard dictionaries, MUSE (Conneau et al., 2017), are criticised for occurring errors and disproportional part-of-speech distribution (Kementchedjhieva et al., 2019; Denisová & Rychlý, 2021).

On top of that, articles dealing with CMs and the BLI task do not consider the utilisation in the lexicography field. They focus on the computational side of the problem and simple word-to-word extraction without reflecting on various aspects of translation and tailoring the evaluation process and gold-standard dictionaries to the lexicographers' needs.

In this paper, we investigate various aspects that influence the training of the CMs. We propose the most suitable parameters for the evaluation dataset based on these aspects while allowing for a lexicography perspective. We show that having a strong evaluation dataset and a clear evaluation process is crucial for setting appropriate training parameters. We assess and discuss the quality of the most common benchmark models on a distant language pair, Estonian-Slovak, a close language pair, Czech-Slovak, and language pair that do not share a script, English-Korean.

Our motivation is to determine important aspects when evaluating CMs on the BLI task and construct a reliable, high-quality evaluation dataset that addresses the above-stated issues. Our contribution is manifold:

1. We set crucial parameters of the evaluation dataset for the BLI task that are reproducible for further research and unifying for the evaluation process.
2. We provide an evaluation of three diverse language pairs on the most cited benchmark CMs in various settings.
3. We link the NLP and lexicography sides of the CMs' evaluation.

This paper is structured as follows. In Section 2, we outline the background information and the benchmark models used for the evaluation. In Section 3, we describe the experimental setup used in training CMs. In Section 4, we list important aspects of the evaluation and provide reasoning for each of them. In Section 5, we offer concluding remarks.

2. Background

In this paper, we train and evaluate three methods: MUSE (Conneau et al., 2017; Lample et al., 2017: two equal articles), VecMap (Artetxe et al., 2016, 2017, 2018a,b), and FastText

for multilingual alignment (Joulin et al., 2018). These methods are frequently cited and utilised as benchmarks.

MUSE. This method connects domain-adversarial training with iterative Procrustes alignment. Moreover, it proposes a novel method for matching translation equivalents candidates, cross-domain similarity local scaling (CSLS). MUSE involves supervised and unsupervised training and training that relies on identical strings. Their code, pre-trained multilingual word embeddings and datasets used for training and evaluation are available on their GitHub repository.² Evaluation datasets were made automatically for 110 languages.

VecMap. VecMap is a robust self-learning framework with multiple steps and iterative learning depending on the setting. It can be trained in a supervised, semi-supervised, and unsupervised manner or uses identical strings as supervision signals. Similarly to the MUSE framework, it has an open-source GitHub repository.³

FastText. The FastText method proposes to optimise the CSLS retrieval criterion used in the MUSE framework. This method provides a supervised setting for training. Their pre-trained aligned models are freely available⁴, and their code is published on the GitHub repository.⁵

3. Experimental Setup

CMs require comparable corpora for training. In this case, comparable means non-aligned and similar in size and text genres (Kovář et al., 2016). The comparable corpora are used to train monolingual word embeddings (MEs) incorporated in CM training (Ruder et al., 2019).

In this paper, we experimented with two types of MEs. We used pre-trained FastText MEs (Bojanowski et al., 2017)⁶ for Estonian, Slovak, Czech, English, and Korean, which were trained on Wikipedia⁷ with dimension 300. The second pre-trained MEs were provided by SketchEngine (Herman, 2021).⁸ These embeddings were trained on web corpora using the same method as FastText, with dimensions 100 for Estonian-Slovak and English-Korean, and 300 for Czech-Slovak.

Additionally, the training involves a different level of supervision: supervised, identical-string-relying, or unsupervised (Ruder et al., 2019). In this paper, for MUSE and VecMap, we set supervised and unsupervised settings and mode that relies on identical strings. For FastText, we selected supervised training only.

Methods trained in supervised mode require a word-to-word dataset called a seed lexicon. Word-to-word means one single-word unit to one or multiple single-word units. The size of the seed lexicon usually varies up to 5K word pairs. Exceeding this limit does not influence the resulting quality (Vulić & Korhonen, 2016). In identical-string-relying mode, the seed lexicon consists of identical strings and numerals that occur in MEs of both languages.

² <https://github.com/facebookresearch/MUSE>

³ <https://github.com/artetxem/vecmap>

⁴ <https://fasttext.cc/docs/en/aligned-vectors.html>

⁵ <https://github.com/facebookresearch/fastText/tree/master/alignment/>

⁶ <https://fasttext.cc/>

⁷ <https://www.wikipedia.org/>

⁸ <https://embeddings.sketchengine.eu/>

Seed lexicons were from various resources. We used the Estonian-Slovak database that Denisová (2021) constructed for the Estonian-Slovak language combination. We selected 5,000 word-to-word translation equivalents from this database, where the source and target language word occurred in the first 300K words of ME files. As this database’s accuracy is only 40%, we manually post-processed selected translation equivalents.

Czech and Slovak are very close languages containing a lot of identical words. Therefore, we matched 5K word-to-word identically spelt translation equivalents which occurred in the first 300K words of ME files. Lastly, we used the MUSE English-Korean training dataset provided by Conneau et al. (2017) as a seed lexicon for the last language pair, English-Korean.

The last crucial parameter in the training setup is the number of word embeddings loaded during training. This parameter influences the vocabulary coverage of the resulting aligned translation equivalents candidates and, therefore, the vocabulary selection for the evaluation dataset. We reason this in Section 4.1.

Among researchers, the standard is to load the first 200K embeddings. In this paper, we experimented with different numbers of loaded embeddings, which we describe in Section 4.1 in further detail.

Additionally, when assessing the models in the evaluation process, we utilise two metrics, i.e., precision and recall. Precision at k ($P@k$) is the proportion of the number of correct translation equivalents to the number of all extracted translation equivalents’ candidates, where k is the amount of extracted target language words for each source language word (Kementchedjheva et al., 2019). In this paper, the most common is $P@10$, meaning we extract ten target language words for each source language word from the evaluation dataset.

The recall is the proportion of the correct translation equivalents found to the number of all translation equivalents from the evaluation dataset. In this article, we focus mainly on computing recall since this is a more important metric in lexicography. Therefore, the most common number for the induced target language words is ten. However, when assessing the models for language learners, precision is preferred.

4. Parameters of the Evaluation Dictionary

In this section, we investigate which factors significantly influence the evaluation and training processes. Each subsection discusses different aspects.

4.1 Vocabulary

MEs used in training significantly influence the resulting quality of the aligned cross-lingual spaces (Artetxe et al., 2018c; Vulić et al., 2020). From the vocabulary perspective, the MEs impact the nature of the words that embeddings contain and the size of the resulting dictionary.

These two factors depend on the domain (Søgaard et al., 2018) and the size of the monolingual corpus that MEs have been trained on. Since we do not assess the quality of

the MEs, we are restricted to the words they contain. Therefore, we should include only these words in the evaluation dataset to avoid out-of-the-vocabulary (OOV) words, i.e., words that do not occur in the MEs.

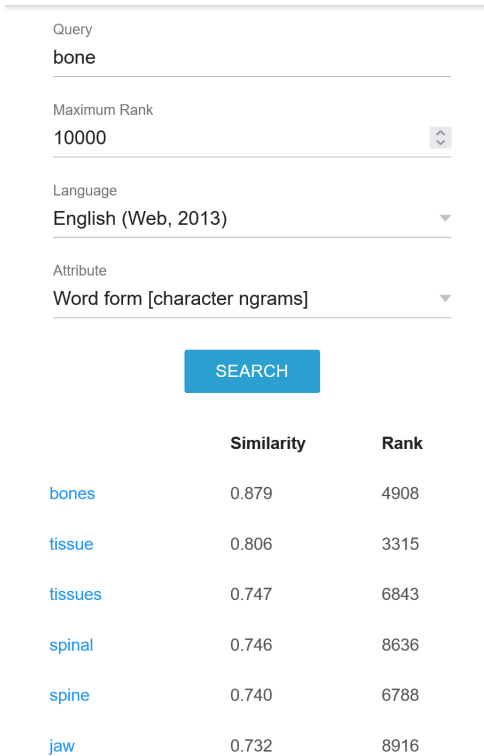


Figure 1: Search for the word *bone* with the SketchEngine tool for monolingual word embeddings with a word rank of 10,000.

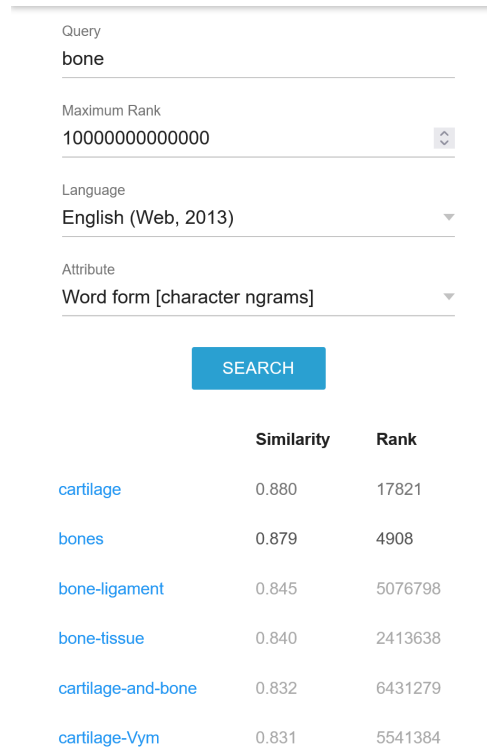


Figure 2: Search for the word *bone* with the SketchEngine tool for monolingual word embeddings with a word rank of 10,000,000,000,000.

OOV words play a significant role during the evaluation process. They are words that do not occur in the shared cross-lingual word embedding space or the MEs. The reasons behind this are various: they are excluded because they had very low or zero occurrences in the monolingual corpus, the MEs contain only the most frequent words, they consist of multiple words (multi-word expressions), or the CM loaded and aligned only a certain amount of the MEs during the training.

Importantly, the MEs and cross-lingual word embedding models do not handle multi-word expressions and words that do not have a one-word equivalent in the target language (e.g., German word *Grundschule* - *primary school, elementary school*). Therefore, we should not include such words in the evaluation dataset.

However, setting the number of loaded embeddings parameter allows us to increase vocabulary coverage for the evaluation dataset and better reflect the resulting quality of the model. Specifically, we have various numbers of the words in the MEs we utilised for training. FastText contains 329,987 Estonian words, 316,098 Slovak words, 627,841 Czech words, 2,519,370 English words, and 879,129 Korean words. For SketchEngine, the Estonian model has 3,307,785 words, Slovak 1,611,402 words, Czech 3,900,455 words, English 6,658,558 words, and Korean 2,949,340 words.

Naturally, a more extensive corpus produces more words, and more words mean greater coverage. However, our goal is not to have as many words as possible at the expense of the quality of the aligned word embeddings. The disadvantage of training with such huge monolingual embeddings is that it is computationally expensive and time-consuming. Moreover, including less frequent words (words with higher rank) does not necessarily mean better results when extracting translation equivalent candidates based on their cosine similarity. Using the word *bone* as an example, Fig. 1 and 2 show that we get more relevant searches if we limit the word rank to a smaller number.

For demonstration purposes and to be able to compare MEs with each other, we constructed the evaluation datasets for Estonian-Slovak and Czech-Slovak that include words occurring in the MEs and OOV words together. The Estonian-Slovak evaluation dataset was compiled using the Estonian-Slovak dictionary from Denisová (2021), similarly to the seed lexicon’s compilation. The evaluation dataset for Czech-Slovak was constructed manually using exclusively words that are different in both languages (e.g., *želva - korytnačka, turtle*). Notably, the evaluation datasets need to differ from the seed lexicons.

For English-Korean, we used the open-source evaluation dataset MUSE (Conneau et al., 2017), which includes only words occurring in the MEs. While aware of this dataset’s drawbacks (Kementchedjheva et al., 2019; Denisová & Rychlý, 2021), we chose it intentionally to help us demonstrate the crucial parameters of the evaluation dataset. Each evaluation dataset contains 1,500 headwords.

We trained the models using the default (or standard) loaded embedding parameter in this experiment. Afterwards, we adjusted it to be optimal considering computational time, the resulting quality and vocabulary coverage. The recall for default and adjusted training is displayed in Table 1.

Given Table 1, each model improved recall for Estonian-Slovak and Czech-Slovak by 10-20% if we increase the number of loaded embeddings from 50K to 300-400K. Generally, the SketchEngine embeddings for Estonian-Slovak appeared to perform worse than FastText when 50K embeddings were loaded. However, after adjusting the loaded embeddings’ parameter, their recall increased, surpassing the models trained with FastText embeddings.

Although the English-Korean evaluation dataset contained words from the first 50K loaded embeddings, the recall for the models trained with FastText embeddings decreased in most cases. It shows that enlarging the number of loaded embeddings in this particular scenario can have a negative impact on recall. As mentioned above, increasing the word rank can include more noise from the MEs and, thus, lower the resulting quality. This is the indicator of the quality of the MEs, not the CMs.

Additionally, the SketchEngine embeddings outperformed FastText embeddings in the majority of cases, except for English-Korean, where FastText embeddings were significantly better. This could be due to the uneven part-of-speech distribution (Kementchedjheva et al., 2019; Denisová & Rychlý, 2021). Therefore, we constructed a new evaluation dataset for English-Korean to compare the results. We discuss this problem in Section 4.3 in further detail.

Although we changed the parameter, some OOV words from our evaluation dataset remain, except for the English-Korean language, where all selected words for the evaluation were among the first 50K words in the monolingual embeddings.

FastText/ SketchEngine (%)	50K loaded			300-400K loaded		
	ET-SK	CZ-SK	EN-KO	ET-SK	CZ-SK	EN-KO
MUSE-S	<u>19.33</u>	<u>57.84</u>	<u>39.97</u>	<u>27.86</u>	<u>68.73</u>	<u>29.98</u>
	20.00	70.94	31.01	42.40	78.95	34.14
MUSE-I	<u>19.26</u>	<u>57.91</u>	<u>39.00</u>	<u>25.80</u>	<u>68.73</u>	<u>23.17</u>
	19.40	71.00	28.41	38.93	79.02	29.65
MUSE-U	<u>19.80</u>	<u>58.58</u>	<u>36.46</u>	<u>24.46</u>	<u>69.13</u>	<u>24.58</u>
	18.80	71.00	26.14	34.80	79.02	25.33
VecMap-S	<u>20.73</u>	<u>58.24</u>	<u>50.51</u>	<u>34.93</u>	<u>69.73</u>	<u>49.00</u>
	20.33	70.67	32.52	51.86	79.02	35.44
VecMap-I	<u>21.00</u>	<u>59.05</u>	<u>41.59</u>	<u>34.73</u>	<u>71.87</u>	<u>33.98</u>
	19.20	71.00	28.63	46.00	80.09	29.87
VecMap-U	21.20	<u>58.98</u>	<u>36.35</u>	<u>33.53</u>	<u>71.94</u>	<u>29.76</u>
	18.86	70.67	21.93	44.80	80.09	12.42
FastText	<u>20.60</u>	<u>57.51</u>	<u>50.40</u>	<u>31.93</u>	<u>67.93</u>	51.91
	21.06	70.54	26.90	49.33	78.28	37.60

Table 1: The recall of models before and after changing the parameter for loaded embeddings (Supervised: MUSE-S, VecMap-S, FastText; Identical: MUSE-I, VecMap-I; Unsupervised: MUSE-U, VecMap-U).

There were 331 Estonian OOV words in the Estonian-Slovak language combination trained with FastText (e.g., *aedvili* – *vegetable*, *ellu jääma* – *to stay alive*, *mürgitama* – *to poison*, etc.) and 40 when trained with SketchEngine (e.g., *enne kui* – *before*, *buteen* – *butene*, *uusik* – *newcomer*, etc.). In Czech-Slovak trained with FastText, there were 11 Czech OOV words (e.g., *cáklý* – *crazy*, *mlsný* – *sweet tooth*, *slušet* – *to suit*, etc.). For SketchEngine, there was 1 OOV word containing a spelling mistake: *onemocnět*, correctly *onemocnět* (*to get ill*). Fig. 3 and 4 show the frequency distribution in the monolingual corpus of the Estonian and Czech OOV words, respectively. The number of occurrences represents the frequency of the OOV words from the monolingual corpus, and word pair rank corresponds to the number of the OOV words.

Finally, we showed that selecting a vocabulary for the evaluation dataset is crucial for setting the number of loaded embeddings during training and mirroring the model’s quality more accurately. The evaluation dataset should consist of the words occurring in the loaded word embeddings from MEs. Moreover, the number of loaded MEs should not exceed the highest word rank in the evaluation dataset. And it should omit OOV words and multi-word expressions, as we do not assess the quality of the MEs.

4.2 Inflected Word Forms

Another essential factor to allow for when constructing an evaluation dataset is inflected word forms. MEs are trained on the corpus where words occur in context, not necessarily

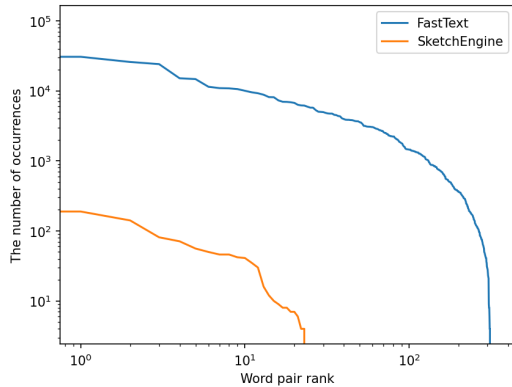


Figure 3: Comparison of the Estonian OOV word pair rank from FastText and SketchEngine MEs and the number of their occurrences from Estonian National Corpus 2017.

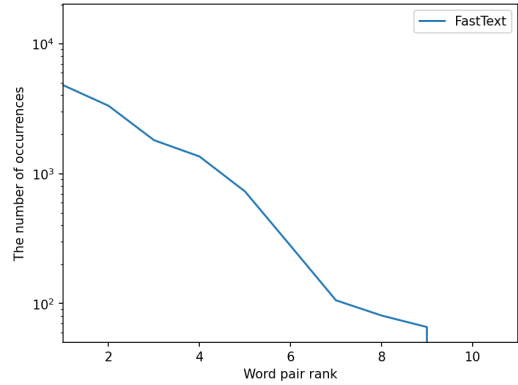


Figure 4: Comparison of the Czech OOV word pair rank from FastText and the number of their occurrences from Czech National Corpus 2017 (SketchEngine had only one OOV with 0 occurrences in the corpus).

in their basic form. Therefore, more common word forms from the text appear in the MEs and subsequently in the CMs.

For example, suppose we extract translation equivalent candidates for the Estonian word *tund* (*hour*) with Slovak as a target language. In that case, we get various word forms such as *hodiny*, *hodinu*, *hodín*, and *hodina*, from which only the last one is the basic word form of this word.

We often seek the word’s basic form for the dictionary, although it is not the most common form that appears in the texts. However, if we do not consider morphological variance and include only basic word forms of the source and target language word pair, i.e., *tund* = *hodina*, all other forms would be counted as an error even though their meaning is the same. Moreover, ignoring the morphological variance results in an inaccurate model recall and precision.

In our experiments, we applied Slovak lemmatiser Majka⁹ on the extracted translation equivalent candidates for Estonian-Slovak, and Czech-Slovak language combinations to create the basic word form of each word. This caused duplicate translation equivalent candidates. For instance, instead of *hodiny*, *hodinu*, *hodín*, and *hodina*, we had four times the word *hodina*. When counting the recall, we counted this as one translation equivalent candidate. We utilised the same evaluation datasets from Section 4.1.

Afterwards, we compared the recall before and after lemmatisation. The results are displayed in Table 2.

Given Table 2, we can see that the recall for each model increased by approximately 1-7%. Thus, we get more accurate results for the model when allowing for morphological variance either by lemmatising the results or including various word forms in the evaluation dataset.

⁹ <https://nlp.fi.muni.cz/czech-morphology-analyser/>

Non-lemmatised/ Lemmed (%)	FastText MEs		SketchEngine MEs	
	ET	CZ	ET	CZ
MUSE-S	27.86/ 29.40	68.73/ 70.80	42.40/ 45.00	78.95/ 79.75
MUSE-I	25.80/ 27.86	68.73/ 70.94	38.93/ 42.60	79.02/ 79.89
MUSE-U	24.46/ 28.80	69.13/ 71.20	34.80/ 41.93	79.02/ 79.82
VecMap-S	34.93/ 35.93	69.73/ 71.74	51.86/ 52.93	79.02/ 79.82
VecMap-I	34.73/ 36.06	71.87/ 72.94	46.00/ 50.86	80.09/ 80.96
VecMap-U	33.53/ 35.20	71.94/ 73.01	44.80/ 50.26	80.09/ 80.96
FastText	31.93/ 32.06	67.93 /70.14	49.33/ 50.53	78.28/ 79.09

Table 2: The recall of models before and after lemmatisation.

4.3 Part of Speech (POS)

In this section, we discuss whether including various word pairs from all POS groups is necessary or whether a more relevant POS can adequately mirror the models’ performance. Moreover, we show the proportion of the POS in the evaluation datasets we used and the performance of the models on various POS.

The selection of POS of the word pairs is the central topic of the articles that critically examine the evaluation datasets for the BLI task. For instance, the analysis of the POS distribution in the MUSE evaluation datasets (Conneau et al., 2017) conducted by Kementchedjhieva et al. (2019) revealed that these datasets contain a large number of proper nouns. The authors saw this as a problem since proper nouns do not carry any meaning; therefore, they are not suitable for reflecting the models’ performance.

Another effort provided by Izbicki (2022) compiled evaluation datasets for 298 languages with as similar POS distribution as possible across the datasets to make results between models and language pairs comparable. However, every POS was represented in their datasets.

We argue that some POS are less relevant to involve in the evaluation dataset than others. Except for proper nouns, such categories as pronouns, conjunctions, articles, and prepositions cannot accurately reflect the models’ performance since they play a syntactic role and their meaning changes within the context or is phrase-dependent. Moreover, in many cases, they do not correspond to each other across the languages and are either not translated or translated with more than one word. Therefore, these POS do not suit for evaluating word-to-word translations.

In this section, we examined the POS distribution in the evaluation datasets we used. We automatically annotated the evaluation datasets to analyse the POS distribution. For Estonian, we used EstNLTK¹⁰, an open-source tool for processing the Estonian language. We tagged the Czech dataset with Majka and utilised the NLP tool Polyglot¹¹ for English. Importantly, these tools are designed to tag words in the context that our datasets were

¹⁰ <https://estnltk.github.io/>

¹¹ <https://polyglot.readthedocs.io/en/latest/>

lacking. Moreover, even for human annotators is challenging to determine the POS of the word, especially when the context is missing. Thus, the results might contain discrepancies.

When we look at the POS distribution in our evaluation datasets, the dataset for Estonian-Slovak was disproportional as it contains many nouns, while other POS have significantly smaller representations. This dataset was derived from the Estonian-Slovak dictionary (Denisová, 2021), which consists mainly of nouns. On the other hand, the POS was distributed more evenly in the Czech-Slovak evaluation dataset, which was constructed manually. Fig. 5 and 6 display the graphs of the POS distribution in these datasets.¹²

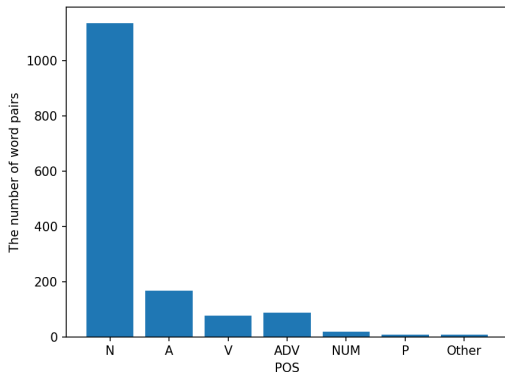


Figure 5: The POS distribution in the Estonian-Slovak evaluation dataset.

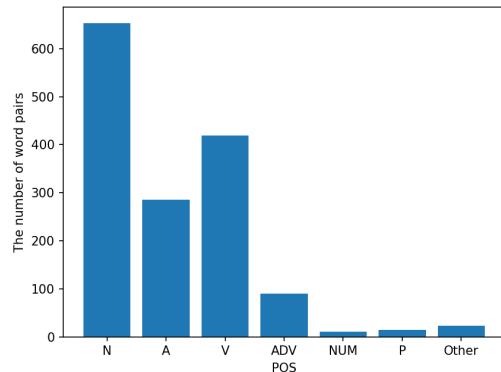


Figure 6: The POS distribution in the Czech-Slovak evaluation dataset.

We utilised the MUSE evaluation dataset (Conneau et al., 2017) to assess the English-Korean language combination. These datasets contain many proper nouns, which is confirmed by the graph in Fig. 7. However, after a manual check, we discovered that some nouns were incorrectly tagged as proper nouns. Moreover, a significant group of words was tagged with the symbol X (in the graph marked as “other”) when the tagger could not identify the POS of the current word.

We compiled a new English-Korean dataset with different POS distributions. We sampled word pairs from the English-Korean dictionary. This dictionary was created with the bilingual SketchEngine tool and post-processed manually (Kovář et al., 2016). This assumes the correctness of the translation equivalents (in contrast to the automatically compiled evaluation dataset as MUSE is). We intentionally avoided involving proper nouns, pronouns, articles, conjunctions, and prepositions. Fig. 8 provides the graph of the POS distribution.

In the next step, we computed recall for both evaluation datasets. We set the same conditions for both datasets: no OOV words, around 1,500 headwords in the dataset, and 400K loaded embeddings. Table 3 outlines the results.

Table 3 shows that recall for models trained with FastText MEs dropped drastically. On the other hand, the models trained with SketchEngine MEs did not decrease significantly.

¹² N = nouns; A = adjectives; V = verbs; ADV = adverbs; NUM = numerals; P = pronouns; PN = proper nouns; Other = conjunctions, interjections, prepositions, unknown, etc.

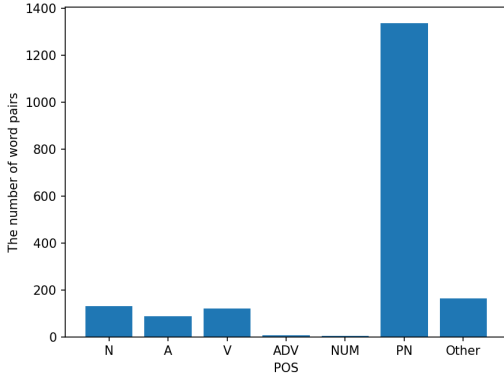


Figure 7: The POS distribution in the English-Korean MUSE evaluation dataset.

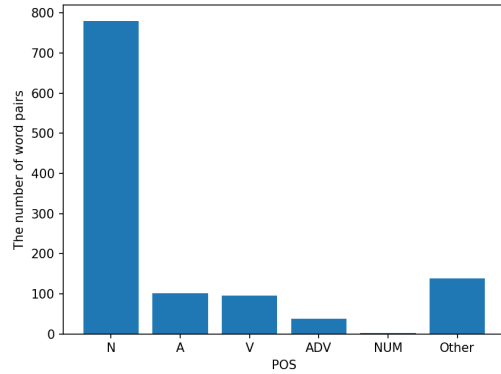


Figure 8: The POS distribution in the English-Korean SketchEngine evaluation dataset.

($\%$)	MUSE dataset		SketchEngine dataset	
	FastText MEs	SketchEngine MEs	FastText MEs	SketchEngine MEs
MUSE-S	29.98	34.14	20.58	31.83
MUSE-I	23.17	29.65	19.89	24.56
MUSE-U	24.58	25.33	19.37	23.18
VecMap-S	49.00	35.44	29.41	36.24
VecMap-I	33.98	29.87	23.44	25.25
VecMap-U	29.76	12.42	22.31	15.22
FastText	51.91	37.60	28.37	37.80

Table 3: The comparison between the results when using the MUSE and SketchEngine evaluation datasets for English-Korean.

However, this changing recall could also result from removing other errors that the MUSE evaluation dataset contains, such as words from different languages, abbreviations, nonsense words, etc. (Denisová & Rychlý, 2021).

The gap between the VecMap trained in a supervised mode evaluated on MUSE and SketchEngine datasets is almost 20%. We investigated some examples of the found and not found word pairs from the evaluation dataset.

A closer look revealed that VecMap likely found a correct equivalent for proper nouns, such as *Abdullah*, *Alexandra*, *Cambodia*, *Cameroon*, *Helsinki*, etc., which made up a significant group in the MUSE dataset, but they had no representation in the SketchEngine dataset.

Furthermore, the VecMap was good at finding equivalents for international words, for example, *algebra*, *alias*, *android*, *idol*, *email*, etc. As in the previous example, these words occurred more frequently in the MUSE than in the SketchEngine dataset.

When looking at the not found words, some were caused by mistakes in the evaluation dataset. For example, in the MUSE dataset were word pairs that consisted of the same word,

i.e., *android* – *android*. In the MUSE dataset occurred words with multiple translation equivalents, from which one was either wrong (*Yemen* translated as *South Yemen*) or VecMap could find only one of them (*fence*).

In the SketchEngine dataset, we observed various words for foods, animals, or numbers, such as *bean*, *tea*, *mudfish*, *rooster*, *two*, *fifty*, etc., for which VecMap could not find an appropriate equivalent but found word that had a similar lexical-semantic relationship (e.g., *tea* – *coffee*, *fifty* – *fourteen*, etc.). On the other hand, the MUSE dataset lacks such words. Moreover, the SketchEngine dataset contains more verbs than the MUSE dataset, a problematic group for VecMap to find an equivalent for (see Table 4).

Finally, we computed the recall for the VecMap supervised model for each POS separately and for all language pairs to observe how the results change. Table 4 displays the results.

VecMap-S (%) FT/SE	ET-SK	CZ-SK	EN-KO
Nouns	31.89/ 48.54	76.87/ 86.21	48.46/ 50.00
Adjectives	48.21/ 63.09	73.07/ 77.97	47.19/ 43.82
Verbs	35.52/ 64.47	63.48/ 67.30	35.00/ 21.66
Adverbs	41.37/ 56.32	70.00/ 81.11	75.00 / 37.50
Numerals	61.11/ 77.77	81.81 / 81.81	25.00/ 25.00
P/ PN	62.50 / 75.00	80.00/ 100	51.00/ 34.25
Others	25.00/ 37.50	69.56/ 100	43.55/ 39.26

Table 4: The recall of the supervised VecMap for each POS in each language.

According to Table 4, the results for different POS and MEs varied from each other greatly. As for distant language pairs, they achieved relatively high recall for adjectives, adverbs, and pronouns/ proper nouns. Furthermore, both language combinations gained low results for verbs.

However, they differed in the outcomes for numerals and nouns. Estonian-Slovak achieved the highest recall on numerals, whereas in a language that does not share a script, English-Korean, it was the lowest. The other way around it was by the results for nouns.

The close language pair, Czech-Slovak, was able to find all the equivalents from the evaluation dataset for pronouns, and small POS groups (in the table as *Others*), such as conjunctions, interjections, prepositions, etc. The explanation for this is that these two groups have a small representation in the evaluation dataset and a high word rank in MEs, so it was easier for the model to find an equivalent. High results were also achieved for nouns and numerals. Similarly to the distant language pairs, verbs were the weakest group.

The reasons behind these diverse results are manifold. For instance, recall depends on OOV words, so if nouns are the biggest group in the Estonian-Slovak evaluation dataset, they also contain a high number of OOV words. Thus, their recall is relatively low compared to

the nouns in the English-Korean evaluation dataset, with no OOV words. Other factors are: how many senses of the word are included in the evaluation dataset (the discussion is provided in Section 4.4), the word rank of the source language words in the MEs, the quality of the MEs and alignment, and the quality of the tagging tool.

Importantly, Table 4 demonstrates the significant impact of the POS distribution in the evaluation dataset on the resulting quality of the model.

4.4 Senses

Another important component when constructing an evaluation dataset is how many senses of one word to include. For example, the English word *band* as a noun has several meanings, such as *musical group*, *piece of cloth*, *range of values*, etc., or it can be a verb. Therefore, if we want the model to find all meanings in the target language, we should induce the same number of translation equivalents' candidates. However, with more extracted translation equivalents candidates comes much noise in the form of various errors, such as words with different POS, words with other lexical-semantic relationships, shortcuts, etc. (Denisová, 2022).

Thus, the precision decreases when the number of extracted target language words is extended, and we need to find the right amount depending on our goal (higher precision or higher recall). In this section, we investigate how the number of extracted target language words impacts precision and recall.

In our experiments, we measured the precision $P@1$, $P@5$, and $P@10$. Moreover, we computed the recall for each stage to compare the results. All models were trained and evaluated under the same conditions as in Section 4.1. Tables 5 and 6 show the outcomes.

Tables 5 and 6 confirm that as the precision increases, the recall drops; reversely, the higher recall, the lower the precision. This means that the more target language words we induce, the higher recall we achieve. However, the precision of our model declines. Therefore, we should set our aim beforehand, whether to use the resulting induced translation equivalents candidates for lexicography, which requires higher recall, or language acquisition that favours precision.

On top of that, the end user is also essential when selecting the nature of the words. For example, when constructing the Estonian-Slovak or English-Korean dictionary for language students, we should focus on frequently used words or words from the basic vocabulary. However, we should select different words rather than mutual when dealing with a close language pair, such as Czech-Slovak. Especially when assessing identical training mode.

5. Conclusion

In this paper, we have evaluated three benchmark CMs in various settings from different points of view on the BLI task. We have used three language pairs for the demonstration, i.e., a distant language pair, Estonian-Slovak, a close language pair, Czech-Slovak, and language pair that does not share a script, English-Korean. We have discussed various parameters that an evaluation dataset should allow for. We showed that these parameters are crucial for reflecting the model's performance precisely and accurately. Moreover, they are vital for setting the parameters in training, such as the number of loaded MEs.

FT MEs	P@1 (%)			P@5 (%)			P@10 (%)		
	Pre./ Rec.	ET-SK	CZ-SK	EN-KO	ET-SK	CZ-SK	EN-KO	ET-SK	CZ-SK
MUSE-S	<u>15.06</u>	<u>62.00</u>	<u>15.06</u>	<u>6.05</u>	<u>17.02</u>	<u>6.08</u>	<u>3.61</u>	<u>8.96</u>	<u>3.88</u>
	<u>14.60</u>	<u>47.96</u>	<u>11.61</u>	<u>24.00</u>	<u>65.33</u>	<u>23.44</u>	<u>27.86</u>	<u>68.73</u>	<u>29.98</u>
MUSE-I	<u>12.10</u>	<u>62.52</u>	<u>12.61</u>	<u>5.21</u>	<u>16.95</u>	<u>4.77</u>	<u>3.34</u>	<u>8.96</u>	<u>3.00</u>
	<u>11.73</u>	<u>48.36</u>	<u>9.72</u>	<u>20.66</u>	<u>65.06</u>	<u>18.42</u>	<u>25.80</u>	<u>68.73</u>	<u>23.17</u>
MUSE-U	<u>10.11</u>	<u>63.38</u>	<u>13.03</u>	<u>5.21</u>	<u>16.95</u>	<u>4.84</u>	<u>3.17</u>	<u>9.01</u>	<u>3.18</u>
	<u>9.80</u>	<u>49.03</u>	<u>10.04</u>	<u>20.66</u>	<u>65.06</u>	<u>18.69</u>	<u>24.46</u>	<u>69.13</u>	<u>24.58</u>
VecMap-S	<u>21.52</u>	<u>61.22</u>	<u>31.18</u>	<u>7.95</u>	<u>17.15</u>	<u>10.94</u>	<u>4.53</u>	<u>9.09</u>	<u>6.35</u>
	<u>20.86</u>	<u>47.36</u>	<u>24.04</u>	<u>31.35</u>	<u>65.86</u>	<u>42.19</u>	<u>34.93</u>	<u>69.73</u>	<u>49.00</u>
VecMap-I	<u>18.70</u>	<u>65.63</u>	<u>19.83</u>	<u>7.66</u>	<u>17.73</u>	<u>7.31</u>	<u>4.50</u>	<u>9.37</u>	<u>4.40</u>
	<u>18.13</u>	<u>50.76</u>	<u>15.28</u>	<u>30.40</u>	<u>68.06</u>	<u>28.20</u>	<u>34.73</u>	<u>71.78</u>	<u>33.98</u>
VecMap-U	<u>16.29</u>	<u>65.63</u>	<u>15.76</u>	<u>7.23</u>	<u>17.75</u>	<u>6.36</u>	<u>4.35</u>	<u>9.38</u>	<u>3.86</u>
	<u>15.80</u>	<u>50.76</u>	<u>12.15</u>	<u>28.66</u>	<u>68.13</u>	<u>24.52</u>	<u>33.53</u>	<u>71.94</u>	<u>29.76</u>
FastText	<u>17.05</u>	<u>59.93</u>	<u>31.74</u>	<u>7.02</u>	<u>16.55</u>	<u>11.46</u>	<u>4.14</u>	<u>8.85</u>	<u>6.73</u>
	<u>16.63</u>	<u>46.35</u>	<u>24.44</u>	<u>27.86</u>	<u>63.52</u>	<u>44.19</u>	<u>31.93</u>	<u>67.93</u>	<u>51.91</u>

Table 5: The precision (pre.) P@1, P@5, and P@10 and recall (rec.) for the models trained with FastText (FT) MEs.

SE MEs	P@1 (%)			P@5 (%)			P@10 (%)		
	Pre./ Rec.	ET-SK	CZ-SK	EN-KO	ET-SK	CZ-SK	EN-KO	ET-SK	CZ-SK
MUSE-S	<u>25.51</u>	<u>72.88</u>	<u>18.78</u>	<u>7.96</u>	<u>19.82</u>	<u>7.17</u>	<u>4.48</u>	<u>10.24</u>	<u>4.42</u>
	<u>24.73</u>	<u>56.37</u>	<u>14.47</u>	<u>37.73</u>	<u>76.41</u>	<u>27.66</u>	<u>42.40</u>	<u>78.95</u>	<u>34.14</u>
MUSE-I	<u>22.55</u>	<u>72.62</u>	<u>12.89</u>	<u>7.24</u>	<u>19.80</u>	<u>6.16</u>	<u>4.11</u>	<u>10.25</u>	<u>3.84</u>
	<u>21.86</u>	<u>56.17</u>	<u>9.94</u>	<u>34.33</u>	<u>76.35</u>	<u>23.77</u>	<u>38.93</u>	<u>79.02</u>	<u>29.65</u>
MUSE-U	<u>19.60</u>	<u>72.71</u>	<u>10.44</u>	<u>6.52</u>	<u>19.82</u>	<u>5.26</u>	<u>3.68</u>	<u>10.25</u>	<u>3.28</u>
	<u>19.00</u>	<u>56.24</u>	<u>8.04</u>	<u>30.93</u>	<u>75.41</u>	<u>20.31</u>	<u>34.80</u>	<u>79.02</u>	<u>25.33</u>
VecMap-S	<u>32.11</u>	<u>72.19</u>	<u>21.30</u>	<u>9.74</u>	<u>19.68</u>	<u>7.61</u>	<u>5.48</u>	<u>10.25</u>	<u>4.59</u>
	<u>31.13</u>	<u>55.84</u>	<u>16.42</u>	<u>46.20</u>	<u>75.88</u>	<u>29.33</u>	<u>51.86</u>	<u>79.02</u>	<u>35.44</u>
VecMap-I	<u>24.33</u>	<u>72.36</u>	<u>13.66</u>	<u>8.35</u>	<u>19.91</u>	<u>6.12</u>	<u>4.86</u>	<u>10.39</u>	<u>3.87</u>
	<u>23.60</u>	<u>55.97</u>	<u>10.53</u>	<u>39.60</u>	<u>76.75</u>	<u>23.60</u>	<u>46.00</u>	<u>80.09</u>	<u>29.87</u>
VecMap-U	<u>24.20</u>	<u>72.45</u>	<u>2.91</u>	<u>8.17</u>	<u>19.91</u>	<u>2.27</u>	<u>4.73</u>	<u>10.39</u>	<u>1.61</u>
	<u>23.46</u>	<u>56.04</u>	<u>3.78</u>	<u>38.73</u>	<u>76.75</u>	<u>8.75</u>	<u>44.80</u>	<u>80.09</u>	<u>12.42</u>
FastText	<u>28.74</u>	<u>72.79</u>	<u>22.21</u>	<u>9.01</u>	<u>19.67</u>	<u>8.08</u>	<u>5.21</u>	<u>10.16</u>	<u>4.87</u>
	<u>27.86</u>	<u>56.31</u>	<u>17.12</u>	<u>42.73</u>	<u>75.81</u>	<u>31.17</u>	<u>49.33</u>	<u>78.28</u>	<u>37.60</u>

Table 6: The precision P@1, P@5, and P@10 and recall for the models trained with SketchEngine (SE) MEs.

To sum up, the high-quality evaluation dataset for the BLI task should contain words that occur in the MEs used in training and omit OOV words and multi-word expressions. It should take inflected word forms into account or lemmatise the results. Moreover, it should prefer nouns, verbs, adjectives, adverbs, and numerals over pronouns, proper nouns, articles, prepositions, and conjunctions. It should determine the number of the extracted target language words based on the final purpose and the number of senses one headword possesses. Finally, when selecting words and evaluation metrics, we should always consider the language pairs, the end user, and the purpose of the dictionary.

We have provided reproducible criteria applicable for evaluating any model or language pair on the BLI task. These criteria help unify the future evaluation process and make the results comparable and transparent. On top of that, we have made the CMs more approachable for the lexicography field by bringing the lexicography perspective into the evaluation.

Moreover, when observing Table 1, we notice that the results for a close language pair are always better when the unsupervised or identical mode is used. Regardless of the data or MEs utilised during training, the results for the Czech-Slovak language combination are constant and predictable favouring identical or unsupervised mode.

On the other hand, the distant language pairs achieve better results when supervision signals are involved in training. In most cases, the models trained on a distant language pair in a supervised mode surpassed their identical or unsupervised counterparts.

Additionally, the performance of the models trained with SketchEngine MEs exceeded the FastText MEs in many instances. Therefore, high-quality MEs are a key component of the resulting CM.

When looking at the models' recall in Table 1 or Tables 5 and 6, we can conclude that these models cannot be used as a standalone resource in lexicography yet. However, they offer an alternative as supplementary data (e.g., frequently occurring words in the corpus) to parallel-data-based methods for small languages or rare language pairs. Also, they are a good source of lexical-semantically related words in the target language. On top of that, they can be valuable in compiling technical dictionaries, especially when MEs are trained on the domain-specific corpus.

6. Acknowledgements

The research in this paper was supported by the Internal Grant Agency of Masaryk University, project CZ.02.2.69/0.0/0.0/19_073/0016943.

7. References

- Artetxe, M., Labaka, G. & Agirre, E. (2016). Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 2289–2294. URL <https://aclanthology.org/D16-1250>.
- Artetxe, M., Labaka, G. & Agirre, E. (2017). Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association*

- for *Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pp. 451–462. URL <https://aclanthology.org/P17-1042>.
- Artetxe, M., Labaka, G. & Agirre, E. (2018a). Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*. pp. 5012–5019. URL <https://doi.org/10.1609/aaai.v32i1.11992>.
- Artetxe, M., Labaka, G. & Agirre, E. (2018b). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pp. 789–798. URL <https://aclanthology.org/P18-1073>.
- Artetxe, M., Labaka, G., Lopez-Gazpio, I. & Agirre, E. (2018c). Uncovering Divergent Linguistic Information in Word Embeddings with Lessons for Intrinsic and Extrinsic Evaluation. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*. Association for Computational Linguistics, pp. 282–291. URL <https://aclanthology.org/K18-1028>.
- 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. URL <https://aclanthology.org/Q17-1010>.
- Conneau, A., Lample, G., Ranzato, M., Denoyer, L. & J’egou, H. (2017). Word Translation Without Parallel Data. *ArXiv*, abs/1710.04087. URL <https://arxiv.org/abs/1710.04087>.
- Denisová, M. (2021). Compiling an Estonian-Slovak Dictionary with English as a Binder. In *Proceedings of the eLex 2021 conference*. Lexical Computing CZ, s.r.o., pp. 107–120. URL https://elex.link/elex2021/wp-content/uploads/2021/08/eLex_2021_06_pp107-120.pdf.
- Denisová, M. (2022). Parallel, or Comparable? That Is the Question: The Comparison of Parallel and Comparable Data-based Methods for Bilingual Lexicon Induction. In *Proceedings of the Sixteenth Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2022*. Tribun EU, pp. 4–13. URL <https://nlp.fi.muni.cz/raslan/raslan22.pdf#page=13>.
- Denisová, M. & Rychlý, P. (2021). When Word Pairs Matter: Analysis of the English-Slovak Evaluation Dataset. In *Recent Advances in Slavonic Natural Language Processing (RASLAN 2021)*. Brno: Tribun EU, pp. 141–149. URL <https://nlp.fi.muni.cz/raslan/2021/paper3.pdf>.
- Glavaš, G., Litschko, R., Ruder, S. & Vulić, I. (2019). How to (Properly) Evaluate Cross-Lingual Word Embeddings: On Strong Baselines, Comparative Analyses, and Some Misconceptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 710–721. URL <https://aclanthology.org/P19-1070>.
- Glavaš, G. & Vulić, I. (2020). Non-Linear Instance-Based Cross-Lingual Mapping for Non-Isomorphic Embedding Spaces. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 7548–7555. URL <https://aclanthology.org/2020.acl-main.675>.
- Herman, O. (2021). Precomputed Word Embeddings for 15+ Languages. *RASLAN 2021 Recent Advances in Slavonic Natural Language Processing*, pp. 41–46. URL <https://nlp.fi.muni.cz/raslan/raslan21.pdf#page=49>.
- Izbicki, M. (2022). Aligning Word Vectors on Low-Resource Languages with Wiktionary. In *Proceedings of the Fifth Workshop on Technologies for Machine Translation of Low-*

- Resource Languages (LoResMT 2022)*. Association for Computational Linguistics, pp. 107–117. URL <https://aclanthology.org/2022.loresmt-1.14>.
- Joulin, A., Bojanowski, P., Mikolov, T., Jégou, H. & Grave, E. (2018). Loss in Translation: Learning Bilingual Word Mapping with a Retrieval Criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 2979–2984. URL <https://aclanthology.org/D18-1330>.
- Karan, M., Vulić, I., Korhonen, A. & Glavaš, G. (2020). Classification-Based Self-Learning for Weakly Supervised Bilingual Lexicon Induction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 6915–6922. URL <https://aclanthology.org/2020.acl-main.618>.
- Kementchedjheva, Y., Hartmann, M. & Søgaard, A. (2019). Lost in Evaluation: Misleading Benchmarks for Bilingual Dictionary Induction. 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)*. Association for Computational Linguistics, pp. 3336–3341. URL <https://aclanthology.org/D19-1328>.
- Kovář, V., Baisa, V. & Jakubíček, M. (2016). Sketch Engine for Bilingual Lexicography. *International Journal of Lexicography*, 29(3), p. 339–352. URL <https://doi.org/10.1093/ijl/ecw029>.
- Lample, G., Denoyer, L. & Ranzato, M. (2017). Unsupervised Machine Translation Using Monolingual Corpora Only. *ArXiv*, abs/1711.00043. URL <https://arxiv.org/abs/1711.00043>.
- Parizi, A.H. & Cook, P. (2021). Evaluating a Joint Training Approach for Learning Cross-lingual Embeddings with Sub-word Information without Parallel Corpora on Lower-resource Languages. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*. Association for Computational Linguistics, pp. 302–307. URL <https://aclanthology.org/2021.starsem-1.29>.
- Ren, S., Liu, S., Zhou, M. & Ma, S. (2020). A Graph-based Coarse-to-fine Method for Unsupervised Bilingual Lexicon Induction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 3476–3485. URL <https://aclanthology.org/2020.acl-main.318>.
- Ruder, S., Vulić, I. & Søgaard, A. (2019). A Survey of Cross-lingual Word Embedding Models. *The Journal of Artificial Intelligence Research*, 65, pp. 569–631. URL <https://doi.org/10.1613/jair.1.11640>.
- Severini, S., Hangya, V., Jalili Sabet, M., Fraser, A. & Schütze, H. (2022). Don’t Forget Cheap Training Signals Before Building Unsupervised Bilingual Word Embeddings. In *Proceedings of the BUCC Workshop within LREC 2022*. European Language Resources Association, pp. 15–22. URL <https://aclanthology.org/2022.bucc-1.3>.
- Søgaard, A., Ruder, S. & Vulić, I. (2018). On the Limitations of Unsupervised Bilingual Dictionary Induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pp. 778–788. URL <https://aclanthology.org/P18-1072>.
- Tian, Z., Li, C., Ren, S., Zuo, Z., Wen, Z., Hu, X., Han, X., Huang, H., Deng, D., Zhang, Q. & Xie, X. (2022). RAPO: An Adaptive Ranking Paradigm for Bilingual Lexicon Induction. *ArXiv*, abs/2210.09926. URL <https://arxiv.org/abs/2210.09926>.
- Vulić, I., Glavaš, G., Reichart, R. & Korhonen, A. (2019). Do We Really Need Fully Unsupervised Cross-Lingual Embeddings? 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)*. Association for Computational Linguistics, pp. 4407–4418. URL <https://aclanthology.org/D19-1449>.

- Vulić, I. & Korhonen, A. (2016). On the Role of Seed Lexicons in Learning Bilingual Word Embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pp. 247–257. URL <https://aclanthology.org/P16-1024>.
- Vulić, I., Korhonen, A. & Glavaš, G. (2020). Improving Bilingual Lexicon Induction with Unsupervised Post-Processing of Monolingual Word Vector Spaces. In *Proceedings of the 5th Workshop on Representation Learning for NLP*. Association for Computational Linguistics, pp. 45–54. URL <https://aclanthology.org/2020.repl4nlp-1.7>.
- Woller, L., Hangya, V. & Fraser, A. (2021). Do not neglect related languages: The case of low-resource Occitan cross-lingual word embeddings. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*. Association for Computational Linguistics, pp. 41–50. URL <https://aclanthology.org/2021.mrl-1.4>.

Definition Extraction for Slovene: Patterns, Transformer Classifiers and ChatGPT

Hanh Thi Hong Tran^{1,2,3}, Vid Podpečan¹,
Mateja Jemec Tomazin⁴, Senja Pollak¹

¹Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia

²Jožef Stefan International Postgraduate School, Jamova cesta 39, 1000 Ljubljana,
Slovenia

³University of La Rochelle, 23 Av. Albert Einstein, La Rochelle, France

⁴ZRC SAZU, Novi trg 2, 1000 Ljubljana, Slovenia

E-mail: tran.hanh@ijs.si

Abstract

Definition Extraction is a Natural Language Processing task that automatically identifies definitions from unstructured text sequences. In our research, we frame this problem as a binary classification task, aiming to detect whether a given sentence is a definition or not, using text sequences in Slovene. The main contributions of our work are two-fold. First, we introduce a novel Slovene corpus for the evaluation of Definition Extraction named *RSDO-def*. The dataset contains labeled sentences from specialized corpora using two different extraction processes: random sampling and pattern-based extraction. Both sets contain manual annotations by linguists with three labels: Definition, Weak definition, and Non-definition. Second, we propose the benchmarks for Slovene Definition Extraction systems that use (1) rule-based techniques; (2) Transformers-based models as binary classifiers; (3) ChatGPT prompting, and evaluate them on both sets of *RSDO-def* corpus. When only the small sample *RSDO-def-random* is considered, the pattern-based rules surpassed the performance of language models classifiers or ChatGPT in terms of F1 on definition class in the strict evaluation setting (considering Weak definition as Non-definition). Meanwhile, language models (classifiers and ChatGPT) outperformed rule-based approaches when applied to the data with a higher number of definitions and more relaxed evaluation scenarios (considering Weak definition as Definition). Comparing ChatGPT and language models classifiers on the definition class of *RSDO-def-random* and *RSDO-def-large*, we observe that higher precision was obtained with classifiers, but higher Recall with ChatGPT.

Keywords: Definition Extraction; RSDO-DEFT; Rule-based; Transformers; ChatGPT

1. Introduction

Definition Extraction is a Natural Language Processing (NLP) task that extracts textual definitions from naturally occurring texts (Navigli & Velardi, 2010). While extracting the definitions of words or phrases from general language corpora is needed for the creation of general dictionaries and lexical databases, extracting definitions of terms from specialized domain corpora can serve for creating specialized dictionaries and glossaries. In our paper, we are interested in the latter.

Definition Extraction is closely tied to the needs of lexicography and terminography. For example, in our recently developed terminological portal¹ supporting also Term Extraction, the Definition Extraction module is used to select selecting examples of use and support the user in the manual definition construction process (currently, the pattern-based extractor is implemented in the portal). Definition Extraction is applied in many other NLP downstream tasks as well, including glossary creation (Klavans & Muresan, 2002; Park et al., 2002), question answering (Cui et al., 2005; Saggion & Gaizauskas, 2004), taxonomy learning (Velardi et al., 2013; Espinosa-Anke et al., 2016), knowledge base generation (Bovi et al., 2015), to cite a few.

SemEval-2020 Task 6: DeftEval: Extracting Term-Definition Pairs in the Free Text (Spala et al., 2020) recently introduced the novel human-annotated English dataset, namely Definition Extraction from Texts (DEFT) corpus and formulated the task as three consecutive subtasks: (1) classification of sentences as definition or non-definition, which is also the task addressed in our work (2) labeling of definitional sentences, and (3) relation classification. However, when it comes to lesser-resourced languages like Slovene, there is no dedicated publicly available annotated collection designed for the development and evaluation of Definition Extraction.

The contribution of this paper is threefold: (1) The creation of a novel Slovene corpus for Definition Extraction evaluation (consisting of *RSDO-def-random* and *RSDO-def-larger*) with three labels: not a definition, weak definition, and strong definition; (2) Filling the research gap in the Definition Extraction for Slovene by experimenting with different neural approaches (3) An empirical evaluation of rule-based, language model based binary classifier, and ChatGPT prompting benchmarks for Definition Extraction task.

This paper is organized as follows: Section 2 presents the related work in Definition Extraction and addresses also the gap between the approaches developed for rich-resourced (e.g., English) and less-resourced (e.g., Slovene) languages. In Section 3, we briefly introduce the novel Slovene corpus for Definition Extraction with two different sample acquisition strategies: random sampling and pattern-based methods. Next, we present the methods and implementation details (Section 4), followed by the description of experimental results (Section 5) and error analysis (Section 6). Section 7 presents the conclusion and our plans for further work.

2. Related Work

Most of the early approaches to Definition Extraction were inspired by the work of Hearst (1992) on lexico-syntactic patterns. The cues of definition sentences include lexical cues (e.g., “*is*”, “*means*”, “*are*”, “*is defined as*”), which are often combined with grammatical rules and syntactic patterns (Klavans & Muresan, 2001; Cui et al., 2004, 2005; Sarmiento et al., 2006; Storrer & Wellingshoff, 2006). As these approaches are only able to detect explicit and structured definitions, they fail to detect sentences containing definitions without predefined linguistic clues, and thus, often suffer from low Recall.

With the advent of machine learning (ML) methods, several supervised and semi-supervised models have been proposed (Fahmi & Bouma, 2006; Westerhout, 2009; Reiplinger et al., 2012; Jin et al., 2013; Espinosa-Anke & Saggion, 2014; Espinosa-Anke et al., 2015). The

¹ <https://terminoloski.slovenscina.eu>

task was then considered as a binary sentence classifier (Fahmi & Bouma, 2006) where they took advantage of features based on bag-of-words (BoW), n-grams, and syntactic information, to mention a few. A hybrid approach (Westerhout, 2009) that combined the rule-based and ML-based classifier was then introduced with further exploration and exploitation in both linguistic and structural information, while Borg et al. (2009) proposed the use of genetic algorithms to learn distinguishing features of definitions and non-definitions and weight the individual features. Different bootstrapping algorithms (Reiplinger et al., 2012; De Benedictis et al., 2013) were also experimented with to boost the performance of the extraction. However, the mentioned methods often depend heavily on manual rules and handcrafted features, which are time- and effort-consuming to design as well as domain-specific. As a result, they are hard to adapt and generalize to a new domain or another specific task. An interesting approach by Navigli & Velardi (2010) proposed automatically learned Word Class Lattices (WCLs), a generalization of word lattices, to model textual definitions, where lattices are learned from an annotated dataset of definitions from Wikipedia.

Recent years have witnessed a shift toward neural network-oriented solutions to prevent the issues from rule-based and traditional ML approaches and better capture a large variety of possible definition realizations. Li et al. (2016) proposed a Long Short-Term Memory (LSTM) classifier, where the features were automatically generated from the raw input sentences and part-of-speech (PoS) sequences. Meanwhile, various neural hybrid methods have been released regarding either the combination between two models (e.g., Anke & Schockaert (2018) combined Convolutional Neural Network (CNN) and bi-LSTM) or between two different representations as an input for the neural model (e.g., Veyseh et al. (2020) leveraged Graph Convolutional Neural (GCN) by concatenating both syntactic and semantic information). Furthermore, Kannan & Ponnusamy (2020) presented a combination of both hybrid strategies by concatenating GloVe and on-the-fly PoS information as an input and feeding them to the bi-LSTM with an additional 1-dimensional Convolution and MaxPool layer on top of that. Meanwhile, Kaparina & Soboleva (2020) made use of both global and contextual information by ensembling FastText and ELMo word embeddings to a Recurrent Neural Network (RNN) architecture.

However, until recently, the existing methods have not yet benefited from the large pretrained language models and the transfer learning paradigm (Kenton & Toutanova, 2019), which is today a standard for developing state-of-the-art (SOTA) solutions to a large variety of NLP downstream tasks. This has changed with the introduction of *SemEval-2020 Shared Task DeftEval: Extracting Term-Definition Pairs in the Free Text* (Spala et al., 2020) and its novel human-annotated English dataset, namely Definition Extraction from Texts (DEFT) corpus. There, several solutions based on transfer learning and Transformer architecture, have been proposed. While several participants opted to simply fine-tune BERT (Davletov et al., 2020; Jeawak et al., 2020; Singh et al., 2020), RoBERTa (Avram et al., 2020) or XLNet (Ranasinghe et al., 2020), others opted for more specific approaches. For example, Caspani et al. (2020) captured contextual information from the input sentence using RoBERTa and applied the Stochastic Weight Averaging (Izmailov et al., 2018) to combine weights of the same network at different stages of training, whereas Zhang & Ren (2020) incorporated several LSTM layers into different Transformer architectures to boost the performance of their definition extractor.

Whilst recent Definition Extraction methods are leveraging Transformers-based language models (see e.g., winning approaches in the *SemEval-2020 Task 6* (Spala et al., 2020)), these methods were not yet sufficiently applied to lesser-resourced language such as Slovene. Here, the related work is limited to rule-based approaches (Pollak, 2014b,a; Pollak et al., 2012) or feature-based ML methods (Fišer et al., 2010). The rule-based approaches have been applied to the specialized corpora in various domains, including karstology (Pollak et al., 2019; Vintar & Martinc, 2022). However, no language-model-based approaches have been tested on the task of Definition Extraction for Slovene.

3. RSDO-def Datasets

One of the contributions of this paper is the creation of the Slovene Definition Extraction evaluation datasets, *RSDO-def* (Jemec Tomazin et al., 2023), publicly available through the CLARIN.SI data sharing repository: <http://hdl.handle.net/11356/1841>. The dataset was annotated in the scope of the project Development of Slovene in Digital Environment (RSDO). The sentences were extracted from the Slovene domain-specific corpora (Jemec Tomazin et al., 2021), collected in the scope of the project Development of Slovene in a Digital Environment, containing texts with annotated terms from four different domains: biomechanics, linguistics, chemistry, and veterinary science. Two different sampling strategies were used to create two different sub-corpora: random sampling (*RSDO-def-random*) and pattern-based selection (*RSDO-def-larger*). While random sampling represents the most realistic evaluation scenario and allows for assessment of Recall of various methods, the number of definitions is very small and therefore represents a too small sample to support a reliable quantitative evaluation of methods. Therefore, in order to increase the number of definitions, we added pattern-based sampling, where a pattern-based Definition Extraction approach (Pollak, 2014a) was used for sentence selection. This approach results in a larger sample of definitions, but on the other hand, as the pattern-based method was used in the data collection process, we had to exclude the method from the evaluation phase.

Both sets were manually annotated by five terminographers, where after individual assessments in case of discrepancies between annotators, a consensus was reached and the final label was confirmed by all five annotators. In the resulting dataset, the sentences were annotated with one of the three labels: Definition, Weak definition, and Non-definition. The criteria for annotation are based on the standard *ISO 1087-1:2000 (E/F) Terminology Work - Vocabulary, Part 1, Theory and Application*, which explains the *definition* as follows: “*Representation of a concept by a descriptive statement which serves to differentiate it from related concepts*”. The most common are intensional definitions, which state the superordinate concept and the delimiting characteristics. Such definitions are often provided in student handbooks and specialized manuals. *Weak definition* labels were assigned if the extracted sentences contained a term and at least one delimiting feature without a superordinate concept, or sentences consisting of superordinate concepts without delimiting features but with some typical examples. Instances were labeled as *Non-definition* if the sentence with the extracted concept did not contain any information about the concept or its delimiting features. Such sentences are also more common in scientific texts, so the imbalance is not unexpected.

The label distribution statistics (after removing a small number of duplicates) are presented in Figure 1. For evaluation (see Section 5), we consider two scenarios: Weak definition is

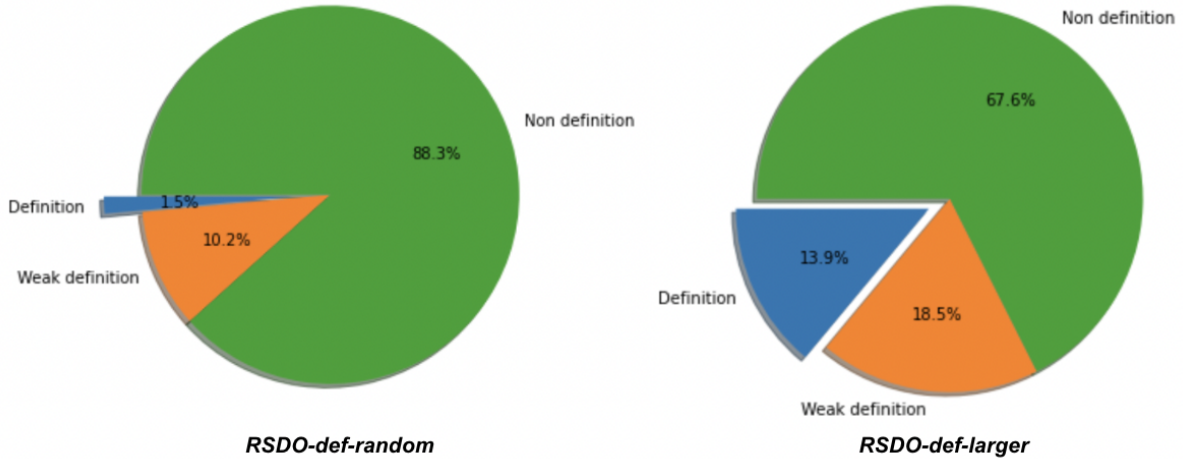


Figure 1: The distribution of Definition, Weak definition, and Non-definition category in each corpus.

considered as (1) Non-definition (Strict evaluation scenario); and (2) Definition (Relaxed evaluation scenario). In both scenarios, there is an imbalance between the number of samples belonging to the positive and negative classes, which reflects also the real-life Definition Extraction settings. Note that these samples are useful for the evaluation of the Definition Extraction methods, but we consider them too small for use as datasets for training definition extractor systems.

3.1 RSDO-def-random

In *RSDO-def-random* corpus, where sentences were selected by random sampling techniques, 961 sentences were manually annotated, out of which 14 examples were assigned the Definition class, 849 examples were assigned the Non-definition class, and 98 examples were assigned the Weak definition class. For the strict evaluation scenario (Weak definitions considered as Non-Definitions) the class distribution is 1.5% (Definitions) vs. 98.5% (Non-definitions), while for the relaxed evaluation scenario (Weak definitions considered as Definitions), the class distribution is 11.7% (Definitions) vs. 88.3% (Non-definitions).

3.2 RSDO-def-larger

In order to increase the number of definitions that represent only a small fraction of the randomly sampled *RSDO-def-random* set, we extended this initial corpus using pattern-based extraction methods. The pattern-based extractor was run on the whole corpus and the results were manually labeled. The resulting *RSDO-def-larger* dataset contains 175 Definitions and 848 Non-definitions while the rest 232 examples are Weak definitions. For the strict evaluation scenario (Weak definitions considered as Non-Definitions) the class distribution is 13.9% (Definitions) vs. 86.1% (Non-definitions), while for the relaxed evaluation scenario (Weak definitions considered as Definitions), the class distribution is 32.4% (Definitions) vs. 67.6% (Non-definitions).

4. Methodology

The main goal of the paper is to evaluate different Definition Extraction approaches. We compare the baseline pattern-based approach (Pollak, 2014a) (with two variants, *is a* pattern type and extended pattern list), with four newly implemented deep-learning Transformer classifiers, and a ChatGPT-based solution. For training the Transformer models, we used Wikipedia as a training set, as in Fišer et al. (2010). The experimental workflow is presented in Figure 1. In this section, we present the methods for the three approaches—Rule-based ones in Section 4.1, Transformer classifiers in Section 4.2, and the ChatGPT-based ones in Section 4.3—followed by the evaluation metrics used in for experiments.

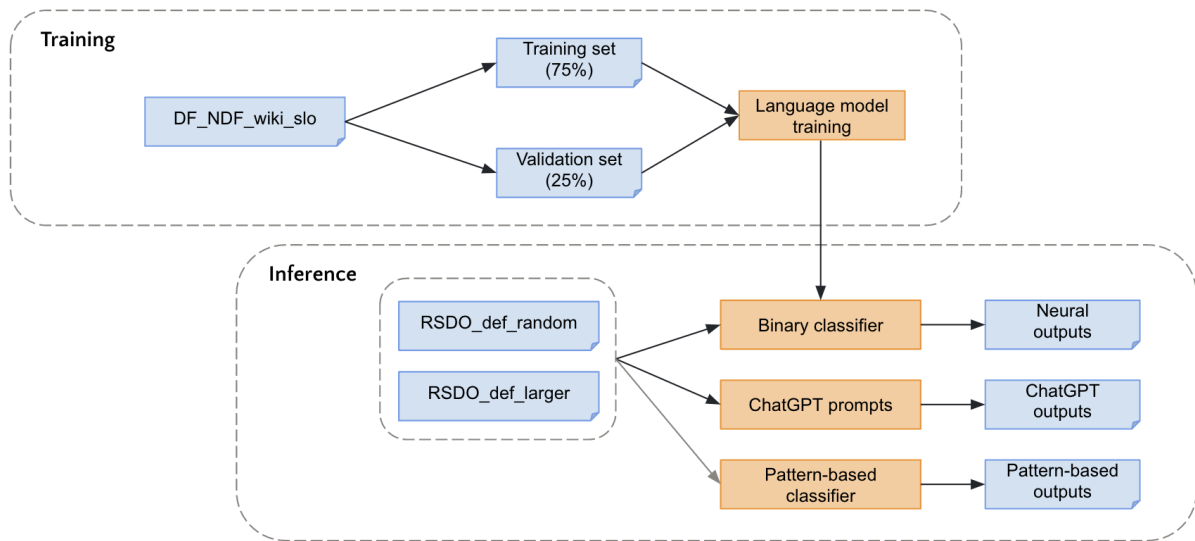


Figure 2: The experimental workflow.

4.1 Rule-based approach

We apply the pattern-based approach, based on a set of soft lexico-syntactic patterns, which was developed by Pollak (2014a,b). The patterns contain a combination of lexical cues (either on the lemma or token level) and information from morphosyntactic descriptions, more specifically part-of-speech and case information (e.g., to detect nominative case forms). In total, 12 soft patterns are defined to extract sentences of type (e.g., *NP-nom je* [Eng. *is*] *NP-nom*, *NP-nom se nanaša na* [Eng. *refers to*] *NP-nom*, *NP-nom pomeni* [Eng. *denotes*] *NP-nom*).

The simplest pattern is “*X je Y*” [“*X is Y*”], where X is the term (noun phrase) to be defined and Y is another noun phrase (usually its hypernym), followed by the differentia (part listing the differences from other types belonging to this class of entities). This corresponds to the genus and differentia definition type, meaning that if we have the term X to be defined, we define it by using its hypernym (Y) and by listing the differences from other types belonging to this class of entities (“X is Y that...”). Since Slovene is a highly inflected language, we can add the condition that the noun phrases should agree in

case and that the case should be nominative (i.e., “*NP-nom is NP-nom*”), where “*NP*” means noun phrase and “*NP-nom*” stands for noun phrase in the nominative case. There can be several variations, for example, including an English translation of a Slovene term. While the majority of the patterns are designed to extract genus and differentia type of definitions, other patterns are targeting Weak definitions, where patterns are designed to find paraphrases and synonym relations or cues of specific functional definitions (e.g., *Naloga [The task of] NP-gen is*). For a detailed description of the pattern-based approach, see Pollak (2014b).

In the evaluation, we consider two different pattern lists:

- *PatternAll* is the list of all 12 patterns;
- *JeStaSoPatterns* is a subset of the entire pattern set, containing only the patterns of “*X is a Y*” type and its variations.

4.2 Transformer classifiers

For the experiments described in this section, we compare different pretrained language models, one monolingual (SloBERTa) and three multilingual ones (mBERT, mDistilBERT, and XLMR)², which are fine-tuned for the definition classification task. Given that the size of our RSDO-def datasets is too small to use them for training a classifier, we use a dataset created from Wikipedia as training data, based on the work by Fišer et al. (2010).

4.2.1 Training datasets

For training our Transformers-based classifiers, we use the dataset *DF_NDF_wiki_slo* created from Wikipedia by Fišer et al. (2010). In the data construction process, the authors considered the first sentence of a Wikipedia article containing an entry term as a definition (Y) and other sentences with the same entry term as non-definitions (N). While this is not the ground truth, it can be considered a silver standard to be used as training data for ML approaches. We prepared the dataset for public release (Podpečan et al., 2023), available via CLARIN.SI: <http://hdl.handle.net/11356/1840>.

For our experiments, we compare two different samples for the negative class. In the dataset, all the sentences with the article key term but the first one, the non-definitions are labeled as N1, and in the version where the key term is not at the beginning of a sentence, these are labeled as N. The rationale of testing also the second approach, is that it is not impossible that non-first sentences in the Wikipedia articles are also definitions, and terms at the beginning of a sentence could indicate such examples, which would introduce noise when treated as a negative class. In total, 34,084 examples were collected, out of which 3,251 belong to the positive (definition) class Y, and 20,684 to the N1 class (non-definitions). In the second scenario, excluding the terms at the beginning of non-initial sentences, the distribution is 3,251 vs. 14,678, for definitions and non-definitions (N), respectively. When training a classifier, we compare both scenarios, one with the negative class of all non-first sentences with the term (N1) and one with only those that do not contain terms at the beginning of the sentence (N). The labeling ratio is presented in Figure 3 with free-text examples.

² Available in the HuggingFace library: <https://huggingface.co/models>

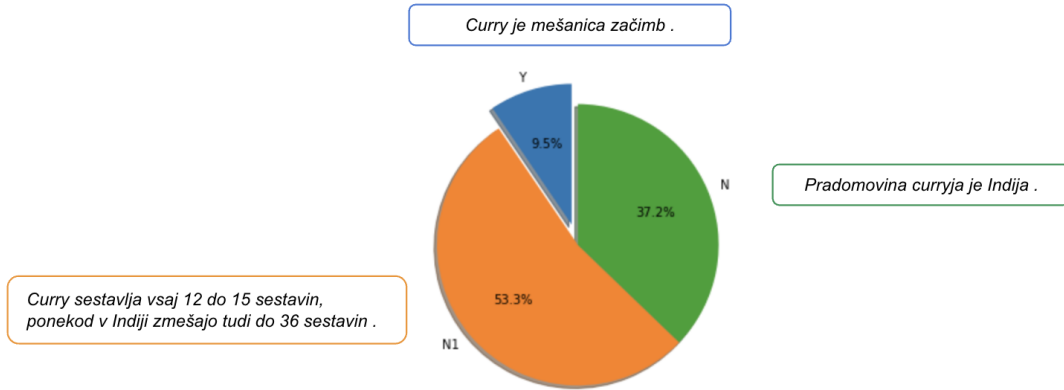


Figure 3: The distribution of each label in the DF_NDF_wiki_slo dataset.

4.2.2 Models

We consider one monolingual and three multilingual pretrained models that we fine-tune for the definition classification task using the training data described in Section 4.2.1.

SloBERTa³ (Ulčar & Robnik-Šikonja, 2021) is a cutting-edge Slovene language model trained as a masked language model, using *fairseq*⁴ toolkit. The corpora used for training the model have 3.47 billion tokens in total with a subword vocabulary of 32,000 tokens, making it a comprehensive resource for Slovene NLP research and development. The model’s performance on benchmark tests highlights its effectiveness for a wide range of NLP downstream tasks, especially with Slovene data.

mBERT⁵ (Kenton & Toutanova, 2019) is a multilingual Transformer-based model pretrained in a self-supervised regime on a massive corpus consisting of 104 languages using two objectives: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). While MLM randomly masks 15% of the words in the input sentences and then fed the entire masked sentence to the model to predict the masked words so that the model can learn the bidirectional representation of the given input sentences, NSP concatenates two masked sentences as input during pretraining and predicts if two sentences were following each other or not so that the model can an inner representation of the languages in the training set that can then be used to extract features useful for downstream tasks.

mDistilBERT⁶ (Sanh et al., 2019) is a distilled (smaller and faster) version of BERT, which uses BERT base model as a teacher and was pretrained on the same corpus as BERT in a self-supervised regime with three objectives: MLM as BERT, Distillation loss where the model was trained to return the same probabilities as BERT, and Cosine embedding loss where the model was also trained to generate hidden states as close as possible as the BERT. Similar to mBERT, the model was trained on a concatenated corpus of 104 different languages from Wikipedia. With only six layers, 768 dimensions, and 12 heads, mDistilBERT used 134M parameters in total, which is significantly fewer than the 177M

³ <https://huggingface.co/EMBEDDIA/sloberta>

⁴ <https://github.com/facebookresearch/fairseq>

⁵ <https://huggingface.co/bert-base-multilingual-uncased>

⁶ <https://huggingface.co/distilbert-base-uncased>

parameters used in mBERT-base. As a result, mDistilBERT is twice as fast as mBERT while maintaining a certain level of performance on various benchmark tests

XLMR⁷ (Conneau et al., 2019) is a multilingual RoBERTa-based version (Liu et al., 2019), which was pretrained on a vast corpus of 2.5TB of filtered CommonCrawl data, spanning 100 different languages. It uses the same training procedure as RoBERTa Liu et al. (2019) which used only the Masked Language Model (MLM) technique without using Next Sentence Prediction (NSP) technique. The model was demonstrated to outperform other pretrained models on a variety of natural language understanding tasks (e.g., question answering, and natural language inference). This open-sourced model is also designed to be fine-tuned on specific tasks (e.g., NER, term extraction), making it a versatile tool for a wide range of NLP applications.

4.2.3 Implementation Details

We divide *DF_NDF_wiki_slo* into two parts: a training set and a test set with a ratio of 0.75: 0.25 in a stratified fashion, respectively. Class weighting is applied for each class in order that the classifier is aware of how to treat each class well in the cost function to improve the performance of the target.

As three labels are proposed in the training and test corpus, we build two separate classifiers: (1) one that predicts definitional sentences (Y) and non-definitional sentences which may also contain the term at the beginning of the sentence (N1); and (2) one that predicts between definitional sentences (Y) and non-definitional sentences that do not start with the key term (N). For each classifier, we evaluate the performance of our two evaluation datasets, *RSDO-def-random* and *RSDO-def-larger* with two distinctive settings: (1) in the strict evaluation scenario, considering Weak definitions as Non-definition (negative class); and (2) in the relaxed scenario, considering Weak definitions as Definitions (positive class).

The training and validation samples were binarized to the desired format. We fine-tuned weight decay and dropout coefficients due to high-performance costs. The learning rate was set equal to 1e-05. All models were trained for 5 epochs with a batch size of 8 and validation occurred at the end of each epoch.

4.3 ChatGPT

Introduced by OpenAI⁸ at The 36th Conference on Neural Information Processing Systems (NeurIPS⁹ 2022), ChatGPT quickly gained immense popularity with more than 1 million users in less than a week due to its ability to generate human-like and convincing responses. The underlying architecture of this conversational agent is GPT-3.5, a large generative pretrained Transformer model containing over 175 billion parameters. Figure 4 demonstrates how ChatGPT describes its capability in Definition Extraction.

We followed a straightforward zero-shot mechanism to classify a given Slovene sequence as a definition or not by accessing the ChatGPT via the official web interface¹⁰ with *ChatGPT*

⁷ <https://huggingface.co/xlm-roberta-base>

⁸ <https://openai.com/>

⁹ <https://nips.cc/>

¹⁰ <https://chat.openai.com>

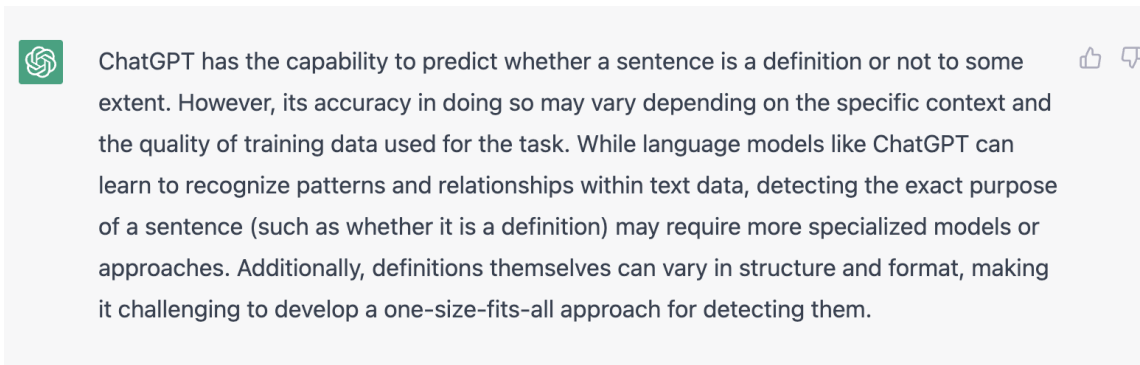


Figure 4: ChatGPT’s capability in Definition Extraction defined by ChatGPT.

Mar 23 Version¹¹ between 5th and 10th April 2023. We defined the vanilla prompt for both of our Slovene subsets as demonstrated in Figure 5.

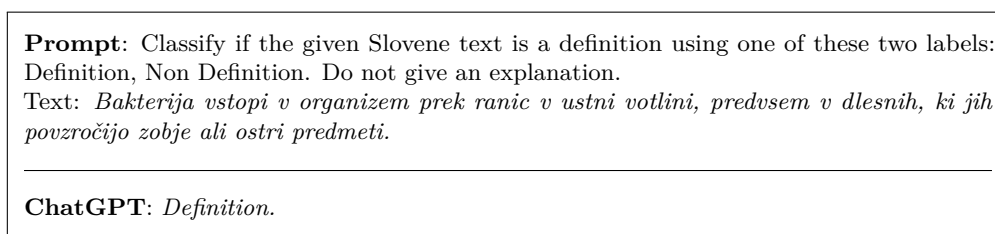


Figure 5: Slovene dataset prompt scenario

Besides the clear instruction on the specific task in our specific language, we also require ChatGPT not to explain in detail the reason why ChatGPT gives the labels for the given sequence so that we can capture only the necessary answer for the output format consistency.

4.4 Evaluation settings

Different approaches are evaluated on the two *RSDO-def* corpora (*RSDO-def-random* and *RSDO-def-larger*). We use two different evaluation scenarios. In the first, stricter setting, we consider Weak definitions as a negative class (Non-definitions), while in the second, relaxed setting, we consider Weak definitions as a positive class (Definitions), as they still produce very relevant content, but, for example, are not formulated as an intensional definition with the superordinate concept. For *RSDO-def-random*, we compare the performance among all the approaches—pattern-based, Transformer classifiers, and Chat-GPT—, whereas for *RSDO-def-larger*, we only consider Transformer classifiers and Chat-GPT results, as the pattern-based approach was used in the dataset construction phase.

As both corpora are imbalanced, we evaluate the performance of the classifier separately for each class by classifying all examples and comparing the predictions with the groundtruth using Precision (P), Recall (R), and F1-score (F1) for both the minority and majority

¹¹ <https://help.openai.com/en/articles/6825453-chatgpt-release-notes>

classes in each setting. Note that the minority (definition) class is the most important for our task. In addition to evaluating each class separately, we also calculate the macro-averaging of these three mentioned evaluation metrics, where no weights are applied for aggregation. Therefore, the evaluation metrics will have a bigger penalization when our classifier does not perform well with the minority classes.

From the terminographical perspective, we consider Precision more important than Recall. With possible integration into the Terminology portal in mind, where different features serve as a tool for compiling new terminological resources for human users, better Precision provides users with sentences that can be turned almost directly into definitions. Our quick survey among professional users revealed that they would only include definitions if they were informative enough, otherwise, they would settle for terms in the source language and terms in the target language without definitions, and not consider automated retrieval of sentences as a useful feature. Higher Precision of extracted sentences facilitates and simplifies the automatization of terminology management, as only with sufficient quality these tools are to be adopted by the users.

5. Results

We report the performance of different setups on *RSDO-def-random* and *RSDO-def-larger* dataset using P, R, and F1 for each class and macro-average —with best results in bold for each of them— in Tables 1 and 2, for the strict and the relaxed evaluation scenarios with regard to the weak definitions, respectively. We highlight the definition class results in each table, as this is the category of our main interest.

The results demonstrate that when the number of definitional samples is higher, as in the relaxed evaluation scenario, the language model tends to capture definitional sequences better than rule-based ones, while if the amount of definitions is very small (strict evaluation scenario), well-structured linguistic patterns used in the rule-based approaches have the advantage if we consider F1-score on the definition class.

Table 1: Comparative evaluation in Precision, Recall, and F1-score in the *strict* evaluation scenario, where we consider Weak definition as Non-definition.

Methods		RSDO-def-random									RSDO-def-large								
		Definition			Non-definition			Macro avg.			Definition			Non-definition			Macro avg.		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Pattern-based	JeStaSo	0.11	0.31	0.17	0.99	0.97	0.98	0.55	0.64	0.57	-	-	-	-	-	-	-	-	-
	Patterns all	0.11	0.31	0.16	0.99	0.96	0.98	0.55	0.64	0.57	-	-	-	-	-	-	-	-	-
Transformers Y/N classifier	SloBERTa	0.09	0.15	0.11	0.99	0.98	0.98	0.54	0.57	0.55	0.64	0.40	0.49	0.91	0.96	0.93	0.77	0.68	0.71
	mBERT	0.12	0.15	0.13	0.99	0.98	0.99	0.55	0.57	0.56	0.67	0.37	0.48	0.90	0.97	0.94	0.79	0.67	0.71
Y/N1 classifier	DistilBERT	0.08	0.15	0.11	0.99	0.97	0.98	0.53	0.56	0.54	0.60	0.39	0.47	0.90	0.96	0.93	0.75	0.67	0.70
	XLM-R	0.09	0.15	0.11	0.99	0.98	0.98	0.54	0.57	0.55	0.65	0.41	0.50	0.91	0.96	0.94	0.78	0.69	0.72
Transformers Y/N1 classifier	SloBERTa*	0.14	0.15	0.15	0.99	0.99	0.99	0.57	0.57	0.57	0.68	0.27	0.39	0.89	0.98	0.93	0.79	0.63	0.66
	mBERT*	0.00	0.00	0.00	0.99	0.98	0.98	0.49	0.49	0.49	0.66	0.23	0.34	0.89	0.98	0.93	0.77	0.60	0.64
	DistilBERT*	0.00	0.00	0.00	0.99	0.98	0.98	0.49	0.49	0.49	0.64	0.25	0.35	0.89	0.98	0.93	0.76	0.61	0.64
	XLMR*	0.10	0.15	0.12	0.99	0.98	0.98	0.54	0.57	0.55	0.64	0.26	0.37	0.89	0.98	0.93	0.77	0.62	0.65
ChatGPT		0.03	0.93	0.06	1.00	0.54	0.70	0.51	0.73	0.38	0.22	0.78	0.34	0.94	0.54	0.68	0.58	0.66	0.51

In the strict evaluation scenario, where we consider Weak definition as Non-definition, the pattern-based JeStaso approach (Precision: 0.11, Recall: 0.31, F1-score: 0.17) surpasses the performance of Transformer classifiers (Y/N and Y/N1) that we proposed in Definition class for *RSDO-def-random* dataset. Despite the lower Recall and F1-score, the Precision of the SloBERTa model is higher, which is also the most important metric in the opinion of the terminographers involved in the terminological portal development. One should notice that this test set suffers from a significant lack of definitions (only 14 instances), which makes the quantitative results non-reliable. Meanwhile, in *RSDO-def-large*, pattern-based approaches were used for data preparation in the annotation process. Therefore, we do not consider the evaluation of the pattern-based methods. The best single neural classifier is XLMR with a Precision of 0.65, Recall of 0.41, and F1-score of 0.50 in predicting the Definition class. Despite lower Precision and F1-score in comparison with language models, ChatGPT dominates Recall with nine times higher in *RSDO-def-random* and three times higher in *RSDO-def-large*, but it strongly underperforms in terms of Precision.

In the relaxed evaluation scenario (see Table 2, where we consider Weak definition as definition, the mBERT Y/N classifier presents the best performance if we consider Precision as the most important metrics for the terminographers (Precision: 0.47, Recall: 0.08, F1-score: 0.13 in *RSDO-def-random*; Precision: 0.89, Recall: 0.22, F1-score: 0.35 in *RSDO-def-large*). Despite mBERT being the classifier with the best Precision, ChatGPT provides the best Recall and F1-score on the definition class, which is twice as high as other classifiers in both *RSDO-def-random* and *RSDO-def-large*. Note that both corpora, regardless of sharing the same characteristics of imbalance, have different proportions of Definitions, and the results are therefore not expected to be comparable.

Table 2: Comparative evaluation in Precision, Recall, and F1-score in the *relaxed* evaluation scenario, where we consider Weak definition as Definition.

Methods		RSDO-def-random									RSDO-def-large								
		Definition			Non-definition			Macro avg.			Definition			Non-definition			Macro avg.		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Pattern-based	JeStaSo	0.23	0.08	0.12	0.89	0.97	0.93	0.56	0.52	0.52	-	-	-	-	-	-	-	-	-
	Patterns all	0.21	0.08	0.11	0.89	0.96	0.93	0.55	0.52	0.52	-	-	-	-	-	-	-	-	-
Transformers	SloBERTa	0.36	0.08	0.13	0.90	0.98	0.94	0.63	0.53	0.53	0.86	0.24	0.38	0.74	0.98	0.84	0.80	0.61	0.61
	mBERT	0.47	0.08	0.13	0.90	0.99	0.94	0.68	0.53	0.54	0.89	0.22	0.35	0.73	0.99	0.84	0.81	0.60	0.60
Y/N classifier	DistilBERT	0.40	0.10	0.16	0.90	0.98	0.94	0.65	0.54	0.55	0.85	0.24	0.38	0.74	0.98	0.84	0.79	0.61	0.61
	XLM-R	0.36	0.08	0.13	0.90	0.98	0.94	0.63	0.53	0.53	0.87	0.24	0.38	0.73	0.98	0.84	0.80	0.61	0.61
Transformers	SloBERTa*	0.43	0.06	0.10	0.89	0.99	0.94	0.66	0.52	0.52	0.86	0.15	0.26	0.71	0.99	0.83	0.79	0.57	0.54
	mBERT*	0.19	0.03	0.05	0.89	0.98	0.93	0.54	0.51	0.49	0.79	0.12	0.21	0.71	0.99	0.82	0.75	0.55	0.52
Y/N1 classifier	DistilBERT*	0.00	0.00	0.00	0.89	0.99	0.93	0.44	0.49	0.47	0.83	0.14	0.24	0.71	0.99	0.83	0.77	0.56	0.53
	XLMR*	0.30	0.06	0.10	0.89	0.98	0.94	0.60	0.52	0.52	0.78	0.14	0.24	0.71	0.98	0.82	0.75	0.56	0.53
ChatGPT		0.17	0.68	0.27	0.93	0.56	0.70	0.55	0.62	0.49	0.43	0.68	0.53	0.79	0.57	0.67	0.61	0.63	0.60

6. Error Analysis

6.1 Pattern-based classifier

In the analysis, we observed several sources of errors in the pattern-based approaches.

First, when sequences are too long, contain lists of items, or have wrong sentence segmentation, the segments rarely contain definitions. Multiple sentences instead of one pose also a problem with regard to evaluation.

Second, there are several examples formulated as questions, but still matching the patterns (e.g., “*Kolikšen delež besed je enopojavnic?*”). These typical syntax errors contained no definitions at all and were also the most common error type. These kinds of errors could be easily removed in the future adaptation of a pattern-based approach. There are examples, where insufficient context is provided, or when a definition is in fact correct, but not related to the term of our interest.

Other sources of errors related to insufficient context for the extracted term, too general phrases, and errors in automated morpho-syntactic annotations as a source of wrongly extracted sentences.

6.2 XLMR classifier

In this section, we discuss the behavior of our best single classifier (XLMR) trained on *DF_NDF_wiki_slo* where Weak definition is considered as Non-definition, regarding F1-score of Definition label.

Figure 6 presents the Kernel Distribution Estimation (KDE) plot, which is often used for depicting the probability density function of the continuous or non-parametric data variables. Here, we plot the density distribution of each sequence regarding the sequence length where the left side of the figure refers to wrong predictions and the right side to the true prediction of the XLMR classifier. The horizontal or x-axis presents the range of values for sequence length in the data set while the vertical or y-axis in a density plot is the probability density function for the kernel density estimation.

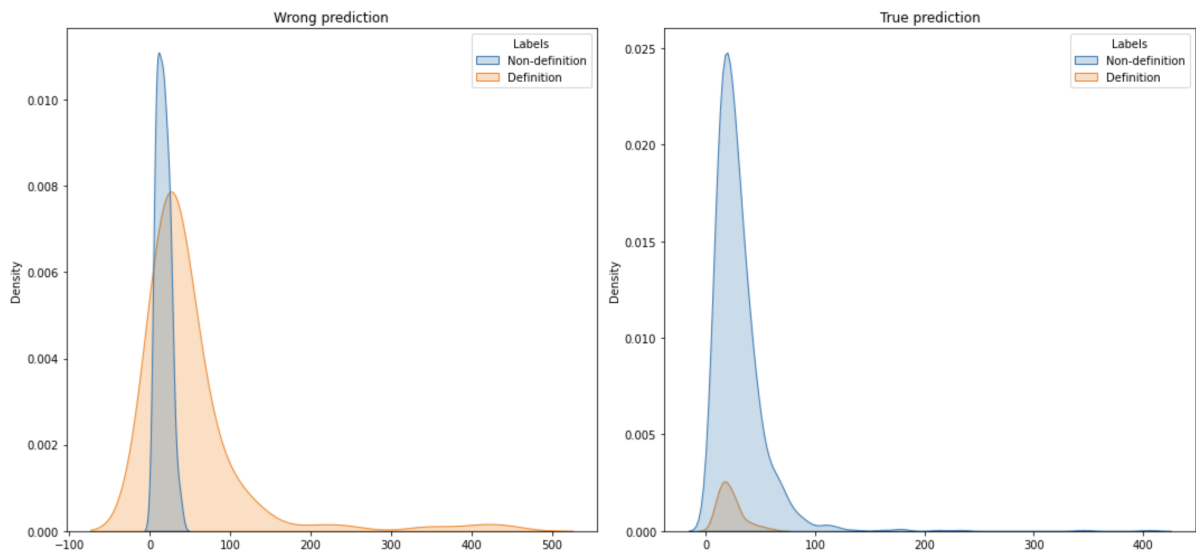


Figure 6: The distribution of sequence length each label wrongly predicted when using XLMR classifier trained on *DF_NDF_wiki_slo* in relaxed definition evaluation scenario.

Despite the infrequency of definitions being present in lengthy sequences, the classifier’s performance was impeded by data poisoning, resulting in the misclassification of instances in both classes. This anomaly is particularly evident in the misidentification of definition classes, where the model erroneously assigned examples with a long right tail to the definition class.

Table 3 lists randomly selected examples of the incorrectly classified sentences from *RSDO-def-large* when we use XLMR classifier trained on *DF_NDF_wiki_slo* where Weak definition is considered as non-definition. Looking at these sequences, we can see that some gold standard definitions are either incorrectly labeled or very difficult to classify even for a human.

For instance, the sentence “*Steklina je zoonoza , prisotna na vseh celinah in ugotovljena že v najmanj 150 državah sveta .*” (translated version: *Rabies is a zoonosis, present on all continents and found in at least 150 countries of the world.*) contains a term and its hypernym, but is not in the definition class, as the differentia part is not well expressed.

Table 3: Examples of the incorrect classified short sentence from *RSDO-def-large* using XLMR classifier trained on *DF_NDF_wiki_slo* where Weak definition is considered as non-definition.

Sentences	Labels	Predictions
Bruceloza je nalezljiva bolezen , ki se pogosto pojavlja pri kozah , ovce so nekoliko manj dovzetne .	0	1
Terminologija in sodobna terminografija (ur . Nina Ledinek , Mojca Žagar Karer in Marjeta Humar) .	0	1
Hitrost je fizikalna količina , ki nam pove , koliko se telo premakne na časovno enoto , in ima enoto [m/s] .	0	1
Prenašalci so okuženi konji in žrebci .	0	1
Steklina je zoonoza , prisotna na vseh celinah in ugotovljena že v najmanj 150 državah sveta .	0	1
Bakterijska bolezen , ki se pojavlja v obliki septikemije , hemoglobinurije , ikterusa in abortusa .	0	1
Slovenija je država , uradno prosta bruceloze .	0	1
...
4 . Amonijak je plin značilnega neprijetnega vonja , ki je dobro topen v vodi .	1	0
V glavi je onaglašena iztočnica z morebitnim izgovorom in slovničnimi podatki .	1	0
Terminov je v praktičnostrokovnih besedilih manj kot v drugih skupinah besedil .	1	0
Sila podlage Fp je reakcija , ki jo povzroči podlaga na opazovano telo .	1	0
Na primer , pojmovni sistem elektrotehnike je množica pojmov , ki določajo področje .	1	0
Izjema je → Toplarna Ljubljana , ki je namenjena tudi oskrbovanju mesta s toplo vodo .	1	0
Glavni predstavnik didaktičnostrokovnih besedil je učbenik (osnovnošolski , srednješolski , univerzitetni) .	1	0
Ponazorimo jih lahko s stavkom a je vrsta b . Gre za nadredne , podredne in priredne odnose .	1	0
3 . Najbolj pričakovano okolje terminov so strokovna besedila .	1	0
Izobraževalni proces je osebni odnos , didaktična in socialna interakcija .	1	0
Obe spojini sta klorida – soli klorovodikove kisline HCl .	1	0
Teoretična in empirična podlaga vprašalnika so sodobne motivacijske teorije (Juriševič , 2006) .	1	0
Izkustveno učenje je proces , ki poteka celo življenje .	1	0
Predilecijska mesta za razmnoževanje so možgani , stena prebavil in uterus .	1	0
Klinični znaki so anoreksija , depresija , povišana telesna temperatura .	1	0
Značilen znak je anemija , smrtnost pa je nizka .	1	0
Najpogostejša metoda za omamljanje kopitarjev je uporaba strelne naprave s penetrirnim klinom .	1	0
Najprimernejše mesto strela je vertikalna sredinska linija čela od 1 do 2 cm nad očmi .	1	0
...

Beside, above all the misclassified instances where the model failed to predict the true labels for definition, 24.5 % of the false negative belongs to the instances which constitutes from multiple sentences. As an example, this is a multiple-sentence instance: “*Imenujemo ga udarni zvok. « (52) ; » Kontaktor je mehanski stikalni aparat , ki ima samo en*

*mirovni položaj , ki ga ne upravljamo ročno , in je sposoben vklapljeti , prevajati in izklapljeti tok v normalnih pogojih obratovanja , upoštevajoč tudi preobremenitve. « (53) ; » Rele je električna naprava , ki povzroči v odvisnosti od spremenljive električne ali druge neelektrične veličine določeno spremembo v istem ali v drugih električnih tokokrogih. « (54) ; » Talilna varovalka je v bistvu namenoma ustvarjena šibka točka (močno zmanjšan prerez vodnika , poskus 4.5.2) na dostopovnem mestu električnega kroga , njeno delovanje pa temelji na odvisnosti toplotnega učinka od gostote toka. « (55) ; » Žirator je vezje , ki omogoča pretvorbo poljubne znane impedance v njeno dualno obliko ali inverzno vrednost. « (56) ... » (translated version: *We call it a percussive sound. « (52) ; » A contactor is a mechanical switching device that has only one rest position, which is not operated manually, and is capable of switching on, transferring and switching off the current under normal operating conditions, taking into account overloads as well. « (53) ; » A relay is an electrical device that causes, depending on a variable electrical or other non-electric quantity, a certain change in the same or other electrical circuits. « (54) ; » The fuse is basically a purposely created weak point (highly reduced conductor cross-section, experiment 4.5.2) at the access point of the electric circuit, and its operation is based on the dependence of the thermal effect on the current density. « (55) ; » A gyrator is a circuit that allows the conversion of any known impedance into its dual form or inverse value. « (56) ... »).**

This type of error could be removed by post-processing rules, based on the sequence length.

6.3 ChatGPT's prompting

Similar to the problems faced by sequence classifier, there exist multiple instances of ambiguity, which affect the model performance. For example, in Figure 7, the sentence *“Danes uporabljamo pretežno tripolne (ali kot jih običajno imenujemo, dvopolne z zascitnim kontaktom) . « (53) ; » Kontaktor je stikalni aparat, ki ga vklapljamo daljinsko (zato ga pogosto imenujemo tudi daljinsko stikalo) . « (53) ; » Oporniki ali izolatorniki so jekleni nosilci okrogle ali pravokotne oblike , vroce pocinkani, s pomocjo katerih pritrjujemo izolatorje na konzole , lesene drogove ali jeklene konstrukcije. « (54) ; » Mocnostni ali smerni releii”* (translated: *“Today we mostly use three-pole (or as they are usually called, two-pole with a protective contact). ” (53) ; » A contactor is a switching device that can be turned on remotely (that’s why it is often called a remote switch). ”(53) ; » Supports or insulators are steel supports of round or rectangular shape, hot-dip galvanized, with the help of which insulators are attached to consoles, wooden poles or steel structures. ” (54) ; » Power or direction relays”*) was detected by ChatGPT as a group of different sentences separated the full stop, and therefore, was tagged with the predicted labels separately. Additional major voting is applied for this scenario to create a comparative result with pattern-based and token classification approaches.

Although we do not require ChatGPT to explain in detail the reason behind the label choice, the chatbot still gave some explanation as demonstrated in Figure 8.

7. Conclusion and Further Work

We presented a novel evaluation dataset for Slovene Definition Extraction and proposed the benchmarks given three mechanisms: rule-based extractor, Transformer-based classifier, and prompt-based extractor with ChatGPT. Regarding the newly introduced corpus,

An example of ChatGPT’s response for the ambiguous cases.

Prompt: Classify if the given Slovene text is a definition using one of these two labels: Definition, Non Definition. Do not give an explanation. Text: *Danes uporabljamo pretežno tripolne (ali kot jih obicajno imenujemo, dvopolne z zascitnim kontaktom) . « (53) ; » Kontaktor je stikalni aparat, ki ga vklapljamu daljinsko (zato ga pogosto imenujemo tudi daljinsko stikalo) . «(53) ; » Oporniki ali izolatorniki so jekleni nosilci okrogle ali pravokotne oblike , vroce pocinkani, s pomocjo katerih pritrjujemo izolatorje na konzole , lesene drogove ali jeklene konstrukcije. « (54) ; » Mocnostni ali smerni releii*

ChatGPT: *Definition, Definition, Definition, Non Definition.*

Groundtruth: *Definition.*

Figure 7: Slovene dataset prompt gave multiple labels for a given text.

An example of ChatGPT’s response for the fail cases with explanation.

Prompt: Classify if the given Slovene text is a definition using one of these two labels: Definition, Non Definition. Do not give an explanation. Text: *Prav tako je za tipizacijo virusov stekline z monoklonalnimi protitelesi pomembna sestava aminokislinskega zaporedja nukleoproteina, saj imajo razlike v aminokislinskem zaporedju lahko za posledico tudi razlike v epitopih, kamor se vezejo monoklonalna protitelesa (Flamand in sod., 1980a, 1980b; Dietzschold in sod., 1987; Smith, 1989).*

ChatGPT: *Definition. The text explains the importance of the amino acid sequence composition of the nucleoprotein for the typing of rabies viruses using monoclonal antibodies, as differences in the amino acid sequence can result in differences in the epitopes to which monoclonal antibodies bind.*

Groundtruth: *Non Definition.*

Figure 8: Slovene dataset prompt answers with explanation.

RSDO-def-random and *RSDO-def-larger* were collected using random sampling techniques for the first, and including also the pattern-based sentence selection method for the second. Regarding the methods, if there are only a few well-structured instances of definitions that have clear linguistic characteristics (e.g., in the strict evaluation scenario, where Weak definitions are considered as non-definitions), a rule-based technique performed better in terms of F1-score (on the Definition class) than language models or prompting. However, for less structured examples (relaxed evaluation scenario with Weak definitions considered as definitions), ChatGPT prompting and language models were more effective than classical rule-based approaches. When comparing prompting and language model classifiers, for the Definition class, classifiers lead to higher Precision, while in terms of Recall, ChatGPT has better results.

The usefulness of the Definition Extraction process in practical applications is severalfold. In our work, we focused more specifically on terminological applications, as the definitions are extracted from the domain-specific corpora. The pattern-based approach is already implemented as part of the Slovene terminological portal as a tool for providing good examples and inspiring the users in their manual Definition Extraction process. In this paper, we also show the potential of language-model-based classifiers, that will be considered for inclusion after additional analysis.

There are also several points that have to be considered in using automated processes in terminographical work. First, for obtaining usable and applicable results, the user must

prepare a specialized corpus for the given domain. If the texts are not representative and include noisy material, also the Definition Extraction process will likely be not relevant. However, approaches that are not considering the task as extraction but as generation could be considered to overcome this limitation.

Next, we should also note that although Definition Extraction can be very useful for the manual definition process, as the Precision of the systems is still far from perfect, the users might be tempted to use the unmodified examples as final definitions and not use the tool as support. Therefore, in our project, we consider the output of Definition Extraction systems as good examples and do not call them definitions.

Our research provides the first study in neural Definition Extraction for Slovene and a multi-domain dataset for Definition Extraction evaluation. In future work, we plan to investigate several directions. First, increasing the size and diversity of the evaluation dataset to improve the reliability of the quantitative results is of crucial importance. Second, one of our goals is to develop a larger collection of labeled sentences that could be used not only for evaluation but also for training Definition Extraction systems. Next, as the annotation is often time- and effort-consuming, active learning can be considered to more efficiently use labeled examples and reduce human efforts. Furthermore, investigating the use of ensemble methods regarding the combination of different meaning information (e.g., local, global, contextual) and/or the combination of multiple models may help improve the overall performance. Last but not least, while in our study, we used ChatGPT for classifying sentences, we plan to leverage the large generative language models also for definition generation.

The definition evaluation gold standard dataset Jemec Tomazin et al. (2023) is publicly available at <http://hdl.handle.net/11356/1841>, the pattern-based approach from Pollak (2014a) is available at https://github.com/vpodpecan/definition_extraction and the code for the language models classifiers at https://github.com/honghanhh/definition_extraction. We also released the silver standard training data (Podpečan et al., 2023) by Fišer et al. (2010), which is now available via CLARIN.SI: <http://hdl.handle.net/11356/1840>.

8. Acknowledgements

The work was partially supported by the Slovene Research Agency (ARRS) core research programme Knowledge Technologies (P2-0103), and the Ministry of Culture of the Republic of Slovenia through the project Development of Slovene in Digital Environment (RSDO), as well as the project Formant combinatorics in Slovenian (J6-3131). The first author was partly funded by Region Nouvelle Aquitaine. This work has also been supported by the TERMITRAD (2020-2019-8510010) project funded by the Nouvelle-Aquitaine Region, France.

9. References

Anke, L.E. & Schockaert, S. (2018). Syntactically Aware Neural Architectures for Definition Extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. pp. 378–385.

- Avram, A.M., Cercel, D.C. & Chiru, C. (2020). UPB at SemEval-2020 Task 6: Pretrained Language Models for Definition Extraction. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. pp. 737–745.
- Borg, C., Rosner, M. & Pace, G. (2009). Evolutionary algorithms for definition extraction. In *Proceedings of the 1st Workshop on Definition Extraction*. pp. 26–32.
- Bovi, C.D., Telesca, L. & Navigli, R. (2015). Large-scale information extraction from textual definitions through deep syntactic and semantic analysis. *Transactions of the Association for Computational Linguistics*, 3, pp. 529–543.
- Caspani, F., Ratnamogan, P., Linger, M. & Hajaiej, M. (2020). ACNLP at SemEval-2020 task 6: A supervised approach for definition extraction. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. pp. 479–486.
- Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L. & Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.
- Cui, H., Kan, M.Y. & Chua, T.S. (2004). Unsupervised learning of soft patterns for generating definitions from online news. In *Proceedings of the 13th international conference on World Wide Web*. pp. 90–99.
- Cui, H., Kan, M.Y. & Chua, T.S. (2005). Generic soft pattern models for definitional question answering. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*. pp. 384–391.
- Davletov, A., Arefyev, N., Shatilov, A., Gordeev, D. & Rey, A. (2020). Gorynych Transformer at SemEval-2020 Task 6: Multi-task Learning for Definition Extraction. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. pp. 487–493.
- De Benedictis, F., Faralli, S. & Navigli, R. (2013). Glossboot: Bootstrapping multilingual domain glossaries from the web. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. pp. 528–538.
- Espinosa-Anke, L. & Saggion, H. (2014). Applying dependency relations to definition extraction. In *International Conference on Applications of Natural Language to Data Bases/Information Systems*. Springer, pp. 63–74.
- Espinosa-Anke, L., Saggion, H. & Delli Bovi, C. (2015). Definition extraction using sense-based embeddings. In *Gupta P, Banchs RE, Rosso P, editors. International Workshop on Embeddings and Semantics (IWES'15); 2015 Sept 15; Alicante, Spain. [Place unknown]:[CEUR]; 2015.[6 p.]*. CEUR.
- Espinosa-Anke, L., Saggion, H., Ronzano, F. & Navigli, R. (2016). Extasem! extending, taxonomizing and semantifying domain terminologies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.
- Fahmi, I. & Bouma, G. (2006). Learning to identify definitions using syntactic features. In *Proceedings of the Workshop on Learning Structured Information in Natural Language Applications*.
- Fišer, D., Pollak, S. & Vintar, S. (2010). Learning to Mine Definitions from Slovene Structured and Unstructured Knowledge-Rich Resources. In *LREC*. Citeseer.
- Hearst, M.A. (1992). Automatic acquisition of hyponyms from large text corpora. In *COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics*.
- Izmailov, P., Wilson, A., Podoprikin, D., Vetrov, D. & Garipov, T. (2018). Averaging weights leads to wider optima and better generalization. In *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018*. pp. 876–885.
- Jeawak, S., Anke, L.E. & Schockaert, S. (2020). Cardiff university at semeval-2020 task 6: Fine-tuning bert for domain-specific definition classification. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. pp. 361–366.

- Jemec Tomazin, M., Podpečan, V., Pollak, S., Thi Hong Tran, H., Fajfar, T., Atelšek, S., Sitar, J. & Žagar Karer, M. (2023). Slovenian Definition Extraction evaluation datasets RSDO-def 1.0. URL <http://hdl.handle.net/11356/1841>. Slovenian language resource repository CLARIN.SI.
- Jemec Tomazin, M., Trojar, M., Žagar, M., Atelšek, S., Fajfar, T. & Erjavec, T. (2021). Corpus of term-annotated texts RSDO5 1.0.
- Jin, Y., Kan, M.Y., Ng, J.P. & He, X. (2013). Mining scientific terms and their definitions: A study of the ACL anthology. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. pp. 780–790.
- Kannan, M. & Ponnusamy, H.S. (2020). TüKaPo at SemEval-2020 Task 6: Def (n) tly Not BERT: Definition Extraction Using pre-BERT Methods in a post-BERT World. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. pp. 724–729.
- Kaparina, J. & Soboleva, A. (2020). DeftPunk at SemEval-2020 Task 6: Using RNN-ensemble for the Sentence Classification. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. pp. 697–703.
- Kenton, J.D.M.W.C. & Toutanova, L.K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*. pp. 4171–4186.
- Klavans, J.L. & Muresan, S. (2001). Evaluation of the DEFINDER system for fully automatic glossary construction. In *Proceedings of the AMIA Symposium*. American Medical Informatics Association, p. 324.
- Klavans, J.L. & Muresan, S. (2002). A method for automatically building and evaluating dictionary resources.
- Li, S., Xu, B. & Chung, T.L. (2016). Definition extraction with lstm recurrent neural networks. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*. Springer, pp. 177–189.
- 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. *arXiv preprint arXiv:1907.11692*.
- Navigli, R. & Velardi, P. (2010). Learning word-class lattices for definition and hypernym extraction. In *Proceedings of the 48th annual meeting of the association for computational linguistics*. pp. 1318–1327.
- Park, Y., Byrd, R.J. & Boguraev, B. (2002). Automatic Glossary Extraction: Beyond Terminology Identification. In *COLING*, volume 10. pp. 1072228–1072370.
- Podpečan, V., Pollak, S., Fišer, D., Vintar, Š. & Tran, T.H.H. (2023). Slovenian Definition Extraction training dataset DF_NDF_wiki_slo 1.0. URL <http://hdl.handle.net/11356/1840>. Slovenian language resource repository CLARIN.SI.
- Pollak, S. (2014a). Extracting definition candidates from specialized corpora. *Slovenščina 2.0: empirical, applied and interdisciplinary research*, 2(1), pp. 1–40.
- Pollak, S. (2014b). Semi-automatic domain modeling from multilingual corpora. *PhD thesis*. Ljubljana: Department of Translation Studies, Faculty of Arts.
- Pollak, S., Repar, A., Martinc, M. & Podpečan, V. (2019). Karst exploration: extracting terms and definitions from karst domain corpus. *Proceedings of eLex*, 2019, pp. 934–956.
- Pollak, S., Vavpetic, A., Kranjc, J., Lavrac, N. & Vintar, S. (2012). NLP workflow for on-line definition extraction from English and Slovene text corpora. In J. Jancsary (ed.) *11th Conference on Natural Language Processing, KONVENS 2012, Empirical Methods in Natural Language Processing, Vienna, Austria, September 19-21, 2012*, volume 5 of *Scientific series of the ÖGAI*. ÖGAI, Wien, Österreich, pp. 53–60. URL http://www.oegai.at/konvens2012/proceedings/10_pollak12o/.

- Ranasinghe, T., Plum, A., Orašan, C. & Mitkov, R. (2020). RGCL at SemEval-2020 Task 6: Neural Approaches to Definition Extraction. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. pp. 717–723.
- Reiplinger, M., Schäfer, U. & Wolska, M. (2012). Extracting glossary sentences from scholarly articles: A comparative evaluation of pattern bootstrapping and deep analysis. In *Proceedings of the ACL-2012 special workshop on rediscovering 50 years of discoveries*. pp. 55–65.
- Saggion, H. & Gaizauskas, R.J. (2004). Mining On-line Sources for Definition Knowledge. In *FLAIRS Conference*. pp. 61–66.
- Sanh, V., Debut, L., Chaumond, J. & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.
- Sarmiento, L., Maia, B., Santos, D., Pinto, A. & Cabral, L. (2006). Corpógrafo v3: From terminological aid to semi-automatic knowledge engine. In *quot; In Nicoletta Calzolari; Khalid Choukri; Aldo Gangemi; Bente Maegaard; Joseph Mariani; Jan Odjik; Daniel Tapias (ed) Proceedings of the 5 th International Conference on Language Resources and Evaluation (LREC'2006)(Genoa Italy 22-28 May 2006)*.
- Singh, A., Kumar, P. & Sinha, A. (2020). DSC IIT-ISM at SemEval-2020 Task 6: Boosting BERT with Dependencies for Definition Extraction. In *SemEval@ COLING*. pp. 710–716.
- Spala, S., Miller, N.A., Deroncourt, F. & Dockhorn, C. (2020). Semeval-2020 task 6: Definition extraction from free text with the deft corpus. *arXiv preprint arXiv:2008.13694*.
- Storrer, A. & Wellinghoff, S. (2006). Automated detection and annotation of term definitions in German text corpora. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*.
- Ulčar, M. & Robnik-Šikonja, M. (2021). SloBERTa: Slovene monolingual large pretrained masked language model.
- Velardi, P., Faralli, S. & Navigli, R. (2013). Ontolearn reloaded: A graph-based algorithm for taxonomy induction. *Computational Linguistics*, 39(3), pp. 665–707.
- Veyseh, A., Deroncourt, F., Dou, D. & Nguyen, T. (2020). A joint model for definition extraction with syntactic connection and semantic consistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34. pp. 9098–9105.
- Vintar, Š. & Martinc, M. (2022). Framing karstology: From definitions to knowledge structures and automatic frame population. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 28.
- Westerhout, E. (2009). Definition extraction using linguistic and structural features. In *Proceedings of the 1st Workshop on Definition Extraction*. pp. 61–67.
- Zhang, H. & Ren, F. (2020). Bertatde at semeval-2020 task 6: Extracting term-definition pairs in free text using pre-trained model. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. pp. 690–696.

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

Sabine Tittel¹

¹Heidelberg Academy of Sciences and Humanities, Seminarstraße 3, D-69117 Heidelberg
E-mail: sabine.tittel@hadw-bw.de

Abstract

Historical lexicography of the Romance languages currently finds itself in a difficult place since the funding of some important dictionaries ended. The newly launched project ALMA will contribute to the future of these dictionaries' content. ALMA combines methods of historical lexicography, text philology, corpus linguistics, and the history of sciences with a Linked Data approach and ontology development. It adopts a Pan-Romance perspective focusing on medieval Italian, French, and Occitan / Gascon within two knowledge domains, 'medicine' and 'law'. ALMA's goals include re-using, extending, further processing, and disseminating lexicographical data by integrating it into its work pipeline. This makes for benefits on both sides: Pivotal for the ALMA project is the anchoring of its philological and lexicological work within the framework of the entire languages examined by the dictionaries. The dictionaries, most notably those whose funding ended, profit by seeing their linguistic, textual, and historico-cultural knowledge put into new formats—e.g., Linked Data—, contexts—e.g., Pan-Romance—, and correlations—e.g., through linking to the historicized domain ontologies ALMA will develop. This introduces the valuable dictionary contents to a knowledge circulation that goes beyond their original scope and ensures its long-term re-use in a somewhat concealed way.

Keywords: Historical Lexicography; Medieval Romance Languages; Corpus Linguistics; Ontology; Linked Data

1. Introduction

Dictionaries of historical language stages are at the core of historical lexicology, text philology, and historiography. They provide the means for grounding research on stable knowledge of the language for generations of researchers. For example, to create a scholarly text edition, the permanent consultation of the dictionaries pertinent to the language and language stage of the respective text is vital. For only the development of a text glossary with the inventory of the lexemes and their meanings can enable and further the editor's understanding of the text. The development of this glossary is not only a philological but also lexicological process and thus needs dictionaries: The text represents, albeit recorded in one or several manuscripts and thus potentially modified, the *parole* of the author while the dictionaries analyze the language on the level of the *langue*.¹ The meaning of a word in a text—as part of the *parole*—cannot properly be grasped without its embedding in the semantic scope of the same word whose senses and sub-senses, its uses as metaphor, metonymy, or other figures of speech, are explained in the dictionaries.

¹ Ferdinand de Saussure, *Cours de linguistique générale*, dichotomy of *langue* and *parole*, edition for example by Wunderli (2013).

Hence, dictionaries both are and facilitate foundational research. Within our context of the historical Romance languages, these include the *Französisches Etymologisches Wörterbuch* (FEW, von Wartburg, 1922–) which presents the diachronic development of the French language until present-day, the *Lessico Etimologico Italiano* (LEI, Pfister, 1979–), the *Dictionnaire étymologique de l'ancien français* (DEAF, Baldinger, 1971–2021), the *Dictionnaire de l'occitan médiéval* (DOM, Stempel, 1996–2013), the *Dictionnaire onomasiologique de l'ancien gascon* (DAG, Baldinger, 1975–2021), the *Dictionnaire du moyen français* (DMF²). These are comprehensive, long-term, and internationally well-known dictionary projects of the medieval Romance languages that provide synopses of the knowledge of the particular historical language stage and regional specification.

Despite being important resources in the field of lexicography, in 2020 and 2021, the funding of several of these long-standing endeavors has come to an untimely end leading to the situation that the dictionaries «find themselves currently at a difficult juncture», Selig et al. (2023: 296). This concerns, amongst others, the DEAF, the DOM, and the DAG. Fortunately, the data, dictionary writing system, and digital framework of the latter have been adopted by the University of Zurich where it will be merged into the *Dictionnaire étymologique d'ancien gascon* (DEAG).³ However, the DAG as a printed oeuvre has come to an end. Much earlier already, in 2006, the financial support for the *Dictionnaire onomasiologique de l'ancien occitan* (DAO) also ended (Glessgen & Tittel, 2018) and in 2007, the *Diccionario del español medieval* (DEM, Müller, 1994–2005) was discontinued.⁴ Furthermore, the DMF, while still hosted by the ATILF institute⁵, has been solely edited in a while by its director Robert Martin after his retirement.

The newly launched long-term project ALMA (*Wissensnetze in der mittelalterlichen Romania / Knowledge Networks in Medieval Romance Speaking Europe*) takes this development into account. ALMA is an inter-institutional project with a funding period of 22 years carried out by the Heidelberg Academy of Sciences and Humanities (HAdW), the Bavarian Academy of Sciences and Humanities (BAdW), and the Academy of Sciences and Literature Mainz (ADW Mainz).⁶ One of ALMA's goals is to integrate four lexicographical state-of-the-art dictionaries—DAG, DEAF, DOM, and LEI—into its scientific concept and work pipeline to produce a highly comprehensive and valuable resource that far exceeds the typical consultation and quotation of dictionaries. This benefits both the ALMA project and the dictionaries, especially those whose funding has been cut.

In this paper, we introduce the ALMA project (Chapter 2) and show the manifold relations to lexicography: re-use, extension, further processing, and dissemination (Chapter 3). We continue with an evaluation of these relations that promise benefits for both the ALMA project and the re-used dictionaries through a minimal case study of the Middle French medical term *addicion* f. (Chapter 4) and close with a short conclusion (Chapter 5).

² Version 2020, ATILF – CNRS & Université de Lorraine, <http://www.atilf.fr/dmf/>; this and all following web publications have been accessed 2023-05-24.

³ And integrated into a larger, Pan-Romance dictionary *Lexique étymologique de la Galloromania médiévale* (LEGaMe) Glessgen (2023).

⁴ In a more promising development, the *Diccionario del español medieval electrónico*, DEMel (Arnold & Langenbacher-Liebgott, 2022–), has resumed the work on medieval Spanish by making the slip inventory (33,000 lemmata with about 900,000 attestations) and prospectively, the printed DEM digitally accessible.

⁵ *Analyses et traitement informatique de la langue française*, Nancy, <https://www.atilf.fr/>.

⁶ <https://www.hadw-bw.de/alma>; directed by Elton Prifti / Wolfgang Schweickard (ADW Mainz), Maria Selig (BAdW), and Sabine Tittel (HAdW).

2. Introducing ALMA

The ALMA project aims to investigate the interaction between language, knowledge and scholarship in the Middle Ages. The field of observation is the Romance cultural sphere that sees the emergence of new knowledge networks expressed in vernacular languages in the time period between around 1100 and 1500 AD. The project traces how medieval Italian, French, Occitan/Gascon⁷, and also Catalan and Spanish are developed into languages of knowledge and scholarship within new functional areas of language that are technically and conceptually complex. This will be exemplified by two knowledge domains, namely ‘medicine’ and ‘law’. These technical, ‘scientific’ languages depicting knowledge and scholarship are a particularly important part of the intellectual and cultural heritage of Europe. Concurrently, the Romance languages are major carriers of a cultural exchange in the Middle Ages that starts to establish the European identity as a knowledge society.

ALMA’s concept combines methods of linguistics, text philology, and the history of sciences with technologies of the digital humanities and ontology engineering.

Romance languages have a rich textual tradition.⁸ To make—a part of—this textual tradition accessible and to establish its empirical research basis at the same time, ALMA will compile credible, domain-specific corpora for ‘medicine’ and ‘law’. These corpora will consist of newly established text editions and of digitized works covering medieval Italian, Old French, Old Occitan, and Old Gascon, giving access to an important cultural sphere of medieval Romance-speaking Europe. The text selection rests on the substantial, decades-long experience gathered by the dictionaries DAG, DEAF, DOM, and LEI, which directly benefits ALMA.⁹ The corpus texts lay the foundation for the reconstruction of the main concepts and concept networks of the two knowledge domains. Applying quantitative methods of corpus linguistics (Hirschmann, 2019)—in particular absolute and relative frequency analyses and co-occurrence analyses—will help carve out these concepts. They provide the basis for the lexical-semantic studies that analyze the internal structure of the concept networks and the depth of their linguistic representation. The lexical-semantic studies also discuss the origin and dissemination of lexical innovations together with the new matters denoted, deepening the knowledge about medieval communication channels. Thus, these studies will evaluate the empirical, quantitative methods of corpus linguistics in a historico-linguistic context combined with the hermeneutical, qualitative approach of historical linguistics. It is at this stage that the lexicographic resources come into play.

3. Relations to Lexicography

The project’s relations to lexicography are manifold: ALMA (1) re-uses, (2) extends, (3) processes, and (4) disseminates existing and well-proven dictionary data.

⁷ For the long-lasting discussion of the differentiation of Occitan and Gascon, see Glessgen (2021); Selig et al. (2023: 266).

⁸ DEAFBibleI (Möhren, 2022), <https://alma.hadw-bw.de/deafbible/>, lists >80 (large and small) texts with medical content and >30 with law-related content for Old French alone. The text corpus of Old French legal documents, *Documents linguistiques galloromans* (DocLing, <http://www.rose.uzh.ch/docling>), comprises >2,200 medieval French charters (deeds of donation, contracts of purchase, inheritance matter, etc.) dating between 1205 and ca. 1450 AD.

⁹ Additionally, the work is supported by the analysis of complementary (Medieval) Latin and vernacular text corpora that are already available for digital research, e.g., the many editions of small legal documents provided by DocLing. This is particularly relevant for the Spanish and the Catalan textual traditions where ALMA will not create its own corpora.

3.1 Re-use

The workflow combining quantitative machine-driven with qualitative competence-linguistic methods is controlled by drawing on the state-of-the-art dictionaries: the lexicography of the Gallo- and Italo-romania, i.e., DAG, DEAF, DOM, LEI primarily, but also FEW and DMF, flanked by the dictionaries of the Iberoromania and of (medieval) Latin.

The cognitive step from a given lexeme, its absolute and relative frequency, and its co-occurrences—the result of the corpus analysis—to one or several concepts, is made by analyzing the meaning(s) of the lexeme in all of the text passages and in constant confrontation with the language system documented in the dictionaries. Self-evidently, ALMA follows standards of quoting dictionary entries. Furthermore, it has the unique advantage of being able to re-use—through database access—the published as well as raw data of the DEAF, DOM, LEI, as well as the DAG (depending on its ongoing integration into LEGaMe). This means being able to evaluate the source materials in the *fichiers* (slip inventories) of the dictionary resources, containing millions of paper slips with text references. Since the funding of these dictionaries, apart from the LEI, has recently ended, re-use through ALMA is an excellent means to keep the valuable data alive as part of an innovative workflow.

A second significant aspect of dictionary re-use concerns the bibliographical supplements of DEAF (DEAFBiblEI), DOM (DOMBibl¹⁰), and LEI (*Bibliografia Generale online / BiG*¹¹), all of which are reference works with immense value for studies on historical linguistics of the Romance languages and text philology. Based on the DEAFBiblEI model, which became the state-of-the-art work used by many monographs, journals, and other dictionaries, ALMA will create a critical research bibliography assessing primary literature—for and beyond the corpus texts—, secondary literature, and dictionaries. This will serve as the bibliographical groundwork entangled with the corpus texts and the lexical-semantic analyses, and also facilitate validation and enrichment of corpus text information. It will also be published as a stand-alone research instrument. While the original bibliographical works will be preserved as such, the pertinent data of DEAFBiblEI (for ‘medicine’ and ‘law’ in Old French) and of DOMBibl (Old Occitan / Gascon) will be fully integrated into the ALMA database and extended therein. The comprehensive LEI BiG (Italian) will be closely interlinked on the level of each mentioned siglum (through APIs for database communication), but will remain an external, independent resource since it is a vital module of an ongoing dictionary project.

3.2 Extension

As mentioned earlier, the lexical-semantic studies carried out by ALMA build on corpus material that has only been partly considered by dictionaries. Here, the ALMA corpora will enlarge the material basis for lexicography in a significant way. An example is the text edition of the *Chirurgia magna* by Gui de Chauliac, written after 1363 and translated into many, Romance and non-Romance languages (Tittel, 2004: 17-29). This key text of the field of medicine provided the foundation for didactic surgery and became very influential until the 17th c. The first treatise of the text in the oldest French manuscript (Montpellier, Ecole de Médecine H 184 [2nd third 15th c], f^{os} 14v^o-36v^o) is accessible through

¹⁰ <http://www.dom-en-ligne.de/>.

¹¹ <https://lei-digitale.it/>.

GuiChaulMT (Tittel, 2004) and its terminology found its way into DEAF, DMF, and FEW. Despite its importance, the French text as a whole (with 254 f^{os}) still lacks an edition (as well as translations). ALMA will fill this gap and in doing so, provide valuable data with great potential for research into the development of the language of medicine.¹²

Within the two domains ‘medicine’ and ‘law’, the lexical-semantic studies will add to, advance, or even replace entries of the four dictionaries DAG, DEAF, DOM, and LEI: The confrontation of the dictionary data with the new, comprehensive material accessible through the corpora will substantially extend the lexicographical knowledge documented so far. Merging corpus texts and dictionaries’ information will realize a vital communication between the *parole* of a text (of many texts, respectively) and the description of the *langue* by the lexicographical resources. Also, the dictionary data that is typically focused on a single language will be put into a multilingual, Pan-Romance context. This will shed new light on the terminology: the semantic scope of the lexemes in each language, the history of the lexemes and their etymology (*histoire du mot*), and the history of the designated concepts (*histoire du concept*) across the languages. It will thus enhance the comprehension of the inter-relatedness of the medieval languages stages of Italian, French, Occitan, and Gascon, which is hitherto scattered among the individual dictionary publications.

The lexical-semantic studies have many features similar to a dictionary:

1. (Multilingual) lexemes as the heads of lexical-semantic analyses; lemmatized in each language,
2. Registration of senses and sub-senses in a hierarchical, tree-like structure reflecting semantic shift,
3. State-of-the-art genus-differentia definitions of the senses following Möhren (2015: 407–417),
4. An apparatus—separated from the semantics—documenting the dated and regionally classified graphical realizations of the lexemes,
5. Contexts (taken from the corpus material) for encyclopedic illustration of the senses,
6. Discussion of the etymology and *histoire du mot* typical for many historical dictionaries,
7. Close-knit interlinking with other lexicographic resources and text corpora.

Since the ALMA project will have access to the online publications of DEAF, DOM, LEI, and potentially DAG (DEAG, respectively), it will be possible to indicate within these publications that a given dictionary entry is incorporated and advanced within a lexical-semantic study; we will return to this with our case study in Chapter 4.2.

3.3 Further Processing

An innovation of the ALMA project is the combination of the philological and linguistic approach with Semantic Web technologies. ALMA’s goals include modeling the project’s results as Linked Open Data (LOD¹³) in *Resource Description Framework* (RDF, Cyganiak et al., 2004–2014) using standard vocabularies such as OntoLex-Lemon (Cimiano et al., 2016). The advantages of modeling data as LOD comprise structural and conceptual

¹² See Tittel (2004: 53–58) for an evaluation of the findings of the first, French treatise.

¹³ <https://www.w3.org/DesignIssues/LinkedData.html>.

interoperability (through same format and shared vocabularies such as OntoLex-Lemon), accessibility (via standard Web protocols), and resource integration (through interlinking data), resulting in cross-resource access (Chiarcos et al., 2013). A pivotal aspect for establishing cross-language access to the content of historical linguistic resources—to words and their meanings—is lexico-semantic mapping: the mapping of concepts (of things) expressed through representations in historical languages (words) to an entity of an external, language-independent knowledge base of the Semantic Web (Tittel, accepted). To enhance this lexico-semantic mapping, ALMA will develop domain-specific ontologies for medicine and law. These ontologies will be historicized, taking into account the specificity of medieval explanation patterns. This bridges the historical semantic gap between historical concepts and entities of modern ontologies—for example, of modern physiology that differs significantly from the medieval humoral pathology and doctrine of pneumata—and prevents anachronistic classifications.

The LOD modeling covers the lexical-semantic studies (as well as text editions and bibliographical data), including the incorporated material of the four dictionaries. Also, the project takes a step forward in that it extends modeling to the original dictionary articles in their entirety, thus feeding full DOM, DEAF, LEI (and possibly DAG) entries as RDF resources into the Semantic Web. Preparatory work on the RDF-modeling of DEAF and DAG (Tittel & Chiarcos, 2018; Tittel, forthcoming) and LEI (Nannini, forthcoming) has already been successfully performed. The concepts represented by the lexical units of the dictionaries will be mapped to the entities of the historicized domain ontologies for medicine and law developed by ALMA.

The following code example shows an extract of a DEAF entry as LOD/RDF serialized in Turtle (Prud'hommeaux & Carothers, 2014) and automatically created from XML with XSLT and Python scripts:¹⁴

```

1 @prefix dbr:      <https://dbpedia.org/resource/> .
2 @prefix dct:      <http://purl.org/dc/terms/> .
3 @prefix deaf:     <https://deaf.ub.uni-heidelberg.de/lemme/> .
4 @prefix decomp:   <http://www.w3.org/ns/lemon/decomp#> .
5 @prefix lexinfo:  <https://lexinfo.net/ontology/3.0/lexinfo#> .
6 @prefix olia:     <http://purl.org/olia/olia.owl#> .
7 @prefix ontolex:  <http://www.w3.org/ns/lemon/ontolex#> .
8 @prefix rdfs:     <http://www.w3.org/2000/01/rdf-schema#> .
9 @prefix skos:     <http://www.w3.org/2004/02/skos/core#> .
10 @prefix vartrans: <http://www.w3.org/ns/lemon/vartrans#> .
11
12 # --- lexical entry -----
13 deaf:fiel a ontolex:LexicalEntry , ontolex:Word ;
14   lexinfo:partOfSpeech "m."@fr ,
15   lexinfo:Noun ;
16   lexinfo:gender lexinfo:male ;
17   ontolex:canonicalForm deaf:fiel_form_fiel .
18 deaf:fiel_form_fiel a ontolex:Form ;
19   ontolex:writtenRep "fiel"@fro .
20
21 # graphical variant
22 deaf:fiel ontolex:otherForm deaf:fiel_form_fel .
23 deaf:fiel_form_fel a ontolex:Form ;
24   ontolex:writtenRep "fel"@fro .
25

```

¹⁴ Examples of complete DEAF entries modeled as LOD in RDF can be found on GitHub, <https://github.com/SabineTittel/LexSemMapping/tree/main/results>.

```

26 # collocation "fiel de terre", sense sense1.h
27 deaf:fiel_de_terre a ontolex:LexicalEntry , ontolex:MultiwordExpression ;
28   decomp:subterm deaf:fiel ;
29   vartrans:lexicalRel lexinfo:collocation ;
30   rdfs:label "fiel de terre"@fro .
31
32 deaf:fiel_de_terre ontolex:sense deaf:fiel_sense1.h ;
33   ontolex:evokes deaf:fiel_sense1.h_lexConcept .
34
35 deaf:fiel_sense1.h a ontolex:LexicalSense ;
36   ontolex:isLexicalizedSenseOf deaf:fiel_sense1.h_lexConcept ;
37   ontolex:usage dbr:Metonymy ;
38   olia:hasRegister olia:TechnicalRegister ;
39   dct:subject dbr:Botany .
40
41 deaf:fiel_sense1.h_lexConcept a ontolex:LexicalConcept ;
42   skos:definition "plante herbacée [...], petite centaurée"@fr ;
43   ontolex:isConceptOf dbr:Centaurium_erythraea ;
44   ontolex:lexicalizedSense deaf:fiel_sense1.h .

```

3.4 Dissemination

ALMA will disseminate the pertinent dictionary articles in the form of RDF resources and make them accessible for semantic research driven by Semantic Web technologies, a contribution well beyond ALMA's core focus.

4. Evaluation: More than Old Wine in New Bottles

We see contributions leading to significant advancements on both sides: ALMA and lexicography.

4.1 Contribution of the Lexicographical Data to ALMA

4.1.1 Reflections on Corpus Integration and why Dictionaries Help

Limiting lexicological research to the material of a self-contained corpus is a problematic approach. All results generated by the analysis of a corpus, irrespective of its composition and size can only be relevant to the subset of the language represented in that corpus. This is because creating a corpus means drawing corpus borders which generate absences. E.g., studying a corpus of the works of a single, particular author, such as Chrétien de Troyes, the famous French poet and founder of the textual genre of the chivalric romance¹⁵ is interesting but does not reveal how his language differs from that of other authors. The insights gained from such a study will be limited (cp. Filatkina, 2009: 79).

ALMA chooses a discourse tradition—technical texts—as the constitutive feature of its corpus because it is necessary to limit the research material. And yet, it is incorrect to presume that only texts that are assumed to belong to this discourse tradition are relevant for the study of the terminology in question. In order to capture the essence of a term, the entire literature (texts of other technical domains, historiography, belles

¹⁵ <https://viaf.org/viaf/87681171/>. For his language, see the *Dictionnaire Électronique de Chrétien de Troyes*, <http://www.atilf.fr/dect/>.

lettres, etc.) is relevant in shedding light on the quality of the word in the language as a whole. For this reason, a corpus that included all—but only—the technical texts of the given domain, per se, would not be able to make valid statements about the existence and meaning of technical vocabulary. An example is the Old French nomenclature of navigation that occurs in the *Vie seint Edmond le rei*, an Anglo-Norman hagiographic poetry from ca. 1193: *dromunz, chalanz, esnekes, hallos* etc., Kjellman (1935: 2029-2034), all designations for particular ships. On the other hand, the internal differentiation of the corpus must be supported by comparison with other literary or technical texts that are accessible through dictionaries. E.g., deeds are more often about sales than is a *chanson de geste*, the medieval epic poem. This naturally leads to a more frequent use of lexemes like Old French *achat* “purchase” or *vente* “sale” in deeds than in a *chanson de geste*. However, this does not imply that they necessarily have to be diaphasically bound, Glessgen (2005: 226).¹⁶

Consequently, interpreting corpus material must be done with great caution, and conclusions about the language must be made only after a recontextualization within the language as a whole. This can be achieved by matching them against broader lexicographical works.¹⁷

4.1.2 Lemmatization

A crucial part of corpus constitution is the tokenization of the text involving lemmatization and part-of-speech annotation. This process can adapt the models already established by the dictionaries building on the standards of the disciplines. For Old French, for example, this is the lemma list of the DEAF that accords to current rules of lemmatization.

4.1.3 Anchoring the Lexical-Semantic Studies

The lexical-semantic studies will be written based on lexicographical grounding. While ALMA focuses on the languages of two particular domains, the dictionaries examine and describe the languages as a system with all functional areas, beyond the technical vocabulary in question. Thus, lexicography enables anchoring the findings within the framework of entire languages. This is crucial for a proper grasp not only of the technical terms to be analyzed but also of the words of the textual contexts, as well as of the contexts of each term. If the technical texts often reveal that a lexeme is of special interest because it designates a special thing and thus its sense definition makes for a new sense in the dictionaries, the consultation of the lexicographical resources might show the opposite: the history of the concept and of the lexeme with its etymology often makes clear that it is really only one concept and no new sense: «Comme réflexion de contrôle face à un ‘nouveau’ sens, on peut se dire que tout sens insolite est un sens erroné», Möhren (2015: 416).

¹⁶ See Coseriu (1980) for diaphasical, diastratical, and diatopical aspects of language.

¹⁷ Cp. Kabatek (2016: 4): «El corpus contiene lengua, naturalmente, pero el corpus no contiene *la lengua*, ni como objeto abstracto, ni como objeto concreto y mental. El corpus [...] nos ofrece una ventana que permite acceder a una parte de esta, pero no al todo, y deja, por tanto, abierta la especulación acerca de lo que no se puede ver». Also Oesterreicher (2006: 485-490) for examples of how, for 16th-century American Spanish, corpus linguistic research yields results that do not stand up to competence-linguistic scrutiny.

4.1.4 Shedding Light on Cross-Domain Relations

The lexical-semantic studies will also concentrate on identifying connections between the various knowledge domains. For medicine, the connection to the domain of astronomy can be seen, for example, in the polysemy of Old French *mirac* and its cognates: The lexemes represent the concept ‘abdominal wall’ and also designate the star Beta Andromedae. In the metabolic-pathological field particularly, such connections can often be observed and must be considered when analyzing the concepts of ‘healthy’ and ‘sick’. The comparison with lexicographical data from all knowledge domains helps clarify such connections.¹⁸

4.2 Contribution of ALMA to Lexicography

We argue that the ALMA project will make two major contributions to the lexicographical resources mentioned above: (1) The enlargement of the lexeme inventory and the enhancement of existing entries through a multilingual perspective and through new findings in new texts that are, furthermore, exploited in a way supported by the machine, (2) the modeling as LOD.

4.2.1 Enlargement of the Lexeme Inventory and Enhancement of Entries

In the following, we describe a minimal case study for our first argument, the impact on lexeme inventory and entries of the dictionaries: *Addicion* f. is—to the best of our current knowledge—a Middle French medical terminus expressing the concept ‘protuberance of an osseous or cartilaginous structure’. We find the terminus attested in Middle French GuiChaulMT 316; 318; 390; etc., defined as “éminence à la surface d’une structure osseuse ou cartilagineuse” in Tittel (2004: 285). Presuming that ALMA will study this concept, we look into the four dictionaries DAG, DEAF, DOM, and LEI:

DAG The dictionary was founded in 1955 by Kurt Baldinger¹⁹ (who also founded the DEAF) and was printed from 1975 to 2021; the preparation of the online version DAGél began in 2014 (Glessgen & Tittel, 2018: 805-808). The DAG has a checkered history due to changes in finances with several concept shifts and alterations in its material base with respect to the time span treated in the dictionary (originally the Gascon from ca. 1100 AD until the end of the 16th c, then until ca. 1300, then—for DAGél—until ca. 1500). DAGél was never made open to the public and is—since 2021—being turned into the DEAG and integrated into the Pan-Romance endeavor LEGaMe. As concerns our data quest for *addicion*, the data of the DAGél does not include a Gascon cognate of this term.²⁰ However, the medieval Gascon scripturality is almost exclusively limited to the text genre of documents (testaments, charters, court records, etc.). Hence, it is unsurprising to not find the term with a medical sense. Nevertheless, it is notable that no attestation at all (with whatever meaning) can be found.

¹⁸ The study on *mirac* also shows the difficulty in capturing the precise scope of a concept in a knowledge area that is still under development: the different authors use the lexeme *mirac* (in its realizations in the individual languages, respectively) in slightly different ways. Comparing the preliminary findings with dictionary knowledge will be helpful.

¹⁹ <https://viaf.org/viaf/109932631/>.

²⁰ We thank M. Glessgen for the database search.

- DEAF The DEAF was edited between the 1960s and 2021 and published in print (since 1974) and online (since 2010). The online edition DEAF*él* consists of a two-tier system consisting of DEAF*plus*, the scholarly dictionary edited for decades but then limited to letters D-K (approx. 10,000 entries), and DEAF*prés*, the dictionary's raw data of the remaining letters of the alphabet which is published only online: 1.5 million slips with references to more than 10 million attestations, pre-structured into >70,000 preliminary dictionary entries with maximum assistance by the machine. Since the letter A is not part of DEAF*plus*, we turn to DEAF*prés* with its valuable yet unverified material which has: ADDICION, <https://deaf.ub.uni-heidelberg.de/lemme/addition>²¹: “action d'ajouter qch., ce qu'on ajoute à qch; accroissement”. This ‘sense definition’ is developed from the slip material, the word family, and the etymon, and does not comply with good definition rules (accounting for the preliminary nature and very limited time spent on editing this DEAF*prés* entry), cp. above. A sub-sense denoting the anatomical concept attested in GuiChaulMT²² is missing.
- DOM The DOM, whose preparatory work began in the 1960s, was published from 1996 to 2013 in printed form; since then, it has only been accessible online as DOM*él*. DOM*él* integrates its own research and existing dictionaries into its publication to create a lemma list covering the whole alphabet. The entry production of DOM*él* follows the concept of «cumulative development» (editing the dictionary) and «incremental functionality» (creating access), Klein (2004: 28f.). As of 2023, DOM*él* combines 1,845 digitized DOM entries, 37,998 entries uniting entries of two other dictionaries under DOM lemmata, and 9,509 *mots nouveaux*, that is, lemmata not previously recorded in lexicography (Selig et al., 2023: 267). The latter represents a significant expansion of Old Occitan vocabulary in both quantitative and qualitative terms (ib. 270). It is obvious that, e.g., significant words of the domain of medicine such as *arteria* f., *arterial* adj., *vena* f., etc. are not treated (properly) by DOM*él* due to its time-frame, concept, and its focus mainly on troubadour lyrics and the pragmatic scripturality of the legal-administrative domain. In DOM 1, 175b ADICION f.²³, we find several sense definitions of which all but one are attested primarily in one medical text.²⁴ However, none of the listed senses corresponds to the concept we find in GuiChaulMT.
- LEI *LEI digitale* is the digital representation of the LEI, advanced through the benefits typically derived from the digital processing and concerning, e.g., entry editing and publication versioning (Prifti, 2022). In LEI 1, 627 ADDITIO²⁵, *addizione* f., we find a sense definition compiled of four approximate translations in modern Italian («équivalents (ou gloses traductives)», Möhren, 2015: 408) depicting two concepts, 「addition」 and 「supplement」 (“aggiunta, complemento; supplemento, integrazione”). The sense denoted by our *addicion* is not listed.

With *addicion* as a minimal case study and based on the current situations of the dictionaries, we argue that a significant enhancement of all dictionaries in question will be achieved by ALMA's contributions, both with respect to improving existing entries and to

²¹ Nota bene: DEAF*él* is currently moving to this address and will be accessible shortly.

²² And possibly in Middle French (1365) AmphYpL² 360 (Lafeuille, 1964), see DEAF*prés*: to be examined.

²³ See also online on <http://www.dom-en-ligne.de/dom.php?lhid=3f9nCHBVSBMrNwNkGSfoMr>.

²⁴ The lexeme is also attested in the *Leys d'Amor*, a treatise of Toulouse poetry from the middle of the 14thc, and in a document from Auvergne.

²⁵ See also on <https://online.lei-digitale.it/> without an entry-specific URL.

filling gaps in their lexeme inventory. In this case, this will be achieved by conducting a lexical-semantic study of the concept ‘protuberance of an osseous or cartilaginous structure’ represented by Old French *addicion* and possibly its cognates. Beyond ALMA’s own publication channels, the new findings can either be indicated by inserting a link into the respective dictionary entries (e.g., into the DEAF*pré* article ADDITION) pointing to the lexical-semantic study published by ALMA at <https://alma.hadw-bw.de>, or by integrating the research results with respect to each language directly into an extended version of the respective dictionary article.²⁶

ALMA foresees lexical-semantic studies for approx. 1,000 lemmata with all senses relevant for the two domains and a large number of graphical realizations in the four languages. Studies will be comprehensive: philological, linguistic, lexicological, lexicographical, and concatenated with the entities of the extralinguistic ontologies. Therefore, they have great potential for new findings and for significant enhancement of the lexicographical resources. A substantial part of the lexeme inventory covered by DEAF*pré* and DOMél will be expanded into valuable and well-researched articles; lexemes, senses, and more data will be added to DAG (DEAG, respectively) and LEI. All will be linked through ALMA.

4.2.2 Modeling as Linked Open Data

The modeling of the lexicographical resources as LOD creates a new publication channel for printed and digitally published works. We envisage the modeling of those dictionaries’ articles that are relevant for the lexical-semantic studies of ALMA (cp. Chapter 3.3). However, once data models and modeling workflows have been installed, one could consider extending the modeling to more dictionary articles. Thus, ALMA’s contribution could go beyond the scope of its own lexeme inventory, and dictionaries in their entirety could benefit from this approach. Offering the lexicographical data as LOD, the linguistic, textual, and historico-cultural knowledge documented therein will be placed within new contexts and correlations, and the dictionary contents will be introduced to a knowledge circulation wider than that of historical lexicography and linguistics. Naturally, the transfer of the dictionary contents to the new formats also includes linking to the extralinguistic, historicized domain ontologies developed by ALMA, as mentioned above. Through the lexico-semantic mapping to the ontologies, the dictionaries will be extended by an onomasiological-ontological component and the availability of their content will be improved by semantic access options. Publishing the dictionary resources as LOD will allow for their exploitation with the benefits of LOD. This is a significant enhancement of their visibility and re-usability within a global research context independent from their original publication form and place, language, and language stage. Overall, the LOD approach will create a frame-like architecture of historical Semantic Web resources fostering the prospective ALMA and dictionary LOD resources and simultaneously enforcing the historical resources of the LOD landscape.

²⁶ Following the example of the entries of ‘DEAF*pré* - Version révisée’, e.g., <https://deaf.ub.uni-heidelberg.de/lemme/alcothedem>.

5. Conclusion

Ceci n'est pas un dictionnaire, ALMA is *not a real dictionary* but it can be interpreted as a particular—*real*—representation of a dictionary, much like Magritte's pipe²⁷ but on another abstraction level. The project's scope includes an elaboration of a dictionary in the form of multilingual, concept-driven lexical-semantic studies—a quasi-monography for each concept—that is deeply rooted in long-approved approaches to lexicography. ALMA is less and more than a dictionary at the same time: It is less because it focuses only on a part of the language covered by the comprehensive dictionaries, i.e., on medical and juridical terminology; and it is more because the conceptual entanglement significantly benefits from the combination of re-using well-trying dictionary knowledge, with the addition of new corpus material, integrating machine-driven methods, and introducing a Pan-Romance perspective. This combination allows statements about (1) the extent to which communication spheres in the vernaculars had already been developed for expert cultures and (2) the extent to which these are connected to the Latin-dominated knowledge networks.

A further substantial and expanding aspect is the extralinguistic facet of the studies: the *histoire du concept* next to the *histoire(s) du mot*. This is not only expressed through textual discussion within the lexico-semantic studies but also through Linked Data modeling and ontology engineering. Two interdependent elements are crucial: (i) the development of hitherto non-existent historicized ontologies for medicine and law and (ii) the lexico-semantic mapping to entities of these and other extra-linguistic knowledge bases of the Semantic Web landscape.

6. Acknowledgements

We would like to thank the anonymous reviewers and Ragini Menon, Maria Selig, and Wolfgang Schweickard (ALMA) for helpful comments and feedback.

7. References

- Arnold, R. & Langenbacher-Liebgott, J. (2022–). *Diccionario del Español Medieval electrónico* (DEMel). Directed by Rafael Arnold and Jutta Langenbacher-Liebgott on the basis of the Fichero of the *Diccionario del español medieval* by Bodo Müller (Heidelberg), in collaboration with Anna-Susan Franke, Karsten Labahn, Caroline Müller, Martin Reiter, Stefan Serafin, and Robert Stephan. University of Rostock und University of Paderborn. URL <https://demel.uni-rostock.de>.
- Baldinger, K. (1971–2021). *DEAF*. *Dictionnaire étymologique de l'ancien français*, *founded by Kurt Baldinger, continued by Frankwalt Möhren and Thomas Städtler*. Québec / Tübingen / Berlin: Presses de L'Université Laval / Niemeyer / De Gruyter. *DEAFél*: <https://deaf.ub.uni-heidelberg.de>].
- Baldinger, K. (1975–2021). *DAG*. *Dictionnaire onomasiologique de l'ancien gascon*, *founded by Kurt Baldinger, directed in collaboration with Inge Popelar, Noline Winkler, continued by Martin Glessgen*. Tübingen / Berlin: Niemeyer / De Gruyter.
- Chiarcos, C., McCrae, J., Cimiano, P. & Fellbaum, C. (2013). Towards Open Data for Linguistics: Lexical Linked Data. In A. Oltramari, P. Vossen, L. Qin & E. Hovy (eds.)

²⁷ <https://collections.lacma.org/node/239578>.

- New Trends of Research in Ontologies and Lexical Resources: Ideas, Projects, Systems.* Berlin, Heidelberg: Springer, pp. 7–25.
- Cimiano, P., McCrae, J. & Buitelaar, P. (2016). *Lexicon Model for Ontologies: Community Report*, 10 May 2016. URL <https://www.w3.org/2016/05/ontolex/>.
- Coseriu, E. (1980). ‘Historische Sprache’ und ‘Dialekt’. In J. Göschel (ed.) *Dialekt und Dialektologie. Ergebnisse des Internationalen Symposions “Zur Theorie des Dialekts” Marburg/Lahn, 5.–10. Sept. 1977.* Franz Steiner Verlag, pp. 106–122.
- Cyганиак, R., Wood, D. & Lanthaler, M. (2004–2014). *RDF 1.1. concepts and abstract syntax: W3C recommendation 25 February 2014.* URL <https://www.w3.org/TR/rdf11-concepts/>.
- Filatkina, N. (2009). Historische formelhafte Sprache als “harte Nuss” der Korpus- und Computerlinguistik. Ihre Annotation und Analyse im HiFoS-Projekt. *Linguistik online*, 39, 3, pp. 75–95.
- Glessgen, M.D. (2005). Diskurstraditionen zwischen pragmatischen Vorgaben und sprachlichen Varietäten. Methodische Überlegungen zur historischen Korpuslinguistik. In A. Schrott & H. Völker (eds.) *Historische Pragmatik und historische Varietätenlinguistik in den romanischen Sprachen.* Göttingen: Universitätsverlag, pp. 207–228.
- Glessgen, M.D. (2021). Pour une histoire textuelle du gascon médiéval. *Revue de Linguistique Romane*, 85(339-340), pp. 325–384.
- Glessgen, M.D. (2023). Les Documents et analyses linguistiques de la Galloromania médiévale (GallRom): structure et potentiel interprétatif. In D. Corbella, Josefa Dorta & Rafael Padrón (eds.) *Perspectives de recherche en linguistique et philologie romanes.* Strasbourg: ELiPhi, pp. 1025–1044.
- Glessgen, M.D. & Tittel, S. (2018). Le Dictionnaire d’ancien gascon électronique (DAGél). In *Atti del XXVIII Congresso internazionale di linguistica e filologia romanza (Roma, 18-23 luglio 2016).* Strasbourg: ELiPi, pp. 805–818.
- Hirschmann, H. (2019). *Korpuslinguistik. Eine Einführung.* Metzler.
- Kabatek, J. (2016). *Lingüística de corpus y lingüística histórica iberorrománica.* Berlin: De Gruyter.
- Kjellman, H. (1935). *La Vie seint Edmund le rei.* Elander.
- Klein, W. (2004). Vom Wörterbuch zum Digitalen Lexikalischen System. *Zeitschrift für Literaturwissenschaft und Linguistik*, 34, pp. 10–55.
- Lafeuille, G. (1964). *Les commentaires de Martin de Saint-Gille sur les Amphorismes Ypocras.* Droz.
- Möhren, F. (2015). L’art du glossaire d’édition. In D. Trotter (ed.) *Manuel de la philologie de l’édition.* Berlin: De Gruyter, pp. 397–437.
- Möhren, F. (2022). *Complément bibliographique 2021.* Berlin / Boston: De Gruyter Akademie Forschung.
- Müller, B. (1994–2005). *Diccionario del español medieval*, volume 1-3. Winter: Heidelberg.
- Nannini, A. (forthcoming). La mappatura semantica del Lessico Etimologico Italiano (LEI). Possibilità, metodi e prospettive. In E. Prifti, L. Becker, J. Kuhn, C. Ossenkop & C. Polzin-Haumann (eds.) *Digitale romanistische Sprachwissenschaft. Stand und Perspektiven. Romanistisches Kolloquium XXXIV, Wien, November 2019.* Berlin: De Gruyter.
- Oesterreicher, W. (2006). Korpuslinguistik und diachronische Lexikologie. Fallbeispiel aus dem amerikanischen Spanisch des 16. Jahrhunderts. In W. Dietrich, U. Hoinkes, B. Roviró & M. Warnecke (eds.) *Lexikalische Semantik und Korpuslinguistik.* Tübingen: Narr, pp. 479–498.

- Pfister, M. (1979–). *LEI*. Lessico Etimologico Italiano, *founded by Max Pfister, directed by Elton Prifti and Wolfgang Schweickard*. Wiesbaden: Reichert.
- Prifti, E. (2022). Il LEI digitale. Un resoconto, con particolare attenzione alla dialettologia. In M. Cortelazzo, S. Morgana & M. Prada (eds.) *Lessicografia storica dialettale e regionale. Atti del XIV Convegno ASLI (Associazione per la Storia della Lingua Italiana) (Milano, 5-7 novembre 2020)*. Firenze: Franco Cesati Editore, pp. 293–314.
- Prud’hommeaux, E. & Carothers, G. (2014). RDF 1.1 Turtle: Terse RDF Triple Language. URL: <http://www.w3.org/TR/turtle/>.
- Selig, M., Reichle, E. & Schöffel, M. (2023). New Entries in the *Dictionnaire de l’ancien occitan*: some preliminary remarks on methodological and historical aspects. In N. Pomino, E.M. Remberger & J. Zwink (eds.) *From Formal Linguistic Theory to the Art of Historical Editions. The Multifaceted Dimensions of Romance Linguistics*. Göttingen: Brill / Vandenhoeck & Ruprecht Verlage, pp. 263–280.
- Stempel, W.D. (1996–2013). *DOM*. Dictionnaire de l’occitan medieval, *founded by Wolf-Dieter Stempel, continued by Maria Selig*. Berlin [i. a.]: De Gruyter.
- Tittel, S. (2004). *Die Anathomie in der Grande Chirurgie des Gui de Chauliac: Wort- und sachgeschichtliche Untersuchungen und Edition*. Tübingen: Niemeyer.
- Tittel, S. (accepted). Lexico-Semantic Mapping of a Historical Dictionary: An Automated Approach with DBpedia, *Proceedings of 4th Conference on Language, Data and Knowledge (LDK 2023)*.
- Tittel, S. (forthcoming). *Integration von historischer lexikalischer Semantik und Ontologien in den Digital Humanities*.
- Tittel, S. & Chiarcos, C. (2018). Historical Lexicography of Old French and Linked Open Data: Transforming the resources of the *Dictionnaire étymologique de l’ancien français* with OntoLex-Lemon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). GLOBALEX Workshop (GLOBALEX-2018)*. Paris: ELRA, pp. 58–66.
- von Wartburg, W. (1922–). *Französisches Etymologisches Wörterbuch (FEW)*. Bonn, Heidelberg, Leipzig/Berlin, Basel: ATILF. [Continued by O. Jänicke, C. T. Gossen, J.-P. Chambon, J.-P. Chauveau, and Yan Greub].
- Wunderli, P. (2013). *Ferdinand de Saussure: Cours de linguistique générale. Zweisprachige Ausgabe französisch-deutsch mit Einleitung, Anmerkungen und Kommentar*. Tübingen: Narr.

Towards a lexical database of Dutch taboo language

Gerhard B van Huyssteen¹, Carole Tiberius²

¹ Centre for Text Technology (CtexT), North-West University, Potchefstroom, South Africa

² Instituut voor de Nederlandse Taal, Leiden, The Netherlands

E-mail: Gerhard.VanHuyssteen@nwu.ac.za, Carole.Tiberius@ivdnt.org

Abstract

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

Keywords: Dutch; lexical database; swearword; taboo language

Warning: This paper contains content that may be offensive or upsetting.

1. Introduction

Despite giant strides that have been made over the past thirty years in digitalising and automating lexicographic work, resources for specialised purposes and non-mainstream languages are still often neglected. As a case in point, even though at least eighteen Dutch paper-based dictionaries of taboo words (see 2.1 for a definition) have been published over the past 45 years (i.e., as visible works of lexicography), none of these are available as (linked) lexical data that could be integrated in natural language processing (NLP) tools and applications (i.e., as invisible works of lexicography).

Lexical databases of taboo language (LDTLs) are specialised digital resources that could be used as sources of linguistic and extralinguistic knowledge in many natural language processing (NLP) systems (see 2.2). Although such an LDTL could be simply a wordlist, for our purposes we consider an LDTL a digital collection of linguistic constructions that has been annotated or enriched in some way (e.g., with part-of-speech information, offensiveness ratings, meanings), and that is structured (e.g., encoded in XML). Most often, the primary use of LDTLs is to recognise words that could be potentially offensive to a specified community of language users (e.g., children). Despite their immediate practical value, and despite the fact that “much work has been done on abusive language detection in general”, much remains to be learned about “lexical knowledge for the detection of abusive language” (Wiegand et al., 2018), as

well as about the development and implementation of LDTLs for languages other than English.

In this paper we will report on the first phase of a project¹ to develop a Dutch LDTL (**TaboeLex**) consisting of potentially offensive constructions (words, word groups, expressions) as linked open data (i.e., a freely available, structured, annotated lexicon that could be linked to other data in future). In section 2, we will give a definition of what we mean by taboo language, and we will set the scope of TaboeLex. Section 3 then describes the design of the database. Section 4 concludes the paper, outlining future work.

2. Definition and scope of TaboeLex

2.1 Taboo language

Referring to the term *swearing*, Stapleton et al. (2022: 2) point out that “precise definitions and criteria are sometimes difficult to pin down [..., e.g.,] whether swear words can be used with literal (as opposed to figurative) meaning”. For purposes of this project, we define *taboo language* as linguistic constructions that are potentially offensive to some users in some contexts; constructions are form-meaning pairings on a morphological, lexical or syntactic level (see Goldberg (2006) for an extended view). We therefore use *taboo language* as a hypernym to include other phenomena and/or synonyms like *swearing*, *cursing/cussing*, *maledicta*, *profanity*, *blasphemy*, *obscenity*, *vulgarity*, *euphemisms and dysphemisms*, *verbal abuse*, *verbal sparring*, *(racial) slurs*, *terms of abuse*, *insults*, *offensive language*, *dirty language*, etc.

Our definitions and categories are all based on an extensive review of literature from various disciplines that aim to define taboo language, identify types of taboo language, sources of taboo language, etc. Most influential were Hirsch (1985), Hoeksema (2019), Jay (2018), Jay and Janschewitz (2008), Lewandowska-Tomaszczyk et al. (2021), Ljung (2011), Ruitenbeek et al. (2022) and Van Sterkenburg (2019), while the following books were also formative in our thinking about taboo language: Andersson and Trudgill (1990); Jay (1992, 2000); McEnery (2006); Montagu (1967); Pinker (2007). To inform us on the values of attributes, we also scrutinised the tags and definitions in GSW (2007) and Van Sterkenburg (2001), in order to create curated lists of possible values (see 3.2).

Some features of taboo words that are relevant to this project, include the following:

¹ Ethical clearance for the research project was obtained through the Language Matters Ethics Committee of the North-West University (ethics number: NWU-00632-19-A7).

- **Morphosyntactic type:** Taboo constructions include linguistic material on various morphosyntactic levels of independence and compositionality; these types are implemented in TaboeLex as an element `<headwordType>`. In addition to words, it also provides for sub-word items (like affixes), reduced forms (like initialisms), and multiword expressions (MWEs) (see 3.1 for values and examples).
- **Taboo domain:** Much work has been done to identify and delineate the source or reference domains of taboo language, such as religion, sex, scatology, animals, death, disease, etc. Within the scope of this paper, suffice to note that a taboo ontology will be declared as part of the `<denotatum>` element, which is a child element of the `<sense>` element (see 3.2).
- **Taboo type:** While the literal vs. figurative meaning requirement for taboo constructions are still being debated, we take the stance that both constructions with literal meanings, and constructions with figurative meanings could be taboo. For example, while neutral, scientific terms (i.e., orthophemisms) like *penis* and *vagina* could be considered by most people in most contexts as non-taboo, they could still be offensive to some people in some contexts, e.g., they might be dysphemistic in front of one's grandparents at a Christmas dinner, or in a geography class for grade 5 learners.

This adds a layer of complexity to the development of LDTLs, since homonymous and polysemous constructions need to be handled appropriately. For example, *emmer* refers mostly to ‘bucket’ (container) – see for example the abridged Dutch dictionary, and the multilingual dictionaries in VDO (2021). However, in some rather obscure cases *emmer* could also refer to ‘an inferior person, specifically a prostitute’ (i.e., as an abusive term), or ‘female genitalia’ (i.e., as an obscenity), as reflected in the more comprehensive, unabridged Dikke Van Dale (DVD Online, 2022). This feature of taboo language is practically resolved by introducing the element `<tabooType>` that can be added to any sense of an entry (see 3.2).

- **Tabooness:** Tabooness ratings of constructions will differ between different social groups and are subject to change over time. It is therefore not only essential that constructions should be rated in terms of their observed tabooness in or for certain groups, but also that such ratings should be re-evaluated regularly. For example, it is the task of the British public regulator for communication services, Ofcom, to determine public attitudes towards offensive language on TV and radio, specifically when children are particularly likely to be listening (roughly speaking between 06:00 and 19:00) (Ipsos MORI, 2021a: 3). To this effect, they commission research reports roundabout every five years (Ipsos MORI, 2016, 2021b; Synovate UK, 2010; The Fuse Group, 2005) to determine which words are to be considered mild, moderate, or strong (Ipsos

MORI, 2021a: 4). Similar (but not necessarily comparable) investigations have been done for Dutch in 1998, 2001, 2007, and 2018 (Van Sterkenburg, 2001, 2008, 2019). The element `<tabooValue>` will capture this knowledge with attribute values on a scale ranging from `highlyTaboo` to `notTaboo`; see section 3.2 for other potential values.

- **Context dependence:** Whether a construction is taboo or not, is not only dependent on the situational and/or textual contexts (e.g., whether the derogatory meanings of *emmer* are activated or not), but also on the social context. The word *rambam* (‘undefined, imaginary illness’) appears only in taboo constructions, like *krijg de rambam* (‘get an illness’), but is not considered taboo in most social contexts. The prototypicality rating (`<tabooPrototypicality>`) will – to a large extent – account for situational, textual, and social contextual dependence of taboo constructions. Words that are taboo in all contexts (e.g., *oetlul* ‘jerk, wanker’) will get the value `alwaysTaboo`, while words that are rarely used in the taboo sense (like *emmer*), will have the value `rarelyTaboo` – see 3.2 for other potential values.
- **Intention and effects:** From a sociopragmatic point of view, taboo language is often defined as language with an expressive/emotive function (Jay, 2020: 39). Hirsch (1985) therefore made a strong case that a taxonomy of taboo language should be based first and foremost on the speech acts (Austin, 1962; Searle, 1969, 1979) in which expressions occur. Following this general approach, we therefore provide for three pragmatic-specific elements, viz. `<speechAct>` for the type of speech act, `<illocution>` for the speaker’s intention, and `<perlocution>` for the effect on the hearer (see 3.2).

2.2 Lexical databases of taboo language

We define LDTLs as digital, structured, enriched collections of linguistic constructions that are potentially offensive to some users in some contexts (e.g., in children’s books). When implemented in NLP systems as simple look-up lists (gazetteers) for filtering of results, they might sometimes also be called *blacklists*, *greylists*, *swearword stop lists*, or *profanity filters* (e.g., Shutterstock, 2020). Two prominent examples of LDTLs are the following:

- Hurtlex is a lexicon of 1,156 Italian “hate words” that were “linked to synset-based computational lexical resources such as MultiWordNet and BabelNet” (Bassignana et al., 2018).
- Taboo Wordnet is an online, synset-based Japanese resource that could “help detection systems regulate and curb the use of offensive words online” (Choo & Bond, 2021). It consists of 2,095 words with 912 synsets, and it is linked to the Open Multilingual Wordnet.

Besides proprietary lists that are not accessible in the open-data domain, there are also numerous data sets for various taboo-related domains available (see Nakov et al., 2021; Rosenthal et al., 2020; Wiegand et al., 2021; Wiegand et al., 2019; Wiegand et al., 2018; Zampieri et al., 2019b; Zampieri et al., 2020 for overviews of available material). The different tagging schemas of more than 60 such data sets have been compared by Lewandowska-Tomaszczyk et al. (2021), with the aim to create an ontology basis for offensive language identification, while also getting insight in how the concept *offensive* is understood across different projects. They use the term *offensive language* similar to how we use *taboo language* (see 2.1) as a superordinate term for all kinds of language phenomena (Lewandowska-Tomaszczyk et al., 2021: 7). Their proposed ontology of offensive language, together with their methodology for the detection of such language, hold the potential to play an important standardisation role with regards to the treatment of taboo language in the context of Linguistic Linked Open Data (LLOD). In the next phase of our project, their ontology will therefore be the first point of reference to which we will compare our own ontology.

Of utmost importance is that re-usability should be a compulsory design requirement of any LTDL. To make the data re-usable for multiple purposes in several different applications, the database should ideally be rich with as much information as possible – either in the database itself, or otherwise through links to other existing resources. By using subsets of data, or a selection of elements, attributes and/or values, the data could be used in a variety of practical NLP applications like some of the following:

- Offensive language identification (Zampieri et al., 2020) has been a prevailing topic in NLP for a number of years, especially with a view on hate speech, cyber-bullying and abuse detection on social media platforms (Akiwowo et al., 2020; Davidson et al., 2017; Fišer et al., 2018; Jarquín-Vásquez et al., 2020; Korotkova & Chung, 2023; Li et al., 2023; Mostafazadeh Davani et al., 2021; Nakov et al., 2021; Narang et al., 2022; Pradhan et al., 2020; Roberts et al., 2019; Rosenthal et al., 2020; Schmidt & Wiegand, 2017; Teh et al., 2018; Waseem et al., 2017; Zampieri et al., 2019a). The identification of taboo language is also an important aspect of sentiment analysis (Byrne & Corney, 2014; Cachola et al., 2018), especially since the speech acts and language associated with sentiment analysis can oftentimes be more subtle or indirect, e.g., by using humour (Ahuja, 2019; Ahuja et al., 2018; Bansal et al., 2020; Meaney et al., 2021), or irony and sarcasm (Frenda et al., 2022; Husain & Uzuner, 2021).
- More recently the evaluation of large language models for biased and toxic language (Osoba & Welser IV, 2017; Schäfer, 2023; Wiegand et al., 2019) have been pushed to the fore with the public availability of OpenAI’s GPT-4 and ChatGPT models. However, from a linguistic and user interface design perspective, our understanding of the implementation of these models in conversational artificial intelligent agents (e.g., speech assistants and chatbots), and especially the relation with taboo language, is still in its infancy.

- LDTLs have been used for many years in applications of text filtering; see Zhou (2019) for an elaborate evaluation of some of these, as well as his own improved implementation. These include, inter alia:
 - **predictive text filtering**, e.g., for search engines, keyboards on mobile phones, online text editors, etc.;
 - **suggestion filtering**, e.g., for spelling checkers and electronic dictionaries (especially dictionary apps for children) that should not suggest swearwords as corrections for ordinary typos;
 - **taboo language censoring**, i.e., redacting, modifying, replacing or removing a word in a text that matches a word in the LDTL; implemented typically as part of parental control software for text, audio, and video (see Porutiu (2023) for an overview and marketing reviews of a number of these applications);
 - **content filtering**, e.g., social media algorithms that (semi-)automatically delete posts or ban users, like Facebook’s profanity filter for Facebook Page, or spam filters used in email applications. Other examples of content filtering include e-lexicography tools for choosing good dictionary examples (Kilgarriff et al., 2008), or computer-assisted language learning systems that automatically selects suitable texts for learners (Belaid, 2016).

2.3 Dutch resources of taboo language

Dutch has a rather long tradition in taboo language research, going back to at least 1834 with an history-focused article by J.F. Willems titled *On some old Dutch curses, oaths and exclamations* [translated – the authors] (Willems, 1834). However, the first specialised printed dictionary focusing on language from a taboo domain only appeared in 1977 (EW, 1977). Since then, at least seventeen other printed dictionaries (or dictionary-like books) on various aspects of taboo language have been published (DBG, 1991/2021; GSW, 2007; GT, 1997; HEW, 1988; KDV, 1998; LNS, 1989; LOS, 1990; Lutz-van Elburg, 1990; Lutz-van Elburg & Jager, 1989; NSW, 1984; Van der Gucht et al., 2018; Van der Meulen et al., 2018; Van Lichtenvoorde & Van Lichtenvoorde, 1993; Van Sterkenburg, 2001; WAON, 2013; WEPCT, 2001; WPTG, 2020-2023). Of these, only three are available as digital data: GSW (2007); Van Sterkenburg (2001); WPTG (2020-2023). Since WPTG (2020-2023) is a general dictionary of slang, and therefore also contains many non-taboo constructions, we only use data from the other two dictionaries as candidate taboo constructions for TaboeLex.

One of the most prominent or most used look-up lists of Dutch taboo words (so to see), is the Dutch version of the *List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words*

(RolfBly, 2020).² This list was derived from the Dutch section of *The Alternative Dictionaries* (TAD, 2004), although it is not clear when this was done, and by whom it was done. RolfBly (2020) consists of 190 constructions: 165 one-word constructions, and 25 MWEs. While this list will be used in a next phase of the project as one of the baselines for evaluation, several potential problems with the list could already be identified:

- The list is not free of linguistic errors. These include:
 - four spelling errors (i.e., **johny* > *johnny*; **pijpbekkieg* > *pijpbekkie*; **tongzoeng* > *tongzoen*; **triootjeg* > *triootje*);
 - six errors related to obsolete orthographic forms due to spelling reforms in Dutch (i.e., **boerelul* > *boerenlul*; **bokkelul* > *bokkenlul*; **krentekakker* > *krentenkakker*; **kuttelikkertje* > *kutlikkertje*; **paardekop* > *paardenkop*; **paardelul* > *paardenlul*);
 - one compound that should be written as one word (i.e., **trottoir prostituée* > *trottoirprostituée*);
 - an ephemeral word that only exists in TAD (2004) and its derivatives (i.e., *hoempert*, apparently meaning ‘hard excrement’).
- The list contains only lemmas, e.g., *op+sodemieter · en* (up+tumble · INF ‘to fuck off’), and no other word forms, e.g., *op+ge · sodemieter · d* (PTCP). This is particularly problematic for purposes of look-up lists in applications using predictive text filtering, and suggestion filtering (see 2.2). In such applications, the input text cannot be lemmatised first, since filtering needs to happen in real-time and on the fly.
- The MWEs are only presented as lemmas, e.g., *op z’n sodemieter gev · en* (on his carcass give · INF ‘to beat the hell out of him’). There is therefore no indication of:
 - orthographic variants, e.g., related to the example above, *zijn/zn/zun* instead of *z’n*, the latter of which does not appear in the 5.9-billion-word nlTenTen20 corpus (Sketch Engine, 2020);
 - morphosyntactic variants, e.g., again related to the above example, *op zijn* (3SG.M) *sodemieter* accounts for only roughly half the cases in the nlTenTen20 corpus; *zijn* is followed by *hun* (3PL), *mijn* (1SG), *ons* (1PL), *de* (DET), and *her* (3SG.F);

² An older version (2014) of the list is available at <https://github.com/chucknorris-io/swear-words/blob/master/nl>, while the list is also reproduced elsewhere on the web.

- lexical variants, e.g., *krijg · en* (‘to get’) occurs more frequently than *gev · en* (‘to give’) on the righthand side of *sodemieter* in the nlTenTen20 corpus (Sketch Engine, 2020); or
 - syntactic variants, e.g., *geeft hem op zijn sodemieter* instead of *hem op zijn sodemieter geeft*.
- In addition, the MWEs are not always presented uniformly. Compare for instance the lemma *op z'n sodemieter geven* that is presented as a prepositional phrase [_{PP} *op_{PREP} z'n_{PN} sodemieter_N*], followed by the verb [*geven_V*]. However, the lemma *reet trappen, voor zijn* has the same [PP V] structure as the former example (i.e., [_{PP} *voor_{PREP} zijn_{PN} reet_N*] [*trappen_V*]), but is presented here as [*reet_N trappen_V*, *voor_{PREP} zijn_{PN}*]. Also, in most cases in the list, only bare verbs are added as lemmas, e.g., *bedonderen* or *belazeren* (both meaning ‘to swindle, take someone for a ride’). However, in the case of [*besodemieteren_V*] (also meaning ‘to swindle, take someone for a ride’) a copula verb phrase [*besodemieterd_{PTCP} zijn_{COP}*] (‘to have been swindled, taken for a ride’) is provided additionally as a separate lemma.
 - Numerous polysemous constructions that are most frequently used in a non-taboo way, are included in the list. Compare for instance *achter het raam zitten*, which is an ordinary phrase for ‘to sit in a window (looking at what’s happening outside)’. However, it is also rarely used with the meaning ‘to work as a prostitute’ (TAD, 2004), or ‘to present oneself in a prostitute-like manner’ (DVD Online, 2022). Also compare *welzijn · s+mafia* (welfare · LK+mafia ‘ineffective and meddling social workers corps’) in the list, which is always used unmarked in the Dutch mainstream media.
 - Many of the examples are general slang that is not taboo at all. Compare for instance *buffelen* (‘to hit; to work hard; to wolf down food’), *huisdealer* (‘drugs dealer associated with a certain establishment’), or *kanen* (‘to eat’; associated with slang in The Hague).
 - Many others are euphemisms, like *de hond uitlaten* (‘to let the dog out’), but which can also be used as a euphemism for ‘to urinate’. Another example is *de koffer induiken* (‘to jump in one’s bed’), which is mostly used euphemistically with the meaning ‘to have sex’.
 - Numerous expected candidates, i.e., highly frequent, highly taboo constructions, are not included in the list. These include words like *debiel* (‘mentally deficient’), *trut* (‘twat, cunt’), *kanker+wijf* (cancer+woman ‘stupid bitch’), and many racial slurs. The list also excludes many English taboo words that are used frequently in Dutch, like *bitch*, *fuck*, and *bullshit*.

A much better and unproblematic list is the *GRoninger OFFensive Lexicon* (GrofLex) (Van der Veen, 2020), a Dutch lexicon of abusive lemmas based on version 1.2 of the Dutch section of HurtLex (Basile, 2020) (see below for more details on Hurtlex). It consists of 847 one-word constructions only (no MWEs). The list has been annotated with part-of speech information, as well as the offensive category (what we call *denotatum* – see 3 below) of each lemma (e.g., ethnic slurs, physical disabilities and diversity, words related to religion, male genitalia, etc.). While the list still contains polysemous constructions (like *kuiken* ‘chicken’; *kalf* ‘calf’; *druif* ‘grape’), and orthophemisms (like *pretentieus* ‘pretentious’, *fascistisch* ‘fascist’, *snob* id.), it could be used fruitfully in a next phase of the project as another baseline for evaluation.

3. Design of the TaboeLex lexical database

Our goal is to design an LDTL for Dutch, of which the data can be integrated into various NLP applications and tools, but which can potentially also be useful for human users, or for linguistic research. The general principles and structure of TaboeLex is in line with most existing standards and encoding formats such as Ontolex-Lemon (Cimiano et al., 2016), DMLex (Měchura et al., 2023), LMF,³ and TEI Lex-0 (Tasovac et al., 2018). General aspects are briefly discussed in section 3.1, followed by those aspects that relates specifically to a LDTL in section 3.2. Figure 1 presents an illustrative example, with LDTL-specific information marked in red. The complete XML schema and documentation, plus eventually all the TaboeLex data, will be made available under a CC BY-SA 4.0 license.

³ <https://www.iso.org/standard/68516.html>

```

<lexicographicResource title="TaboeLex" language="ndl">
  <entry id="debiel-word-n">
    <headword>debiel</headword>
    <headwordType>word</headwordType>
    <partOfSpeech tag="noun" />
    <variantForm>dubiel</variantForm>
    <patternForm />
    <linkExternal gigantMolex="12324" />
    <sense>
      <denotatum>entity [person] [mental ability/health]</denotatum>
      <definition language="eng">mentally deficient person</definition>
      <example>
        <text>Mensen laat je toch niet zo opnaaien door die achterlijke
          debiel.</text>
        <source>nlTenTen20-23694165</source>
      </example>
      <tabooType value="dysphemism">epithet</tabooType>
      <tabooValue value="highlyTaboo"></tabooValue>
      <tabooPrototypicality value="alwaysTaboo"></tabooPrototypicality>
      <speechAct>
        <member value="insult">
          <member value="name-calling">
            <member value="abuse">
              </member>
            </member>
          </member>
        </member>
      </speechAct>
      <illocution>
        <member value="anger">
          <member value="disrespect">
            <member value="contempt">
              </member>
            </member>
          </member>
        </member>
      </illocution>
      <perlocution>
        <member value="offensive">
          <member value="derogatory">
            <member value="insulting">
              </member>
            </member>
          </member>
        </member>
      </perlocution>
      <relation type="synonym">
        <member idref="debiel-word-n" />
        <member idref="idiot-word-n" />
      </relation>
    </sense>
  </entry>

```

Figure 1: Sample entry for *debiel* ('retard; retarded')

3.1 General design

Following our definition of constructions as form-meaning pairings, each taboo construction in the database is defined by aspects related to form, and aspects related to meaning. Regarding form, we use common elements like <headword>, <headwordType>, <partOfSpeech> (of the headword), and <variantForm> (e.g., for variants like *f*ck*, *f@ck*, *fark*, etc. for the English loanword *fuck*). The element <headwordType> could be extended in future to provide more detailed subcategories, but currently has the following primary values (with Dutch examples):

- subword: for affixes (e.g., *·erik* in *bang·erik* (scared·NMLZ ‘coward’)), and affixoids (e.g., *kanker÷* ‘cancer’ used as an intensifier in *kanker÷homo* ‘bad gay man’)⁴;
- reductionForm: for initialisms like *WTF*;
- word: for the uninflected form of words, e.g., *neuk·en* (fuck·INF ‘to fuck’); and
- MWE: for multiword expressions like:
 - word groups, e.g., *kwark blaffen* (‘to ejaculate (male)’), where neither *kwark* (‘curd’), nor *blaffen* (‘to bark’) is taboo, but their combination in a word group is;
 - construction idiom, e.g., *krijg X* (‘get X’), used as an imprecation, where X can be various illnesses; and
 - fixed expression, e.g., *Ik kan kakken en pissen en u gemakkelijk missen* (‘I can shit and piss without missing you at all’).

The rationale behind the element `<patternForm>` is to include some kind of pattern representation for each headword: on the one hand to allow for the automatic identification of the headword in corpus data (cf. Gantar & Krek, 2022; Odiijk, to appear), and on the other hand to deal with the flexibility and variation that many MWEs exhibit. For single words (see Figure 1), the pattern representation is the same as `<headword>`. For verbal MWEs, the pattern representation is a finite sentence, similar to the way in which patterns are being described in the Corpus Pattern Analysis approach of Hanks (2013). However, rather than using semantic types in the argument slots, we use dummies such as *iemand* ‘someone’, and *iets* ‘something’. See also the recently compiled DUCAME⁵ (*DUtch CAnonicalised Multiword Expressions*) resource, and the pattern descriptions in the project *Woordcombinaties*⁶.

The last aspect related to the form of an entry, involves the representation of all related word forms of a lemma, e.g., the verb *neuk·en* (‘to fuck’) has the grammatical forms *neuk* (1SG), *neuk·t* (2/3SG), *neuk·te* (SG.PST), *neuk·ten* (PL.PST), and *ge·neuk·t* (PTCP). Moreover, a comprehensive LDTL should ideally not only include grammatical forms, but also compounds (like *vuist+neuk·en* (fist+fuck·INF ‘to fist fuck’)), and derivations (like *neuk·er* ‘fucker’). This morphological information will be resolved in TaboeLex by means of links (`<linkExternal>`) to another lexical database, viz.

⁴ We use the following notations: middle dot (`·`) for affix boundaries; divide symbol (`÷`) for affixoid boundaries; plus symbol (`+`) for compound boundaries.

⁵ <https://surfdrive.surf.nl/files/index.php/s/2Maw8O0QTPH0oBP>

⁶ <https://woordcombinaties.ivdnt.org>

GiGaNT-Molex⁷, the modern part of the computational lexicon of the Dutch language, compiled by the Dutch Language Institute. Because it is linked to GiGaNT-Molex, the full inflectional paradigms and word-formation families of the headwords need not be stored in TaboeLex itself. Instead, this information can be retrieved dynamically from GiGaNT-Molex, if required. This also pertains to MWEs, which are included in GiGaNT-Molex as a whole, and with individual components linked to the appropriate lemmas. This element could also be used in future to link TaboeLex data to other resources, such as thesauri, translation dictionaries, etc.

All information related to the meaning side of a construction are accommodated under the <sense> element. While most of its children elements are taboo-specific (see 3.2), three common elements are included, viz. <definition> (in English); <example>, including the <text> and reference to the <source>; and <relation> to represent lexical relations like synonyms and antonyms.

3.2 LDTL-specific design feature

Various elements, attributes, and/or values that are specific to LDTLs have been added to the design. These are all part of the <sense> element since their values can vary depending on which sense of the word is involved; see the information in red in Figure 1. The taboo-specific elements are the following:⁸

- <denotatum>: The denotata on a superordinate level are: event; relation; state; entity; locale; process. Subtypes provide for constructions related to specific domains; for example, the exonymic epithet *kaas+kop* (cheese+head ‘Dutch person’) will have the value entity [person] [inhabitant, citizen], while a euphemistic verb like *drukk · en* (press · INF ‘to defecate’) will be process [body] [substance] [excretion].
- <tabooType>: We distinguish four main taboo types on lexicopragmatic grounds, viz.:
 - orthophemism (e.g., *penis*);
 - euphemism (e.g., *klok-en-hamer-spel* clock-and-hammer-game ‘penis’);
 - dysphemism (e.g., *paal* pole ‘penis’); and

⁷ <https://ivdnt.org/corpora-lexica/gigant/>

⁸ Since it is impossible in terms of space restrictions to list all possible values for all elements or attributes here, these will be made available as part of the XML schema and documentation; suffice to present here some illustrative examples.

- witticism, i.e., constructions that were created originally with the purpose to be humorous (e.g., *sperma+spuiter* sperm+gusher ‘penis’).

Additionally, we also provide for constructions that can be both euphemistic or dysphemistic, like *aap* (monkey ‘penis’); these are eu-/dysphemism. Following Hoeksema (2019), we also have a category *rudeImperative*, for expressions like *sterf aan bloedpoep* (‘die of bloody diarrhoea’). Including subcategories (not discussed here) and a category *other* (for miscellaneous cases), `<tabooType>` has a total of 16 values.

- `<tabooValue>`: To indicate to what degree the construction is generally considered to be taboo, a Likert-like scale of values are available: *highlyTaboo*; *moderatelyTaboo*; *slightlyTaboo*; *notTaboo* (e.g., for orthophemisms). Since assignment of these values will be based on empirical research by Van Sterkenburg (2019), an additional value, *unspecified*, is required for constructions for which such empirical data is not available.
- `<tabooPrototypicality>`: The prototypicality of a taboo construction is expressed here as a value of its prominence in multiple sources as an exclusively taboo construction (more prototypical), or not (less prototypical). These values are also expressed on a Likert-like scale: *alwaysTaboo*; *oftenTaboo*; *sometimesTaboo*; *rarelyTaboo*. In addition to an *unspecified* value like above, a sixth value is required for constructions that are euphemistic.
- `<speechAct>`: We distinguish 32 values that can be used to complete the leading sentence: “This lemma is mostly/often used in/as an act of ...”. These values range from very specific (e.g., *blasphemy* or *self-malediction*), to general (e.g., *expressivenessNegative*), and include also values for “positive” speech acts (e.g., *expressionPhysicalSensationPositive*). A sense can be assigned multiple values.
- `<illocution>`: A total of 60 illocutionary intentions have been identified in the literature. They complete the leading sentence: “This lemma is often used to express ...”, with values like *dislike*, *disgust*, *admiration*, *aestheticAppreciation*, *horror*, etc. Again, a sense can be assigned multiple values.
- `<perlocution>`: To complete the leading sentence: “This lemma is often used to be / perceived as being ...”, we distinguish 16 values like *offensive*, *politicallyCorrect*, *racist*, *jocular*, etc. As with the previous two elements, multiple values can be assigned to a sense.

4. First evaluation, conclusions, and future work

To populate TaboeLex as far as possible automatically in the next phase of the project, we will compile a combined list of candidate taboo constructions based on existing Dutch taboo-language dictionaries, which will then be curated based on corpus material. In addition, we will use the labels that are used for taboo constructions in these dictionaries as a second set of seed terms in a bootstrapping fashion to extract increasingly more data from existing resources, specifically dictionaries developed and maintained by the Dutch Language Institute. Thus, we divide the labour between two separate sets of seed terms: a list with macrostructure seed terms, and one with microstructure seed terms.

The list of macrostructure seed terms (or candidate <headword> list) consists of two primary sources, and several secondary sources. The two primary sources are two well-known, published dictionaries that are available as open, unparsed, digital data,⁹ viz. GSW (2007) and Van Sterkenburg (2001). We consider them primary, since they are assumed to be authoritative on whether a given construction is taboo or not. Our secondary sources are considered less authoritative, since they are lists that are generally available (and editable) from the internet. These include a list of lemmas tagged as *pejoratief* (pejorative) and *scheldwoord* (swearword) from Wiktionary (Wiktionary (NL), 2023), and a list of Dutch sexual slang and euphemisms from Wikipedia (Wikipedia (NL), 2023). RolfBly (2020) and Van der Veen (2020) will not be included in the candidate list, so that we can use them as part of our quality assessment.

An initial seed list shows that most taboo constructions only occur in one of the primary or secondary sources – see Table 1. The small overlap between the two printed dictionaries (GSW, 2007; Van Sterkenburg, 2001) can be ascribed to their different coverage of semantic domains: while GSW (2007) includes all kinds of taboo words from a variety of domains, Van Sterkenburg (2001) is more focused on taboo constructions related to oaths, curses, and (self-)maledictions. Similarly, the relatively small overlap between the other lists may also be due to a difference in focus, scope or aim of the respective lists, with the greatest overlap (37,4%) between GSW (2007) and (Wiktionary (NL), 2023).

⁹ We distinguished between *parsed digital data* (e.g., a lexical/lexicographic database); *unparsed digital data* (e.g., a word document with systematic formatting); and *digital documents/files* (e.g., scanned PDF documents). All these types can be *open* (i.e., available for research and development), or *proprietary* (i.e., not available).

	n	Van Sterkenburg (2001)	Wiktionary (NL) (2023)	Wikipedia (NL) (2023)
GSW (2007)	2,619	82 (4,2%)	382 (37,4%)	68 (6,7%)
Van Sterkenburg (2001)	1,973		31 (3,0%)	74 (7,3%)
Wiktionary (NL) (2023)	1,022			44 (4,3%)
Wikipedia (NL) (2023)	1,015			
Total	6,629			
Unique	5,295			

Table 1: Lemma lists, with number and percentage of shared lemmas between lists

The resulting candidate list will be further populated by extending it with headwords from other lexical resources that are labelled with one of the microstructure seed terms, i.e., constructions that occur either as tags in existing dictionaries (not only taboo dictionaries), or in the definitions of such dictionaries. In English, these would include stylistic tags like *vulgar* or *obscene*, and orthophemisms like *male genitalia* or *faeces*. Our initial list of microstructure seed terms is based on the tags and definitions used in GSW (2007) and Van Sterkenburg (2001). Initial results show that some tags do indeed result in new candidates, but that manual inspection of the results is needed. For example, a label such as *straattaal* (lit. street language 'bad language') produced one result in the ANW¹⁰, i.e., *straatbijbel* (lit. street Bible, a version of the Bible meant for young people), which is indeed written in a type of informal, street language, but which is clearly not a potential taboo construction.

To check the validity of the taboo constructions in our candidate list, we will check the constructions on the list against corpus data (and this information will be included in the database). A small pilot study shows that simply checking for occurrences in a corpus is not enough. The frequency counts require manual inspection of the data as some candidate constructions do occur in the corpus, but not as taboo constructions. For instance, all occurrences of *God in de hoge hemel* ('God in the highest heaven') and *God vergeef me* ('God forgive me') in the nlTenTen20 corpus can be considered as non-offensive. Furthermore, (normalised) frequencies can differ substantially between different types of corpora. As taboo constructions may be more likely to occur in certain types of texts than others, this is not unexpected but needs to be considered when interpreting frequency data. Moreover, the fact that a taboo construction does not occur in the corpus data does not automatically imply that it should be removed from the list.

¹⁰ <https://anw.ivdnt.org/search>

Once TaboeLex is populated with the curated list of taboo constructions, the lexicographic editing process will start. The very first step will be to validate the ontology of our annotation schema against (a) other similar ontologies, notably the one of (Lewandowska-Tomaszczyk et al., 2021); and (b) real-world data. Editing will therefore be done in a modular way, gradually refining not only the annotation schema, but also the amount of information for each taboo construction in the database.

5. References

- Ahuja, V. (2019). *Computational Analysis of Humour*. Master of Science. Hyderabad: International Institute of Information Technology.
- Ahuja, V., Mamidi, R., & Singh, N. (2018). From Humour to Hatred: A Computational Analysis of Off-Colour Humour. In M. Zhang, V. Ng, D. Zhao, S. Li, & H. Zan (eds.) *Natural Language Processing and Chinese Computing*. Cham: Springer International Publishing, pp. 144–153. https://doi.org/10.1007/978-3-319-99501-4_12.
- Akiwowo, S., Vidgen, B., Prabhakaran, V., & Waseem, Z. (eds.) (2020). *Proceedings of the Fourth Workshop on Online Abuse and Harms*. Association for Computational Linguistics. <https://aclanthology.org/2020.alw-1.0>.
- Andersson, L. G., & Trudgill, P. (1990). *Bad Language*. London: Penguin.
- Austin, J. L. (1962). *How to do things with words*. Oxford.
- Bansal, S., Garimella, V., Suhane, A., Patro, J., & Mukherjee, A. (2020). Code-switching patterns can be an effective route to improve performance of downstream NLP applications: A case study of humour, sarcasm and hate speech detection. *arXiv pre-print server*(arXiv:2005.02295v1 [cs.CL]). <https://doi.org/10.48550/arXiv.2005.02295>. (24 May 2023)
- Basile, V. (2020). *Hurtlex NL lexicon version 1.2* [GitHub repository]. <https://doi.org/https://github.com/valeriobasile/hurtlex/tree/master/lexica/NL/1.2>. (24 May 2023)
- Bassignana, E., Basile, V., & Patti, V. (2018). Hurtlex: A Multilingual Lexicon of Words to Hurt. In *CLiC-it*. <https://doi.org/10.4000/BOOKS.AACCADEMIA.3085>.
- Belaid, A. M. (2016). *The Localisation of the PARSNIP Model and Authentic Materials*. Chisinau: Scholars' Press.
- Byrne, E., & Corney, D. (2014). Sweet FA: Sentiment, Swearing and Soccer. SoMuS ICMR 2014 Workshop, Glasgow, Scotland, 01 April.
- Cachola, I., Holgate, E., Preoțiu-Pietro, D., & Li, J. J. (2018). Expressively vulgar: The socio-dynamics of vulgarity and its effects on sentiment analysis in social media. Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, New Mexico, August 20-26.
- Choo, Y. H. M., & Bond, F. (2021). Taboo Wordnet. 11th GlobalWordnet Conference, Potchefstroom, South Africa, 18-21 January.
- Cimiano, P., McCrae, J. P., & Buitelaar, P. (2016). *Lexicon model for ontologies:*

Community report.

- Davidson, T., Warmesley, D., Macy, M., & Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. International AAAI Conference on Web and Social Media (ICWSM), Montréal, Québec, Canada, May 15–18.
- DBG: *Duizend bommen en granaten: Scheldwoordenboek van Haddock*. (1991/2021). Edited by A. Algoud. Brussel: Casterman.
- DVD Online: *Dikke Van Dale Online (Van Dale Groot woordenboek van de Nederlandse taal)*. (2022). Edited by. Utrecht: Van Dale Uitgevers.
- EW: *Erotisch woordenboek*. (1977). Edited by H. Heestermans, P. Van Sterkenburg, & J. v. d. V. Van der Kleij. Baarn: Erven Thomas Rap.
- Fišer, D., Huang, R., Prabhakaran, V., Voigt, R., Waseem, Z., & Wernimont, J. (eds.) (2018). *Proceedings of the Second Workshop on Abusive Language Online (ALW2)*. Brussels, Belgium: Association for Computational Linguistics. <https://aclanthology.org/W18-5100>.
- Frenda, S., Cignarella, A. T., Basile, V., Bosco, C., Patti, V., & Rosso, P. (2022). The unbearable hurtfulness of sarcasm. *Expert Systems with Applications*, 193, 116398. <https://doi.org/https://doi.org/10.1016/j.eswa.2021.116398>. (24 May 2023)
- Gantar, P., & Krek, S. (2022). Creating the lexicon of multi-word expressions for Slovene. Methodology and structure. XX EURALEX International Congress, Mannheim.
- Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press.
- GSW: *Groot scheldwoordenboek: van apenkant tot zweefteef*. (2007). Edited by M. De Coster. Antwerpen: Standaard.
- GT: *Gespierde taal. Verbaal geweld voor in het buitenland: Beknopt scheldwoordenboek Nederlands-Engels, -Duits, -Frans en -Spaans*. (1997). Edited by J. Frijters. Zutphen: Alpha.
- Hanks, P. (2013). *Lexical analysis: Norms and exploitations*. Cambridge: The MIT Press.
- HEW: *Homo-erotisch woordenboek*. (1988). Edited by A. Joustra. Amsterdam: Thomas Rap. https://dbnl.org/tekst/jous008homo01_01/. (24 May 2023)
- Hirsch, R. (1985). Taxonomies of swearing. In L.-G. Andersson & R. Hirsch (eds.) *Perspectives on Swearing*. Göteborg: Department of Linguistics, University of Göteborg, pp. 37–59.
- Hoeksema, J. (2019). Taboo terms and their grammar. In K. Allan (ed.) *The Oxford Handbook of Taboo Words and Language*. Oxford: Oxford University Press, pp. 160–179.
- Husain, F., & Uzuner, O. (2021). Leveraging Offensive Language for Sarcasm and Sentiment Detection in Arabic. 6th Arabic Natural Language Processing Workshop, Kyiv, Ukraine, 19 April.
- Ipsos MORI. (2016). *Attitudes to potentially offensive language and gestures on TV and radio*. <http://stakeholders.ofcom.org.uk/binaries/research/tv->

- research/Offensive-language/Offensive-Language-2016-report.pdf. (24 May 2023)
- Ipsos MORI. (2021a). *Public attitudes towards offensive language on TV and Radio: Quick Reference Guide*.
https://www.ofcom.org.uk/__data/assets/pdf_file/0020/225335/offensive-language-quick-reference-guide.pdf. (24 May 2023)
- Ipsos MORI. (2021b). *Public attitudes towards offensive language on TV and Radio: Summary Report*.
https://www.ofcom.org.uk/__data/assets/pdf_file/0021/225336/offensive-language-summary-report.pdf. (24 May 2023)
- Jarquín-Vásquez, H. J., Montes-y-Gómez, M., & Villaseñor-Pineda, L. (2020). Not All Swear Words Are Used Equal: Attention over Word n-grams for Abusive Language Identification. In K. M. Figueroa Mora, J. Anzures Marín, J. Cerda, J. A. Carrasco-Ochoa, J. F. Martínez-Trinidad, & J. A. Olvera-López, *Pattern Recognition*. Cham. https://doi.org/10.1007/978-3-030-49076-8_27.
- Jay, T. B. (1992). *Cursing in America: A psycholinguistic study of dirty language in the courts, in the movies, in the schoolyards and on the streets*. Amsterdam: John Benjamins.
- Jay, T. B. (2000). *Why we curse: A neuro-psycho-social theory of speech*. Amsterdam: John Benjamins.
- Jay, T. B. (2018). The psychology of expressing and interpreting linguistic taboos. In *The Oxford Handbook of Taboo Words and Language*. pp. 76–95. <https://doi.org/10.1093/oxfordhb/9780198808190.013.5>.
- Jay, T. B. (2020). Ten issues facing taboo word scholars. In N. Nassenstein & A. Storch (eds.) *Swearing and Cursing*. Berlin: De Gruyter Mouton, pp. 37–52. <https://doi.org/10.1515/9781501511202-002>.
- Jay, T. B., & Janschewitz, K. (2008). The pragmatics of swearing. *Journal of Politeness Research. Language, Behaviour, Culture*, 4(2). <https://doi.org/10.1515/jplr.2008.013>.
- KDV: *Krijg de vinkentering! 1001 Nederlandse en Vlaamse verwensingen*. (1998). Edited by E. Sanders & R. Tempelaars. Amsterdam: Uitgeverij Contact.
- Kilgarriff, A., Husák, M., McAdam, K., Rundell, M., & Rychlý, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus. 13th EURALEX International Congress, Spain, July.
- Korotkova, E., & Chung, I. K. Y. (2023). Beyond Toxic: Toxicity Detection Datasets are Not Enough for Brand Safety. *arXiv preprint*(arXiv:2303.15110v1 [cs.CL]). <https://doi.org/10.48550/arXiv.2303.15110>. (24 May 2023)
- Lewandowska-Tomaszczyk, B., Žitnik, S., Bączkowska, A., Liebeskind, C., Mitrović, J., & Oleškevičiene, G. V. (2021). LOD-connected offensive language ontology and tagset enrichment. 42nd Conference on Very Important Topics (CVIT 2016), RWTH Aachen.
- Li, Z., Cabello, L., Yong, C., & Hershovich, D. (2023). Cross-Cultural Transfer Learning for Chinese Offensive Language Detection. *arXiv pre-print server*. <https://doi.org/arXiv:2303.17927v1> [cs.CL]. (24 May 2023)

- Ljung, M. (2011). *Swearing: A Cross-Cultural Linguistic Study*. Palgrave Macmillan.
- LNS: *Luilebol! Het Nederlands scheldwoordenboek*. (1989). Edited by H. Heestermans. Amsterdam: Thomas Rap.
- LOS: *Lik op stuk: Nieuw Nederlands woordenboek van agressief taalgebruik*. (1990). Edited by D. De Bleecker, P. Thomas, & J. Van Haver. Tiel: Lannoo.
- Lutz-van Elburg, I. (1990). *More Dutch you won't learn in class: not for hypocrites*. Rotterdam: Wilkerdon.
- Lutz-van Elburg, I., & Jager, P. C. W. (1989). *Dutch you won't learn in class (not for hypocrites)*. Lelystad: Zander Media Service.
- McEnery, T. (2006). *Swearing in English: Bad Language, Purity and Power from 1586 to the Present*. London: Routledge.
- Meaney, J. A., Wilson, S., Chiruzzo, L., Lopez, A., & Magdy, W. (2021). SemEval-2021 Task 7: HaHackathon, Detecting and Rating Humor and Offense. 15th International Workshop on Semantic Evaluation (SemEval-2021), Online, August.
- Měchura, M., Filip, D., & Krek, S. (2023). *Data Model for Lexicography Version 1.0* <https://docs.oasis-open.org/lexidma/dmlex/v1.0/wd01/dmlex-v1.0-wd01.html>. (24 May 2023)
- Montagu, A. (1967). *The anatomy of swearing*. New York: The Macmillan Company.
- Mostafazadeh Davani, A., Kiela, D., Lambert, M., Vidgen, B., Prabhakaran, V., & Waseem, Z. (eds.) (2021). *Proceedings of the Fifth Workshop on Online Abuse and Harms (WOAH 2021)*. Association for Computational Linguistics. <https://aclanthology.org/2021.woah-1.0>.
- Nakov, P., Nayak, V., Dent, K., Bhatawdekar, A., Sheikh, Hardalov, M., Dinkov, Y., Zlatkova, D., Bouchard, G., & Augenstein, I. (2021). Detecting Abusive Language on Online Platforms: A Critical Analysis. *arXiv pre-print server*. <https://doi.org/10.26434/chemrxiv-2021-00153>. (24 May 2023)
- Narang, K., Mostafazadeh Davani, A., Mathias, L., Vidgen, B., & Talat, Z. (eds.) (2022). *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*. Seattle, Washington (Hybrid): Association for Computational Linguistics. <https://aclanthology.org/2022.woah-1.0>.
- NSW: *Nationaal scheldwoordenboek: Schelden van de Schelde tot Terschelling*. (1984). Edited by K. Laps. Amsterdam: Ploegsma.
- Odiijk, J. (to appear). *MWE-Finder: Querying for multiword expressions in large Dutch text corpora*. Berlin: Language Science Press.
- Osoba, O. A., & Welser IV, W. (2017). *An intelligence in our image: The risks of bias and errors in artificial intelligence*. Santa Monica: Rand Corporation. https://www.rand.org/pubs/research_reports/RR1744.html. (24 May 2023)
- Pinker, S. (2007). *The Stuff of Thought : Language as a Window Into Human Nature*. New York: Viking.
- Porutiu, T. (2023, 06 January). Profanity Filters: Everything You Need to Know + Our Top 5 Picks. <https://VPNOverview.com>. (24 May 2023)
- Pradhan, R., Chaturvedi, A., Tripathi, A., & Sharma, D. K. (2020). A Review on Offensive Language Detection. In M. L. Kolhe, S. Tiwari, M. C. Trivedi, & K. K.

- Mishra, *Advances in Data and Information Sciences* Singapore.
- Roberts, S. T., Tetreault, J., Prabhakaran, V., & Waseem, Z. (eds.) (2019). *Proceedings of the Third Workshop on Abusive Language Online*. Florence, Italy: Association for Computational Linguistics. <https://aclanthology.org/W19-3500>.
- RolfBly. (2020). *Dutch LDNOOBW (List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words)* [GitHub repository]. <https://doi.org/https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words/blob/5faf2ba42d7b1c0977169ec3611df25a3c08eb13/nl>. (24 May 2023)
- Rosenthal, S., Atanasova, P., Karadzhov, G., Zampieri, M., & Nakov, P. (2020). SOLID: A Large-Scale Semi-Supervised Dataset for Offensive Language Identification. *14th International Workshop on Semantic Evaluation (SemEval-2020)*. <https://doi.org/https://doi.org/10.48550/arXiv.2004.14454>.
- Ruitenbeek, W., Zwart, V., Van Der Noord, R., Gnezdilov, Z., & Caselli, T. (2022). “Zo Grof !”: A Comprehensive Corpus for Offensive and Abusive Language in Dutch. *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)* Seattle, Washington (Hybrid), July.
- Schäfer, J. (2023). Bias Mitigation for Capturing Potentially Illegal Hate Speech. *Datenbank-Spektrum*. <https://doi.org/10.1007/s13222-023-00439-0>.
- Schmidt, A., & Wiegand, M. (2017). A Survey on Hate Speech Detection using Natural Language Processing. Fifth International Workshop on Natural Language Processing for Social Media, Valencia, Spain, 3 April.
- Searle, J. R. (1969). *Speech acts*. Cambridge: Cambridge University Press.
- Searle, J. R. (1979). *Expression and meaning: Studies in the theory of speech acts*. Cambridge: Cambridge University Press.
- Shutterstock. (2020). *List of Dirty, Naughty, Obscene, and Otherwise Bad Words* [GitHub repository]. <https://doi.org/https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words/blob/master/README.md>. (24 May 2023)
- Sketch Engine. (2020). *Dutch Web corpus 2020 (nlTenTen20)*. <https://www.sketchengine.eu/>. (24 May 2023)
- Stapleton, K., Beers Fägersten, K., Stephens, R., & Loveday, C. (2022). The power of swearing: What we know and what we don’t. *Lingua*, 277. <https://doi.org/10.1016/j.lingua.2022.103406>.
- Synovate UK. (2010). *Audience attitudes towards offensive language on television and radio*. https://www.ofcom.org.uk/__data/assets/pdf_file/0017/27260/offensive-lang.pdf. (24 May 2023)
- TAD: *The Alternative Dictionaries*. (2004). Edited by H.-C. Holm. <http://www.notam02.no/~hcholm/altlang/>. (24 May 2023)
- Tasovac, T., Romary, L., Banski, P., Bowers, J., De Does, J., Depuydt, K., Erjavec, T., Geyken, A., Herold, A., Hildenbrandt, V., Khemakhem, M., Lehečka, B., Petrović, S., Salgado, A., & Witt, A. (2018). *TEI Lex-0: A baseline encoding for lexicographic data. Version 0.9.1* [GitHub repository].

- <https://doi.org/https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html>. (24 May 2023)
- Teh, P. L., Cheng, C.-B., & Chee, W. M. (2018). *Identifying and Categorising Profane Words in Hate Speech*. Proceedings of the 2nd International Conference on Compute and Data Analysis - ICCDA 2018.
- The Fuse Group. (2005). *Language and Sexual Imagery in Broadcasting: A Contextual Investigation*.
https://www.ofcom.org.uk/___data/assets/pdf_file/0012/24015/language.pdf. (24 May 2023)
- Van der Gucht, F., Van der Meulen, M., Verlinde, R., & Vanbeylen, W. (2018). *Het groot Vlaams vloekboek: Slimmer schelden en vaardiger vloeken*. Tielt: Lannoo.
- Van der Meulen, M., Van der Gucht, F., Verlinde, R., & Vanbeylen, W. (2018). *Het groot Nederlands vloekboek: Slimmer schelden en vaardiger vloeken*. Tielt: Lannoo.
- Van der Veen, H. (2020). *GRoninger OFFensive Lexicon (GrofLex)* [GitHub repository].
<https://doi.org/https://github.com/hylkevdeveen/GrofLex>. (24 May 2023)
- Van Lichtenvoorde, M., & Van Lichtenvoorde, M. (1993). *Scheldwoorden van de jaren negentig*. Helmond: Michon.
- Van Sterkenburg, P. G. J. (2001). *Vloeken. Een cultuurbepaalde reactie op woede, irritatie en frustratie* 2e ed. Den Haag: Sdu Uitgevers.
- Van Sterkenburg, P. G. J. (2008). *Krachttermen*. Schiedam: Scriptum.
- Van Sterkenburg, P. G. J. (2019). *Rot lekker zelf op: Over politiek incorrect en ander ongepast taalgebruik*. Schiedam: Scriptum.
- VDO: *Van Dale Online*. (2021). Utrecht: Van Dale Uitgevers. <https://www.vandale.nl/>. (24 May 2023)
- WAON: *Woordenboek van het Algemeen Onbeschaafd Nederlands*. (2013). Edited by H. Aalbrecht & P. Wagenaar. Houten & Antwerpen: Uitgeverij Unieboek | Het Spectrum bv.
- Waseem, Z., Chung, W. H. K., Hovy, D., & Tetreault, J. (eds.) (2017). *Proceedings of the First Workshop on Abusive Language Online*. Vancouver, BC, Canada: Association for Computational Linguistics. <https://doi.org/10.18653/v1/W17-30>.
- WEPT: *Woordenboek van eufemismen en politiek correct taalgebruik*. (2001). Edited by M. De Coster. Amsterdam: Veen/Het Taalfonds.
- Wiegand, M., Ruppenhofer, J., & Eder, E. (2021). Implicitly Abusive Language – What does it actually look like and why are we not getting there? *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL): Human Language Technologies*, 576–587.
<https://doi.org/10.18653/V1/2021.NAACL-MAIN.48>.
- Wiegand, M., Ruppenhofer, J., & Kleinbauer, T. (2019). Detection of Abusive Language: the Problem of Biased Datasets. *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 602–608.
<https://doi.org/10.18653/v1/N19-1060>.
- Wiegand, M., Ruppenhofer, J., Schmidt, A., & Greenberg, C. (2018). Inducing a

- Lexicon of Abusive Words – A Feature-Based Approach. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technology, New Orleans, Louisiana.
- Wikipedia (NL). (2023). *Seksuele volkstaal en eufemismen*. Retrieved 10 April 2023 from <https://nl.wikipedia.org>.
- Wiktionary (NL). (2023). *Lemmas tagged <pejoratief> and <scheldwoord>*. Retrieved 10 April 2023 from <https://nl.wiktionary.org>.
- Willems, J. F. (1834). Over eenige oude Nederlandsche vloeken, eeden en uitroepingen. *Nederduitsche Letteroefeningen*, 218–230.
- WPTG: *Woordenboek van Populair Taalgebruik*. (2020-2023). Edited by M. De Coster. <https://www.ensie.nl/woordenboek-van-populair-taalgebruik#>. (24 May 2023)
- Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019a). Predicting the Type and Target of Offensive Posts in Social Media. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019), Minneapolis, Minnesota, 2-7 June.
- Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019b). SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). 13th International Workshop on Semantic Evaluation (SemEval-2019), Minneapolis, USA, 6-7 June.
- Zampieri, M., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., Derczynski, L., Pitenis, Z., & Çöltekin, Ç. (2020). SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). *14th International Workshop on Semantic Evaluation (SemEval-2020)*, 1425–1447. <https://doi.org/10.18653/v1/2020.semeval-1.188>.
- Zhou, V. (2019, 4 February). Building a Better Profanity Detection Library with scikit-learn. <https://VictorZhou.com>. (24 May 2023)

The Czechoslovak Word of the Week. Re-joining Czechs and Slovaks together in an example of invisible lexicography work

Peter Malčovský¹, Michal Škrabal², Vladimír Benko¹, Jan Kocek²

¹ Slovak Academy of Sciences, L. Štúr Institute of Linguistics,
Panská 26, 811 01 Bratislava, Slovakia

² Charles University in Prague, Institute of the Czech National Corpus,
Panská 7, 110 00 Praha 1, Czech Republic

E-mail: peter.malcovsky@juls.savba.sk, michal.skrabal@ff.cuni.cz,
vladimir.benko@juls.savba.sk, jan.kocek@ff.cuni.cz

Abstract

The Czecho-Slovak Word of the Week is a joint popularization project of Czech and Slovak linguists. Throughout the year, each and every week, we are publishing a new entry on the website <https://slovo.juls.savba.sk>, written parallelly in Czech and Slovak, the central part being a language feuilleton supplemented with data drawn from language corpora and quotations from contemporary and historical monolingual and translation dictionaries. In a way, we see the website as a dictionary, with a fixed macrostructure of 52 weekly published entries, and a microstructure, determined by the order of the individual components. Thus, our project could be considered a good example of “invisible lexicography” in practice. The target audience is presented with various kinds of lexicographic information unobtrusively, covertly, and invisibly, usually not even feeling that they are “leafing through” a dictionary. At this year’s eLex, we plan to present not only the website but also the database behind it. Our solution uses modern web technologies: the *JHipster* application generator in combination with the *Vue* front-end framework and the *PostgreSQL* database. The application allows the administrator to easily enter content, including importing and formatting texts from various sources, and to use audio samples from spoken corpora as well.

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

1. Introduction

The Czecho-Slovak Word of the Week is a joint year-long popularization project of the Institute of the Czech National Corpus and the L. Štúr Institute of Linguistics of the Slovak Academy of Sciences, that was inaugurated on the occasion of the 30th anniversary of the dissolution of Czechoslovakia (January 1, 1993). Throughout the year, each and every week, a new entry, written in parallel in Czech and Slovak, is published on the project website (<https://slovo.juls.savba.sk>). We intend to draw the

attention of both the Czech and the Slovak publics (especially the younger generation, for whom the former mutual intelligibility between the two languages no longer holds) to the interesting parallels, but chiefly the differences, between our two languages. We try to do so in a user-friendly and entertaining way, the central part of each entry being a language feuilleton (a very popular genre in Czechia and Slovakia), supplemented with data drawn from language corpora (SYN2015, SYN2020, and ORAL v1 for Czech; prim-10.0-public-all and s-hovor-7.0 for Slovak) and the respective entries from some older monolingual and bilingual dictionaries (Bernolák 1825, Jungmann 1835-1839, SSJČ 1960-1971, SSJ 1959-1968, KSSJ 2003, ČSS 1981, SČS 1967). In a way, we see the website as being a dictionary with a fixed macrostructure (52 entries including some multi-word units¹) and a microstructure determined by the order of the individual components (described in Škrabal & Benko, 2019: 475-476). Thus, our project could be considered a good example of “invisible lexicography” in practice. The target audience is presented with various lexicographic information – be it frequency statistics for various text types, examples from both written and spoken corpora, or quotes from older dictionaries – unobtrusively, covertly, and “invisibly”, usually without them having the feeling that they are “leafing through” a dictionary.

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

In this paper, we want to focus mainly on the technical background of the whole project. In the following chapter, both the backend and the frontend will be described as well as the specific work with the database, i.e., the way to add a new entry to it. Other aspects of the project (project team setup², workflow, promotion, etc.) are left aside on purpose, as we plan to devote a separate article to them after the project is finished (December 2023).

2. Technical overview

We had several options for the technical implementation of the planned project website.

¹ The choice of the list of entries was more or less random and influenced by our personal preferences.

² Our team is largely made up of external writers of feuilletons, mostly linguists. Their texts are edited, proofread, and supplemented with information from corpora and dictionaries, for each entry has the same microstructure. In addition, a programmer and a graphic designer were necessary for the successful implementation of the project.

We rejected the simplest solution: static HTML pages, mainly because the content could be filled and changed only by a technician. A reasonable option was also to modify the existing *Word at Glance* website (<https://www.korpus.cz/slovo-v-kostce/>), which is visually based on a similar tile system. Alternatively, an established content management system, such as *WordPress*, could be used too. Considering various factors (e.g., technical limitations of our infrastructure with respect to *WordPress*, the original estimate of the scope – a word for each day, thus, up to 365 episodes of the series which would require massive automation³), we decided to base our own solution on the technology with which we already had experience. Naturally, this approach has its pros (control over every aspect of the website, no need to limit to existing templates, pay for external hosting, etc.) and cons (more overall work, inability to get “free” features for the blog that are common in established systems). An important requirement for our website – after taking the broad target audience, promotion via social networks, and overall trends into account – was to be mobile-friendly (see also Fig. 11 below).

2.1 *JHipster* application generator

Our web application was generated by *JHipster* generator. *JHipster* is a development platform that can quickly generate, develop, and deploy modern web applications and microservice architectures. The generator has been around since 2013 and is well established and popular. It generates *Spring Boot* based Java web server along with web frontend application (*Angular*, *React* or *Vue* based). Generating an application skeleton using *JHipster* is simple and straightforward, requiring only the answering of a dozen questions (application name, monolith or microservices application, database type and brand, etc.)

³ This original idea was abandoned during the preparations, as it turned out that the daily frequency would require disproportionately more time and energy, as well as a larger team, which was not possible due to the limited budget. Furthermore, we supposed that a daily periodicity would not have been beneficial in maintaining the interest of the audience; in fact, it might have had quite the opposite effect.

```
Administrator: Windows PowerShell
PS C:\work\csl_slovo> jhipster
INFO! Using JHipster version installed globally

JHIPSTER

https://www.jhipster.tech

Welcome to JHipster v7.4.1
Application files will be generated in folder: C:\work\csl_slovo

Documentation for creating an application is at https://www.jhipster.tech/creating-an-app/
If you find JHipster useful, consider sponsoring the project at https://opencollective.com/generator-jhipster

> Which *type* of application would you like to create? Monolithic application (recommended for simple projects)
> What is the base name of your application? csl_slovo
> Do you want to make it reactive with Spring WebFlux? No
> What is your default Java package name? com.juls.cslslovo
> Which *type* of authentication would you like to use? JWT authentication (stateless, with a token)
> Which *type* of database would you like to use? SQL (H2, PostgreSQL, MySQL, MariaDB, Oracle, MSSQL)
> Which *production* database would you like to use? PostgreSQL
> Which *development* database would you like to use? PostgreSQL
> Which cache do you want to use? (Spring cache abstraction) Ehcache (local cache, for a single node)
> Do you want to use Hibernate 2nd level cache? Yes
> Would you like to use Maven or Gradle for building the backend? Maven
> Do you want to use the JHipster Registry to configure, monitor and scale your application? No
> Which other technologies would you like to use?
> Which *Framework* would you like to use for the client? Vue
> Do you want to generate the admin UI? Yes
> Would you like to use a Bootswatch theme (https://bootswatch.com/)? Cerulean
> Choose a Bootswatch variant navbar theme (https://bootswatch.com/)? Primary
> Would you like to enable internationalization support? Yes
> Please choose the native language of the application English
> Please choose additional languages to install Slovak
> Besides JUnit and Jest, which testing frameworks would you like to use?
> Would you like to install other generators from the JHipster Marketplace? No
```

Figure 1: *JHipster* questionnaire

Immediately after answering these questions, the generator creates the first version of an application, both backend server and frontend web. The application is already executable at this point, obviously, with no business logic yet.

Features list includes:

- User management: frontend & backend for creating and editing users with roles (Admin, User);
- Metrics: a smart console for displaying runtime characteristics of the running server (memory, CPU load, number of server threads, number of requests and their result codes);
- Health page, Configuration page, Logs settings page: further server diagnostics and settings.

2.2 Modelling and generating entities

In this step we populated the web application with data. *JHipster* comes with a handy tool – *JDL-Studio* – where data entities can be modelled and visualized, along with relations between the entities.

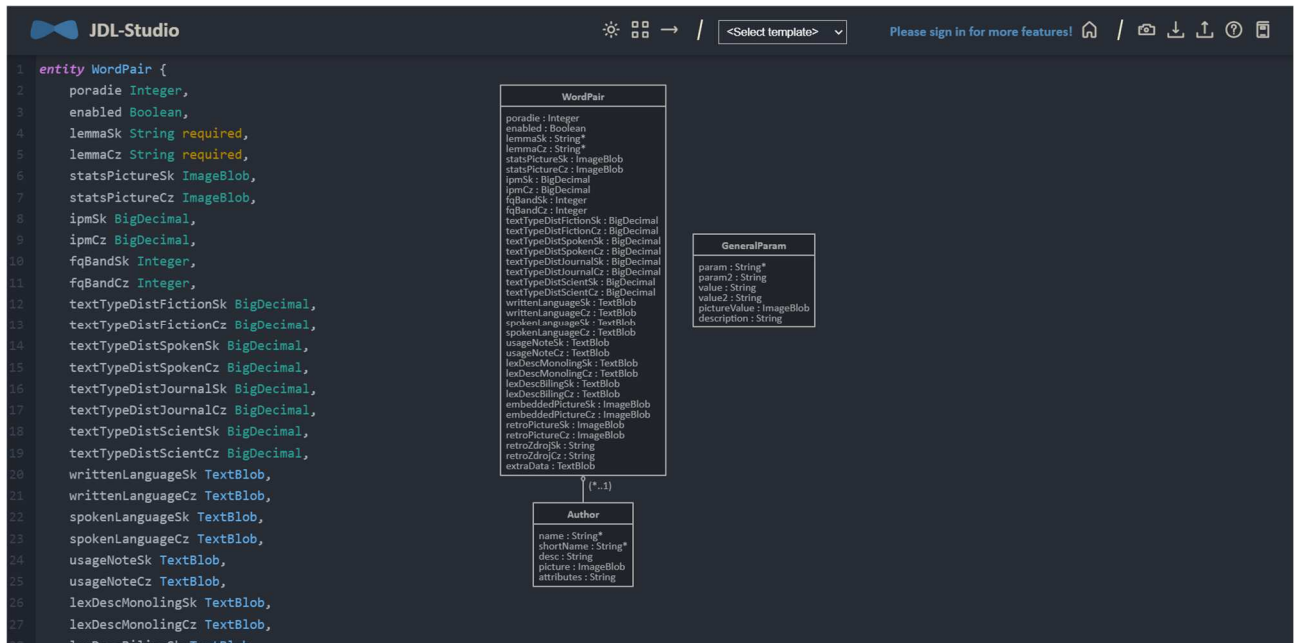


Figure 2: JDL-Studio tool

Three entities were created:

- The **WordPair** entity is a crucial entity, storing all the data necessary to display language posts. All entries have separate Czech (CZ) and Slovak (SK) items and can be mapped in a straightforward way to the user interface. Every record in the WordPair entity is used for rendering just one post (the entire “Word of the Week” article). The entity consists of text items (LemmaCZ, LemmaSK, etc.), long text items – TextBlob (UsageNoteCZ, UsageNoteSK, etc.), and image items – ImageBlob (StatsPictureCZ, StatsPictureSK), along with special items like Order (validity date for a post in numeric format), Enabled flag, ID (unique numeric value for each record), etc.
- The **Author** entity stores information data about authors of feuilletons, namely: author’s name, photograph, and short biography. Each record in the WordPair entity is linked with two records in the Author entity: AuthorCZ and AuthorSK.
- The **GeneralParam** entity stores general-purpose parameters and data for the web application, e.g., various text templates.

Once the entities scripts are ready, we include them in the application, using *JHipster* command. This results in actual database tables being created on the backend (*Liquibase* scripts to create the tables). Besides, we get Java objects representing the entities and Repository and Resource Java Beans to access and manipulate the entities via *JPA* framework. Lastly, CRUD operations (Create, Read, Update, Delete) are completely implemented for all our entities, both on the back- and frontend side. REST API endpoints are created on the server so that the *Vue* frontend can access them.

Populated user interface for entities editing looks like this:

ID	PORADIE	ENABLED	LEMMA SK	LEMMA CZ	AUTHOR SK	AUTHOR CZ	
1	20230109	true	řava	velbloud	Martina Blažeková	Michal Škrabal	View Edit Delete
2	20230220	true	omrvinky	drobky	Martina Blažeková	Michal Škrabal	View Edit Delete
1101	20230123	true	vankuš	polštář	Peter Malčovský	Gabriel Pleska	View Edit Delete
1251	20230130	true	fúrik	kolečko	Martina Blažeková	Jan Nejedlý	View Edit Delete
1351	20230515	true	zimomriavky	husí kůže	Zuzana Klučárová	Gabriel Pleska	View Edit Delete
1451	20230424	true	perý	řty	Natália Kolenčíková	Gabriel Pleska	View Edit Delete

Figure 3: Word Pairs entity





ID	NAME	SHORT NAME	DESC	PICTURE	ATTRIBUTES	
1001	Michal Škrabal	mis	Český lexikograf a korpusový lingvista, toho času filatelista a ředitel Ústavu Českého národního korpusu.	 image/jpeg, 7599 bytes	FK	View Edit Delete
1002	Martina Blažeková	martina	Československá komparatistka, doktorandka FF UK a autorka jedné básnické sbírky.	 image/jpeg, 16694 bytes	F	View Edit Delete
1151	Gabriel Pleska	gabriel	Edituje a píše pro Peníze.cz, Finmag a Heroine. Sbírá neúžitečné informace a roste jak dříví v lese. Jen stromy přes zimu přirůstají pomaleji.	 image/jpeg, 12149 bytes	F	View Edit Delete
1201	Peter Malčovský	malco	Odborný pracovník, Jazykovedný ústav L. Štúra SAV. Moderátor galérie www.instagram.com/dedina_bnw/	 image/jpeg, 19490 bytes	FC	View Edit Delete

Figure 4: Authors entity

Česko(slovenské slovo
týdne / týždňa

Domů / Domov Archiv V médiích / médiách Info Entities Administration


Create or edit a Author

ID
1001

Name
Michal Škrabal

Short Name
mis

Desc
Český lexikograf a korpusový lingvista, toho času ředitel Ústavu Českého národního korpusu.

Picture

 image/jpeg, 7 599 bytes x
 Choose File No file chosen

Attributes
FK

Cancel Save

Figure 5: Create/edit an author form

2.3 Creating an application homepage and other pages

With the application skeleton generated by *JHipster*, one can finalize the application by the manual creation of the homepage and other pages. Obviously, this part took most of the entirety of the development time. We have created:

- Home page – word pairs “posts” viewer with navigation (previous week, next week);
- About page – information about the project purpose and team, contacts, and credits;
- Media page – memorable promo actions for the project in various media (TV, radio, blogs);
- Archive page – a timeline with all the published posts.

2.4 Technology stack overview

2.4.1 PostgreSQL

PostgreSQL is a powerful, open-source object-relational database system with over 35 years of active development that has earned a strong reputation for reliability, feature robustness, and performance. Nowadays, the *PostgreSQL* is used even in enterprise

solutions, competing with legendary systems like *Oracle DB*. It also incorporates full search features, including stemming (Czech language stemmer is available, Slovak language stemmer not yet), removing stop words during search.

2.4.2 JPA (Spring Data)

JPA stands for Jakarta Persistence API, a set of concepts for Java objects persistence and object-relational mapping (ORM). On our server, this layer is used for accessing the physical database. *JPA* allows various conceptual approaches to handle data in the application (Code First, Model First, Database First). On a practical level, Repository objects are created in our server for DB access purposes. Each entity has its own Repository object. The framework tries to help with DB queries as much as possible – for trivial ones like *findById*, a coder does not need to write any code. Simple queries can be written just by query method name (e.g., *findAllByNameLike*), *JPQL* database-agnostic query language, or native DB query. Combined with pagination and ordering support, it is not complicated to create backend queries for various frontend grids.

Code examples:

Query used to pick current Word of the Week record, given current system date as parameter. The *JPA* translates method name to actual query by itself:

```
// current
public List<WordPair> findTop1ByEnabledTrueAndPoradieLessThanEqualOrderByPoradieDesc(Integer
currDateNumber);
```

Figure 6: Query used to pick current Word of the Week record, given current system date as a parameter

```
//findAllTiny
@Query(
    "Select new com.peto.wotd.service.dto.WordPairTinyDTO(w.id, w.poradie, w.enabled,
w.lemmaSk, w.lemmaCz, ask.name, acz.name, ask.shortName, acz.shortName)" +
    " from WordPair w " +
    " left join Author ask on w.authorSk.id = ask.id " +
    " left join Author acz on w.authorCZ.id = acz.id "
)
public Page<WordPairTinyDTO> findAllTiny(Pageable pageable);
```

Figure 7: A more complex query for obtaining all word pairs list, using reduced DTO object for effective transfer. The query gets pagination settings from UI as a parameter (e.g., page 3, ordered by lemmaCZ).

2.4.3 *Vue* frontend framework

Vue is an approachable, performant, and versatile framework for building web user interfaces. Introduced in 2014, it has gained popularity and user base since then. Given the tremendous development rate in this web frontend area, we can look at the *Vue* as “just another web framework”. Nevertheless, the *Vue* belongs to state-of-art ones as of 2023.

From a developer’s point of view, *Vue* is similar to the *React* framework, yet improved and simplified in many ways. Unlike in *React*, *Vue* comes with handy HTML tags for if-then-else constructs, loop constructs, etc. so there is no need to combine HTML code with JavaScript code, producing a hard-to-read, hard-to-maintain mess.

Code examples:

```
<div v-if="isMobile" class="col-md-12" style="padding-left: 0; padding-right: 0">
  <h4 class="centered">
    <a @click="navigatePrev()"><font-awesome-icon icon="chevron-circle-left" size="1x" /></a>
    {{ currDate }}
    <a @click="navigateNext()"><font-awesome-icon icon="chevron-circle-right" size="1x" /></a>
  </h4>
</div>
<div v-else class="col-md-12">
  <h2 class="centered">
    <v-tooltip top>
```

Figure 8: Usage of *v-if* and *v-else* *Vue* tags to render different content for mobile and desktop web

```
<v-timeline :dense="isMobile">
  <v-timeline-item v-for="item in items" :key="item.id" color="#c7c7c7" small>
    <a v-bind:href="'/' + item.id" style="color: #0e5a9d">
      {{ getDateFormattedFromPoradie(item.poradie) }} <br />
    </a>
    <h5 class="centered" style="margin-bottom: 0px; margin-top: 0.2em">
      <span class="magenta">{{ item.lemmaCz }} / {{ item.lemmaSk }}</span>
    </h5>
```

Figure 9: *Vue* tag *v-for* in action to render all timeline items for Archive page

2.5 Visual identity

The project’s website as well as the accompanying graphic material is based on the visual identity created by Jan Kocek from the Institute of the Czech National Corpus. The homepage uses a tile system, with each tile containing a different type of content

(feuilleton, frequency statistics, dictionary data, etc.). The colour scheme is pale blue/black/purple, along with the red and blue of the Czech and Slovak flags. Combined with the modern Roboto font, this is a simple and fresh design.

The graphic designer also created an icon for the site, a logo, templates for metadata for social networks, a template for the quotes that appear at the end of the page, and a template for the side events related to the project.

◀ 17. 4. – 23. 4. 2023 ▶

Česko(-)slovenské



slovo týdne

tchyně



slovo týždňa

svokra

Pár slov o slově tchyně

Předně si slíbme: dnes žádně vtípy o tchyních, ano? To je seriózní sloupek!

A protože už jsem ve svém oboru už léta honěný a současnou kodifikací (čti: to, jak bychom měli psát podle ÚJČ) znám vcelku obstojně, už jen máloco mě překvapí. [Na rozdíl od jiných](#). Ale ruku na srdce, mluvíte-li o tchyni, vyslovujete ji krátce? Nelžete! Respektive takto: znáte vůbec někoho, kdo ji krátce vyslovuje? (Ostraváci se nepočítají!) Solva, protože takových lidí bude asi tolik jako těch, co vyslovují *matjesy* s oním [j]. A přesto tuto nedélku držíme uměle při životě. Tedy jak kde: v oficiálních psaných textech, nad nimiž bdí oko korektora, se to ještě ctí (byť i zde se občas najde dlouhá, tj. nekodifikovaná varianta). Ovšem v textech internetových, kde se plebs vyjadřuje volně a nad pravopisem si příliš hlavu neláme, je situace jiná. A proto: ještě než někoho po gramanácovsku osočíte z neznalosti normy – neřkuli celé mateřštiny, pozor! jste totiž sami v menšině... A teď mi řekněte, vážení kolegové z ÚJČ, co by vám udělalo, zrovnoprávnit dlouhou variantu (a neuvádět tak prostý lid v překvapení). Když už jste zrovnoprávnili tripletu [čurat/čúrat/čúrat](#)...

A ještě pro jednu věc mám tohleto slovo rád. Vždy jsem záviděl jiným jazykům aspiraci, neboli příděch, v první řadě romštině, na jehož fonetickém svérázu se tento jev výrazně spolupodílí. My takových (pseudo)aspirovaných slov mnoho nemáme, o to s větší rozkoší je vypouštím z úst, ať už je to [tʰi:ně], [tʰa:n] nebo [tʰoʃ].

Poche (čti [pʰe]) – a pak že prý se nedá sloupek o tchyních napsat bez vtípu o těchto. (Jen se bojím, že má sparringpartnerka tomuto lákadlu neodolá...)

Čech o slově svokra

No dobře, my máme sexy příděch, zato *svokra* zní starobyleji, vznešeněji, slavnostněji... A přitom jde o popisnost až prozaickou: 'svě krve žena'. Však jsme ji dříve taky mívali, ještě Jungmann uvádí heslo *swekra* a jeho řidiší varianty *swekry* a *swekrew*; ba co víc, rozlišuje *swekruši* (tchyně) od *swekrušny* (švagrová). Ale mamá sláva, někdejší *swekru* postupně vytlačila *tchyně*, ana se zjevila prý už za času Komenského namísto původního staročeského tvaru *tšče* (srov. mužský protějšek *test*, tchán'). Anebo snad *tchyně*...?



Michal Škrabal

Český lexikograf a korpusový lingvista, toho času filatelista a ředitel Ústavu Českého národního korpusu.

Niečo o slove svokra

O svokrách len dobre. (Hlavne ak máme nábeh čoskoro sa svokrou stať, dokonca viacnásobnou.) Slovo *svokra* (*svokrička*, *svokruša*, *svokruška*), označujúce manželovu matku, je ženským náprotivkom k mužskému pomenovaniu *svokor*, manželov otec; spoločne sa nazývajú *svokrovcami*. Na označenie matky manželky vo vzťahu k manželovi máme v slovenčine menej používané slovo *testiná* (*testinká*), s príslušným mužským pendantom *test*, spoločne *testovci*. Paralelné používanie týchto slov je však dlhodobá na ústupe, označenie *svokor*, *svokra*, *svokrovcami* sa postupne zaužívalo jednotne pre ženiných i mužových rodičov. Pri ich priamom oslovení sa však často nahrádzajú rovnakými pomenovaniami ako pre vlastných rodičov – teda *mama* a *otec*. Zatiaľ čo niekto rieši *svokrovské* problémy (najmä vo vzťahu svokra – nevesta, svokra – zať), Rudo Sloboda v románe *Krv* (1991) spomína aj iné: „*Takí zaťovia, ktorých príliš nútite prijať svokrovské móresy, odchádzajú do krčiem, aby si oddýchli.*“

Takmer rovnako ako *svokry* sa v bežnej komunikácii skloňuje *svokrin jazyk* (po latinsky *Sansevieria trifasciata*, v češtine *tchynin jazyk* alebo *tenura*, v angličtine *mother-in-law's tongue*) – kvietok, ktorý dostal meno podľa tvaru svojich listov: sú totiž dlhé a špicaté (mečovité), nelichotivo pripomínajúce jazyk svokry. Najznámejšia a najodolnejšia izbová rastlina – *izbovka* (po česky tak pohodovo – *pokojočka*), akýsi nezmar medzi nezmarami, mnohým z nás dobre známa z parapetných dosiek československých škôl, škôlok a úradov, má dnes napriek občasnému prívlastku *socialistická* stále viac obdivovateľov. Pre svoju absolútnu nenáročnosť a zároveň dekoratívnosť a dlhovekosť je objektom čoraz väčšieho zborateľského záujmu sukulentárov. Rýchlo sa rozmnožuje a jej listy majú protizápalové liečivé účinky, podobne ako aloa pravá (*Aloe vera*). A pretože funguje ako perfektná prírodná čistička vzduchu, výborne sa hodí aj do spálne. (Tu si prímyslíme veľkého smajlika.)

Či už vnímame *svokru* ako milé, rodinne založené stvorenie, ktoré chce každému len dobre, predovšetkým svojmu už dospelému dieťaťu (a vnúčatám), alebo ako vďačný terč [vtípy](#), dokonca až čierneho humoru, každý z nás sa rád zasmieje:

– „Viete, prečo sa svokry pochovávajú do pol pása?“ – „Aby sa mal kto starať o hrob!“

– „Svoju svokru volám vegeta.“ – „Prečo?“ – „Pretože sa mieša do všetkého.“

Mladomanžel príde domov podnapitý a vo dverách stretne svokru s metlou v ruke. Prižmúri oči a povie: „Mamička, zametáte alebo odlietate?“

Slovenka o slove tchyně

Zatiaľ čo slovenské slovo *svokra* nás baví možným etymologickým východiskom v indoeurópskom **swē-* (svoj) + **ker-* (hlava/krava), malý prieskum medzi českými internetovými vtípkármi ma doviedol až k ich uletenému tvrdeniu, že slovo *tchyně* vzniklo zložením slov *tchoř* a *svině* – odpadávam a končím.



Monika Kapustová

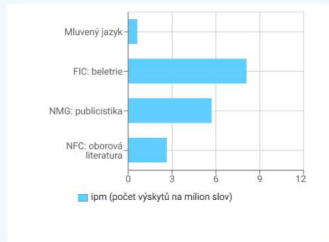
Venuje sa jazykovede, píše, prekladá, učí. Miluje vtípy o svokrách, babkách, dedkoch, deťoch a hodinách slovenčiny.

Frekvenční charakteristika

kolikrát v milionu slov:

3,78

frekvenční pásmo:



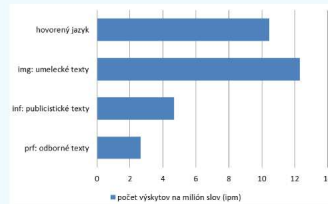
Zdroj: syn2015 + oral_v1

Frekvenčná charakteristika

koľkokrát na milión slov:

7,37

frekvenčné pásmo:



Zdroj: prim-10.0-public-all + s-hovor-7.0

Podobně používaná slova



Zdroj: SYN v8

Podobne používané slová



Zdroj: Sémantická podobnosť slov

Ukázky z psaného jazyka

- Zimu strávil v Lübecku se strýčkovou rodinou, tchánem a tchyní.
- Její budoucí tchyně, stará žena, byla známa svou zlomyslností.
- Stačil jeden pohled a pochopil, že na návštěvě u tchyně zrovna dří Mladá královna a že obě jsou nyní v královských komnatách.
- Ve stresu z předvánočního shonu se snadno můžete chytit s tchyní, manželem, kolegyní, šéfem ...
- Tchán s tchyní matky pěti dětí a neúspěšné prezidentské kandidátky pocházejí z kantonu Thurgau.

Další příklady užití slova v psaném jazyce
Zdroj: SYN2020

Ukázky psaného jazyka

- „Dáš si guláš?“ spýtala sa mama. „Čo varila Vlasta?“ Zasmiala som sa. „Otázka by mala znieť: Čo varila svokra?“
- So svojou matkou mal dobrý vzťah, chodil k nej na návštevy. So svokrou vychádzal dobre.
- Preto si myslíte, že sused bol opitý? - Pretože vykrikoval, že sa nebojí ani manželky ani svokry .
- Uvedomila si však, že pani Wang Šanpaová je špehúňka a donášačka jej svokry .

Zdroj: prim-10.0-public-all

Ukázka z mluveného jazyka



Šárka_7576
.. kreten ale hroznej . to viš že Jolka ta je úplně hotová protože von . (mlasknutí) Jolky Jolka jakoby viš ta jak tam s ní
Iveta_7577
no
Šárka_7576
Márovo maminka .. tak @ . vona jakoby podniká s ze svoji tchyní jako z Martinovo mámu voni maj jakoby pole .

Další příklady užití slova v mluvené řeči
Zdroj: Oral v1

Ukázka hovoreného jazyka



Brekvim_Briock
Občas na premiéru príde , no . So svokrou .
Poedua_Pohinoová
A najlepšie , že budúca svokra jej teraz vybavuje nejakého vizážistu , ktorý maskuje a spolu týchto , akože nejaký strašne známy .
Vredum_Becagi
Mamka zase rodičov , mamku mala Češku , takže to sa .
Svokra Polka . Slovensko - poľsko - český rozprávala . Takže to už jednoducho človek chytí niečo .

Zdroj: s-hovor-7.0

tchyně v českých slovnících

Slovník spisovného jazyka českého
tchyně (ob. tchyně), -ě ž. (2. mn. -) matka manžela n. manželky; tchán a t.; bála se své tchyně (Podl.)

Zdroj: Slovník spisovného jazyka českého
Další zdroje: Wikipedie, Wikislovník, Čestina 2.0, Forvo

svokra v slovenských slovníkoch

Slovník slovenského jazyka
svokra, -y, -kier ž manželova matka
Krátky slovník slovenského jazyka
svokra -y -kier ž
1. manželova matka
2. hovor. manželkina matka, testiná

Zdroj: Slovník slovenského jazyka, Krátky slovník slovenského jazyka
Dalšie zdroje: Slovníkový portál, Aranea, Wikipédia, Wikislovník, Forvo



Figure 10: Screenshot of the web page⁴



Figure 11: Screenshot of the mobile version

⁴ A short feuilletton is followed by frequency information (among various text types), similarly used words, examples from written and spoken data, and excerpts from older dictionaries.

2.6 Workflow

A logged-in user with the admin role can create and edit data in the system. A new pair of words is entered into the system as follows:

Via the menu Entities / Word Pair, we get to an overview of all word pairs (see Figure 3 above). Pressing the “Create a new word pair” button will open the “Create/edit a word pair” form. There we can enter the basic data: the publication date of the post in the YYYYMMDD form and the Czech and Slovak lemma form. Pressing the “Populate” button will fill in some of the following items using the templates defined in the General Params entity. These are templates for corpus samples and dictionary data with links. Other items can be added subsequently: frequency graphs and numeric items for frequencies, word-clouds for the “Similarly used words” tile, or screenshots from older monolingual dictionaries. We get the data from the “Word at Glance” portal, the JÚLŠ dictionary portal, and other tools.

Create or edit a WordPair

ID

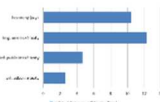
Poradie

Enabled

Lemma Sk

Lemma Cz

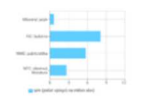
Stats Picture Sk



image/png, 3 412 bytes ✕

 No file chosen

Stats Picture Cz



image/png, 9 732 bytes ✕

 No file chosen

Figure 11: Create/edit a WordPair form

We save the document with a feuilleton in HTML format. Having opened the HTML file in a text editor, we copy-paste it into the “Usage Note Sk” and “Usage Note Cz” entries. We press the button below the item to clear the text and manually edit the section headings. Corpus entries “Written Language Sk”, “Written Language Cz”, “Spoken Language Sk”, “Spoken Language Cz” supplement the template with apposite examples from the corpora. We highlight the keyword pressing the relevant button. We equip the dictionary entries “Lex Desc Monoling Sk” and “Lex Desc Monoling Cz” by taking the formatted dictionary text from the dictionary portals page code (via the browser function “View page source”). After the copy-paste we press the button below the entry. The rich dictionary formatting will be preserved. The links in the “Additional Resources” section are already filled in. We add the audio recordings for the spoken language samples, in JSON format, to the “Extra Data” entry. We get this audio material from spoken corpora, e.g.:

```
"audioSK" : "https://korpus.juls.savba.sk/hovor-7.0-web/2008-07-26-Briock/Briock_00573.691.ogg#t=5,10"
```

Finally, at the bottom of the form, we enter the name of the writers of the Czech and Slovak feuilleton according to the authors’ list and save the form via the “Save” button. After returning to the overview list of word pairs, we can check the new entry via the “View” button.

3. Future work

The database described above is fine-tuned now: it appears to be both robust and flexible enough for further use. At least three possible uses can be imagined: 1) another, follow-up project created by the users themselves (user-generated content supervised by professional editors); 2) other language pairs (Czech-German/Polish, Slovak-Hungarian/Polish); 3) adding another language(s) (e.g., those of the Visegrad area: Czech, Slovak, Polish, Hungarian). The last option is certainly the most implementation-intensive, but even that seems relatively straightforward, adding entries for the new language(s) to the Word Pair data entity and modifying the main website to display the language data in 3(+) columns instead of the current two. After adding the new features to the data model, the *JHipster* generator can be re-run to re-generate the code for the entire updated system. Some caution is necessary here, however, as the manual edits we made to the code may be lost in this process.

4. Acknowledgements

This work has been, in part, funded by the Slovak VEGA Grant Agency, Project No. 2/0016/21. It was also supported by the European Regional Development Fund-Project “Creativity and Adaptability as Conditions of the Success of Europe in an Interrelated World” (No. CZ.02.1.01/0.0/0.0/16_019/0000734). During its creation we used the tools developed within the Czech National Corpus project (LM2015044)

funded by the Ministry of Education, Youth and Sports of the Czech Republic within the framework of Large Research, Development and Innovation Infrastructures.

5. References

- Bernolák, Antoninus (1825). *Slowár Slowenský Čěsko-Latinsko-Ňemecko-Uherský seu Lexicon Slavicum Bohemico-Latino-Germanico-Ungaricum*. Budae: Typis et Sumtibus Typogr. Reg. Univers. Hungaricae.
- [ČSS] *Čěsko-slovenský slovník* (1981). Bratislava: Veda.
- Jungmann, Josef (1835-1839). *Slownjk čěsko-ňemecký*. Praha: Knížecí arcibiskupská knihtiskárna.
- [KSSJ] *Krátký slovník slovenského jazyka*. Bratislava: Veda.
- Machálek, T. (2019). Slovo v kostce – agregátor slovních profilů. Praha: FF UK. Accessed at: <http://korpus.cz/slovo-v-kostce/>. (21 April 2023)
- Machálek, T. (2020). Word at a Glance: Modular Word Profile Aggregator. In N. Calzolari et al. (eds.) *Proceedings of The 12th Language Resources and Evaluation Conference*. Marseille: ELRA, pp. 7011-7016.
- [SČS] Gašparíková, Želmíra & Kamiš, Adolf (1967). *Slovensko-čěský slovník*. Praha: SPN 1967.
- [SSJ] *Slovník slovenského jazyka* (1959-1968). Bratislava: Vydavateľstvo SAV.
- [SSJČ] *Slovník spisovného jazyka čěského* (1960-1971). Praha: ČSAV.
- Škrabal, M. & Benko, V. (2019). Make my (Czechoslovak word of the) day. In I. Kosem et al. (eds.). *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference. 1-3 October 2019, Sintra, Portugal*. Brno: Lexical Computing CZ, s.r.o., pp. 467-477.

The Central Word Register of the Danish language

Thomas Widmann

The Danish Language Council

Abstract

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

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

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

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

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

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

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

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

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

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

1. Introduction

A common challenge when working with lexicographic and computational linguistic resources is the lack of compatibility. Each resource has its own approach to issues such as homonym resolution, part-of-speech tags and lexical coverage. Furthermore, licensing issues can make it exceedingly difficult to determine which resources can be legally reused in a project.

This problem is particularly pronounced for smaller languages, as the initial cost of undertaking any computational linguistic project becomes increasingly prohibitive for smaller actors to initiate.

Although numerous electronic resources for Danish exist—including machine-readable dictionaries, corpora, and taggers—reusing them can be challenging because they are not based on the same fundamental resources, nor do they share database keys or similar attributes. Consequently, the development of language technology for Danish has become more difficult than necessary.

The solution has been known in other areas for years: using a shared database key that facilitates the merging of diverse databases. For example, Denmark has a system called the *Centrale Personregister* (CPR), which assigns a unique identification number to each resident of Denmark. This system offers significant practical benefits; for instance, when an individual changes their address, they only need to inform the local council, and all relevant parties (e.g., the tax authorities, the health system, and the bank) are notified automatically.

Inspired by the CPR, we decided to address this issue by creating a new resource framework: The Central Word Register (Danish: *Det Centrale Ordregister*: COR).

The COR was supported by the Danish Agency for Digitisation, and the project involves the Danish Language Council, the Society for Danish Language and Literature, and the Centre for Language Technology at the University of Copenhagen. It assigns unique identification numbers to all lemmas and word forms in Danish. The Danish Language Council is responsible for the basic register, comprising orthography and morphology for the vocabulary covered by *Retskrivningsordbogen*, the Danish Orthographic Dictionary (Dansk Sprognævn, 2012). This basic register, which we will call COR₁ in the following, was launched in September 2022 and is accessible at ordregister.dk.

In this article, we will first describe the structure of the COR, outline the basic resource’s structure, and demonstrate how new COR resources can be added. We will then explore various lexicographic applications, with a particular focus on the Danish Language Council’s website RO^{hist}, which enables comparisons of different historical orthographic dictionaries. Subsequently, we will discuss COR linkers (programs that automatically assign COR identification numbers to all words in a running text) and, finally, examine invisible

lexicography and crowdsourcing. Our aim is to provide readers with both the motivation to begin utilising the COR and the practical skills to do so.

2. Structure and Components of the COR

2.1 The Orthographical Foundation: *Retskrivningsordbogen*

Retskrivningsordbogen (Dansk Sprognævn, 2012) is the official reference for Danish language orthography. Published by the Danish Language Council (Dansk Sprognævn), it serves as the primary authority on Danish orthography in accordance with the Danish Orthography Act (LBK 332).

The dictionary is regularly updated to reflect the latest changes in Danish orthography, ensuring it remains current and accurate. The most recent edition was published in 2012; however, new words are added annually to its online version, keeping it up-to-date with contemporary language usage. The latest update was in November 2022.

The categorisation of the basic vocabulary in COR₁ into lemmas is based on *Retskrivningsordbogen*. As a result, it follows the same principle for what constitutes a lemma (Dansk Sprognævn, 2012: 13f):

Opdelingen i opslagsord er principielt uafhængig af ordenes betydning. Det bevirker at ord med forskellig betydning er slået sammen i ét opslagsord hvis de i øvrigt har samme stavemåde, udtale, ordklasse og bøjning, og hvis de indgår i sammensætninger på samme måde.¹

From this quotation, it is evident that neither the semantics nor the etymology is considered when determining what a lemma is.

The COR can be regarded as an enhanced and optimised version of *Retskrivningsordbogen*, specifically tailored for natural language processing purposes. Building upon the foundation provided by the dictionary, the COR aims to facilitate and improve the development and utilisation of Danish language technologies. However, there are some key differences between the two:

1. *Retskrivningsordbogen* is designed for humans; the COR is designed for easy use by computer programs.
2. *Retskrivningsordbogen* does not include all inflected forms (and of the ones that are present in the data, only a few are displayed in the book); the COR offers more comprehensive coverage.
3. *Retskrivningsordbogen* comes with the restriction that it cannot be used to create dictionaries; the COR can be used without any restrictions.
4. *Retskrivningsordbogen* contains a good number of usage examples; the COR has none.
5. *Retskrivningsordbogen* has references to its rule appendix; the COR has none.
6. *Retskrivningsordbogen* has more and longer glosses than the COR.²

¹ “In principle, the division into headwords is independent of the words’ meaning. This leads to the merger of words with distinct meanings into a single headword, provided they share the same spelling, pronunciation, word class, and inflection, and if they are employed in compounds in the same way.”

² The COR provides glosses solely for the purpose of disambiguating homographs. For instance, COR.70558 *kalk* carries the gloss “et mineral” (“a mineral”), whereas COR.77824 *kalk* is glossed “krus el. bæger”

2.2 The structure of the COR

2.2.1 Unique Identification Numbers

In the following, we will first describe the structure of COR₁ and then discuss the general structure for other COR resources.

In COR₁, all lemmas found in *Retskrivningsordbogen* and their forms are assigned unique identification numbers. These consist of the prefix ‘COR’ followed by a 5-digit index number indicating the specific lemma. For example, COR.56746 corresponds to the lemma *avocado*.

To specify a particular form of a lemma, a three-digit grammatical code is appended to indicate the part of speech and inflection of the word. For example, COR.56746.111 corresponds to the singular definite form of this common-gender noun, i.e., *avocadoen/avokadoen*. A list of these grammatical codes can be found on ordregister.dk.

In addition to the lemma and grammatical code, a two-digit code is added to indicate orthographical variation. This ensures that each ID number is unique. For example, COR.56746.111.02 corresponds to the form *avokadoen*. (Both forms, *avocadoen* and *avokadoen*, are co-official, and neither is preferred.)

The ID numbers are arbitrary and are not assigned alphabetically. The lemma indices in the *Retskrivningsordbogen* range from 0 to 99,999, and they are not assigned based on alphabetical order. For practical reasons, the interval is divided by word class. For example, adjectives have indices between 15,000 and 29,999, and nouns have indices between 40,000 and 99,999. However, this division is not a formal requirement, and other COR resources are not expected to follow this pattern.

Here are the actual contents of COR₁ for *avocado*:

The grammatical code in column 4 exhibits a one-to-one correspondence with the second part of the numerical code. For instance, 110 consistently translates to *sb.fk.sg.ubest*. The final column displays a 1 if the form is derived from the dataset underpinning *Retskrivningsordbogen* and is consequently part of the official norm. Conversely, a 0 signifies that the form has been auto-generated, and users should exercise caution when utilising these forms.

(“mug or cup”); the former is the same in *Retskrivningsordbogen*, but the latter bears the glos “krus el. bæger til altervin” (“mug or cup for sacramental wine”) in the dictionary – the latter half has been omitted from the COR because it is not needed for disambiguation. Additionally, COR.82322 *kalkbrud* “limestone quarry” has a gloss in the dictionary to aid the user identify the word, but it does not have one in the COR because the lemma has no homographs.

ID	lemma	gloss	gram. code	form	norm
COR.56746.110.01	avocado	–	sb.fk.sg.ubest	avocado	1
COR.56746.110.02	avocado	–	sb.fk.sg.ubest	avokado	1
COR.56746.111.01	avocado	–	sb.fk.sg.best	avocadoen	1
COR.56746.111.02	avocado	–	sb.fk.sg.best	avokadoen	1
COR.56746.112.01	avocado	–	sb.fk.pl.ubest	avocadoer	1
COR.56746.112.02	avocado	–	sb.fk.pl.ubest	avokadoer	1
COR.56746.113.01	avocado	–	sb.fk.pl.best	avocadoerne	1
COR.56746.113.02	avocado	–	sb.fk.pl.best	avokadoerne	1
COR.56746.114.01	avocado	–	sb.fk.sg.ubest.gen	avocados	1
COR.56746.114.02	avocado	–	sb.fk.sg.ubest.gen	avokados	1
COR.56746.115.01	avocado	–	sb.fk.sg.best.gen	avocadoens	1
COR.56746.115.02	avocado	–	sb.fk.sg.best.gen	avokadoens	1
COR.56746.116.01	avocado	–	sb.fk.pl.ubest.gen	avocadoers	1
COR.56746.116.02	avocado	–	sb.fk.pl.ubest.gen	avokadoers	1
COR.56746.117.01	avocado	–	sb.fk.pl.best.gen	avocadoernes	1
COR.56746.117.02	avocado	–	sb.fk.pl.best.gen	avokadoernes	1

Other COR resources ought to adhere to the same general syntax, which includes:

1. The resource name.
2. The lemma id.
3. Any required subdivisions. It is not necessary for these to match COR_1 ; the need for subdivisions, along with their quantity and digit count, is specific to each resource. However, this information must be explicitly detailed on the website.

2.2.2 The COR resource landscape

There are three levels of COR resources:

- Level 1 corresponds to the most recent edition of Retskrivningsordbogen. Prefix: `COR`.
- Level 2 will contain a plethora of resources from professional language environments in Denmark, specifically members of the Danish Language Council’s board of representatives. Additional resources will be included over time. At the present time, it comprises a resource containing supplementary lemmas from the Danish Dictionary (published by the Society for Danish Language and Literature [DSL]); this resource is called `COR.EXT`. Level 2 will also feature a semantic extension to the basic register produced by DSL and the Centre for Language Technology at the University of Copenhagen (CST). For more information on their work developing this semantic component, see Nimb et al. (2022). Prefix: `COR.NAME` (where `NAME` is an alphanumeric identifier).
- Level 3 encompasses all other resources without restrictions. Any relevant project can be assigned a prefix and an ID range if one contacts the Danish Language Council. Prefix: `COR.OPEN.NAME` (where `NAME` follows the same rules as above).

Besides a name, each resource is allocated a series of unique ID numbers. These numbers should be utilised in combination with existing ones in other resources on the same or lower levels. For example, any COR resource that indexes *avocado/avokado* should ideally use the number 56746 for it. A hypothetical dictionary of common spelling errors called `COR.OPEN.ORTHEERROR` should thus define the common misspelling *advokado* as `COR.OPEN.ORTHEERROR.0056746` (preferably padding with extra zeros like this to match other ID numbers in its number series). Existing ID numbers should be used for perfect matches and mere orthographic variation; resources should use their own numbers primarily for non-existing lemmas and ones that do not correspond one-to-one with an existing entry. For example, if a resource needs to index *avocado* and *avokado* separately, it should allocate new numbers to both; the same applies if a resource needs to index *avocado/avokado* together with the antiquated term for this, *advokatpære*.

2.2.3 Relations

In the Central Word Register (COR), *relations* act as a mechanism to establish connections between lemmas and word forms, clarifying their associations with one another.

These relations facilitate the organisation and search for data within COR, enabling users and developers of language technology tools to trace connections between lemmas and word forms in order to identify related linguistic components.

Various types of relations can be defined, including:

Abbreviation	Definition
fus	fusion of two or more COR indexes
rep	replaced by one or more COR indexes
spl	split into two or more COR indexes
sms	compound of two COR indexes (for compound words)
hyr	hypernym (superordinate concept) for two or more COR indexes
hyp	hyponym (subordinate concept) for another COR index
rim	rhyme (for rhyming dictionaries)

The examples above demonstrate the versatility of relations. Each resource can define its own relations; the basic register does not currently use any.

To exemplify this, consider the modern lemmas *fjeder* “(metal) spring” and *fjer* “feather”; they share the same etymology, and as recently as in the orthographical dictionary of 1923 (Glahder, 1923), there was free variation between the forms *Fjeder* and *Fjer* regardless of the meaning. To add this 1923 dictionary to COR, we would resolve this issue by creating a new ID number (in the following 4008020) and adding information about its relation to the two modern entries.

COR-id	lemma	form	relation
COR.70131	fjeder	fjeder	
COR.70759	fjer	fjer	
COR.DR01923.4008020.x.01	Fjeder	Fjeder	fus:70759+70131
COR.DR01923.4008020.x.02	Fjeder	Fjer	fus:70759+70131
COR.DR01923.0070131			rep:4008020
COR.DR01923.0070759			rep:4008020

Here, `rep:4008020` means “this ID number has been replaced by 4008020”, and `fus:70759+70131` means “this ID number is a fusion of 70759 and 70131”.

The same applies if two historical lemmas correspond to one modern one – for example, the current dictionary only has one lemma *skade* for both the bird (the magpie) and the fish (the common skate) because, as mentioned above, neither semantics nor etymology is taken into account when determining what a lemma is; however, in the 1955 dictionary (Dansk Sprognævn, 1955) there were two corresponding lemmas:

COR-id	lemma	glosse	relation
COR.45662	skade	en fugl; en fisk	
COR.R01955.4011080	skade	en fugl	rep:45662
COR.R01955.4011081	skade	en fisk	rep:45662
COR.R01955.0045662			spl:4011080+4011081

By establishing such relations, the COR can effectively manage the connections between historical and modern lemmas, enhancing the overall organisation and retrieval of linguistic data.

3. Accessing and utilising the COR

COR’s master register and certain other resources can be accessed in two ways:

The entire register can be downloaded as a CSV file from `ordregister.dk`. This allows for working with, among other things, the master register offline and integrating it into one’s own systems.

There is also an online interface that can be used to search the master register’s data and access information on lemmas and word forms. This information can be displayed in HTML or accessed from a program in either CSV or JSON format. This interface can be found at the same address: `ordregister.dk`.

For instance, the following three lines of Python will look up the lemma given an ID number:

```
url = "https://ordregister.dk/id/COR." + str(id) + ".json"
data = json.loads(urlopen(url).read())
word = data['lemma']
```

Most other resources will only have lists of defined lemmas and forms available on ordregister.dk. There will then be a link to the URL from which the resource can be accessed (if it is publicly available).

4. Applications of the COR

4.1 Historical Lexicography

The Danish Language Council’s RO^{hist} project (rohist.dk) is a search engine that allows users to compare Danish orthographical dictionaries from 1872 to 2012.

Work is ongoing to expand RO^{hist} with all Danish historical orthographical dictionaries and other orthographic resources, such as normative textbooks. For example, we are currently converting Ove Malling’s textbook *Store og gode Handlinger af Danske, Norske og Holstenere* (Malling, 1777) into a dictionary that can be added to RO^{hist} (cf. Hartling & Widmann (2020)).

The ID numbers in COR₁ correspond to the latest edition of *Retskrivningsordbogen*, but we plan to also assign COR numbers to the historical dictionaries in RO^{hist}. These dictionaries will be level 2 resources and will thus have their own prefix and ID number range.

The same ID number will be reused if the lemma is the same, even if the spelling has changed. A word like *fråse* (Dansk Sprognævn, 2012) will therefore have the same COR number as *frådse* in Dansk Sprognævn (1996) and as *fraadse* in Grundtvig (1872):

COR id	lemma
COR.37337	fråse
COR.RO2001.37337	fråse
COR.RO1996.37337	frådse
COR.DHO1872.37337	fraadse

This will simplify the implementation of RO^{hist} considerably. When searching for *fråse* in the future RO^{hist}, one would simply need to find the COR number (here 37337) and then determine whether this is defined in the historical dictionaries through a simple lookup.

The existing links between the dictionaries in RO^{hist} will form the basis for this work. We will therefore take the relational database underlying RO^{hist}, analyse the data, and assign historical COR ID numbers based on this analysis. This also means that if an error is found in RO^{hist} – for example, if a historical spelling has been linked to the wrong lemma – one would simply need to correct the COR ID number in the historical dictionary where the error occurred.

4.2 COR Linkers

In corpus linguistics and other computational linguistic applications, developing programs that assign the correct COR id (including the grammatical code) to each word in a text is essential. These programs are called COR linkers. With a text COR-linked, generating a part-of-speech-tagged text becomes straightforward since all necessary information is contained in the grammatical code. Moreover, COR linking allows for disambiguation of

homographs and makes the annotated text suitable for various NLP tasks, such as spelling and grammar checking, speech synthesis, machine translation, and dialogue systems.

Consider the Danish noun phrase “to kendte russiske historikere” (“two renowned Russian historians”) as an example of how to COR-link a text. Three of the four words have more than one potential match in COR₁ (in the table below, the correct link (match) is marked in bold):

Token	Meaning in English	Ambiguous COR ids
to	two	COR.01528.600.01 (numeral)
		COR.30835.200.01 (vb. inf.)
		COR.30835.209.01 (vb. imp.)
kendte	renowned (plur.)	COR.18159.302.01 (adj. sing. det.)
		COR.18159.303.01 (adj. plur.)
		COR.30330.206.01 (vb. past act.)
		COR.30330.214.01 (vb. past part. sing. det.)
russiske	Russian (plur.)	COR.30330.215.01 (vb. past part. plur.)
		COR.22261.302.01 (adj. sing. det.)
		COR.22261.303.01 (adj. plur.)
historikere	historians	COR.58774.112.01 (noun plur.)

At the Danish Language Council, a COR linker project called CLINK is currently being developed. It is an input-output automaton that accepts a tokenised text (or a full corpus) as input, expands the input to its maximal COR-linking, filters away irrelevant links (for homographic tokens only), and delivers a minimally linked version as output. CLINK uses several strategies to achieve an output as close to the optimal linking as possible.

The fundamental idea is to use three different analytical strategies: LSYN (local syntax), CTXT (context), and FREQ (frequency), each implemented as a stand-alone module. Input and output formats are the same: Each module reads a well-linked text as input and writes a well-linked text as output. CLINK modules can only *remove* links but cannot alter the input otherwise. Intuitively, each time a well-linked text passes through a CLINK-module, some of its lexical ambiguity is eliminated, altering the decision basis for the following iteration.

The LSYN module makes congruence-based decisions (sentence-internally), while CTXT is based on semantic triggers and long-distance associations. FREQ uses lookups in a frequency table (including bigrams and trigrams); it always outputs a minimally linked text (with a single link per token), guaranteeing a recall of 1.0 (but usually a less than satisfactory precision). In contrast, the recall of the other modules depends critically on the amount of ‘triggering’ contexts in the input text, and they typically show a very high precision at the expense of a low recall. Hence, FREQ is located as the last module in the CLINK pipeline (as the fallback strategy), ensuring that the output is indeed minimally linked.

As the input and output formats are the same in all modules, the modules can be swapped freely. One can also insert new modules. For instance, the possibility of creating an AI-based module (using TensorFlow) is currently being investigated.

4.3 Crowdsourcing and Invisible Lexicography

Companies and individuals can effortlessly contribute their own resources to the COR by applying for a unique prefix and number series, and then incorporating COR ID numbers into their data as previously described. This process will gradually transform COR into a vast, largely crowdsourced resource, particularly if new contributions are distributed under an open-source licence.

We encourage those who have requested a prefix and number series to publish their lemma lists on ordregister.dk, simplifying the process of finding relevant data. For instance, by visiting the website, one can identify who has defined data for ID 56746 (*avocado/avokado*).

The development of COR exemplifies the concept of invisible lexicography (using lexical data without users realising they are employing a "dictionary") by making lexical data machine-readable and integrating it seamlessly into various contexts. By assigning unique identification numbers to every lemma and form in the Danish language, COR provides a common key that facilitates more efficient reuse of language resources.

Fundamentally, COR is a developer-oriented feature with the potential to impact a broad range of user-facing applications, such as spellcheckers, translation services, and search engines. However, many users interacting with these tools may be unaware of the underlying database or the efforts involved in creating it.

COR represents an exciting advancement in the field of lexicography and language technology. By rendering lexical data machine-readable and accessible to developers, COR has the potential to revolutionise the way we process and analyse language. It also offers new opportunities for collaboration and crowdsourcing, as users can contribute their own data and insights to the database.

We hope that many will release COR-linked corpora and additional resources that further enhance the overall utility of COR.

5. Benefits and Applications of the COR

The COR offers a variety of advantages and potential applications within the realm of Danish language research and natural language processing. In this section, we shall outline the primary benefits of the COR and its multifarious applications.

Enhanced Resource Reusability and Collaboration: The COR encourages different resources to use the same lemma ID numbers, potentially adding *relations* to further describe the relationships between them. This approach will hopefully lead to enhanced resource reusability and collaboration.

Support for Historical Lexicography: The Danish Language Council's intention to make previous orthographical dictionaries available in COR format will enable users to trace the evolution of the language over time. This capability allows for more precise study of historical texts and the adaptation of NLP software to work on such texts.

Efficient POS Tagging and Homograph Resolution: The development of COR linkers, exemplified by the Danish Language Council's CLINK project, assigns the

correct COR number to every word in a text, thereby addressing part-of-speech tagging and homograph resolution concomitantly. This development simplifies language analysis and significantly bolsters the accuracy and efficiency of NLP applications.

Crowdsourcing in Lexicography: The COR permits users to contribute their own data and insights by publishing resources with added COR ID numbers. This approach encourages broader community participation in the development of the Danish language, culminating in a more comprehensive and diverse database that benefits both researchers and NLP practitioners.

Uncomplicated Access and Integration: The COR database can be accessed via a downloadable CSV file or an API, allowing developers to effortlessly retrieve ID numbers, lemmas, and forms in either CSV or JSON format. This streamlined access promotes the concept of “invisible lexicography”, enabling seamless integration with a variety of applications and tools.

In conclusion, the COR provides a groundbreaking foundation for the Danish language, augmenting collaboration, streamlining processes, and promoting further research and development. The benefits and applications of the COR extend beyond academia, opening up new possibilities for natural language processing, historical analysis, and the future of Danish language studies.

6. Future Prospects and Conclusion

As we have demonstrated throughout this article, the Central Word Register (COR) offers significant benefits and potential for applications Danish language research and natural language processing. In this concluding section, we will briefly discuss future prospects for the COR and summarise the key points of the article.

6.1 Future Prospects

The future of the COR project promises several exciting developments and prospects, which are outlined below.

Expansion of Semantic Components: As part of the COR project, a semantic component is being developed (Nimb et al., 2022), which will further enrich the database and allow for more sophisticated linguistic analyses and applications.

Development of Additional Tools and Applications: As the COR continues to evolve and expand, new tools and applications are expected to be developed. These may include advanced COR linkers, state-of-the-art natural language processing utilities and other innovative language technologies that will further enhance its utility and encourage its widespread adoption in language research and technology.

More COR Resources: With the ongoing development and promotion of the COR project, we anticipate a significant increase in the number of COR-tagged resources, stemming from both our own efforts and the collaborative contributions of the wider community through crowdsourcing initiatives.

Integration with Other Language Projects: The COR’s potential for integration with parallel projects in other languages offers the possibility of creating shared,

unified linguistic resources with other languages, particularly the other North Germanic languages. Such a resource could significantly advance language research and technology in the region, fostering greater collaboration and understanding among researchers and practitioners working in these languages.

In summary, the future prospects of the COR project are bright, with the potential for significant advancements in linguistic research and natural language processing technologies. The ongoing development of semantic components, tools, applications, and resources will further solidify the COR's position as a vital and innovative resource in the world of language research and technology.

6.2 Conclusion

In conclusion, Det Centrale Ordregister (COR) represents a groundbreaking initiative in the field of Danish language studies, addressing the challenges of resource compatibility and promoting greater collaboration, efficiency, and innovation. Through the establishment of a shared database key and a multi-level structure, the COR has the potential to significantly impact not only academic research but also the broader landscape of natural language processing and language technology.

With promising future prospects, including the addition of semantic components, integration with parallel projects, and the development of new tools and applications, the COR is poised to become an indispensable resource for researchers, practitioners, and enthusiasts alike. By providing both the motivation and practical skills to engage with the COR, we hope to contribute to a vibrant and thriving community of Danish language research and development.

7. Acknowledgements

I am grateful to my colleagues at the Danish Language Council (Dansk Sprognævn), as well as our partners, the Society for Danish Language and Literature (DSL, Det Danske Sprog- og Litteraturselskab) and the Centre for Language Technology (CST, Center for Sprogteknologi, KU) for their collaboration and dedication in tackling the challenge of creating the COR. I would also like to express my sincere appreciation to the Digitalisation Agency (Digitaliseringsstyrelsen) for their generous financial support, which has been instrumental in facilitating the realisation of this ambitious project.

8. References

- Dansk Sprognævn (1955). *Retskrivningsordbog*. Copenhagen: Gyldendal.
- Dansk Sprognævn (1996). *Retskrivningsordbogen*. Copenhagen: Aschehoug, 2 edition.
- Dansk Sprognævn (2012). *Retskrivningsordbogen*. Copenhagen: Alinea, 4 edition.
- Glahder, J. (1923). *Dansk Retskrivningsordbog. Udgivet af Undervisningsministeriets Retskrivningsudvalg*. Copenhagen: Gyldendal.
- Grundtvig, S. (1872). *Dansk Haandordbog med den af Kultusministeriet anbefalede Retskrivning*. Copenhagen: C. A. Reitzel.
- Hartling, A.S. & Widmann, T. (2020). Den første ortografiske rettesnor for dansk – fra læsebog til ordbog: Malling (1777) på <http://rohist.dsn.dk>. In Y. Goldshtein,

- I.S. Hansen & T.T. Hougaard (eds.) *18. Møde om Udforskningen af Dansk Sprog*. Institut for Kommunikation og Kultur, Aarhus University.
- LBK 332 (1997). Lov om dansk retskrivning. LBK nr. 332. URL <https://www.retsinformation.dk/eli/lta/1997/332>.
- Malling, O. (1777). *Store og gode Handlinger af Danske, Norske og Holstenere*. Copenhagen.
- Nimb, S., Pedersen, B.S., Sørensen, N.C.H., Flörke, I., Olsen, S. & Troelsgaard, T. (2022). COR-S – den semantiske del af Det Centrale OrdRegister (COR). *LexicoNordica*, 29, pp. 75–97.

Invisible meaning relations for representing near equivalents

Arvi Tavast¹, Kristina Koppel¹, Margit Langemets¹, Silver Vapper¹, Madis Jürviste¹

¹ Institute of the Estonian language, Roosikrantsi 6, Tallinn, Estonia
E-mail: first.last@eki.ee

Abstract

One of the key design principles of the Ekilex dictionary writing system is its symmetrical many-to-many relationship between word and meaning. Ekilex is currently being used for creating the EKI Combined Dictionary (CombiDic), with a primary goal of increasing coverage of languages beyond Estonian. This paper discusses the pilot project of integrating English, which began with generating a list of candidate equivalents for post-editing. The primary focus of the paper is on how near equivalents (narrower, wider, approximate) are represented in the symmetrical data model. Since meanings are language-independent entities in such a model, and equivalence is essentially about similarity of meanings, the near equivalents are represented using relations between meanings. To the dictionary user, the relations remain invisible and are only queried to retrieve target-language words for display. Transitioning from the traditional flowing text in the target language field to this more structured approach significantly affects the work process. We examine the advantages and disadvantages of this change in the paper.

Keywords: dictionary writing system; data model; multilingual dictionary; near equivalents

1. Introduction

The task of describing the lexical aspect of language has traditionally been assigned to lexical resources such as general and specialised dictionaries, termbases, lexicons, encyclopedias, and so on. The creation and utilization of these resources is a well-established field, boasting traditions dating back thousands of years. Lexicographers can trace their lineage back to the Sumerian-Akkadian bilingual word lists (Boulanger, 2003, p. 76), while terminologists can trace theirs to the Onomasticon of Amenemope (Boulanger, 2003, p. 111).

The format, structure and data model of lexical resources have remained largely unchanged for millennia, due to the restrictions of the publishing medium, which up until very recently has been a flat, two-dimensional, hierarchical, sequentially-accessed format like paper. The enduring influence of the paper mindset also manifests itself in the data models and creation principles of early electronic lexical resources and data exchange standards (e.g. Budin et al., 2012), with the possible exception of Ontolex-Lemon (McCrae et al., 2017). Three aspects of this heritage are now ripe for reevaluation.

1.1 Hierarchical data model

The lexicographic tradition of listing words (alphabetically or otherwise) and providing each with whatever information the lexicographer deems necessary is particularly ingrained among both lexicographers and readers (Atkins & Rundell, 2008; Flinz, 2011)

Indeed, since paper is not searchable, it needed an access structure built into the very organisation of the lexical resource. Two contrasting solutions were employed: onomasiology (concept-orientation, documenting concepts and their designations, mainly used in terminology) and semasiology (word-orientation, documenting words and their meanings, mainly used in lexicography). This is also the reason for the strict distinction between dictionaries and termbases, based on their method of compilation.

Both of these orientations result in a hierarchy rather than a network. In database terms, they are based on a one-to-many (1:n) relationship, either relating one concept to many terms or one word to many meanings. If there is repeated information in the latter side of the relationship (e.g. the same meaning for synonyms), there is no natural way to express that in the model. Such information can be simply repeated, or addressed with a cross-reference. It is worth noting that serial data exchange formats based on XML or JSON are also inherently hierarchical in this regard.

Maintaining consistency, i.e. guaranteeing that all repetitions are handled purposefully without unnecessary duplication or internal conflicts, has been a challenge for even the most diligent lexicographers. Readers are routinely provided with conflicting information within a single lexical resource. (Tavast, 2008; Tavast & Taukar, 2013)

This simplistic type of relationship, and therefore the opposition between the orientations, was natural and necessary on paper. However, with more expressive formats now available, there is no need to uphold it.

1.2 Directionality

The concept of directionality is deeply entrenched in general language lexicography (e.g. Adamska-Sałaciak, 2014). In this model, one language is designated as the source language, with any others considered target languages. This is based on the understanding that exact equivalence between languages is unattainable in a dictionary of any practical size. Consequently, the target language side can't simply consist of a single word. Instead, it must convey the full richness of the source language using a variety of means: typically more than one equivalent, words that merely suggest the meaning of the source word, rarely used words in the target language, words with domain or register qualifiers, extended explanations, and so on.

Directional compilation raises the issue of dictionary reversal (Sierra, 2000). The experience with the Estonian-Russian dictionary (EVS) reveals that high-frequency

target language words were frequently used as equivalents for non-synonymous headwords. This created a misleading impression that these words had as many meanings as they were used as equivalents for. As this dictionary was compiled semasiologically and directionally, importing it into the non-directional Ekilex system revealed a different perspective on equivalences, which the authors found unsatisfactory.

1.3 Authored work

The third tradition we aim to question is the perception of a dictionary as an authored work, reflecting the views of the author(s) rather than serving as a source of objective information about language. Dictionaries are even granted a certain level of copyright protection (see Langemets & Voll, 2008 for a case study of our own experience).

This perspective has been both convenient and beneficial, acting as a shield: given the impossibility of complete objectivity in language description, the author has full discretion over the dictionary’s content. Two authors describing the same language will invariably produce different outputs. The majority of content disagreements can be dismissed by citing the inherent subjectivity of each description.

While the utility of this view on subjectivity is perfectly understandable from the author’s perspective, it may not align with the reader’s expectations. Although the personal insights of authors can be intriguing, it’s reasonable to assume that at least some readers are seeking information about language instead.

2. Background of Ekilex

Since 2017, the Institute of the Estonian Language (EKI) has been developing Ekilex, an in-house dictionary writing system (Tavast et al., 2018, 2021). One of its central design principles is the symmetry of its data model: the many-to-many relationship between word and meaning simultaneously accommodates semasiological and onomasiological resources. It is currently being used for compiling the general dictionary of Estonian – EKI Combined Dictionary (CombiDic) – as well as over 120 termbases. Lexicographers and terminologists are working on the same data, but from opposing viewpoints. Completed resources are accessible to readers via the language portal Sõnaveeb (Koppel et al., 2019).

At the heart of the Ekilex data model is a many-to-many relationship between word and meaning: a word can have multiple meanings and a meaning (or concept) can be designated by multiple words (terms) in several languages.

A word in this model is a language-specific but meaning-agnostic character sequence, containing data elements that do not depend on the meaning, such as language, gender, aspect, morphology, pronunciation, etymology. Conversely, a meaning is a language-agnostic unit of knowledge containing data elements that do not depend on how (or if

at all) this meaning is expressed in any languages, such as domain, semantic type, definition.

To implement a many-to-many relationship between these two main entities in relational database terms, we use a link table. In its purest form, a link table only contains pairs of word IDs and meaning IDs, indicating which word is associated with which meaning. However, during the initial design of the data model, we quickly realised that a substantial proportion of all data categories – ranging from part of speech to example sentences – belong to this link table, rather than to the word or meaning themselves. We defined the link entity as "this word in this meaning as described in this dictionary," and called it a lexeme. A lexeme contains information that depends on the combination of word and meaning, such as part of speech, usage example, collocation, register, and proficiency level.

The number of possible meanings greatly exceeds the number of words in any language. ‘The human brain contains eighty-six billion neurons, each with about ten thousand synaptic contacts whose strength can vary. The space of mental representations that opens up is practically infinite.’ (Dehaene, 2020, p. 10) A fundamental challenge for creating lexical resources is therefore the need to simplify the continuous reality of language into the discrete representation of a dictionary.

There is an important consequence for the dictionary data model, especially one (like Ekilex) where meanings have their own database entities rather than being represented by free-form text. While the database entities for words correspond non-controversially to words in language and are able to represent their relevant properties (orthography, morphology, etymology, etc.) exhaustively, meanings are more difficult first to individuate and then to describe. Decisions regarding how fine sense distinctions should be and what exactly the senses are, depends on various factors including the volume of the dictionary, purpose, target group and even available funding.

Language	Words in CombiDic	Words in termbases	Total words
Estonian	159,891	141,982	301,873
Russian	172,393	57,531	229,924
English	2,945	89,521	92,466
Latin	4,051	21,766	25,817
German	2,116	17,264	19,380
French	7,983	9,184	17,167
Norwegian	19	14,097	14,116

Ukrainian	11,270	2	11,272
Finnish	1,518	7,943	9,461
Spanish	127	2,257	2,384

Table 1: Languages with at least 1000 words in the datasets of Ekilex (as of 11 April 2023).

Currently, all datasets in Ekilex contain Estonian as one of their languages. Termbases are mostly concept-oriented and consequently directionless, but in CombiDic, where directionality is still pertinent, Estonian has thus far maintained a special status as the pivot language. For lexicographers, this means originating dictionary entries from the Estonian side and adding equivalents in other languages. For readers, this implies that the opposite direction (e.g. English-Estonian) and other combinations (e.g. German-French) are accessible if searched for, but might not have been thoroughly reviewed by a lexicographer. Table 1 lists the most widely covered languages in the datasets of Ekilex. The seemingly random variations are due to external factors, including the availability of existing material (Russian, Norwegian), special status of a language (Latin in life sciences) and recent world affairs (Ukrainian).

One of our purposes has been to increase foreign language coverage in CombiDic. We started a new project in 2021 to semi-automatically add English equivalents. The project had a dual goal: to add the foreign language most widely spoken in Estonia, and to design and test the whole process for adding other languages in the future. The remainder of this paper addresses two challenges:

- Generating a list of candidate English equivalents for the Estonian headwords for manual post-editing by lexicographers.
- Integrating multiple bilingual dictionaries into the Ekilex data model and systematically managing their interrelations within the model.

3. Generating candidate equivalents

To add English equivalents to the Estonian headwords in CombiDic using a process of post-editing lexicography (Jakubíček et al., 2018, 2021), a dataset of possible candidates was automatically generated. We used two existing English-Estonian dictionaries: the English-Estonian Machine Translation Dictionary compiled by Indrek Hein of the Institute of the Estonian Language, and the Password Estonian-English Glossary compiled by K Dictionaries in cooperation with the Institute and the publishing house TEA (Langemets et al., 1999; Kernerman, 2015). To ensure wider vocabulary coverage, we gathered possible equivalents from parallel corpora.

Equivalents were gathered by processing sentence pairs and doing word alignments using the ArgMax matching method (Sabet et al., 2021), which were then gathered in frequency lists. We chose this specific algorithm based on a sample ‘gold standard’ parallel data set for Estonian–English word alignments.

Two types of sources based on presumed translation quality were used: a proprietary corpus based on professional translation memories, and publicly available corpora. Detecting potential candidates from publicly available corpora led to a lot of noise in the data, e.g. candidates including numbers, symbols, foreign alphabets and punctuation marks that were all automatically deleted before importing into Ekilex. While frequency lists based on translation memories included less noise than those from public corpora, they still required substantial reductions and filtering.

Most of the additional filtering was based on statistical relevance and heuristics taken from random samples of data. For instance, we removed candidates with a frequency of 1 or 2 from headwords with more than five different equivalent candidates, as these low-frequency matches were almost always incorrect.

When importing the candidate equivalents, we set the threshold from 5 to 30. When a headword had fewer than five candidates, we imported all of them, even if the frequency was 1. Prior to importing the data into Ekilex, we combined the corpus and dictionary data, assigning weights to candidates based on their origin. These weights determined the visible order of equivalent candidates in Ekilex. We also appended metadata—such as part of speech information, example usages, and definitions—to candidates sourced from dictionaries."

4. Bilingual data in a many-to-many data model

In this section, we discuss how bilingual dictionaries fit in the many-to-many data model of Ekilex. Specifically, we detail three key insights this model provides to bilingual dictionary authors, along with their associated costs and benefits.

4.1 Model structure

The bilingual dictionaries under discussion here belong to general language lexicography, which has traditionally employed a semasiological data model. The central entity in such a model is the word, with its senses branching out hierarchically (each word has one or more senses). As the Ekilex data model is symmetrical between word and meaning, and meaning has its own set of language-independent properties, we can transcend this simple hierarchy.

The essence of equivalence is a meaning relation: equivalent words share the same meaning. The Ekilex many-to-many data model represents these situations using a single mechanism, connecting the word entities to the same meaning entity. A meaning

has two words of the same language in the case of full synonymy, and of different languages in the case of full equivalence.

An immediate objection is that in language reality, perfect equivalence between languages or absolute synonymy within a language is an extremely rare and possibly non-existent phenomenon (Lyons, 1981; Cruse, 1986; Murphy, 2003; Pym, 2010). The meaning of a lexical item is not even identical across speakers of a single language, and keeps developing during the lifespan of a single individual as exposure to linguistic input accumulates (Ramskar et al., 2013, 2014).

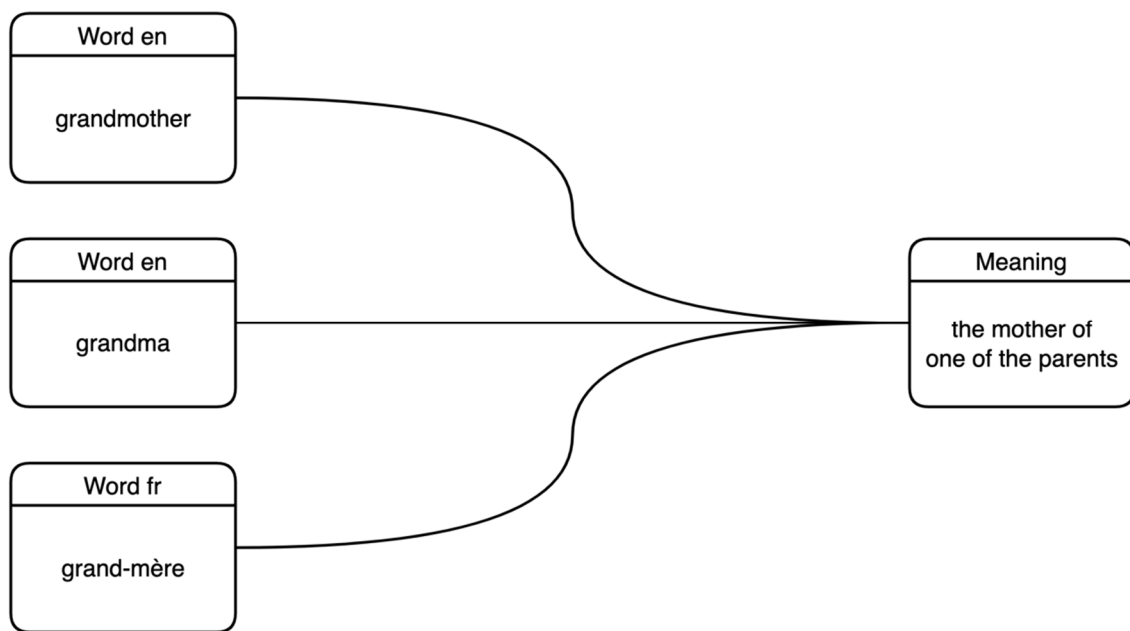


Figure 1: Full synonymy (between *grandmother* and *grandma*) and equivalence (between these and *grand-mère*) represented as all three words having the same meaning.

It is both a lexicographic tradition and an inevitable need to simplify language reality in order to fit it into the finite form of a dictionary. This includes claiming full equivalence or synonymy between lexical items with meanings that the lexicographer considers sufficiently close, as shown on Figure 1. A rule of thumb used in practice is to see full equivalence or synonymy only when the definition is exactly the same. So what we are changing in the case of full equivalents, is only the technical implementation, not the lexicographic principle.

This paper is concerned with the next step: what if the meanings are so different that they can't possibly be simplified into a claim of full equivalence, but still close enough to qualify as candidates for being represented as some sort of equivalents in a bilingual dictionary? Recurrent examples of this include the following:

- Meanings that are not lexicalised in one of the languages, or where the target-language word is too rare for inclusion in the dictionary. Example: '*grandmother*'

in Swedish, where one needs to specify between *mormor* ‘mother’s mother’ and *farmor* ‘father’s mother’ instead.

- Meanings that are culturally different but are still considered to somehow correspond to each other, at least within the precision limits of the dictionary. Example: French *pain*, English *bread* and German *Brot* may be equivalent in a very broad sense, but they are culturally different enough in their shape, colour, texture and taste to warrant a more detailed treatment in a more advanced dictionary.

Our current solution to this situation is that the data model stays the same, each word still has its own meaning (exact meaning, given the level of simplification chosen for the dictionary), and there is a similarity relation between those meanings. So instead of representing that these *words* are similar in their meaning, we represent that these words have *meanings* that are similar, see Figure 2.

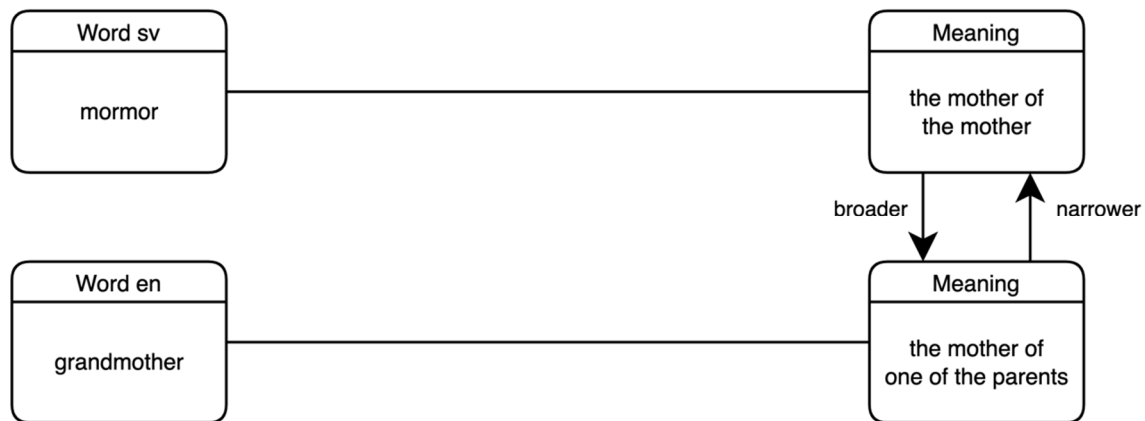


Figure 2: Narrower and wider equivalents represented as a meaning relation.

We have chosen to use three types of meaning relations for representing near equivalents: narrower ($A > B$), wider ($A < B$) and approximately same ($A \approx B$), where A and B designate meanings.

Representing near equivalents with meaning relations has the counter-intuitive consequence that not all meanings have designations in all languages. For the reader, these meaning relations themselves remain invisible, but are traversed in order to retrieve the corresponding target language words and render a habitual presentation of near equivalents.

4.2 Directionality

Cross-linguistic equivalence is symmetrical by nature. From $A = B$, it inevitably follows that $B = A$, and any claim to the contrary is motivated solely by lexicographic tradition.

Outside the dictionary, languages are not inherently “source” or “target”. The same lexical resource can be queried¹ in any direction, and should yield sensible results regardless of the direction, even if the equivalences are not exact. Table 2 lists situations that arise when near equivalence relations are viewed from directions other than the original premise that the lexicographer had in mind.

Premise	Inference	Description
$A \approx B$	$B \approx A$	Approximate equivalence is symmetrical. If <i>bread</i> is almost equivalent to <i>Brot</i> , then <i>Brot</i> is almost equivalent to <i>bread</i> .
$A > B$	$B < A$	Wider and narrower are opposites. If <i>mormor</i> is a narrower equivalent for <i>grandmother</i> , then <i>grandmother</i> is a wider equivalent for <i>mormor</i> .
$A > B$ and $B = C$	$A > C$	Adding more languages requires coordination between all languages. If <i>mormor</i> is a narrower equivalent for <i>grandmother</i> , and <i>grandmother</i> is a full equivalent for <i>grand-mère</i> , then <i>mormor</i> is a narrower equivalent for <i>grand-mère</i> .

Table 2: Types of relations between the meanings of words of three languages: *A*, *B* and *C*.

The possibility of being queried in any direction could also be described as automatic and immediate reversal of the bilingual dictionary, which understandably complicates the lexicographer’s task. It is no longer sufficient or even possible to use the target language field for any explanatory information that comes to mind (equivalent, approximate equivalent, several equivalents based on meaning nuances or usage patterns, explanation in case there is no equivalent, etc.). The following additional tasks need to be considered:

- Separation of data types. Each data element needs to go to its own field, rather than as flowing text in a single large field. The fields may not even belong to the same database entity, e.g. it is important to distinguish between properties of the target word in this meaning (e.g. register) and the meaning itself (e.g. domain, definition). In our experience, this has proven to be difficult already in a monolingual situation, and the situation will be further complicated with

¹ In our own resources, this was the case already before Ekilex, e.g. in the Estonian-Finnish <http://www.eki.ee/dict/efi> or the Estonian-Russian <https://portaal.eki.ee/dict/evs/>.

more languages, as described below.

- In the process of entering an equivalent, the lexicographer needs to immediately consider all properties of the target word, including definitions and example sentences, what other meanings does the target word have, or where else it has been or will be used as an equivalent. Especially the latter potentially creates a rabbit hole for the lexicographer to fall down, in the style of Dyvik’s semantic mirrors (1998, 2004). The work process thus needs to accommodate the following of mirrored chains of equivalence, limit their depth somehow, or include a separate step for cleaning up the opposite language direction.

4.3 Authorship

It has traditionally often been the case that bilingual dictionaries (e.g. Estonian-English and Estonian-French) are separate works authored by non-overlapping groups of lexicographers, even if they share one of the languages. This organisation of work is incompatible with the understanding that equivalence is about meanings: full equivalents share the same meaning and partial equivalents have related meanings. Meanings are independent of languages and especially of language pairs. To continue with the example used above, it is difficult to imagine how the assertion that “mother’s mother is a type of grandmother” could depend on the language(s) in question, so it should be safe to enter it as a language-independent meaning relation.

Consequently, equivalence information entered by the team working on one language pair has an effect on all other language pairs. Here are some situations from our initial experience where this may become an issue:

- An assertion may or may not correspond to facts of life, or its degree of simplification may be debatable. However, both its truth value and the suitability of the degree of simplification remain language-independent. If some factual claim needs correcting, then it needs correcting for all languages, which in turn requires coordination between the teams of all languages.
- The need to express the meaning relation in the first place does depend on specific languages, in this case Swedish. Without Swedish, full equivalence between *grandmother*, *grand-mère*, *Großmutter* etc. would probably be sufficient for a general dictionary. Once Swedish is added, though, the number of related meanings is increased from one to three, and all language teams need to decide whether the added meanings of mother’s mother and father’s mother require a word in their language. For Estonian, as an example, they might, as the words *emaema* ‘mother’s mother’ and *isaema* ‘father’s mother’ do exist, even if used much less frequently than *vanaema* ‘grandmother’.
- Adding even more languages may introduce more distinctions based on

parameters that were unlexicalised in previously added languages, e.g. whether the grandparent is living or deceased. Two issues may arise here: cooperation will be needed between languages making the same distinction, and intersecting multiple distinctions may result in a network of relationships that is difficult to understand as a whole.

- One of the methods for starting a new dictionary project is to import material from existing dictionaries. If these are traditional enough, they will probably contain manually written textual solutions for representing near equivalents, e.g. a sentence explaining that Swedish distinguishes between maternal and paternal grandmother. If this sentence is imported for one language *and* the meaning relations are created for another language, then the same information will appear twice in different wordings for the reader. Again, cooperation is required, and rephrasing or even simply removing such duplication may involve significant amounts of work.
- As the number of authors increases with the number of languages, they will more frequently introduce changes that may affect other languages. Staying on top of the flow of changes will require either an alerting system or periodic “sanity check” queries from the database. In both cases it depends on non-trivial organisational decisions about what kind of changes need the attention of other languages. A balance between overwhelming numbers of notifications and the danger of missing an important change needs to be worked out in practice.

So if we continue with the assumption that lexicographers are human (as opposed to artificial intelligence) and therefore limited in their language proficiency in all the language pairs that may need a bilingual dictionary, the only way forward is cooperation.

A recurring request that EKI receives from potential dictionary teams is to use Ekilex for authoring a stand-alone unidirectional bilingual dictionary, often with Estonian as the target language. While granting such requests would be technically possible in the same way that specialised dictionaries are created as stand-alone works, we have chosen not to. We invite them to cooperate with the CombiDic team to add their language(s) to CombiDic instead.

The objective is to eventually have hundreds of languages in CombiDic, with the consequence that the dictionary will have hundreds, if not thousands of authors contributing to various languages, some of them professionally, but many sporadically. The potential challenge of managing such a huge team, both organisationally and regarding intellectual property rights, is acknowledged, but is outside the scope of this paper.

5. Summary

As we have started to add new languages to CombiDic, the symmetrical data model of Ekilex has brought about a number of changes compared to traditional bilingual lexicography.

Our primary objective has been to develop a data structure intended to unify and formalise relationships of both full and near equivalence. Although the meaning relations proposed in this paper will remain invisible, the dictionary users stand to benefit from these in the form of better considered and coordinated equivalents.

We began by detailing the process of generating candidate equivalents for post-editing lexicography and subsequently explored the costs and benefits of the symmetrical data model for integrating a multitude of languages into CombiDic. Given that equivalence fundamentally pertains to meanings, it is represented at the meaning level in the database. Full equivalents relate to the same meaning, while each near equivalent has its own meaning, with these meanings being interrelated. Currently, we employ three types of meaning relations: wider, narrower, and approximate.

The flip side of the benefits of better coordination and uniformity of lexicographic principles is the required change in the work process. Entering more information or more thoroughly considered information is inevitably more labour-intensive than the habitual approach of entering much less information. The only reason for undertaking such change is to eventually provide a superior dictionary for the user.

As we are in the first phases of adding the pilot language (English), there a lot to learn about the data model and the work process, especially how both unfold in practice.

6. Acknowledgements

This work was supported by the Estonian Research Council grant PRG 1978.

7. References

- Adamska-Sałaciak, A. (2014). Explaining Meaning in a Bilingual Dictionary. In P. Durkin (Ed.), *The Oxford handbook of lexicography*. Oxford: Oxford University Press (forthcoming).
- Atkins, S. B. T., & Rundell, M. (2008). *The Oxford guide to practical lexicography*. Oxford University Press.
- Boulanger, J.-C. (2003). *Les Inventeurs de dictionnaires: De l'eduba des scribes mésopotamiens au scriptorium des moines médiévaux*. Les Presses de l'Université d'Ottawa.
- Budin, G., Majewski, S., & Mörth, K. (2012). Creating Lexical Resources in TEI P5. *Journal of the Text Encoding Initiative, Issue 3*, Article Issue 3.

<https://doi.org/10.4000/jtei.522>

- CombiDic: Langemets, M., Hein, I., Jürviste, M., Kallas, J., Kiisla, O., Koppel, K., Kuusk, K., Leemets, T., Mäearu, S., Paet, T., Päll, P., Raadik, M., Risberg, L., Tammik, H., Tavast, A., Tiits, M., Tsepelina, K., Tuulik, M., Valdre, T., Viks, Ü., Sai, E., Smirnova, A., Tubin, V., (2022). EKI ühendsõnastik 2022 [EKI Combined Dictionary 2022]. Sõnaveeb: Eesti Keele Instituut.
- Cruse, D. A. (1986). *Lexical semantics*. Cambridge university press.
- Dehaene, S. (2020). *How We Learn: The New Science of Education and the Brain*. Penguin UK.
- Dyvik, H. (1998). A translational basis for semantics. *Language and Computers*, 24, 51–86.
- Dyvik, H. (2004). Translations as semantic mirrors: From parallel corpus to wordnet. *Language and Computers*, 49(1), 311–326.
- EVS: Laasi, H., Lagle, T., Leemets, H., Liiv, M., Pärn, H., Simm, L., Viks, Ü., Õim, A., Kallas, J., Kotova, N., Melts, M., Matt, K., Tubin, V., (auth.); Liiv, M., Melts, N., Romet, A., Kallas, J., Riikoja, E., Martoja, I., Smirnov, S., Tetsov, M., Tiits, M., Valdre, T., Veskimägi, E., (ed.) (2019). Eesti-vene sõnaraamat 2019. (1. trükk 1997–2009.) 2., täiendatud ja kohandatud veebiväljaanne [Estonian-Russian dictionary 2019]. Sõnaveeb: Eesti Keele Instituut.
- Flinz, C. (2011). The microstructure of Online Linguistics Dictionaries: Obligatory and facultative elements. *Electronic Lexicography in the 21st Century: New Applications for New Users. Proceedings of ELex 2011, 10-12 November 2011, Bled, Slovenia*, 83–88.
- Jakubíček, M., Kovář, V., & Rychlý, P. (2021). Million-Click Dictionary: Tools and Methods for Automatic Dictionary Drafting and Post-Editing. *EURALEX XIX*.
- Jakubíček, M., Měchura, M., Kovar, V., & Rychly, P. (2018). Practical post-editing lexicography with lexicology and sketch engine. *The XVIII EURALEX International Congress*, 65.
- Kernerman, I. (2015). A multilingual trilogy: Developing three multi-language lexicographic datasets. *Dictionaries: Journal of the Dictionary Society of North America*, 36(1), 136–149.
- Koppel, K., Tavast, A., Langemets, M., & Kallas, J. (2019). Aggregating dictionaries into the language portal Sõnaveeb: Issues with and without a solution. *Electronic Lexicography in the 21st Century. Proceedings of the ELex 2019 Conference. 1-3 October 2019, Sintra, Portugal*, 1–3.
- Langemets, M., Tiits, M., Uibo, U., Pihlak, A., & Kernerman, L. (Eds.). (1999). *Password: Inglise-eesti seletav sõnaraamat = English dictionary for speakers of Estonian* (6. tr). TEA Kirjastus.
- Langemets, M., & Voll, P. (2008). Sõnaraamatu kohtulingvistiline analüüs: Eesti pretsedent [Linguistic forensic analysis of a dictionary: an Estonian precedent]. *Eesti Rakenduslingvistika Ühingu Aastaraamat*, 4, 67–86.
- Lyons, J. (1981). *Language, meaning and context*.
- McCrae, J. P., 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: Proceedings of eLex 2017 Conference, Leiden, 2017* (pp. 587–597). Lexical Computing CZ s.r.o. <https://elex.link/elex2017/proceedings-download/>
- Murphy, M. L. (2003). *Semantic relations and the lexicon: Antonymy, synonymy and other paradigms*. Cambridge University Press.
- Pym, A. (2010). *Exploring Translation Theories*. Routledge.
- Ramscar, M., Hendrix, P., Love, B., & Baayen, R. H. (2013). Learning is not decline: The mental lexicon as a window into cognition across the lifespan. *The Mental Lexicon*, 8(3), 450–481.
- Ramscar, M., Hendrix, P., Shaoul, C., Milin, P., & Baayen, R. H. (2014). The myth of cognitive decline: Non-linear dynamics of lifelong learning. *Topics in Cognitive Science*, 6(1), 5–42.
- Sabet, M. J., Dufter, P., Yvon, F., & Schütze, H. (2021). *SimAlign: High Quality Word Alignments without Parallel Training Data using Static and Contextualized Embeddings* (arXiv:2004.08728). arXiv. <https://doi.org/10.48550/arXiv.2004.08728>
- Sierra, G. (2000). The onomasiological dictionary: A gap in lexicography. *Proceedings of the Ninth Euralex International Congress*, 223–235.
- Tavast, A. (2008). *The Translator is Human Too: A Case for Instrumentalism in Multilingual Specialised Communication*. Tartu Ülikooli Kirjastus.
- Tavast, A., Koppel, K., Langemets, M., & Kallas, J. (2021). Towards the Superdictionary: Layers, Tools and Unidirectional Meaning Relations. *EURALEX XIX*.
- Tavast, A., Langemets, M., Kallas, J., & Koppel, K. (2018). *Unified Data Modelling for Presenting Lexical Data: The Case of EKILEX*.
- Tavast, A., & Taukar, M. (2013). *Mitmekeelne oskussuhtlus*. Valgus.

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

Olena Synchak

Ukrainian Catholic University, 2a Kozelnytska Str., Lviv, 79026, Ukraine

E-mail: o_synchak@ucu.edu.ua

Abstract

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

Keywords: Military feminine personal noun, Web Dictionary of Ukrainian Feminine

Personal Nouns (WDF), r2u.org.ua, General Regionally Annotated Corpus of Ukrainian, GRAC, dictionary entry, subject labels

1. Introduction

Feminine personal nouns, also known as feminine terms, are a vibrant and growing segment of the Ukrainian lexicon. Their dynamic expansion through the derivation of new words has attracted the attention of linguists, including O. Synchak (2022), N. Kostusiak et al. (2020), V. Machek (2020), and N. Klymenko (2019). The linguists primarily investigate feminine terms by examining press articles (Kravets, 2021; Styshov, 2020; Navalna, 2017), works of fiction (Zayets, 2020; Brus, 2017-2018; Kaidash, 2017), and dictionary entries (Synchak & Starko, 2022; Tomilenko, 2021; Puzyrenko, 2012). Their research focuses on revealing the derivational models used (Neliuba, 2011; Semeniuk, 2000), studying the historical development of these terms (Brus, 2019), and exploring the impact of socio-cultural factors on their modern usage (Arkhanhelska, 2019).

Feminine personal nouns are primarily formed by adding a suffix to a masculine form to indicate the female gender. These terms encompass a wide range of female professions, positions, activities, actions, beliefs, community affiliations, places of residence, and other characteristics possessed by women. The usage of feminine terms

in texts has gradually become the subject of corpus-based studies in different languages (Machek, 2022; Koster, 2020; Elmiger, 2009). Researchers have also started to apply corpus data to study the system of Ukrainian feminine personal nouns (Starko & Sychak, 2023; Sychak & Starko, 2022).

One of the earliest studies on Ukrainian feminine personal nouns was carried out by I. Feketa (1968), who classified them into lexico-semantic groups. Recently, M. Brus (2019) has also applied the principle of thematic grouping of feminine terms in her monograph, complementing it with derivational and functional analysis, as well as diachronic and synchronic approaches. However, few researchers have examined the functioning of feminine terms within thematic groups or attempted to define the most dynamic group of feminine derivatives.

Using corpus data over the last 20 years, V. Starko and O. Sychak (2023) argue that the dynamics of using feminine terms in media texts correspond to their thematic groups. Thus, feminine terms from sports and military spheres have different dynamics of usage in the press. If sports feminine terms are produced dynamically throughout a 20-year term (with slight peaks during periods of championships or Olympics), then the number of women's titles in the military sector is growing dramatically, first, after Russia's armed attack on the territory of Ukraine in 2014 and the annexation of Crimea, and especially – after the full-scale invasion of February 24, 2022 (Starko & Sychak, 2023).

The focus of this study is on military feminine terms, which predominantly refer to women serving in the military (*генералка* '(female) general', *снайперка* '(female) sniper' etc.). However, it also includes women involved in logistics (*волонтерка* '(female) volunteer') or non-combat roles (*парамедикиня* '(female) paramedic' etc.). In recognizing the constant danger faced by these women operating in the military amidst war, we utilize military feminine personal nouns to acknowledge their contributions in this specific context. To delve into the examination of military feminine terms, an article by O. Mykhailova and T. Spilnyk (2019) and a brief passage in T. Kravets's dissertation (2019, pp. 154-155) are dedicated to this subject.

The Ukrainian media is actively creating new terms to depict the actions of female soldiers, while simultaneously developing nuanced meanings (Abetka feminityviv, 2022). These processes are happening at such a rapid pace that traditional dictionaries struggle to keep up. Instead, electronic dictionaries with the means of corpus data devise innovative methods of describing feminine personal nouns in dictionary entries.

2. Military feminine terms in the scope of online dictionaries

The *Web Dictionary of Ukrainian Feminine Personal Nouns* (WDUF in what follows) is a corpus-based dictionary. It was compiled by Olena Sychak, with academic editing by Hanna Dydyk-Meush and academic consultation by Vasyl Starko

(WDF, 2022). The dictionary was published on the r2u.org.ua dictionary portal in early 2022. WDF includes approximately 2,000 female terms, with a particular emphasis on derivational variations such as *міністерка* – *міністриня* – *міністреса* meaning ‘(female) minister’. The dictionary provides full definitions, supplies illustrations from a GRAC corpus, and lists other dictionaries that register the feminine noun in question (Synchak & Starko, 2022). The corpus data is enhanced with input from a language panel and recommendations from the compiler. Because of the various methods used to collect and present linguistic data, this resource could be a robust foundation for the standardization of feminine personal nouns in the Ukrainian language (Synchak, 2022).

Military feminine personal nouns make up 5% of the total number of words in the WDF, that includes 76 military terms: *військовослужбовиця* ‘servicewoman’, *лейтенантка* ‘(female) lieutenant’, *воїнка* ‘(female) warrior’, *генералка* ‘(female) general’, *адміралка* ‘(female) admiral’, and others. Most of these words are not registered in explanatory dictionaries of the Ukrainian language, but they have made it into the WDF thanks to usage data discovered in the GRAC corpus (Shvedova et al., 2023) and Google search engine. Entries with derivational variants are nested (see Fig. 1 below), meaning that the definitions and illustrations are provided for each feminine form. All derivational alternatives are presented on an equal basis in order to assist readers in selecting the most appropriate feminine term.

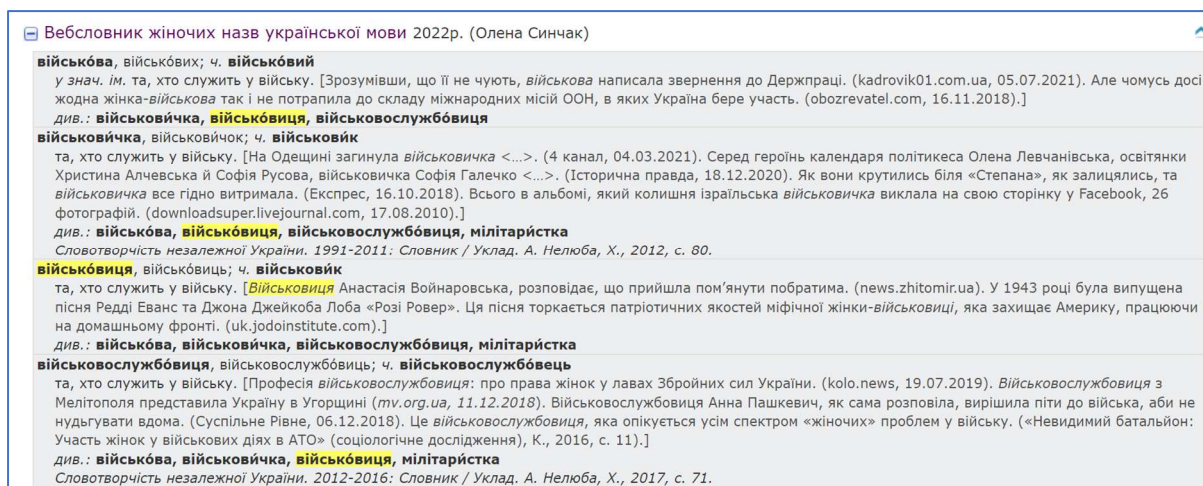


Figure 1: Derivational alternatives in the WDF (r2u.org.ua)

The online dictionary *Alphabet of Feminine Personal Nouns* (‘Alphabet’ in what follows) was created for the needs of the journalistic community, which highlights the war of Russia against Ukraine (Abetka femininityviv, 2022). It consists of 106 lexical items, among which 76 titles were taken from the WDF (r2u.org.ua, Lviv, 2022), and the remaining 30 were elaborated specifically for ‘Behind the Gender’ project, aimed at fair representation of women and men in Ukrainian media.

Despite the fact that the WDUF served as the inspiration for the ‘Alphabet’ dictionary entry, the lexicographic description of feminine terms in the latter has been slightly condensed (Fig. 2). First, only meanings associated with the military, a woman's participation in the war, or her new status as a result of the war are provided in the case of words with several meanings. Second, the dictionary only includes one word, such as *парамедиця* ‘(female) paramedic’, with a reference to other derivational alternatives included in the WDUF if one of the multiple alternative names is already well-known or has advantages over the others. However, if all derivational variants are equal (*миротвориця* // *миротворка* ‘(female) peacemaker’) or none of them has been established so far (*табірниця* // *таборянка* ‘(female) camper’), then all the options are given in the dictionary.

There are usage notes in the ‘Alphabet’ that are formed as a ‘Interesting to know’ section after certain entries, just like the ‘Recommendations’ section of the WDUF. In this section, attention is focused on the nuances of meanings, spelling variants or other aspects of using certain feminine terms in the military sphere. Feminine personal nouns unregistered in the WDUF are specially described for the project ‘Behind the Gender’, but later they will be added to the former.

• **гранатомётниця, гранатометниця; ч. гранатомётник**

Військова, яка стріляє з гранатомета, метає гранати. — Хто, як не ми? Історія *гранатометниці* та радіотелефоністки Сови, яка з 2014 в лавах ЗСУ (vikna.tv, 08.10.2022). *Гранатометницю* з Буковини нагородили Хрестом бойових заслуг (zahid.espresso.tv, 31.07.2022). Військовослужбовиця Збройних Сил України з позивним «Натаха» служить *гранатометницею* на передовій, неподалік Кримського. (Суспільне Донбас, 14.10.2021).

Figure 2: A dictionary entry in the ‘Alphabet of Feminine Personal Nouns’ (behindthenews.ua)

The primary distinction between these two dictionaries is revealed in the functional capabilities of the websites where they are hosted. If the *r2u.org.ua* lexicographic platform enables a full-text search (to search for feminine forms, you need to select the WDUF among other dictionaries on the platform), then on the *behindthenews.ua* website a search is only available by letters of the alphabet.

3. Method and material

The main aim of this study is to collect military feminine personal nouns in order to determine the frequency of their use in the corpus and to analyze changes in their semantics. By applying corpus data, I look for the best way to present military feminine terms in the updated version of the WDUF. For this purpose, 134 military

feminine titles are being analyzed. They were chosen based on the WDUF (76 titles) and the ‘Alphabet’ (30 titles) registries, and they were reinforced by those selected from the GRAC-16's unlemmatized military feminine personal nouns (28 titles). Thus, all the sources utilized in this paper are freely accessible.

GRAC is made up of a wide range of sources and texts and includes the most extensive and detailed metadata (Shvedova et al., 2023). The latest version 16 has recently been released, encompassing the period from 1816 up until 2022. Almost 1.9 billion tokens, or 1.5 billion words, make up the corpus as a whole. Mass media texts have been significantly increased in the most recent version of the corpus, especially for the years 2021–2022. There are many more media outlets now, and some have archives dating back 10–15 years. This makes GRAC a crucial tool for studies involving chronology (Starko & Synchak, 2023). This expansion of the corpus is crucial for the study of the feminine nouns in question since it enables to observe their usage as thoroughly as possible.

GRAC allows the creation of search queries through the use of the Corpus Query Language (CQL), which enables the combination of morphological and semantic tags (Starko & Synchak, 2023). For example, it is possible to search for feminine terms by utilizing the CQL query according to the most productive suffixes, such as *-к(а)*, *-иц(я)*, *-ин(я)*:

```
[tag="noun:anim.*"&lemma=".*ка"]  
[tag="noun:anim.*"&lemma=".*иц(я)"]  
[tag="noun:anim.*"&lemma=".*ин(я)"]
```

Using these CQL queries, unlemmatized feminine lexical items in GRAC-16 were identified and assigned the POS tag “unknown” through a semi-automatic approach. Lexical items were manually inspected and extracted from the list, confirming their usage in the corpus as references to women. This process identified over 2,000 feminine derivatives (more details are provided in: Starko & Synchak, 2023). From this list, 28 military feminine terms were selected through manual inspection.

These extracted from GRAC-16 new military feminine titles were complemented by those extracted from the WDUF and the ‘Alphabet’ registries, and a list of the frequencies of their use in corpus was created in decreasing order (Fig. 3). Although the relative frequency indicators are also provided, the order of the terms in said list is determined by their absolute frequency indications. In some cases, irrelevant contexts had to be seeded out manually. For example, for *капітанка* ‘(female) captain’. GRAC showed 85 contexts, which, in addition to the name of the woman, also contained the title of the headdress and the village's name. After the elimination of excessive uses, the number of contexts was reduced to 57.

The contexts for unlemmatized words were found through CQL queries for regular expressions (more details are provided in: Starko & Synchak, 2023):

[word=".*демінер(ка|ку|кою|ки|ок|ками|кам)"] – search for words with suffix к(а);
 [word=".*мінометниц(я|ю|єю|і|ь|ями|ям)"] – search for words with suffix -иц(я).

▼	слово ▼	▼	абсолют ▼	відносі ▼	
1	героїня		25279	13,48	‘heroine’
2	захисниця		9245	4,93	‘(female) defendant’
3	волонтерка		8441	4,5	‘(female) volunteer’
4	льотчиця		5835	3,11	‘airwoman’
5	заручниця		1862	0,99	‘(female) hostage’
6	полонянка		1754	0,94	‘bondswoman’
7	партизанка		1190	0,64	‘(female) partisan’
8	військовослужбовиц		711	0,38	‘servicewoman’
9	ветеранка		538	0,29	‘(female) veteran’
10	медикиня		508	0,27	‘(female) medic’
11	парамедикиня		460	0,25	‘(female) paramedic’
12	солдатка		425	0,23	‘(female) soldier’
13	снайперка		374	0,2	‘(female) sniper’
14	медичка		316	0,17	‘(female) medic’
15	войовниця		310	0,17	‘warrior woman’
16	офіцерка		267	0,14	‘(female) officer’
17	радистка		239	0,13	‘radiowoman’

Figure 3: Frequency list of military feminine personal nouns (with absolute and relative frequency indicators)

Based on the corpus data, it is possible to compare the frequency of a word’s usage and to track its time dynamics. For instance, this chart (Fig. 4) was created using the relative frequency indicators in news texts in GRAC corpus for the period 2000–2022. It represents the usage dynamics of the nouns *військовослужбовиця* ‘servicewoman’ and *солдатка* ‘(female) soldier’. The chart indicates that the frequency of use of the noun *солдатка*, despite minor peaks in 2004 and 2017, is quite low. However, since 2015, the usage of the noun *військовослужбовиця* has rapidly increased, eventually surpassing the usage of the lexeme *солдатка*.

The rapid increase in the frequency of the noun *військовослужбовиця* is obviously related to the higher involvement of Ukrainian women in military operations brought on by Russia's invasion of Ukraine in 2014 and the media's generally increased focus on women's involvement in the war. However, now and later the unequal coverage of the GRAC corpus throughout the years should also be taken into consideration.

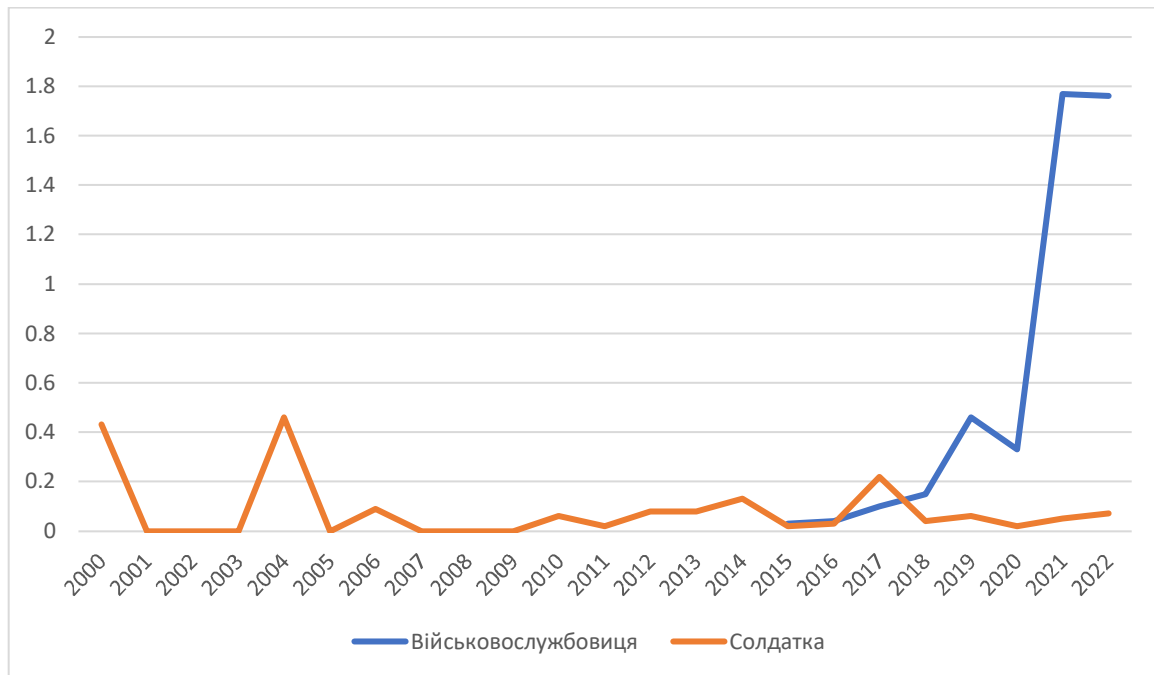


Figure 4: The dynamics of Ukrainian words referring to a female soldier (orange) and servicewoman (blue) in news texts.

In addition, the GRAC corpus provides authentic illustrations for dictionary entries. Thus, the meanings of military feminine terms will be compared based on examples from corpus and from the explanatory dictionaries. Due to this comparison, I hope to determine which words have changed their meaning in the setting of war, and which have lost their stylistic color. Based on this, new dictionary entries for the WDUF will be created.

4. Experiment

As a compiler of the WDUF, the idea of using the corpus to investigate the frequency and dynamics of word usage drives me to explore, if it is worthwhile to display quantitative data from the corpus in its updated edition. Further information is included in the WDUF to describe derivational variants of feminine personal nouns: results of the language panel and author's recommendations. It seems that quantitative data could substantially supplement the author's argument in the recommendations.

Military feminine personal nouns have a lot of derivational variants (Fig. 5), so their

description in the WDUF should be accompanied with recommendations. Could the quantitative data from corpus provide insight into the most advantageous derivational variant? Can high frequency of word usage in the corpus give grounds to recommend one variant among others? Is it worth recommending to use a word created according to a better model despite being not the most frequent?

1. героїня – 13,48	геройка – 0,05	‘heroine’
2. льотчиця – 3, 11	летунка – 0,01	‘airwoman’
3. військовослужбовиця – 0,38	військовослужбовка – 0,01	‘servicewoman’
4. парамедикиня – 0,15	парамедичка – 0,01 парамедиця – 0	‘(female) paramedic’
5. (бойова) медициня – 0,27	(бойова) медичка – 0,17	‘combat (female) medic’
6. доброволиця – 0,05	доброволка – 0,01	‘(female) volunteer’
7. навідниця – 0,05	наводчиця – 0,01	‘(female) aimer’
8. піхотинка – 0,01	піхотиниця – 0,01	‘infantrywoman’
9. миротвориця – 0,01	миротворка – 0	‘(female) peacemaker’
10. ухильянтка – 0,01	ухильниця – 0	‘a woman who evades military service’

Figure 5: Derivational variants of military feminine terms (with relative frequency indicators from GRAK-16)

The majority of the data in Figure 5 supports a quantitative approach. The most frequent variants are created according to better derivational models than less frequent words, for instance, the following pairs: *військовослужбовиця* – *військовослужбовка* ‘servicewoman’, *бойова медициня* – *медичка* ‘combat (female) medic’, *доброволиця* – *доброволка* ‘(female) volunteer’, *піхотинка* – *піхотиниця* ‘infantrywoman’, *миротвориця* – *миротворка* ‘(female) peacekeeper’, *ухильянтка* – *ухильниця* ‘a woman who evades military service’ etc. Also, the second term in the pair *навідниця* – *наводчиця* ‘(female) aimer’ appears as a result of Russification language policy, and it gains low frequency in corpus. But in the pair *льотчиця* – *летунка* ‘airwoman’ the russified term predominates based on frequency of usage. Nevertheless, there are words, the use of which can be justified not so much by quantitative indicators, as by qualitative characteristics. For example, in the triad *парамедикиня* – *парамедичка* – *парамедиця* ‘(female) paramedic’, I prefer the latter term, which despite having no contexts of use in the corpus, occurs in several printed sources, and fits the principle of language economy best. A more detailed argument in favor of this title can be found in the corresponding article of the WDUF.

A twenty-year study of use dynamics for the nouns *медикиня* – *медичка* ‘(female) medic’ in news texts reveals an interesting trend (Fig. 6). From 2019, when the revised Ukrainian Orthography established rules on the creation of feminine personal nouns (Tomilenko, 2021, p. 39), people start using both terms in parallel. However, up to 2018, only the word *медичка* was used. But by 2022 the word *медикиня* acquires an unprecedented frequency in corpus texts. This testifies that people need guidance in feminization—a system of orthography with clear derivational rules. But dictionaries also play a tremendous role in the codification of feminine personal nouns.

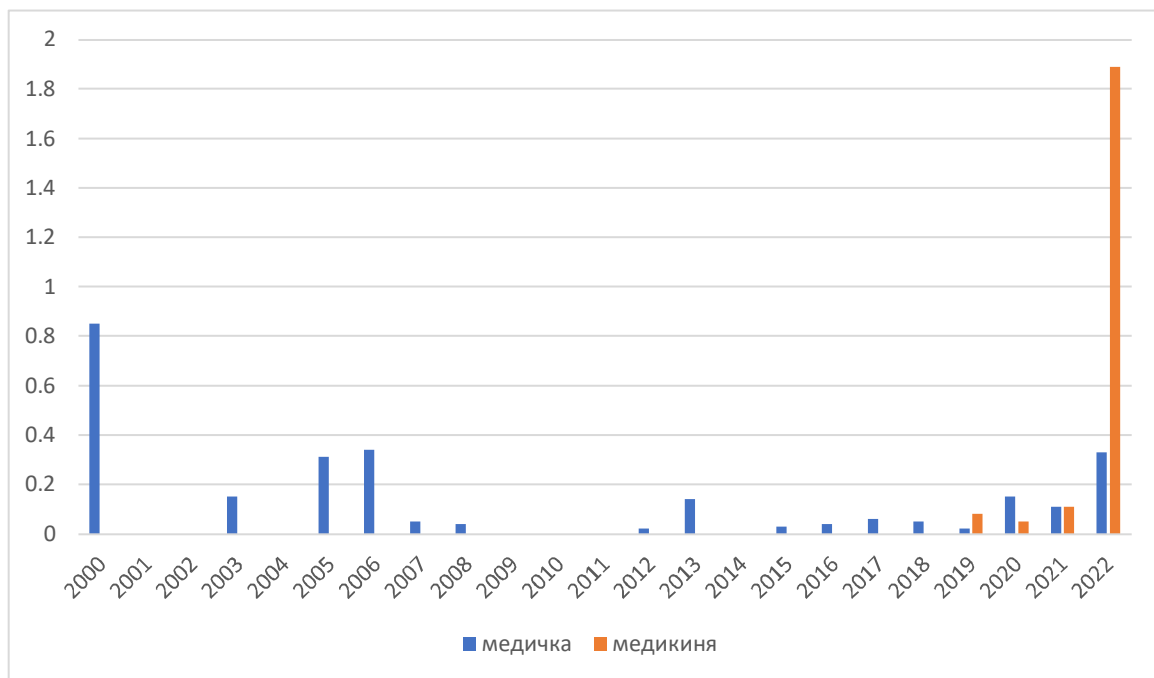


Figure 6: The dynamics of Ukrainian words referring to a female medic

After semantic analysis of the military feminine terms, we can identify five lexico-semantic groups (see Appendix I):

1. Names of women by military rank
2. Names of women according to action or function performed (military *nomen agentis*)
3. Names of women belonging to a military or other group
4. Names of women by characteristics and achieved results
5. Names of women from the enemy side.

On the list of feminine military titles, 4.5% of the words have both military and sporting meanings, including *снайперка* ‘(female) sniper’, *бійчиня* ‘(female) fighter’, *капітанка* ‘(female) captain’, *стрільчиня* ‘(female) shooter’, *резервістка* ‘(female) reservist’. The same ‘military’ in these words creates the semantic core, followed by

the semes ‘fleet’ or ‘police’, depending on the context. The seme ‘sport’ is located closer to the periphery, where additional semes (such as ‘organized group’ – *капітанка* or ‘struggle for something’ – *бійчиня*) are placed (Fig. 7).

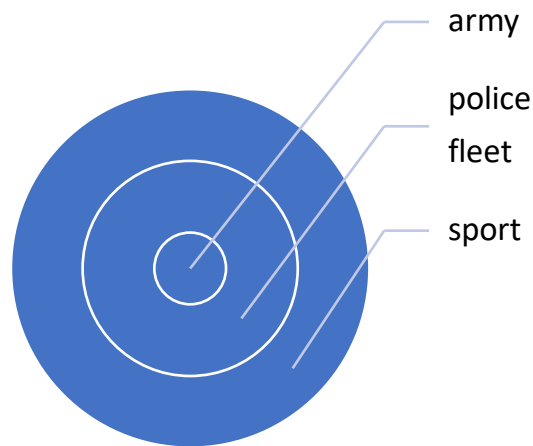


Figure 7: Semantic structure of feminine terms with ambiguous meanings ‘army + sport’

5. Results

This study has proved that the GRAK-16 is becoming a more reliable source for the study of feminine nouns in question than its earlier iterations. In particular, the updated GRAC-16 managed to find contexts for 84% of titles (113 words) from the list, only for 16% of the nouns (21 words) contexts were found in a Google search.

The analyzed material allows to clarify the definition of military feminine terms, which includes nominations used in reference to: 1. a woman directly participating in military actions (*сержантка* ‘(female) sergeant’, *танкістка* ‘tankwoman’ etc.); 2. a woman who serves in the supply and assistance link of the army (*волонтерка* ‘(female) volunteer’, *парамедиця* ‘(female) paramedic’ etc.); and 3. a woman involved in or affected by war (*партизанка* ‘(female) partisan’, *заручниця* ‘hostage’ etc). Most of the terms (97%) refer to Ukrainian women and only 3% denote women engaged in the war from the Russian side: *гауляйтерка* ‘(female) gauleiter’, *окупантка* ‘(female) occupant’, *ополченка* ‘(female) member of a pro-Russian militia or paramilitary group’, *бойовичка* ‘(female) combatant fighting in the side of the Russian occupation’.

The most frequent feminine personal nouns representing women's active participation in the military and at war were identified based on frequency (Fig. 3). Among the top 10 are the words *героїня* ‘heroine’, *захисниця* ‘(female) defender’, *волонтерка* ‘(female) volunteer’, *льотчиця* ‘airwoman’, *заручниця* ‘(female) hostage’, *полонянка* ‘bondswoman’, *партизанка* ‘(female) partisan’, *військовослужбовиця* ‘servicewoman’, *ветеранка* ‘(female) veteran’, *бойова медикиня* ‘combat (female) medic’. Although

героїня and *захисниця* have a long history, their usage in war-related contexts is relatively new. Nouns that depict the Ukrainian woman as a social actor who took on the difficulties of war, rather than, much less frequently, as a victim of war (*заручниця* ‘(female) hostage’, *polonianka* ‘bondswoman’), predominate in the examined list.

Frequency indications are crucial for lexicographic description in the WDUF, they are particularly important for arranging derivational variants and providing usage notes. Thus, it seems reasonable to include information about a word's usage frequency from corpus texts to support the sleeker derivational variant. In this way quantitative data from the corpus can be utilised for supporting the compiler's argumentation.

In addition, military feminine personal nouns have been tested for their recognition in the explanatory dictionaries: SUM-11 (1970-1980) and SUM-20 (2015-2023). As it turned out, barely a third (32.6%) of the 134 terms have a history dating back a century, and these titles are largely represented in explanatory dictionaries (e.g., *автоматниця* ‘(female) sub-machine gunner,’ *воячка* ‘virago,’ *зв'язкова* ‘signalwoman,’ *зенітниця* ‘(female) antiaircrafter,’ *кулеметниця* ‘(female) machine-gunner,’ *фронтівичка* ‘frontwoman,’ etc.).

However, the majority of the military feminine terms analyzed (67.4%) were developed between 2014 and 2022 and are not registered in monolingual Ukrainian dictionaries: *адміралка* ‘female admiral,’ *аеророзвідниця* ‘(female) air-scout,’ *армійка* ‘army woman,’ *військовослужбовиця* ‘servicewoman,’ *мінерка* ‘(female) miner,’ *миротворка* ‘female peacemaker,’ *снайперка* ‘(female) sniper,’ *танкістка* ‘tankwoman,’ etc. This once again proves that feminine terms do not arise as a tribute to fashion, but as a language system's response to social demands (Vplyv suspilnykh zmin, 2017, pp. 382). Since there was a need to name a woman as an active participant in the war, language actively provides such an opportunity.

Among these new coined terms, the feminine titles denoting military ranks, such as *полковниця* ‘(female) colonel’ and *офіцерка* ‘(female) officer’, are considered the fundamental vocabulary for news reports, however they are still uncommon in Ukrainian official military discourse.

Summarizing the lexico-semantic grouping (Fig. 8), we can acknowledge that the quantitatively largest group of military feminine personal nouns refers to women by the action or function they perform in the military (54%). The second largest group consists of names of women belonging to a military or other group (25%). The third group consists of names that refer to women by their military rank (12%).

It is suggested that not all selected feminine terms (see Appendix I) can be accompanied with a military subject label in the WDUF. In fact, this label can mostly be applied to the first three groups, including also nouns *ополчення*,

бойовичка from the fifth group. Except few examples, feminine terms from the last two groups primarily refer to women’s behavior or achieved results, not exclusively in a war context. For example, words like *героїня* ‘heroine’ or *звитяжниця* ‘female conqueror’ are predominantly used in the domains of arts or sports. While these words develop new meanings primarily in the context of war, their usage extends to much broader settings. Other words, like *ухилянтка* // *ухильниця* ‘woman who evades military service’, have a negative connotation as they express the act of avoiding military duty. Consequently, they also cannot be accompanied by a military subject label.

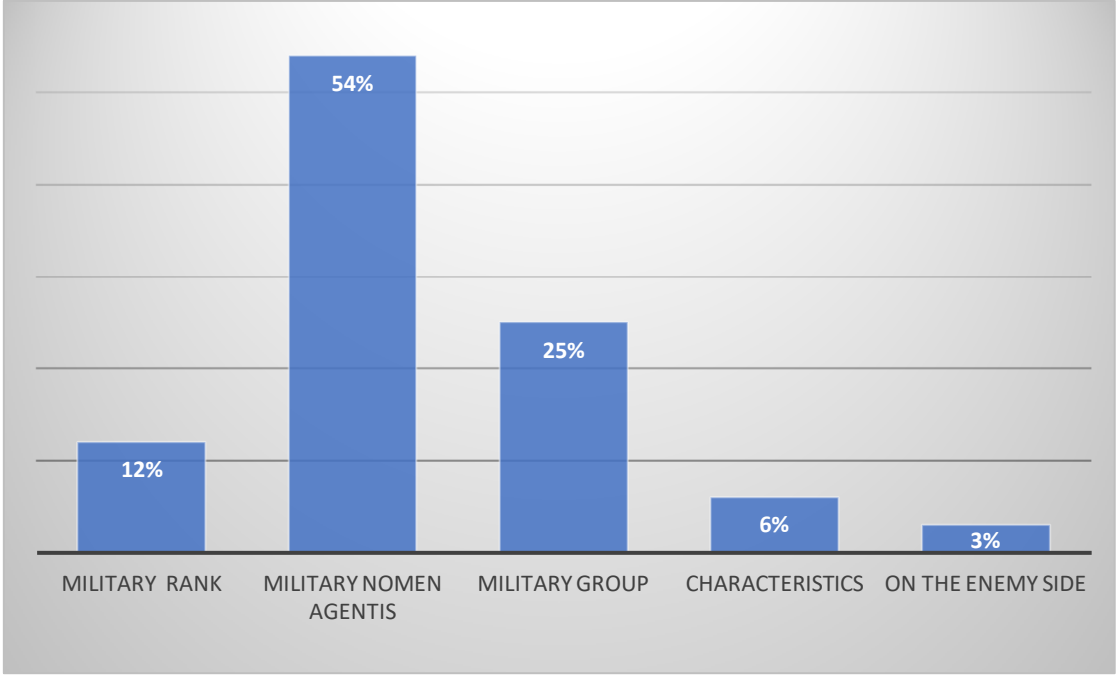


Figure 8: Lexico-semantic groups of military feminine personal nouns

6. Discussion

During one of the most turbulent periods in Ukraine’s modern history, Ukrainian women construct their own subjectivity and agency by taking part in humanitarian, volunteer, and military missions. As of 2022, the proportion of women serving in the Ukrainian military is 22%, which is one of the highest rates compared to other European nations. Women in Ukraine are standing guard for democratic changes as they defend the state’s integrity at the cost of their own lives.

To investigate the evolution of military feminine personal nouns, their definitions from recent dictionaries (WDUF and Alphabet), as well as explanatory dictionaries SUM-11 and SUM-20, were compared with their new contextual usage in GRAK-16. This analysis revealed a shift in meanings and nuances that occurred in several feminine terms following Russia’s full-scale invasion of Ukraine on February 24, 2022.

Semantic change is evident, for example, in the word *героїня* ‘heroine’. In addition to the hitherto common meaning of “woman who distinguished herself in something”, journalistic texts dated 2022 (GRAK-16) have developed the meanings of “woman who performed a heroic deed” and “a woman outstanding in her abilities and activities, who shows courage, dedication and bravery in work and battle.” (Fig. 9).

<p>ГЕРОЇНЯ [heroyinia] ‘heroine’</p>	<p>1. a woman outstanding in her abilities and activities, who shows courage, dedication and bravery in work and battle. – <i>Heroyi ta heroyini ZSU ne vidpuskaiut vorohiv bez boiu. (24 kanal, 2022). Vin podiakuvav kozhnomu heroievi ta heroini, yaki trymaiut oboronu Ukrainy vid tsoho teroru. (24 kanal, 2022). Heroi ta heroini shchodnia vmyraiut za nashu svobodu y sered nykh – zovsim yuni voiny, taki yak Oleksandra Anikieva. (24 kanal, 2022). Spodivaiemos, shcho naiblyzhchym chasom “Ptashka” povernetsia razom iz inshymy ukrainskymy heroiniamy. (www.0352.ua, Ternopil, 2022).</i> // Slava heroyam i heroyiniam! – one of the versions of the response to the greeting “Glory to Ukraine!”; the slogan of the struggle for the independence of Ukraine.</p> <p>2. woman who performed a heroic deed, distinguished herself in something. – <i>Mariupolskyi fotohraf naholosyv, shcho matir khlopchyka – heroinia, adzhe vriatuvala svoikh ditei. (24 kanal, 2022). Ne heroinia ya, ale chyniu tak, yak vchynyla b usiaka normalna zhinka <...>. (O. Pylypenko «Normalni liudy abo Dekameron staroi divy», 2011).</i></p>
---	---

Figure 9: Example of a dictionary entry in the update to the WDUF

In 2022, in news texts, the name *героїнка* ‘heroine’ become widespread to denote a woman who shows courage and dedication either in battle or in fighting the enemy in the rear. Thus, in the first meaning, this word coincides with the meaning of the word *героїня* ‘heroine’, but in the second meaning, it contains the additional seme ‘fearless’ (Fig. 10).

<p>ГЕРОЇНКА [heroyka] ‘heroine’</p>	<p>1. a woman outstanding in her abilities and activities, who shows courage, dedication and bravery in work and battle. – <i>My ne vtomymosia diakuvaty nashym vidvazhnym heroyam i heroykam, yaki shchodnia boroniat i zvilniaiut nashu Ukrainu vid okupanta. (Svoi.City, 2022). Pered vitanniamy khvylynoiu movchannia vshanuvaly pamiat vsikh heroiv ta heroyok, yaki zahynuly, zakhyshchaiuchy Ukrainu vid voroha. (Biliaivka.City, 2022). 84 riatuvalnyky, a zaraz spravzhni heroyi ta heroyky, buly zmusheni</i></p>
--	---

	<p><i>vyzhyvaty pid prytsilom putinskykh voiak bilshe misiatsia. (Apostrof, 2022).</i></p> <p>// Heroyka Ukrayiny – the highest honorary title for exceptional services to the state and people of Ukraine. – <i>Prochytaye yiyi tvir Heroyka Ukrayiny, narodna artystka Ada Rohovtseva. (24 kanal, 2022).</i></p> <p>2. fearless woman who performed a heroic deed. – <i>Heroyka vykhodyla do vorohiv ta rakhuvala vorozhu tekhniku. (24 kanal, 2022).</i> <i>Pid chas aktyvnykh boiovykh dii u Volnovasi Iryna Romanchenko pokazala sebe yak spravzhnia heroyka, khocha sama zhinka sebe takoiu ne vvazhaie. (Volnovakha.City, 2022).</i></p>
--	--

Figure 10: Example of a dictionary entry in the update to the WDF

A new definition of the term **захисниця** ‘(female) defender’ emerges as well, one that explanatory dictionaries have not yet recognized: ‘woman who protects the residents of her country, maintains the territorial integrity of her state’ (Fig. 11). The rest of the meanings of this term coincide with the definitions in dictionaries.

<p>ЗАХИСНИЦЯ [zakhysnytsia] ‘(female) defender’</p>	<p>1. woman who protects the citizens of her country, protects the territorial integrity of her state. – <i>Boroniachy Ukrainu vid rosiiskoho vtorhnennia, zahynula zakhysnytsia Ukrainy Mariana «Kvitka». (Ukrainskyi prostir, 2022).</i> <i>Ta vona zalyshylas, prosyla ne zhality yii, bo vona zakhysnytsia ta ukrainka, yaka vykonuie svii oboviazok — riatuie pobratymiv ta tsyvilnykh. (Ukrainskyi prostir, 2022).</i> <i>Nekhai kozhen ukrainskyi zakhysnyk ta kozhna zakhysnytsia povernutsia dodomu, do svoikh ridnykh ta blyzkykh! (www.0352.ua, Ternopil, 2022).</i> <i>Zakhysnytsia Ukrainy z 26 okremoi artyleriiskoi bryhady ranishe pratsiuvala vchytelkoiu istorii ta vykladala do 2016 roku. (www.0352.ua, Ternopil, 2022).</i></p> <p>// Den zakhysnykiv i zakhysnyts Ukrayiny – the national holiday of Ukraine, which is celebrated on October 14. – <i>4 lypnia 2021 roku Verkhovna Rada pereimenuvala sviato na “Den zakhysnykiv i zakhysnyts Ukrainy”. (www.0352.ua, Ternopil, 2022).</i></p>
--	---

Figure 11: Example of a dictionary entry in the update to the WDF

Another difference appears on the semantic level of the words **солдатка** ‘(female) soldier’, **полковниця** ‘(female) colonel’, **підполковниця** ‘(female) lieutenant-colonel’. Since 2014, these nominations are used to indicate the rank of the women in the

army, however SUM-11 recognizes them as andronymic titles that signify women by military rank of their husbands only.

It is possible to challenge the stylistic labels applied in the explanatory dictionaries for a number of feminine terms by looking at the broader contexts of their usage in corpus texts. In particular, words like *полонянка* ‘bondswoman’ (also designated as folklore), *войовниця* ‘warrior woman’ and *летунка* ‘airwoman’ do not need to be recognized as archaic in current situations. Similarly, the name *ветеранка* ‘(female) veteran’ has lost its colloquial meaning (although SUM-20 continues to label it that way), and the terms *командирка* ‘(female) commander’ and *медичка* ‘(female) medic’, in addition to colloquial usages, have developed neutral meanings.

In some military feminine personal nouns, the corpus texts testify to the development of metaphorical meanings that are not recorded in explanatory dictionaries. For example, the name *кулеметниця* ‘(female) machine-gunner’, in addition to the primary meaning of ‘fighter of a machine gun unit, machine gun shooter’, develops the figurative meaning ‘woman who works very fast’ (Vplyv suspilnykh zmin, 2017). The name *бійчиня* ‘(female) fighter’, in addition to the meanings 1) ‘female combatant during military operations,’ 2) ‘soldier, private,’ 3) ‘athlete who engages in combat sports,’ also expresses the figurative meaning ‘woman who fights to achieve something’ (WDF, 2022). Finally, the term *навідниця* ‘(female) aimer’ refers to both a civilian woman who directs artillery and a warrior who aims a grenade (Abetka feminityviv, 2022).

All these examples confirm that corpus-based dictionaries greatly surpass traditional lexicographic sources in providing a more accurate representation of how words function in context. Hence the corpus also enables users to “capture” the words’ figurative as well as literal meanings.

Military feminine personal nouns also function as part of fixed expression (*Слава героям та героїням!* ‘Glory to heroes and heroines!’, *Національна героїня* ‘National heroine’, *Геройка України* ‘Heroine of Ukraine’, *Захисниця України* ‘(Female) Defender of Ukraine’), collocations (*солдатка-контрактниця* ‘(female) contract soldier’, *військова летунка* ‘military airwoman’, *морська піхотинка* ‘(female) amphibious soldier’, *дешифрувальниця польотів* ‘(female) flight decipherer’, *бригадна генералка* ‘(female) brigadier general’, *заручниця обставин* ‘hostage of circumstances’), as well as idioms (*одна в полі воїнка* ‘alone in the field a warrior’, *звитяжниця духу* ‘(female) conqueror of the spirit’). It is obvious that dictionaries that list military feminine terms should also include these fixed expressions.

7. Conclusion

The proportion of women serving in the Ukrainian army is rising, and as a result, there are more instances of women and men being equally represented in the military discourse (*День захисників та захисниць України* ‘Day of the Male and Female

Defenders of Ukraine’; *Нашим відважним героям і героїкам* ‘to our brave heroes and heroines’), as well as in other areas of social life (*До уваги відвідувачів та відвідувачок мезую!* ‘Museum visitors, attention!’). The sharp rise in the number of military feminine nouns strongly proves that their use is not a tribute to fashion but rather the language system's response to a social demand: the majority of the feminine personal nouns in the military (67.3%) were produced between 2014 and 2022.

If Ukrainian explanatory dictionaries contain only a third (32.6%) of the military feminine personal nouns discussed in this article, then corpus-based dictionaries for describing such novel linguistic material have greater functionality at all stages of lexicographical work, from word selection to register, searching for contexts, up to the identification of metaphorical meanings, fixed phrases, and idioms, as well as the formulation of usage recommendations.

In the updated version of WDUF, the quantitative data obtained from the corpus will be used in the arrangement of the word-forming variant, as well as in the usage note to argue in favor of one of them. Adding subject labels can also enrich the WDUF's lexicographic description of feminine terms. However, mainly three lexical-semantic groups of military feminine personal nouns might be marked with military label: 1. names of women by military rank; 2. names of women according to the performed action or function (military *nomen agentis*); 3. names of women belonging to a (military) group.

In general, incorporating subject labels (such as military, sports, legal, medical, etc.) into the updated WDUF, along with utilizing the r2u.org.ua platform's full-text search, can provide significant assistance. Firstly, it facilitates further exploration of feminine terms within thematic groups. Secondly, it helps unify the lexicographical descriptions of words within the same group. In the upcoming edition of the WDUF, it is fitting to maintain a combination of traditional lexicography approaches and corpus-based methods, as well as to balance description with prescription.

8. Acknowledgements

This research has received funding from the Humanities Faculty of the Ukrainian Catholic University (Lviv, Ukraine). The realization of the ‘Abetka feminityviv’ project was carried out under the ‘Behind the Gender’ project in 2022. The financial support for the WDUF was provided by the ‘Believe in Yourself’ Foundation in 2021.

I extend my heartfelt thanks to Vasyl Starko for his invaluable assistance in selecting non-lemmatized military feminine forms from the GRAK-16 corpus and providing guidance on CQL queries. I am also deeply grateful to my husband, Ingolf Petzold, for his unwavering support throughout the writing process. Foremost, special appreciation goes to the brave defenders of Ukraine who selflessly fight for our

dignity, life, and freedom. *Слава героям і героїням!* ‘Glory to the heroes and heroines!’

Appendix I:

Lexico-semantic groups of military feminine personal nouns

(In descending order of frequency in GRAC-16)

Names of women by military rank	Names of women according to the performed action or function	Names of women belonging to a military group or another group	Names of women by characteristics and achieved results
1. солдатка ‘(female) soldier’	1. захисниця ‘(female) defender’	1. військовослужбовиця ‘servicewoman’	1. войовниця ‘(female) warrior’
2. офіцерка ‘(female) officer’	2. волонтерка ‘(female) volunteer’	2. військовослужбовка ‘servicewoman’	2. героїня ‘heroine’
3. полковниця ‘(female) colonel’	3. льотчиця ‘airwoman’	3. військова ‘servicewoman’	3. героїнка ‘heroine’
4. лейтенантка ‘(female) lieutenant’	4. снайперка ‘(female) sniper’	4. армійка ‘army woman’	4. воїтелька ‘(female) warrior’
5. генералка ‘(female) general’	5. медикія бойова ‘combat (female) medic’	5. заручниця ‘hostage’	5. звияжниця ‘(female) conqueror’
6. сержантка ‘(female) sergeant’	6. медичка бойова ‘combat (female) medic’	6. полонянка ‘bondswoman’	6. ухиянтка ‘a woman who evades military service’
7. капітанка ‘(female) captain’	7. парамедикія ‘(female) paramedic’	7. військовополонена ‘(female) prisoner of war’	7. титанка ‘(female) Titan’
8. хорунжа ‘(female) standard bearer’	8. парамедичка ‘(female) paramedic’	8. партизанка ‘(female) partisan’	8. ухильниця ‘a woman who evades military service’
9. підполковниця ‘(female) lieutenant-colonel’	9. радистка ‘radiowoman’	9. ветеранка ‘(female) veteran’	
10. поручниця ‘(female) guarantor, lieutenant’	10. комісарка ‘(female) commissar’	10. доброволиця ‘(female) volunteer’	
11. підпоручиця ‘second (female) lieutenant’	11. командирка ‘(female) commander’	11. доброволка ‘(female) volunteer’	
12. віцеадміралка ‘(female) vice-	12. командувачка ‘(female) army commander’	12. контрактниця ‘(female) contract soldier’	
	13. десантниця ‘(female) paratrooper’	13. диверсантка ‘(female) saboteur; attacker’	
	14. артилеристка ‘artillerywoman’	14. резервістка ‘(female) reservist’	
	15. танкістка ‘tankwoman’	15. госпітальєрка ‘(female) member of	
			Names of women on the enemy side 1. гауляйтерка ‘(female) head of the

<p>admiral'</p> <p>13. адміралка '(female) admiral'</p> <p>14. контрадміралка '(female) rear-admiral'</p> <p>15. майорка '(female) major'</p> <p>16. прапорщиця '(female) ensign'</p>	<p>16. кулеметниця '(female) machine-gunner'</p> <p>17. бійчиня '(female) fighter'</p> <p>18. воїнка '(female) warrior'</p> <p>19. стрільчиня '(female) shooter'</p> <p>20. навідниця '(female) aimer'</p> <p>21. зенітниця '(female) antiaircraft'</p> <p>22. воячка 'virago'</p> <p>23. постачальниця '(female) supplier, contractor (for the army)'</p> <p>24. фронтовичка 'frontwoman'</p> <p>25. аеророзвідниця '(female) air-scout'</p> <p>26. найманка '(female) mercenary'</p> <p>27. оборонниця '(female) defender'</p> <p>28. летунка 'airwoman'</p> <p>29. наводчиця '(female) aimer'</p> <p>30. підривниця '(female) demolition woman, sapper'</p> <p>31. мінометниця '(female) springer'</p> <p>32. коректувальниця '(female) spotter'</p> <p>33. гранатометниця '(female) bomber'</p> <p>34. дезертирка '(female) deserter'</p> <p>35. радіотелефоністка '(female) radio telephone operator'</p>	<p>the "Hospitaliers" medical battalion'</p> <p>16. нацгвардійка '(female) member of the National Guard'</p> <p>17. гвардійка 'guardswoman'</p> <p>18. айдарівка '(female) member of the "Aydar" military battalion'</p> <p>19. тероборонівка '(female) member of the "Territorial defense" group or battalion'</p> <p>20. азовка '(female) member of the "Aydar" military battalion'</p> <p>21. січовичка 'service woman in the "Ukrayinski Sichovi striltsi" Legion'</p> <p>22. оунівка '(female) member of the "OUN"'</p> <p>23. усуска 'service woman in the "Ukrayinski Sichovi striltsi" Legion'</p> <p>24. усусівка 'service woman in the "Ukrayinski Sichovi striltsi" Legion'</p> <p>25. мілітаристка 'servicewoman'</p> <p>26. військовичка 'servicewoman'</p> <p>27. військовиця 'servicewoman'</p> <p>28. дивізійниця '(female) division officer'</p> <p>29. призовниця '(female) conscript'</p>	<p>administrative-territorial unit occupied by Russia in Ukraine'</p> <p>2. окупантка '(female) occupant, invader'</p> <p>3. ополченка '(female) member of a pro-Russian militia or paramilitary group'</p> <p>4. бойовичка '(female) combatant fighting on the side of Russian occupators'</p>
---	--	--	---

36. миротвориця '(female) peacemaker'	30. воєнчиня 'servicewoman'
37. штурманка '(female) navigator'	31. військовополонянка '(female) prisoner of war'
38. медіаторка '(female) mediator'	32. атошниця 'servicewoman in the "АТО"'
39. авіаторка '(female) aviator'	33. спецназівка 'servicewoman in the special forces unit'
40. демінерка '(female) deminer'	
41. фасилітаторка '(female) facilitator'	
42. капеланка '(female) chaplain'	
43. саперка '(female) sapper'	
44. піхотинка 'infantrywoman'	
45. наємниця '(female) mercenary'	
46. воєнкорка '(female) war correspondent'	
47. воєначальниця '(female) commander'	
48. корегувальниця '(artillery) spotter'	
49. комбатантка '(female) combatant'	
50. чотарка '(female) lieutenant'	
51. автоматниця '(female) sub-machine gunner'	
52. дешифрувальниця '(female) decipherer'	
53. оперативниця '(female) operator'	
54. зв'язківниця 'signalwoman'	
55. гвинтівочниця 'riflewoman'	
56. піхотиниця 'infantrywoman'	
57. екскомбатантка	

	‘(female) ex-combatant’		
58.	медсанбатівка ‘(female) paramedic’		
59.	штурмовичка ‘(female) stormer’		
60.	артрозвідниця ‘(female) artillery scout’		
61.	воякиня ‘virago’		
62.	зв’язкова ‘signalwoman’		
63.	зв’язківка ‘signalwoman’		
64.	командорка ‘(female) commodore’		
65.	медиця бойова ‘combat (female) medic’		
66.	миробудівниця ‘(female) peacemaker’		
67.	миротворка ‘(female) peacemaker’		
68.	мінерка ‘(female) miner’		
69.	парамедиця ‘(female) paramedic’		
70.	пілоткиня ‘airwoman’		
71.	баталістка ‘(female) artist who depicts battles’		

9. References

- Abetka feminivityv* (2022). [Alphabet of Ukrainian Feminine Personal Nouns]. Comp. by O. Synchak, 2022. URL: <https://behindthenews.ua/spetsproiekti/po-toy-bik-genderu/abetka-feminitiviv-358/> (07.11.2022).
- Arkhanhelska, A. (2019). *Femina cognita. Ukrainska zhinka u slovi y slovnyku* [Femina cognita. The Ukrainian Woman in Words and Dictionaries]. Kyiv.
- Brus, M. (2019). *Feminityvy v ukrainskii movi: heneza, evoliutsiia, funktsionuvannia*. [Feminine Personal Nouns in Ukraine: Genesis, Evolution, and Functioning],

Ivano-Frankivsk.

- Brus, M. (2017-18). Feminityvy v poetychnii movi Tarasa Shevchenka [Femininitives in the poetic language of Taras Shevchenko]. *Prykarpatskyi visnyk NTSh. Slovo* [Prykarpatskyi visnyk NTSh. Word], 4-3, pp. 46-59.
- Elmiger, D. (2009). *Sprachliche Gleichbehandlung von Frau und Mann: Eine korpusgestützte Untersuchung über den Sprachwandel*. DOI: 10.13092/lo.39.477.
- Feketa, I. (1968). *Zhinochi osobovi nazvy v ukrainskii movi (tvorennia i vzhyvannia)*: dys. kand. filol. nauk. Uzhhorod 1968 [Feminine Personal Nouns in the Ukrainian Language (formation and use): dissertation candidate of philology science]. Uzhhorod.
- Kaidash, A. (2017). Systema feminityviv khudozhnoho movlennia Liuko Dashvar [The system of feminine personal nouns in the writing of Lyuko Dashvar]. *Mova i mizhkulturna komunikatsiya* [Language and intercultural communication], 1, pp. 199-206.
- Klymenko, N. (2019). Tendentsii rodovoi katehoryzatsii imennykiv u suchasnykh ukrainskii ta novohretskii movakh [Trends in Gender Categorization of Nouns in Modern Ukrainian and Modern Greek Languages]. *Ukrainska mova v konteksti suchasnoi slavistyky: Monohrafiia* [The Ukrainian Language in the Context of Modern Slavic Studies: Monograph]. Kyiv.
- Koster, D. (2020). Do Representations of Gender and Profession Change over Time? Insights from a Longitudinal Corpus Study on Dutch Language Textbooks (1974–2017). *Journal of Gender Studies*. DOI: 10.1080/09589236.2020.1754176
- Kostusiak, N., Navalna, M., Mezhev, O. (2020). The Functional-Cognitive Category of Femininity in Modern Ukrainian. *Cognitive Studies* [Études cognitives], 20, Article 2310. DOI: 10.11649/cs.2310
- Kravets, T. (2021). *Henderni stereotypy v suchasnomu ukrainskomu masmediinomu dyskursi: Dysertatsiia* [Gender Stereotypes in the Discourse of Today's Mass Media in Ukraine. Dissertation], Kyivskiy natsionalnyi universytet imeni Tarasa Shevchenka, Kyiv.
- Machek, V. (2022). Feminityvy v tekstakh slovatskoho narodnoho korpusu 20–30 r. 20 st. [Femininitives in the Texts of the Slovak Peoples Corpus 20–30 th. of the 20th Century]. *Naukovi zapysky Natsionalnoho universytetu "Ostrozka akademiia": seriya "Filolohiya"* [Scientific notes of the National University 'Ostroh Academy': series 'Philology']. Vyd-vo NaUOA, Ostroh, 13 (81), pp. 81-86.
- Machek, V. (2020). Feminityvy v narodnorozmovnii movi zakarpatskykh rusyniv 20–30 rr. XX st. (na materialy periodychnykh vydan) [Feminine Personal Nouns in the Vernacular language of the Transcarpathian Ruthenians of the 20s–30s of the 20th century (based on the media publications)]. *Vcheni zapysky TNU imeni V. I. Vernadskoho. Seriya: Filolohiya. Sotsialni komunikatsiyi* [Scholarly notes of V. I. Vernadskyi's TNU. Series: Philology. Social communications], 31 (70), №4, P. I, pp. 86-92.
- Mykhaylova, O., Spilnyk, T. (2019). Prahmatychnyi aspekt vzhyvannia ta perekladu

- feminityviv u viiskovomu dyskursi [The pragmatic aspect of the use and translation of feminitives in military discourse]. *Naukovyi visnyk Mizhnarodnoho humanitarnoho universytetu. Ser.: Filolohiya* [Scientific Bulletin of the International Humanitarian University. Ser.: Philology], 43, Vol. 4, pp. 78-81.
- Navalna, M. (2017). Imennyky na poznachennia osib zhinochoi stati v movi internet-vydanna "Ukrainska pravda" [Feminine Personal Nouns in the Language of the Online Publication 'Ukrainian Pravda']. *Zbirnyk naukovykh statei. Ceriya "Filolohichni nauky"* [Collection of scientific articles. Series 'Philological sciences'], 4, pp. 13-23.
- Neliuba, A. (2011). Innovatsiini zrushennia v ukrainskomu zhinochomu slovotvori [Innovative Shifts in Ukrainian Women's Vocabulary]. *Linhvistyka* [Linguistics], 2(23), pp. 49–59.
- Puzyrenko, Ya. (2012). Chynnyky vplyvu na leksykohrafuvannia nazv osib zhinochoi stati yak vidobrazhennia suchasnykh problem slovnykarstva [Factors influencing the lexicographical description of feminine personal nouns as a reflection of modern vocabulary problems]. *Visnyk Dnipropetrovskoho universytetu imeni Alfreda Nobelia. Seriya 'Filolohichni nauky'* [Bulletin of the Dnipropetrovsk University named after Alfred Nobel. Series 'Philological sciences'], 2(4), pp. 163-167.
- Semeniuk, S. (2000). *Formuvannia slovotvirnoyi systemy imennykiv z modyfikatsiinym znachenniam zhinochoyi stati v noviy ukrainskiy movi: Dysertatsiia* [Formation of the Derivational System of Nouns with the Modifying Meaning of Feminine Gender in Modern Ukrainian: Thesis], Zaporizhia State University.
- Shvedova, M., Waldenfels, R. von, Yarygin, S., Rysin, A., Starko, V., Nikolajenko, T. et al. (2017-2023). *GRAC: General Regionally Annotated Corpus of Ukrainian*. Electronic resource: Kyiv, Lviv, Jena. URL: <http://uacorpus.org>.
- SUM-20: *Slovnyk ukrayinskoyi movy* (2015–2023). [A Dictionary of the Ukrainian Language]. In 20 vols. Naukova Dumka, Kyiv. Vols. 1–12. URL: <https://services.ulif.org.ua/expl>.
- SUM-11: *Slovnyk ukrayinskoyi movy* (1970-1980). [A Dictionary of the Ukrainian Language]. In 11 vols. Naukova Dumka, Kyiv, 2018–2023. URL: <http://sum.in.ua>.
- Styshov, O. (2020). Osoblyvosti sufiksalnogo slovotvorennia neolohizmiv na poznachennia osib u suchasniy ukrainskiy movi [Peculiarities of Suffixal Word Formation of the Neologisms Denoting Identification of Persons in Modern Ukrainian Language]. *Linhvistychni doslidzhennia: zb. nauk. prats KhNPU imeni H. S. Skovorody* [Linguistic studies: Coll. of Science Proceedings of H.S. Skovoroda KhNPU], 53, pp. 127-140.
- Starko, V., Synchak, O. (2023). Feminine Personal Nouns in Ukrainian: Dynamics in a Corpus. *COLINS-2023: 7th International Conference on Computational Linguistics and Intelligent Systems*, National Technical University "Kharkiv Polytechnic Institute", Kharkiv, Ukraine, pp. 407-425.

- Synchak, O., Starko, V. (2022). Ukrainian Feminine Personal Nouns in Online Dictionaries and Corpora. *COLINS-2022: 6th International Conference on Computational Linguistics and Intelligent Systems*. Gliwice, Poland. URL: <http://ceur-ws.org/Vol-3171/paper58.pdf>.
- Synchak, O. (2022). Vebslovnnyk zhinochykh nazv ukrayinskoyi movy yak novitnie yavyshe v ukrayinskomu slovnykarstvi [Web Dictionary of Feminine Names of the Ukrainian Language as a New Phenomenon in Ukrainian Lexicography]. *Wroclawska Ukrainistyka: Lingua, Litterae, Sermo*, Wroclaw, pp. 251-267.
- Taranenko, O. (2021). *Androtsentryzm u systemi movnykh koordynat i suchasnyi hendernyi rukh*: Monohrafiia [Androcentrism in the System of Linguistic Coordinates and Modern Gender Movement: Monograph]. Vydavnychi dim Dmytra Buraho, Kyiv.
- Tomilenko, L. (2021). Feminityvy v ukraïnskii perekladnii leksykohrafiï pochatku XX stolittia (pisliarevoliutsiyna doba) [Feminine personal nouns in the Ukrainian translated lexicography of the beginning of the 20th century (post-revolutionary era)]. *Slavia – časopis pro slovanskou filologii*, 4, pp. 440-452.
- Ukrayinskyy pravopys* (2019). [Ukrainian Orthography]. Naukova dumka, Kyiv.
- Vplyv suspilnykh zmin na rozvytok ukraïnskoi movy*: Monohrafiya. (2017). [The Influence of Social Changes on the Development of the Ukrainian Language]. Ye. A. Karpilovska, L. P. Kysliuk, N. F. Klymenko, V. I. Krytska, T. K. Puzdyrieva, Yu. V. Romaniuk; Vidp. red. Ye. A. Karpilovska. Vydavnychi dim Dmytra Buraho, Kyiv.
- WDFU: *Vebslovnnyk zhinochykh nazv ukraïnskoi movy*. (2022). [A Web Dictionary of Ukrainian Feminine Personal Nouns]. Comp. by O. Synchak. URL: <https://r2u.org.ua>.
- Zayets, V. (2020). Slovtvirni typy feminityviv suchasnoi prozy ta literaturna norma [Word-forming types of feminine personal nouns in modern prose and the literary norm]. *Colloquium-journal*. Warszawa, 19(71), pp. 17-22.

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

Kate Challis¹, Tom Drusa

¹ Iowa State University, Ames, Iowa, USA

E-mail: kchallis@iastate.edu, t.drusa@gmail.com

Abstract

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

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

1. Introduction

The purpose of this design-based research study is to build a Chrome browser extension that provides second language (L2) readers with the visual attention cues to corpus-based information that can improve their attention to top-down reading strategies by offsetting the burden on working memory. In this section, we present key theoretical concepts explored in prior theoretical research related to L2 reading, visual attention cues, and data-driven vocabulary learning.

1.1 Cognitive demands of second language reading

Prior research has shown that the awareness and use of top-down, i.e. global/holistic reading strategies accounts for 52% of the total variance in L2 reading ability (Song, 1999), suggesting that tasks during reading which help readers to attend to information at the discourse level are beneficial. These activities include having a global view of the reading process, making guesses, taking risks, concentrating on the main idea rather than getting sidetracked by trivialities, reading to confirm/refine/reject hypotheses made about the meaning of the text as a whole, summarizing the main ideas, and focusing less on graphophonic and syntactic accuracy than on accurate global understanding. In other words, proficient readers employ strategies that enable them to attend to meaning at the discourse level. Top-down strategies are consistently found

to be better for L2 reading than bottom-up strategies (Brantmeier 2002).

However, L2 reading places a high cognitive burden on working memory, vocabulary recall, and discourse synthesis strategies of all readers, in particular for those who lack reading proficiency in their first language (L1) and those who are at the novice- or beginning-level in their L2 (Kupermann et al., 2022). This cognitive burden makes it especially difficult for learners to attend to top-down learning strategies, even when explicitly trying to do so. One explanation for the high cognitive load which learners experience while reading is an inability to distinguish between information that is important and that which is redundant and unnecessary for learning (Kalyuga and Sweller, 2005). Additionally, the so-called ‘redundancy effect’ occurs when information is presented through multiple simultaneous modalities without allowing for the learner to attend to prioritization of information; researchers have shown that the redundancy effect hinders learning (Mayer et al., 2001; Diao and Sweller, 2007; Liao et al., 2020). However, since reading is a visual task, visual attention cues provided by an outside stimuli can potentially be used to offload some of the burden that L2 reading places on working memory, in particular in self-paced reading tasks where learning outcomes are closely correlated to time spent looking at written text (Schmidt-Weigand et al., 2010).

1.2 Visual attention cues, working memory, and second language reading

Visual attention is a key component of reading because it allows the brain to identify orthographic units during lexical processing. The visual attention span (VAS), which is the maximum number of distinct visual elements that the brain can process simultaneously at a glance (Bosse et al., 2007; Bosse & Valdois, 2009) has been linked to reading performance in both L1 and L2 (Awadh et al., 2016; Lobier, Peyrin, Le Bas, & Valdois, 2012), especially when the stimuli are alphanumeric (Verhallen & Bus, 2011). This research also suggests that readers attend to visual cues while reading.

The connectionist Multi-Trace-Memory reading model (Ans et al., 1998) suggests that there is a correlation between visual attention capacity and reading performance, and that a reduction in VAS is detrimental to familiar word processing (Adelman, Marquis, & Sabatos de Vito, 2010; Grainger et al., 2016). In other words, visual attention “seems to be modulated by the amount of attentional resources available” (Frey & Bosse, 2018; Lobier et al., 2013), suggesting that L2 readers benefit from any strategy that can be used to shift attentional resources towards visual processing.

Some researchers disagree with the idea that visual attention is directly connected to reading performance (Gori et al., 2014, Gori and Facoetti, 2014, Lorusso et al., 2011, Facoetti et al., 2006), while other researchers have found additional non-VAS based evidence to confirm this connection, for example by measuring attentional blink, visual search, and visuospatial attention (Cirino et al., 2022).

1.3 Corpus-based vocabulary learning

Data-driven learning (DDL) is an effective pedagogical approach in which learners are encouraged to independently analyze and explore corpora. Independent, self-motivated reading from authentic texts causes target vocabulary items to become more salient (Chapelle, 2003), but can be further enriched when empirical, corpus-based word frequency and dispersion data are made easily transparent to learners. In a sea of unfamiliar words, it is difficult for learners to make intelligent decisions about which words to prioritize and which to leave for later, and all texts—authentic or contrived—are composed of words. In essence, DDL approaches to L2 reading are implicitly connected to L2 vocabulary building.

Receptive vocabulary refers to the words that a person can understand when encountered in a context, but may not necessarily be able to actively produce in writing or speech independently. Prior research indicates that a receptive vocabulary of approximately 6–9k word families (in English) is needed to achieve 98% text coverage, the amount considered by many researchers to represent an amount of unknown vocabulary that avoids cognitive overload during L2 reading (Nation, 2006; Hu & Nation, 2000; van Zeeland & Schmitt, 2013).

Useful words for L2 learners to prioritize in their learning are those which occur with high frequency and wide dispersion in the language (Gardner & Davies, 2014; Lei & Liu, 2016) since “actual frequency of occurrence is a more reliable indicator of usefulness than pure intuition” (Garnier & Schmitt, 2015). Although language variation depends on its situational context (Gray & Egbert, 2019), there is evidence that the bulk of a language’s highest frequency words, i.e. its function words, are important for expressing information regardless of the subject matter (Matthews & Cheng, 2015). Other research suggests that a small number of words can account for a large number of possible ideas which learners would be likely to either express or encounter (Laufer, 2013; Agernäs, 2015).

Another feature of word “coreness” is its dispersion, referring to how evenly a word is distributed within a certain text or text type. In corpus-based research, data about lexical dispersion is normally accounted for by using an index of dispersion and a predetermined threshold that must be reached in order for a word to be included (Burch, Egbert & Biber, 2016). There is currently no consensus on the best formula for measuring lexical dispersion within a corpus, and this remains a topic of open debate within the field of corpus linguistics (Burch, Egbert & Biber, 2016). A word’s dispersion across a range of different registers and modalities is usually obtained indirectly by designing a corpus to include texts from a range of different registers and modalities (Davies, 2005; Davies & Gardner, 2010; Brezina & Gablasova, 2015). This is particularly important to attend to when the corpus, a sample of language data, is intended to represent a larger language domain. It is well established that linguistic features of texts, including word choice, differ across registers (Biber, 1989), therefore a corpus aiming to serve as a language model must contain individual texts that are

representative of that variety (Sinclair, 1991; Atkins, Clear, & Ostler, 1992; Biber, 1993; Egbert, Biber and Gray, 2022).

It should be noted that successful vocabulary learning via DDL seems to depend greatly on the individual learner (Lee, Warschauer, & Lee, 2020). Researchers suggest that DDL approaches should make learners aware of both the general characteristics of the corpus being used (i.e. what register does the corpus purport to represent) as well as the underlying text processing methods (Gardner, 2007).

1.4 Research Questions

This design-based research study was motivated by the following research questions:

1.4.1 Research Question 1

How can a Chrome browser extension help L2 learners attend to core vocabulary items while reading authentic texts?

1.4.2 Research Question 2

How can a Chrome browser extension help facilitate data-driven learning for L2 learners?

2. Methods

The current study addresses the above research questions by exploring the specific use case of L2 Czech, hence this section presents the corpus-based Czech resources, such as the CGSL (Challis, 2022), Majka (Šmerk, 2007), Wiktionary (Wiktionary), and Euphonometer (Plecháč, 2017) which supplied the Chrome browser extension Lex-See with its underlying data. It should also be noted that certain design principles of Lex-See were specifically informed by linguistic characteristics of Czech, such as the need to create a bank of word forms for each lemmas, and a disregard for the concept of ‘word family’, which would have comprised so many word forms in Czech as to render this concept mostly useless. However, in principle, the methods outlined here could be applied to any language, limited primarily by corpus availability.

2.1 The Czech General Service List (CGSL) for frequency data

The Czech General Service List (CGSL) (Challis, 2022) is a frequency-ranked list of Czech lemmas (lemma + part-of-speech) with high frequency and wide dispersion across written and spoken Czech. It was built following the quantitative methodology developed for the new-GSL by Brezina & Gablasova (2015) through comparing the first 10 000 most highly ranked (by normalized average reduced frequency) lemmas of five different corpora of written Czech, namely: SYN2020, Koditex, csTenTen17, ORALv1, and ORTOFONv2. These corpora were purposefully chosen for their differences in modality, size, and design in an effort to minimize biases implicit to the design of any

single corpus and account for dispersion of words within the language. It should be noted that these five corpora shared many (but not all) of the underlying text processing methods (Hajič et al., 2007; Jelínek, 2008; Straková, Straka & Hajič, 2014; Suchomel, 2018; Kopřivová et al., 2017). Crucially, most corpora in this study use a very similar underlying tagset. It is currently unknown which of these tools or manual editing processes exerted the most influence on the final outcome.

The overlap of the first 10k ranked items of these five corpora were compared pairwise, and as expected, there was a high percent overlap and rank correlation between items from corpora with the same underlying modality, i.e. lemposes from csTenTen17 were more similar in order and rank correlations to those in SYN2020 and Koditex than to ORALv1 and ORTOFONv2. The final CGSL is the union of the intersection of lemposes common to the written corpora and the intersection of lemposes common to the spoken corpora. Final rank assignments on the CGSL were made by 1) ensuring that each lemos in this union had a rank value assigned to it for each list (missing rank values were assigned an arbitrary value of 10,001 as a penalty for not being common to multiple corpora), 2) combining all items on the CGSL-common, CGSL-written, and CGSL-spoken, as illustrated in Figure 1.

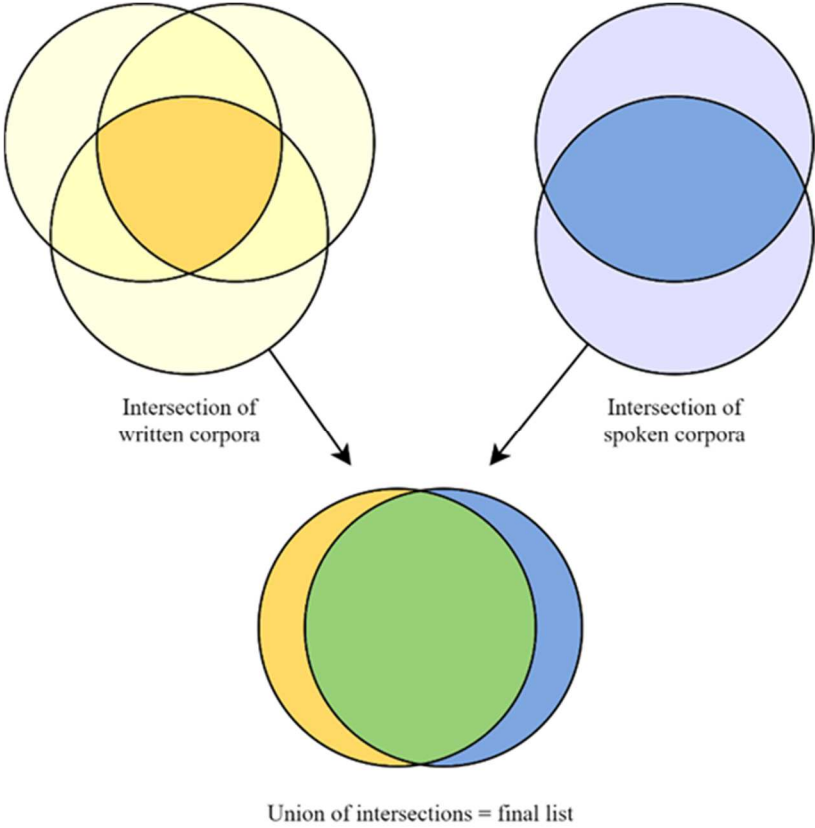


Figure 1. Illustration of CGSL design

Lempos ranks were determined by ordering according to the median, minimum, and product of the ranks across the CGSL-common, CGSL-written, and CGSL-spoken. The median value was useful as a measure of central tendency in the data, but in cases where lemposes shared the same median value, the lempos with the lower minimum value (representing a higher rank, i.e. more frequently occurring) took precedence. Even with both of these measures, there were still a few instances of lemposes “tying” in rank, especially among the highest frequency lemposes. The product of all the scores thus served as a final tiebreaker, since this measure is able to capture effects from the extreme values.

Each item on the CGSL which only occurred in the CGSL-written or CGSL-spoken was labeled as ‘written’ or ‘spoken’, respectively. Thus the final version of the CGSL consists of three main parts: 1) the common lexical core (4,903 lemposes), 2) the lemposes representing spoken Czech (3,048), and 3) the lemposes representing written Czech (2,654). Before the CGSL was compiled, each lempos was manually checked for consistency by a L1 Czech speaker.

2.2 Majka for word form data

Majka (Šmerk, 2007) is a morphological analyzer, a program that can map between the lemma and its associated word forms as well as each of their respective morphological tags. This free tool was designed as a language-agnostic solution to morphological parsing, and is currently available for 15 languages, including Czech, for which it was originally developed. Majka is designed to maximize speed, effectively traversing data precompiled in the form of a finite state automaton – it is therefore language-agnostic, the language and tagset specific data being kept in separate database files.

Lex-See was built by querying Majka’s Czech database to build a list of possible forms for the CGSL lemmata. Of the 10 605 entries, 529 were missing all data, the noun *hospoda* (Eng. ‘pub’) being one of the more curious missing entries, considering that this is a regular, high-frequency word. Apart from several other similarly inexplicable examples, missing entries were generally due to the same issues encountered when building CGSL, which included differences in decisions about the granularity of lemmata, colloquialisms, vulgarisms, interjections, etc. Since the volume was manageable, we were able to fill in the missing forms manually following the patterns and extent produced by Majka based on L1 knowledge.

2.3 Wiktionary for word meaning data

Wiktionary (Wiktionary) is a multilingual crowd-sourced web dictionary of terms, run alongside the well-known Wikipedia encyclopedia. Its openness and semi-structuredness make it suitable for use in various natural language processing tasks, as bots and an application programming interface (API) can be used to read, cross-check, or add data. Entries can typically contain etymology, part-of-speech, word forms depending on

grammatical categories, phonetic transcription, meaning, examples of use, semantically-related terms and translations.

We scraped Wiktionary for the existing entries for CGSL lemmata so that we could provide the translation and possibly an example of its use. Missing data were handled similarly to Majka missing data, i.e. via manual entry by the L1 Czech researcher.

2.4 Euphonometer for pronunciation data

While phonetic data about the base form of lemmas could have also been scraped from Czech Wiktionary, we found that it contained inconsistencies in data formatting and availability. Instead, we were able to use a tool for quantifying euphony of Czech and Slovak texts called Euphonometer (Plecháč, 2017), which features a handy phonetic transcription mode.

In addition to providing this information to the user, we then compiled similar-sounding lemmas using Levenshtein distance as a metric of phonetic similarity. Thus when the user views a lemma, we can present them with a list of the closest possible sound-alikes to be aware of.

While one might consider working with the phonetics of individual forms of the lemma, that would increase the search space by orders of magnitude. Therefore we decided against it, also because similarities between words derived from the same lemma are not especially surprising; we suspect that similarities between forms will also translate into similarities between their lemmas, as these follow a regular pattern.

3. Results

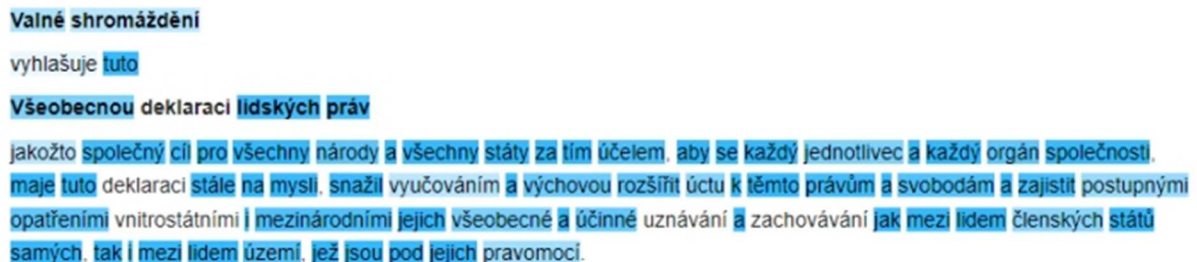
3.1 Lex-See highlighting

The primary design feature of Lex-See is that users have the ability to specify how the background color of a word on a webpage appears, aka its ‘highlighting’. The features which can specify highlighting are whether or not the word is on the CGSL, where (rank-wise) a word falls on the CGSL, what the part-of-speech (POS) of a word is associated with the lowest-rank (i.e. most frequently occurring) item on the CGSL, and if a word is part of CGSL-common, CGSL-written, or CGSL-spoken. This section will now discuss and provide examples for each of these features.

3.1.1 User-defined highlighting

The Lex-See options menu provides the ability for users to specify the color of word highlighting. This can either be a static coloring, or users can specify a range of colors for the lowest and highest rank ends of the scale, which causes words in the middle ranks of the list to appear as gradient shades between the two. For pages with a white background, if blue is chosen as the color to highlight words with the lowest rank (i.e.

the most common words), and white is chosen as the color to highlight words with the highest rank (i.e. the least common words), then all CGSL words appear highlighted on the page in a range of shades of blue, as seen in Figure 2.



Valné shromáždění
vyhlašuje **tuto**
Všeobecnou deklaraci lidských práv
jakožto **společný cíl pro všechny národy a všechny státy za tím účelem, aby se každý jednotlivec a každý orgán společnosti, máje tuto deklaraci stále na mysli, snažil vyučováním a výchovou rozšířit úctu k těmto právům a svobodám a zajistit postupnými opatřeními vnitrostátními i mezinárodními jejich všeobecné a účinné uznávání a zachovávání jak mezi lidem členských států samých, tak i mezi lidem území, jež jsou pod jejich pravomocí.**

Figure 2. Simple highlighting of text

The darker shade of blue is a visual attention cue that intuitively signals to readers which of the words in the text have a relatively stronger importance, which was determined by the frequency and dispersion from the underlying data. Users can choose whether the color distribution follows a linear or logarithmic function, of which the latter differentiates between relative rank differences of words more strongly.

However, if all the high frequency words on a page are highlighted, even in a range of shades, it is almost as ineffective as none being highlighted, since this does not meet the goal of providing differential visual attention cues to words with higher relative importance. Additionally, if every word is highlighted, the redundancy effect is likely to hinder learning. Lex-See solves this problem by allowing users to define the thresholds of the CGSL to either highlight or ignore certain words. For example, if a learner estimates that they already know approximately the first 2k most frequent words on the CGSL, and therefore do not need visual attention cues associated with these words, he or she can specify for highlighting to occur on just the words on the CGSL with rank 2001 or higher.

3.1.2 Part-of-speech highlighting

With lemmas as the underlying unit of analysis, items on the CGSL contain at least a small measure of function information, namely the word class, or POS associated with a word. Lex-See allows users to specify highlighting rules based on a word's POS in the CGSL, and in cases where the form can belong to multiple lemmata, the default POS selection is the one associated with the lowest-ranked (i.e. most frequently occurring) lemma. However, duplicate entries are quite infrequent; the CGSL is composed of lemmoses, but if we consider plain lemmata, only 73 of them contain multiple POS entries, for example rád_A (Engl: 'happy', adjective) and rád_D (Engl: 'happily', adverb). This means that 99.3% of the items on the CGSL have non-overlapping POS

tags, likely making this feature of particular benefit to L2 learners. An average word then can belong to just 1.012 lemma.

The ability to highlight words based on their most likely word class allows visual attention to be directed differently between function words and lexical words. Lex-See allows users to add multiple layers of highlighting rules based on word class, with the ability to group multiple word classes into the same rule; nouns, verbs, adjectives, and adverbs (which are typically lexical, or open-class words) can be highlighted according to one set of user-specified color, rank threshold, and scaling criterion, while numerals, prepositions, conjunctions, particles/unknown, and interjections (which are typically function, or closed-class words) can be highlighted according to a different set of criteria.

When function words are highlighted in, for example, the same static shade of yellow, it becomes a visual attention cue to L2 readers that helps differentiate them from lexical words. While we do not have sufficient empirical evidence about the difference between how L2 learners perceive, acquire, and use function words, we believe that since these kinds of words are less information-dense and occur with different frequency distributions than lexical words, it makes intuitive sense that visual attention cues can help L2 learners differentiate between these categories. Anecdotally, we have found this feature to be of particular benefit in L2 Czech reading thanks to the variety of function words present in Czech, particularly in written modality.

3.1.3 Highlighting of non-CGSL words

One of the user-defined features for Lex-See highlighting is whether or not the word or any of its associated word forms appears on the CGSL at all. Users can specify highlighting of words that do not occur on the CGSL, however these will always appear in a static color shade due to a lack of ranked frequency information. It turns out that infrequent words end up being so-called ‘keywords’, and typically include named entities, register-specific vocabulary, foreign words, and irregular forms of words, such as archaisms and diminutives (which are abundant in Czech literary texts).

It is useful for L2 readers to have distinct visual attention cues for keywords, since these are the main words which provide the ‘aboutness’ of a text. Anecdotally, it seems that top-down reading strategies are easier for L2 readers to apply to keywords than to unknown high-frequency lexical words. Perhaps this is due to the fact that keywords themselves convey information beyond the word-level; peculiar word choice seems to provide information about an author’s broader stance and message that the choice of common, high-frequency words does not.

3.1.4 Modality-based highlighting

Finally, since the CGSL also contains information about whether a word is common to written, spoken, or both registers of Czech, Lex-See is able to highlight words based on

this feature. This can be especially useful to inform how users can create their own lists.

3.2 Organizing words by meaning and sound

The definitions gathered from Czech Wiktionary (Wiktionary) and phonetic information gathered from Euphonometer (Plecháč, 2017) allow Lex-See users to quickly identify information about word meaning and sound during the process of L2 reading. For all words which occur on the CGSL, a bubble with a word definition appears upon mouse-over, saving users considerable time and effort in dictionary lookup. Additionally, Lex-See allows users to inspect specific words in greater detail via a dialog box containing the example sentences scraped from Wiktionary, as well as concordance lines of all the examples in the target text.

Another visualization feature of Lex-See is the ability to view a bar graph illustrating the counts of all word forms within the target text.

This information is especially useful for L2 Czech learners, who lack intuitions about the form frequency of certain words. When verb conjugations and noun declensions are presented in table form, as is typical in L2 Czech textbooks, it is difficult to prioritize learning one form over another and the redundancy effect takes full force, since low frequency word forms are not as salient as high frequency forms. The purpose of allowing users to explore form frequency distributions through visualization is to quickly convey information about which forms are more likely to be important.

One of the most useful features of Lex-See is the ability for users to explore other high frequency words that sound similar to a target word. This information, based on the Levenshtein distances calculated from the Euphonometer data is also displayed in the word inspection window in order of most to least similar.

3.3 List building, filtering and exporting

3.3.1 Building lists

Perhaps the feature that most closely aligns with principles of DDL is Lex-See's list-building functionality. Users can add any word, whether or not it occurs within the CGSL, to one or more lists. User lists are stored locally and persist between reading sessions, with the maximum number of lists based on compute limits. Users have the ability to name each list as well as to add a note in a text box field for each list item. Furthermore, they can define a combination of modifier keys for individual lists, which can then be used to add words to the particular list when combined with a mouse click. The beauty of being able to build a list by clicking directly on the written text is that the reader is able to minimize the shift in visual attention (i.e. distraction) caused by the act of building a list.

3.3.2 Filtering with lists

Once users have built their own Lex-See wordlists, they can then use them to define highlighting rules in addition to the other criteria. This means that it is possible to use a list of words that deserve extra attention, or the opposite, i.e. a list of words that are not necessary to highlight. Learners can use this feature to build a list that approximates their own personal receptive vocabulary of words not to highlight, which we suspect will be more useful than estimating an arbitrary rank threshold of CGSL words to avoid highlighting.

3.3.3 Exporting lists

Finally, Lex-See allows users to export personal wordlists in .csv, .tsv, and .pdf format, including all corresponding data from Lex-See as well as user created notes. This facilitates easy reuse with, for example, third-party flashcard applications.

3.4 Qualitative user data

This project was originally conceptualized as a way to solve a problem one of the researchers experienced first-hand as an L2 Czech learner. After many persistent attempts to read authentic Czech texts, which were often extremely challenging, the L2 Czech learner decided to turn to a translation of text familiar to her in English, *Harry Potter and the Sorcerer's Stone* (Rowling, 1999). While reading aloud with her L1 Czech collaborator, she was observed to have difficulty in differentiating between which new (to her) words deserved attention and which were relatively unimportant. For example, within a single chapter, the L2 Czech learner ascribed equal importance to learning *naráz*, *čest*, *šum*, *síň*, and *palec* (Engl: simultaneously, honor, noise, hall, inch) as the words *jiskrnýma*, *lektvary*, *zmodrat*, and *škrobeně* (Engl: sparkling, potions, to turn blue, starchily); in a world with limited time and attention capacity, the former set of words would be more beneficial to prioritize because they are more frequent and less specific to the content of *Harry Potter*. This real-world observation provided the original impetus to build both the CGSL and Lex-See.

The next book that the researchers read together was *Dášeňka čili život štěněte* [‘Dášeňka, or The Life of a Puppy’] (Čapek, 1935). This was done by means of the earliest versions of Lex-See, and the process informed many of the design features described in this text. For example, it wasn’t until actually using the tool that the researchers understood the need for the user to be able to specify a threshold of high-frequency words to prevent from being highlighted, and thus avoid the redundancy effect.

The researchers continue to explore L2 Czech reading, most recently with a relatively unknown text called *Valchař se směje aneb tutlanci a pozorníci* [‘The Miller Laughs, or Smugglers and Watchmen’] (Četyna, 1958). Anecdotally, the main character in this text is a fictionalized version of one of the L2 Czech learner’s 18th century ancestors.

Figure 3 illustrates Lex-See highlighting on an excerpt of this text, illustrating how visual attention cues can be used to help distinguish between different categories of words; in this example, function words are yellow, non-CGSL (i.e. 'keywords') are red, and CGSL words are shades of blue on a logarithmic scale.

["They're already **in** the pub," guessed the tall one.
They both quickly stood up **and** tried **to make out** **the gable** of the building
which was concealed by the trees.
"That's odd." The skinny **man** **shook** his head.
"What's odd?"
"**That** trees **tend to** grow where they shouldn't."
"You're right."]

Figure 3. Screenshot of highlighted words from *Valchař se směje*

One of the more humorous experiences of reading this text was seeing how Lex-See handled the glossary of archaisms found at the end of the book, shown in Figure 4, in which words not on the CGSL are highlighted in red.

práchno — **troud**
připučit — **přimáčknout**
položnica — **šestinedělka**
suchotnica — **vřes**
světadlo — **buková louč**

Figure 4. Screenshot from a glossary of archaisms.

Although the highlighting of this particular set of words did not help with L2 word prioritization in any meaningful way, the L1 Czech reader could still intuit major usage and register differences.

4. Discussion

4.1 Limitations

In this section we present limitations to the current study as well as avenues for future research.

4.1.1 Limitations in the underlying data

A computational tool is only as good as its underlying data, and there is clearly much room for improvement in all the sources of data used to fuel Lex-See. Perhaps most important to note is that it is not yet known the extent to which the items on the CGSL are actually useful to L2 Czech learners. It is assumed based on prior research in L2 vocabulary acquisition that words with high frequency and wide dispersion in a language will be useful, but this has not yet been attested and thus deserves further research.

In order to sound pleasant and make sense to L1 Czech speakers, L2 Czech learners need to be able to correctly produce names in vocative case. However, following the methodology of Brezina and Gablasova (2015), the CGSL contains no proper nouns, which means that an entire grammatical case of Czech is likely to only have limited highlighting potential in Lex-See.

The creation of the CGSL revealed inconsistencies in lemmatization and tagging in the underlying corpus data which were not immediately apparent from extensive review of the respective corpus documentation, and it is not known the extent to which variability in text processing affected the outcome of the content, rank, and modality labeling of items on the final list.

The Wiktionary definitions and examples data has not been attested for accuracy and scope of meaning, which is a known limitation. Additionally, it is not yet known the extent to which the IPA data scraped from Euphonometer reflects prototypical pronunciation of the base form of Czech words; words are known to have different pronunciations in isolation than within the context of other words in a sentence, such as connected speech. This limitation is probably not a primary concern for lower level L2 Czech learners who must first focus on building their receptive vocabulary, but it may become more problematic as proficiency levels increase.

4.1.2 Limitations in Lex-See

At the moment, the Lex-See uses lists of forms for CGSL lemmata in a plain, unoptimized format. This is somewhat ironic considering they were provided by performance-focused Majka based on an efficient storage format. While today's computers are powerful enough to pull this off, there are efficiency gains to be had in using a more specialized format. Going even further, including a morphological analyzer

directly would allow us to provide more options for words not in the CGSL that have not yet been preprocessed.

While working with live web pages can be useful to the user, it also presents many challenges. Website creators can be creative and web pages vary considerably in both structure and looks, yet the inserted user interface elements should work and blend visually with as many of these as possible. In a future version these should be rewritten to leverage modern Web Components features for better isolation.

4.2 Future Research and Conclusion

The most obvious research objective for future research is to design a user experiment to measure the extent to which Lex-See helps L2 learners to 1) attend to top-down reading strategies, and 2) improve vocabulary, grammar, and pronunciation learning. Also, while research in DDL has been shown to be effective for learning, it is not known the extent to which it is more effective than non-corpus based learning methods. Future experimental research could use Lex-See to control for variations in DDL methodology in order to gain a clearer understanding of how DDL compares to traditional classroom approaches in terms of learner outcomes.

One potential future use of Lex-See would be to use its capacity to direct a reader's visual attention to help corpus builders identify and remediate flaws in underlying data sources, such as the CGSL. While reading a text highlighted through Lex-See, we have observed multiple instances of common word forms which are incorrectly highlighted, perhaps due to inconsistencies in corpus tagging, or missing form data in Majka. In principle, Lex-See could be used to uncover similar inconsistencies in corpora of other languages, making it of value not only to L2 learners but for corpus developers and data scientists.

Lex-See could also potentially be used on texts written by L2 learners themselves to measure a variety of linguistic features, including lexical density and complexity. This information might be useful as a way to measure a learner's progress over time, and to build corpus-informed assessments. A Lex-See user study could also measure the extent to which organization strategies of user-created wordlists impacts top-down reading strategies and/or vocabulary learning.

Although adjustments can undoubtedly be made to improve Lex-See, the tool is immediately useful as a vocabulary learning tool. Visual attention cues built into Lex-See help L2 learners attend to word POS, meaning, relative coreness, modality, and patterns in form that may occur within the target text. Additionally, these cues may offset some degree of the burden on working memory during the process of L2 reading, allowing readers to more effectively apply the top-down strategies which are associated with reading proficiency and improvement. Finally, Lex-See follows the model of DDL by providing users with corpus-based information and tooling that can be applied to authentic texts, but while also allowing learners to make their own choices about how

to prioritize, organize, and explore their own L2 reading and learning experience.

In summary, Lex-See is a language agnostic Chrome browser extension tool designed to facilitate L2 reading by means of visual attention cues fueled by corpus-based data. Currently optimized for Czech, we hope to extend the scope of this tool to other languages, in particular those which lack quality corpus-based L2 learning materials.

5. References

- Adelman, J. S., Marquis, S. J., & Sabatos de Vito, S. G. (2010). Letters in words are read simultaneously, not left-to-right. *Psychological Science*, 21(12), pp. 1799–1801. doi: 10.1177/0956797610387442
- Agernäs, E. (2015). *Vocabulary size and type goals in advanced EFL and ESL classrooms. A review of research on lexical threshold, lexical coverage, reading and listening comprehension.*
- Ans, B., Carbonnel, S., & Valdois, S. (1998). A connectionist multi-trace memory model of polysyllabic word reading. *Psychological Review*, 105, pp. 678–723. doi: 10.1037/0033-295X.105.4.678-723
- Atkins, S., Clear, J., & Ostler, N. (1992). Corpus design criteria. *Literary and linguistic computing*, 7(1), pp. 1–16.
- Awadh, F. H. R., Phénix, T., Antzaka, A., Lallier, M., Carreiras, M., & Valdois, S. (2016). Cross-language modulation of visual attention span: An Arabic-French-Spanish comparison in skilled adult readers. *Frontiers in Psychology*, 7, p. 307. Retrieved from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4779959/>. doi: 10.3389/fpsyg.2016.00307
- Biber, D. (1989). A typology of English texts. *Linguistics*, 27(1).
- Biber, D. (1993). Representativeness in corpus design. *Literary and linguistic computing*, 8(4), pp. 243–257.
- Bosse, M.-L., Tainturier, M. J., & Valdois, S. (2007). Developmental dyslexia: The visual attention span deficit hypothesis. *Cognition*, 104(2), pp. 198–230. doi: 10.1016/j.cognition.2006.05.009
- Bosse, M.-L., & Valdois, S. (2009). Influence of the visual attention span on child reading performance: A cross-sectional study. *Journal of Research in Reading*, 32(2), pp. 230–253. doi: 10.1111/j.1467-9817.2008.01387.x
- Brantmeier, C. (2002). Second language reading strategy research at the secondary and university levels: Variations, disparities, and generalizability. *The Reading Matrix*, 2(3).
- Brezina, V., & Gablasova, D. (2015). Is There a Core General Vocabulary? Introducing the New General Service List. *Applied Linguistics*, 36(1), pp. 1–22.
- Burch, B., Egbert, J., & Biber, D. (2016). Measuring and interpreting lexical dispersion in corpus linguistics. *Journal of Research Design and Statistics in Linguistics and Communication Science*, 3(2), pp. 189–216.
- Čapek, K. (1935). *Dášeňka čili život štěněte.*
- Cirino, P. T., Barnes, M. A., Roberts, G., Miciak, J., & Gioia, A. (2022). Visual

- attention and reading: A test of their relation across paradigms. *Journal of Experimental Child Psychology*, p. 214, 105289.
- Challis, K. (2022). *Is there a core vocabulary for Czech? Introducing the Czech General Service List*. doi:10.13140/RG.2.2.17678.02889.
- Chapelle, C. (2003). *English language learning and technology*.
- Czerepowicka, M. (2021). The structure of a dictionary entry and grammatical properties of multi-word units. *Electronic lexicography in the 21st century (eLex 2021) Post-editing lexicography*, p. 79.
- Četyna, B. (1958). *Valchař se směje aneb tutlanci a pozorníci*. Krajské nakladatelství v Ostravě.
- Čermák, F., & Kren, M. (2011). *A frequency dictionary of Czech: core vocabulary for learners*. Routledge.
- Davies, A. (2005). *An introduction to applied linguistics*.
- Davies, M., & Gardner, D. (2010). *Word frequency list of American English*.
- Diao, Y., & Sweller, J. (2007). Redundancy in foreign language reading comprehension instruction: Concurrent written and spoken presentations. *Learning and instruction*, 17(1), pp. 78–88.
- Egbert, J., Biber, D., & Gray, B. (2022). *Designing and Evaluating Language Corpora: A Practical Framework for Corpus Representativeness*. Cambridge University Press.
- Facoetti, A., Zorzi, M., Cestnick, L., Lorusso, M. L., Molteni, M., Paganoni, P., ... & Mascetti, G. G. (2006). The relationship between visuo-spatial attention and nonword reading in developmental dyslexia. *Cognitive neuropsychology*, 23(6), pp. 841–855.
- Frey, A., & Bosse, M. L. (2018). Perceptual span, visual span, and visual attention span: Three potential ways to quantify limits on visual processing during reading. *Visual Cognition*, 26(6), pp. 412–429.
- Gardner, D. (2007). Validating the construct of word in applied corpus-based vocabulary research: A critical survey. *Applied linguistics*, 28(2), pp. 241–265.
- Gardner, D., & Davies, M. (2014). A new academic vocabulary list. *Applied linguistics*, 35(3), pp. 305–327.
- Garnier, M., & Schmitt, N. (2015). The PHaVE List: A pedagogical list of phrasal verbs and their most frequent meaning senses. *Language Teaching Research*, 19(6), pp. 645–666.
- Gori, S., & Facoetti, A. (2014). Perceptual learning as a possible new approach for remediation and prevention of developmental dyslexia. *Vision research*, 99, pp. 78–87.
- Gori, S., Cecchini, P., Bigoni, A., Molteni, M., & Facoetti, A. (2014). Magnocellular-dorsal pathway and sub-lexical route in developmental dyslexia. *Frontiers in human neuroscience*, 8, p. 460.
- Grainger, J., Dufau, S., & Ziegler, J. C. (2016). A vision of reading. *Trends in Cognitive Sciences*, 20(3), pp. 171–179. doi: 10.1016/j.tics.2015.12.008
- Gray, B., & Egbert, J. (2019). Register and register variation. *Register Studies*, 1(1),

- pp. 1–9.
- Hajič, J., Votrubec, J., Krbec, P., & Květoň, P. (2007, June). The best of two worlds: Cooperation of statistical and rule-based taggers for Czech. In *Proceedings of the workshop on Balto-Slavonic natural language processing*, pp. 67–74.
- Hu, M., & Nation, I.S.P. (2000) Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, 13(1), pp. 403-430.
- Jelínek, T. (2008). Nové značkování v Českém národním korpusu. *Naše řeč*, (1), pp. 13–20. [New Tagging in the Czech National Corpus].
- Kalyuga, S., & Sweller, J. (2005). Rapid dynamic assessment of expertise to improve the efficiency of adaptive e-learning. *Educational Technology Research and Development*, 53(3), pp. 83–93.
- Karlík, P. (2012). *Příruční mluvnice češtiny*. Lidové noviny. [Handbook of Czech]
- 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, pp. 7–36.
- Kopřivová, M., Lukeš, D., Komrsková, Z., Poukarová, P., Waclawičová, M., Benešová, L., Křen, M. (2017). *ORAL: korpus neformální mluvené češtiny, verze 1 z 2. 6. 2017*. Ústav Českého národního korpusu FF UK. Praha. Accessible at <http://www.korpus.cz> [ORAL: A Corpus of Informal Spoken Czech]
- Kopřivová, M., Komrsková, Z., Lukeš, D., & Poukarová, P. (2017). Korpus ORAL: sestavení, lemmatizace a morfologické značkování. *Korpus–gramatika–axiologie*, 15, 47-67. [The ORAL Corpus: Assembly, Lemmatization and Morphological Tagging]
- Korobov, M. (2015). Morphological analyzer and generator for Russian and Ukrainian languages. In *International conference on analysis of images, social networks and texts*, pp. 320–332. Springer, Cham.
- Kuperman, V., Siegelman, N., Schroeder, S., Acartürk, C., Alexeeva, S., Amenta, S., ... & Usal, K. A. (2022). Text reading in English as a second language: Evidence from the Multilingual Eye-Movements Corpus. *Studies in Second Language Acquisition*, pp. 1–35.
- Laufer, B. (2013). Lexical thresholds for reading comprehension: What they are and how they can be used for teaching purposes. *Tesol Quarterly*, 47(4), pp. 867–872.
- Lee, H., Warschauer, M., & Lee, J. H. (2020). Toward the establishment of a data-driven learning model: Role of learner factors in corpus-based second language vocabulary learning. *The Modern Language Journal*, 104(2), pp. 345–362.
- Lei, L., & Liu, D. (2016). A new medical academic word list: A corpus-based study with enhanced methodology. *Journal of English for academic purposes*, 22, pp. 42–53.
- Liao, S., Kruger, J. L., & Doherty, S. (2020). The impact of monolingual and bilingual subtitles on visual attention, cognitive load, and comprehension. *The Journal of Specialised Translation*, 33, pp. 70–98.
- Lobier, M., Dubois, M., & Valdois, S. (2013). The role of visual processing speed in reading speed development. *PLoS ONE*, 8, p. 4. Retrieved from

<http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0058097>

- Lobier, M., Peyrin, C., Le Bas, J. F., & Valdois, S. (2012). Pre-orthographic character string processing and parietal cortex: A role for visual attention in reading? *Neuropsychologia*, 50(9), pp. 2195–2204. doi: 10.1016/j.neuropsychologia.2012.05.023
- Lorusso, M. L., Facoetti, A., & Bakker, D. J. (2011). Neuropsychological treatment of dyslexia: does type of treatment matter? *Journal of learning disabilities*, 44(2), pp. 136–149.
- Matthews, J., & Cheng, J. (2015). Recognition of high frequency words from speech as a predictor of L2 listening comprehension. *System*, 52, pp. 1–13.
- Mayer, R. E., Heiser, J., & Lonn, S. (2001). Cognitive constraints on multimedia learning: When presenting more material results in less understanding. *Journal of educational psychology*, 93(1), p. 187.
- Nation, I. (2006). How large a vocabulary is needed for reading and listening?. *Canadian modern language review*, 63(1), pp. 59–82.
- Plecháč, P. (2017). *Euphonometer 2.0*. Prague: Institute of Czech Literature, CAS. Available at: <http://versologie.cz>.
- Rowling, J. K. (1999). *Harry Potter and the Sorcerer's Stone*. Scholastic.
- Sanosi, A. B. (2018). The effect of Quizlet on vocabulary acquisition. *Asian Journal of Education and e-learning*, 6(4).
- Schmidt-Weigand, F., Kohnert, A., & Glowalla, U. (2010). A closer look at split visual attention in system-and self-paced instruction in multimedia learning. *Learning and instruction*, 20(2), pp. 100–110.
- Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.
- Song, M. J. (1999). Reading strategies and second language reading ability: The magnitude of the relationship. *English Teaching*, 54(3), pp. 73–95.
- Straková, J., Straka, M., & Hajic, J. (2014). Open-source tools for morphology, lemmatization, POS tagging and named entity recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 13–18.
- Suchomel, V. (2018). csTenTen17, a Recent Czech Web Corpus. In *RASLAN*, pp. 111–123.
- Šmerk, P. (2007). Fast Morphological Analysis of Czech. In Petr Sojka and Aleš Horák *Proceedings of Third Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2009*. Brno: Masaryk University, 2007. pp. 13–16. ISBN 978-80-210-5048-8. Available at: <https://nlp.fi.muni.cz/ma/>
- Těšitelová, M. (1987). *O češtině v číslech*. Academia.
- Van Zeeland, H., & Schmitt, N. (2013). Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension?. *Applied Linguistics*, 34(4), pp. 457–479.
- Verhallen, M. J., & Bus, A. G. (2011). Young second language learners' visual attention to illustrations in storybooks. *Journal of Early Childhood Literacy*, 11(4), pp. 480–500.

- Wiktionary: The free dictionary*. Accessed at: <https://wiktionary.org>.
- Wright, B. A. (2016). Transforming vocabulary learning with Quizlet. *Transformation in language education*. Tokyo: JALT, pp. 436–440.
- Ziková, M. (2017). Supletivismus in Petr Karlík, Marek Nekula, Jana Pleskalová (eds.), *CzechEncy - Nový encyklopedický slovník češtiny*. Accessed at: <https://www.czechency.org/slovník/SUPLETIVISMUS>

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

Enikő Héja¹, Noémi Ligeti-Nagy¹, László Simon², Veronika Lipp²

¹Hungarian Research Centre for Linguistics, Language Technology Research Group,
Budapest, Hungary

² Hungarian Research Centre for Linguistics, Lexical Knowledge Representation Research
Group, Budapest, Hungary

E-mail: {surname.forename}@nytud.hu

Abstract

The present paper describes the initial phase of a collaboration between Hungarian lexicographers and computational linguists aimed at compiling the new version of The Explanatory Dictionary of the Hungarian Language. This research thread focuses on the automatic sense induction of Hungarian adjectives in attributive positions, and their salient nominal contexts, with a particular emphasis on polysemies. The proposed methodology is intended to facilitate lexicographers' work in characterizing both the micro- and macrostructure of adjectives in a monolingual setting. A corpus-driven, unsupervised graph-based approach was employed, which, as per our expectations, could potentially reduce the reliance on human intuition, especially in the ambiguous domain of polysemic sense distinctions. Initially, distributional criteria for meaning distinction were introduced, followed by the description of the employed algorithm. The algorithm models adjectival semantics using two unique subgraphs: connected graph components are used to model adjectival semantic domains, while maximally connected subgraphs, so called cliques, model polysemies. Automatically induced meaning distinctions were validated using salient nominal context candidates extracted from corpus data. We expect that while connected graph components aid in characterizing the adjectival macrostructure, cliques provide lexicographers with useful insights for establishing the adjectival microstructure. These hypotheses were also tested: we investigated the extent to which the proposed framework can assist expert lexicographers during the dictionary compilation process by comparing a sample of our automatically obtained results to the previous version of The Explanatory Dictionary of the Hungarian Language.

Keywords: automatic sense induction, monolingual lexicography, polysemy, unsupervised graph-based approach, adjectives

1. Introduction

Although corpus-based methodology is increasingly central in monolingual lexicography, complemented by a variety of software tools and detailed guidelines (cf. Atkins & Rundell, 2008), we are not aware of any lexicographic projects employing a corpus-driven approach. Such an approach could significantly contribute to the field: notably, it could expedite the workflow and reduce the reliance on human intuition during the lexicographic process. This can be particularly useful in the nebulous area of meaning distinctions, thus assisting in the formation of the microstructure, a well-established challenge in both bilingual and monolingual dictionaries (Adamska-Sałaciak, 2006; Hanks, 2012; Véronis, 2003). A

corpus-driven technique should strive to leverage corpus data to the fullest extent, with minimal human intervention. Consequently, establishing operationalizable distributional criteria for sense distinction is crucial. Regrettably, to our knowledge, there is no widely accepted distributional definition of polysemy that would allow for more data-driven, and hence more objective meaning distinctions (cf. Geeraerts, 2010).

This challenge is even more pronounced in the case of adjectives. Adjectives pose a significant difficulty when attempting to divide them into distinct senses (Moon, 1987). It is hard to analyze them in isolation because they essentially constitute an aspect of the modified noun. Furthermore, adjectival lexical semantics represents a relatively under-researched area in linguistics. While several attempts have been made to identify different verbal structures and their associated meanings based on distributional properties (e.g. Levin, 1993; Kipper-Schuler, 2005 and Sass et al., 2010 for Hungarian), we are not aware of any similar initiatives concerning adjectives. This is even more so in the case of Hungarian adjectives: to our knowledge, only Kiefer (2003, 2008) provides a detailed examination of adjectival semantics.

Accordingly, our primary objectives are: (1) to provide sufficient criteria to grasp adjectival sense distinction, including polysemies; (2) to model these criteria and (3) to evaluate the extent to which this technique can aid expert lexicographers to develop the adjectival microstructure of the new version of *The Explanatory Dictionary of the Hungarian Language* (EDHL). The EDHL is an up-to-date online dictionary of contemporary Hungarian (covering 2001–2020) that is being compiled using corpus-driven methods (Lipp & Simon, 2021).

As per our expectations, the automatically extracted adjectival subsenses should provide lexicographers with a ready-to-use adjectival microstructure, significantly facilitating their work. This hypothesis was tested from two distinct angles: First, approximately 60 automatically extracted polysemies were compared to the relevant microstructures of a traditional explanatory dictionary from multiple perspectives, including coverage and, most importantly, the motivatedness of meaning distinctions. In relation to this, special attention was devoted to the nominal contexts of the adjectives. We expect that the detected subsenses subcategorize certain semantic classes. Secondly, approximately 6400 adjectives from the Hungarian Webcorpus 2.0 (Nemeskey, 2020) were partitioned into semantic domains fully automatically. This partition was then compared with the macrostructure of the EDHL to examine the extent to which it could streamline the headword selection process.

2. Motivation

2.1 Lexicographic background

In lexicography, three distinct paradigms are employed: traditional, corpus-based, and corpus-driven approaches (Atkins & Rundell, 2008; Svensén, 2009). Within the traditional approach, lexicographers heavily rely on their linguistic intuition, which results in an imbalanced description of the relevant linguistic phenomena.

The two Hungarian monolingual general-purpose dictionaries of the 20th century, *A magyar nyelv értelmező szótára* [The Explanatory Dictionary of the Hungarian Language; EDHL] (Bárcki & Országh, 1959–1962) and *Értelmező kéziszótár* [Concise Hungarian Explanatory Dictionary; CHDL¹, CHDL²] (Juhász et al., 1972; Pusztai & Csábi, 2003), were compiled

using the traditional method. The editors of EDHL relied on their own mental lexicon throughout the dictionary creation process. As the leading editor asserts, “Our own language knowledge and language sense, which we constantly verified through surveys, served as the natural basis for our work in recording word meaning, usage, and stylistic value” (Ország, 1953: 397). Work on *A magyar nyelv nagyszótára* [Comprehensive Dictionary of Hungarian; CDH] (Ittész, 2006–2021) began in 1985 based on a historical corpus. However, the limited size of the corpus (30 million words) did not provide sufficient data for dictionary writing.

To modernize linguistic research and link Hungarian lexicography to ongoing European projects, a text database of significant size and quality is needed. Databases like the Hungarian National Corpus (Váradi, 2002) (HNC) and the Hungarian Gigaword Corpus (Oravecz et al., 2014), while comparable to prominent corpora like the British National Corpus (Burnard, 2007) and Deutsches Referenzkorpus (Kupietz et al., 2010), are not suitable for lexicographic research due to various limitations. Similarly, web-scraped databases, such as the Hungarian Web Corpus (Jakubíček et al., 2013) are also insufficient due to their imbalanced nature and the limited metadata they provide.

The corpus-based lexicography focuses on word usage patterns and relies on the contexts in which words typically occur (Hanks, 2010). Senses and subsenses are established based on such information, utilizing suitable corpus tools. Taking a step further, the corpus-driven methodology aims to explore the meaning space of a word through fully automatic means, further reducing the reliance on human intuition. One of the significant advantages of this technique is its ability to handle vast data sets. In 2021, the Hungarian Research Centre for Linguistics initiated a project to update the EDHL, originally created in the 1960s, using automatic methods applied to a new, extensive, and representative input corpus. The primary objective is to obtain an objective lexical profile for each dictionary entry, anticipating that this information will expedite the creation of a new explanatory dictionary (Lipp & Simon, 2021).

2.2 Consistent methodology

Our proposed method aligns perfectly with the envisioned framework for creating the new version of EDHL. It not only relies on data but also leverages unlabeled data, apart from the part-of-speech annotation. This means that the algorithm processes data with minimal presuppositions about meanings. Moreover, our methodology is based on a substantial amount of data, especially from a lexicographic standpoint. The adjectival meanings are distilled from a subset of 170 million sentences, extracted from the Webcorpus 2.0 (Nemeskey, 2020). Contextual information is retrieved from the 180-million-word HNC. Furthermore, if needed, the amount of data utilized can be expanded.

The data-driven technique we employ relies on distributional criteria for meaning distinction, which we consider a novel contribution to the field. These criteria, in contrast to previous definitions based on etymology or sense relatedness, offer a more intersubjective approach. Additionally, they can be easily modeled using a simple graph-based approach.

Hopefully, the corpus-driven method can be enhanced through a meticulous lexicographic post-editing phase. The close collaboration between different fields ideally leads to the development of data-oriented, explicit lexicographic editing principles that apply to both the macrostructure and microstructure of the dictionary.

In the next section, we will present the distributional criteria for meaning distinction, followed by an overview of the unsupervised word sense induction experiment conducted on Hungarian monolingual data. The workflow can be conceptually divided into two main stages: i) The detection of subsense candidates for a given adjective, ii) discrimination between the different meanings of the given adjective by extracting relevant context nouns.

3. Distributional criteria for meaning distinction

3.1 Near-synonymy

First, let us recall the notion of near-synonymy (cf. Ploux & Victorri, 1998), a relaxed version of synonymy (cf. Frege, 1892), which is heavily relied upon when formulating the distributional criteria for meaning distinction. That is, two expressions are *near-synonyms* if they are interchangeable in a restricted set of contexts, preserving the meaning of the original sentence. For instance, the Hungarian adjectives *finom* 'fine' and *lágy* 'soft' are synonyms before nouns related to music, such as the Hungarian counterparts of 'music,' 'rhythm,' 'melody,' etc., as *lágy zene* and *finom zene* convey the same meanings. For the sake of the present research, the notion of near-synonymy is further extended: we also consider the members of tight semantic classes to be near-synonyms, as they denote different senses of a word, even though they may not preserve the truth value. This extension aligns with our original purpose of meaning distinction.¹

3.2 Criteria for meaning distinction

Accordingly, an adjective has multiple meanings if:

1. There is (at least) one near-synonym for each sense of the adjective.
2. There is a set of context nouns that form grammatical constructions both with the original adjective and with the near-synonym.
3. The two sets of context nouns that characterize the different senses are non-overlapping.
4. The non-overlapping set of nouns forms a semantic category, reflecting the sub-selectional properties of adjectives (Pustejovsky, 1995).

Example 1 illustrates the four criteria using two automatically extracted senses of the adjective *napfényes* ('sunny'). As observed, there is a near-synonym for each sense: *napsütétes* ('sunshiny') for the first sense and *napsütötte* ('sunlit') for the second sense. The listed nouns below the adjectives are those that form grammatical constructions with the respective near-synonyms, such as *napfényes/napsütétes vasárnap* ('sunny/sunshiny Sunday'), *napfényes/napsütétes nap* ('sunny/sunshiny day'), and *napfényes/napsütötte terület* ('sunny/sunlit area'), *napfényes/napsütötte terasz* ('sunny/sunlit terrace').

Importantly, the two sets of nouns do not overlap; there are no instances like **napsütétes terasz* ('sunshiny terrace') or **napsütötte nap* ('sunlit day'), and the same holds true for all adjective-noun pairs where the noun comes from the context noun set of the other sense. Finally, the nouns that match the above criteria form a semantic category: time periods with the first sense, and areas, places with the second.

¹ For example, *fekete* 'black' may belong to two different near-synonymy sets: one containing surnames and the other containing names of colors.

- (1) **Sense 1:** *napfényes* 'sunny', *napsütéses* 'sunshiny'
Nouns of sense 1: *vasárnap* 'Sunday', *nap* 'day'

Sense 2: *napfényes* 'sunny', *napsütötte* 'sunlit'
Nouns of sense 2: *terület* 'area', *sziget* 'island', *oldal* 'side', *terasz* 'terrace'

4. Representation of the investigated phenomena

The present discussion is confined to a brief overview of the algorithm, possibly from a lexicographic perspective – with only the necessary amount of technical details. For more detailed information, please refer to Héja & Ligeti-Nagy (2022a,b). First, the representation of the input categories will be described, followed by the presentation of the various adjectival meaning representations and the related simple graph-theoretic concepts. Finally, we discuss how the salient nominal contexts were detected. It is important to emphasize that at this stage, the meaning representations are induced *fully automatically* from corpus data.

4.1 Selection of input adjectives

The adjectives of interest were selected based on the 180-million-word HNC. Specifically, we considered all the adjectives that occurred at least 2 times in the HNC.

4.2 Representation of adjectives

In the subsequent step, static vector representations (Mikolov et al., 2013a,b) were generated for the selected adjectives using the first 999 files (21GB of raw texts) from the Webcorpus 2.0 (Nemeskey, 2020). The cc. 170-million sentence training corpus consists of the normalized version of the original texts. To create the vector representations, 300-dimensional vectors were trained using the Gensim Python package (Rehurek & Sojka, 2011). The training was performed using the Continuous Bag-of-Words (CBOW) algorithm with a window size of 6k and a minimum frequency of 3. Roughly 8.5 million word forms were assigned embeddings. The trained language model (LMs) can be accessed at the following link: https://nlp.nytud.hu/word2vec/cbow_3.tar.gz.

While we acknowledge that static word embeddings have become outdated in the field of natural language processing, they still offer several advantages over more recent contextual embeddings. They are easy to train and handle, and importantly, they provide interpretability, which is crucial for lexicography. However, one drawback of this approach is the “meaning conflation deficiency” as described in (Camacho-Collados & Pilehvar, 2018), which states that such representations conflate the various subsenses of a lemma into one point in the semantic space.

In the subsequent sections, we will demonstrate that the meaning conflation deficiency can be effectively addressed through graph representations, particularly in the case of adjectival polysemies. This approach yields highly interpretable results and mitigates the limitations associated with static word embeddings.

4.3 Graph-based representation of adjectival meanings

Our methodology is based on the graph representation of adjectives. A graph is a mathematical structure composed of nodes and edges. In this context, nodes represent adjectives, while edges connecting two nodes represent whether the two adjectives are semantically similar. As can be seen in Figure 1, the ego graph² of *érzékeny* ('sensitive') includes all the adjacent adjectives to *érzékeny*, along with the edges between those adjacent adjectives. It demonstrates that the Hungarian adjective *érzékeny* is semantically similar to *gyengéd* ('gentle'), *törékeny* ('fragile'), and *fogékony* ('receptive'). As these latter nodes are not interconnected, they likely belong to different subsenses of the central adjective *érzékeny*.

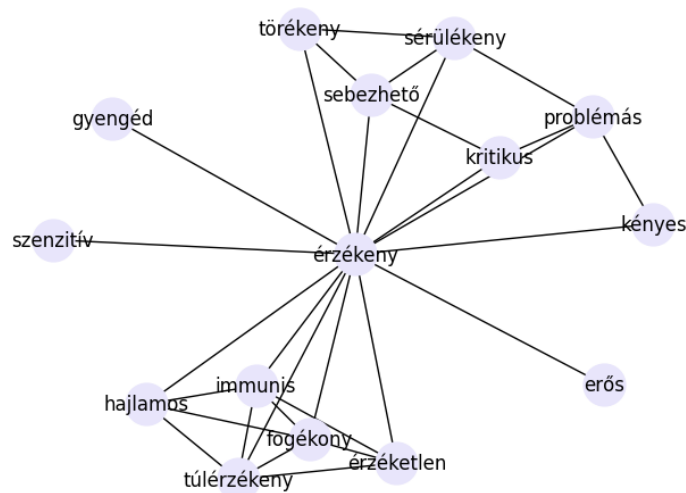


Figure 1: The ego graph of *érzékeny* 'sensitive'³

4.4 Representing near-synonymy classes as cliques

Following the generation of the graph representation of the adjectival semantic space, near-synonymy classes are modeled via maximally connected subgraphs, also known as *cliques*. A clique is a (sub)graph in which every node is connected to every other node in the (sub)graph (cf. Figure 2).

The basic premise of this representation is that in an adjectival clique, the meaning of each element is similar to that of every other element, thus, cliques are strong candidates for near-synonymy classes representing a (sub)sense of an adjective. Indeed, the meanings of *gyönyörű* 'beautiful', *csodaszép* 'stunning', *gyönyörűség* 'gorgeous', *szépséges* 'lovely',

² The ego graph or ego network is a specialized type of graph consisting of a central node (the ego) and all other nodes directly connected to it (the alters). Edges between the alters also form part of the ego graph.

³ *törékeny*: 'fragile', *sérülékeny*: 'vulnerable', *sebezhető*: 'susceptible', *kritikus*: 'critical', *problémás*: 'problematic', *kényes*: 'delicate', *erős*: 'strong', *immunis*: 'immune', *hajlamos*: 'prone', *fogékony*: 'receptive', *érzéketlen*: 'insensitive', *túlérzékeny*: 'oversensitive', *szenzitív*: 'sensitive', *gyengéd*: 'gentle'

⁴ *gyönyörű*: 'beautiful', *csodaszép*: 'stunning', *gyönyörűség*: 'gorgeous', *szépséges*: 'lovely', *mesés*: 'fabulous', *tündéri*: 'adorable'

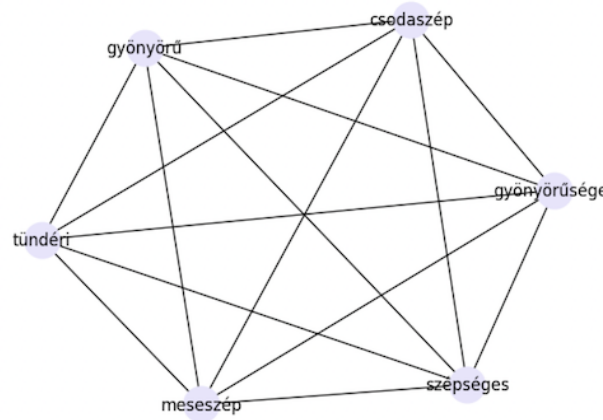


Figure 2: The clique modeling the near-synonymy class of *gyönyörű* 'beautiful'⁴

meseszép 'fabulous', and *tündéri* 'adorable' are highly similar, indicating that these adjectives belong to the very same meaning.

4.5 Meaning distinction: one adjective in multiple cliques

Consequently, in the next step, multiple subsenses of a lemma are to be modeled by multiple cliques. That is, an adjective may have multiple senses, if it belongs to multiple cliques (cf. criterion 1).

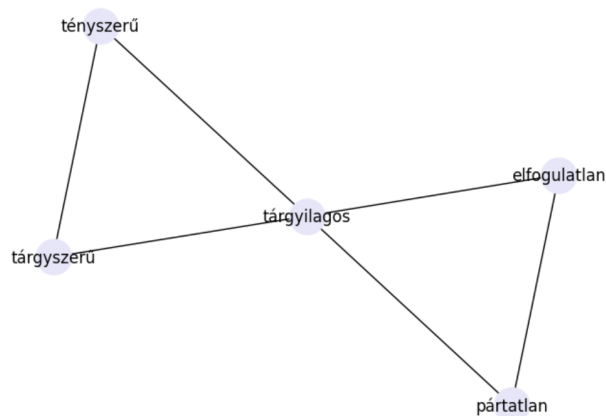


Figure 3: The Hungarian adjective *tárgyilagós* 'objective' belongs to two cliques⁵

For example, as illustrated in Figure 3, the Hungarian adjective *tárgyilagós* 'objective' belongs to two different cliques, indicating two distinct subsenses of the lexeme: clique 1 comprises *tárgyszerű* 'concise' and *tényszerű* 'factual' as near-synonym candidates, whereas clique 2 consists of *pártatlan* 'impartial' and *elfogulatlan* 'unbiased', representing a different subsense. Notably, this sense distinction is further underpinned by the following nouns (cf. criterion 3). The elements of clique 1 co-occur with nouns such as *leírás* 'description', *ismertetés* 'exposé', *vita* 'discussion', while adjectives in clique 2 co-occur with nouns

⁵ *tárgyszerű*: 'concise', *tényszerű*: 'factual', *tárgyilagós*: 'objective', *elfogulatlan*: 'unbiased', *pártatlan* 'impartial'

like *megítélés* 'judgement', *vélemény* 'opinion', and *eljárás* 'procedure'. This outcome supports our intuition according to which the first sense of *tárgyilagós* is more objective corresponding to the facts, while the second sense is used rather in the sense of being impartial.

4.6 Clique validation via the following nouns

4.6.1 Extracting the nominal contexts

Three out of the four criteria for meaning distinction pertain to the nouns modified by the attribute adjectives: there should be (1) a set of (2) non-overlapping context nouns (3) that form coherent semantic classes, reflecting the sub-selectional properties of the adjectival near-synonymy sets. These three *clique validation steps* are vital to our workflow. They align with Levin (1993) and are predicated on the assumption that adjectives, similar to verbs, impose semantic selectional restrictions on their arguments. Consequently, tight nominal semantic classes are required to validate the adjectival subsense candidates. In cases of two meaning candidates, i.e., two shared cliques, criterion (2) and (3) can be expressed more formally as computing the symmetric difference of the nominal sets A and B , where A comprises nouns occurring after all adjectives in clique 1, and B includes nouns occurring after all adjectives in clique 2.

Let's revisit example 1: *napfényes* 'sunny' had two separate subsenses, *napsütéses* 'sunshiny' and *napsütötte* 'sunlit':

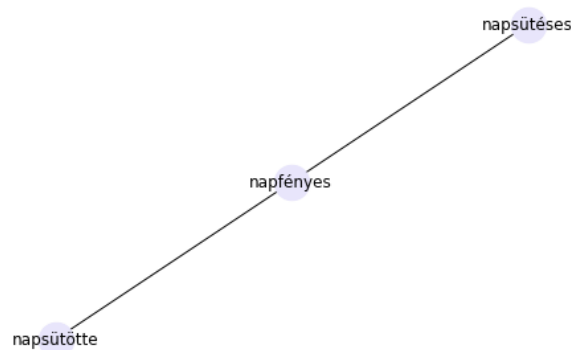


Figure 4: The Hungarian adjective *napfényes* 'sunny' belongs to two cliques⁶

It was also claimed that the two separate submeanings are characterized by two distinct sets of nouns, as follows:

- (2) **Sense 1:** *napfényes* 'sunny', *napsütéses* 'sunshiny'
Nouns of sense 1: *vasárnap* 'Sunday', *nap* 'day'

Sense 2: *napfényes* 'sunny', *napsütötte* 'sunlit'
Nouns of sense 2: *terület* 'area', *sziget* 'island', *oldal* 'side', *terasz* 'terrace'

⁶ *napsütötte*: 'sunlit', *napfényes*: 'sunny', *napsütéses*: 'sunshiny'

It is noteworthy that these nouns form non-overlapping sets: nouns co-occurring with both senses were discarded. The resulting nominal sets were first checked for semantic coherence by automated means, then by meticulous lexicographic inspection.

4.6.2 Detecting the salient nominal contexts via binary trees (dendrograms)

In many cases, the set of retrieved nominal contexts was too large to interpret at a glance. In such instances, the word2vec representations (as described in Section 4.2) of the context nouns were clustered to yield salient semantic categories for the given subsense. The noun vectors were clustered using a hierarchical agglomerative algorithm with cosine distance and average linkage. For instance, *mindennapi* 'common' had been assigned two meanings: *hétköznapi* 'ordinary' and *mindennapos* 'everyday'. On one hand, the respective near-synonyms are rather enlightening with regard to the two senses of the adjective; one of them meaning 'normal' or 'ordinary', while the other refers to regular, everyday activities. However, we still need to know which nouns can induce the relevant meanings. For this purpose, dendrograms are created, yielding information that, for example, language-related things, such as *szóhasználat* 'word usage' and *nyelvhasználat* 'language use', along with *hős* 'hero', *figura* 'character', and *jelenet* 'scene', are more likely to be common or ordinary than periodical. On the other hand, *gyakorlás* 'practice' and *testmozgás* 'exercise' are regular, everyday activities and not necessarily common or ordinary ones. Therefore, the branches of the dendrogram indicate the semantic classes of nouns that the adjectival senses subcategorize.

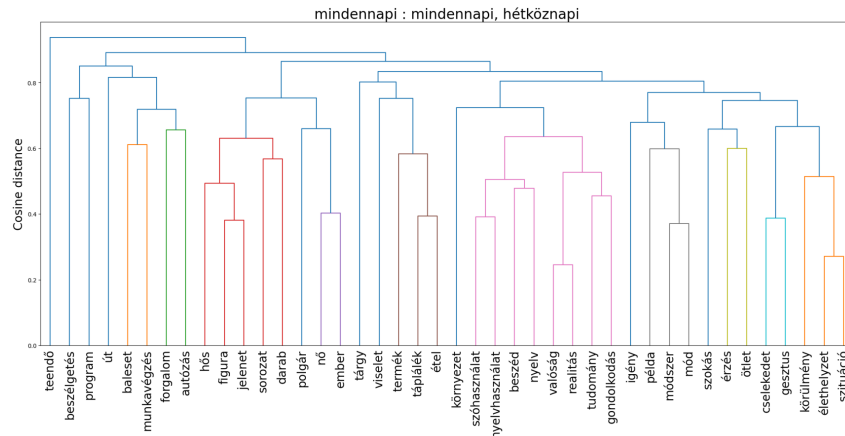


Figure 5: The dendrogram of the adjective *everyday* meaning 'ordinary' with its salient nominal contexts⁷

⁷ *mindennapi*: 'everyday', *hétköznapi*: 'ordinary', *teendő*: 'task', *beszélgetés*: 'conversation', *program*: 'program', *út*: 'road', *baleset*: 'accident', *munkavégzés*: 'work', *forgalom*: 'traffic', *autózás*: 'driving', *hős*: 'hero', *figura*: 'figure', *jelenet*: 'scene', *sorozat*: 'series', *darab*: 'piece', *polgár*: 'citizen', *nő*: 'woman', *ember*: 'human', *tárgy*: 'object', *viselet*: 'clothing', *termék*: 'product', *táplálék*: 'nutrition', *étel*: 'food', *környezet*: 'environment', *szóhasználat*: 'word usage', *nyelvhasználat*: 'language usage', *beszéd*: 'speech', *nyelv*: 'language', *valóság*: 'reality', *realitás*: 'reality', *tudomány*: 'science', *gondolkodás*: 'thinking', *igény*: 'demand', *példa*: 'example', *módszer*: 'method', *mód*: 'way', *szokás*: 'habit', *érzés*: 'feeling', *ötlet*: 'idea', *cselekedet*: 'action', *gesztus*: 'gesture', *körülmény*: 'circumstance', *élethelyzet*: 'life situation', *situáció*: 'situation'.

4.7 Representing semantic domains as connected components

A connected component is a subset of network nodes such that there is a *path* from each node in the subset to any other node in the same subset. As Zinoviev (2018: 129) notes, "The property of connectedness is global and, while important for social and communication networks [...], may not be adequate for semantic, product, and other types of networks". In the light of this assertion, it was quite unexpected that the connected components of the adjectival graph strictly corresponded to non-overlapping, semantically coherent components. The original adjectival graph, consisting of 10,153 adjectives, was dissected into 1,807 components using this technique, yielding a partition over 6,417 adjectives. Each component corresponds to a well-defined semantic domain. Note that one component of such networks is always a giant connected component (GCC), which comprises approximately one-third of the input adjectives (3,736) in this case. Unfortunately, the GCC merges multiple clear-cut semantic domains into one huge conglomerate, thus remaining uninformative about the meaning of the node adjectives as a whole.

Moreover, the adjectival graph components not only keep the various semantic domains separate but also reveal the relations between the inner node adjectives. These relations provide valuable information regarding polysemies and meaning shifts (Figure 6).

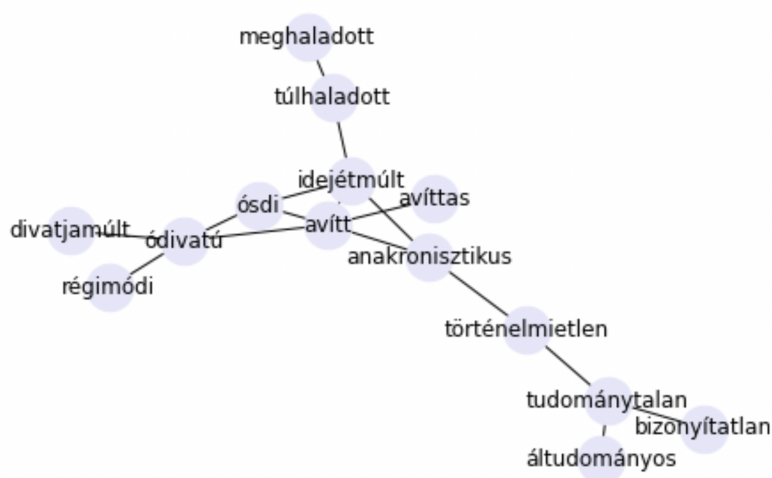


Figure 6: A connected component of the adjectival graph from the semantic domain *outdated*⁸

As Figure 6 indicates, there is an adjectival semantic field corresponding to *idejétmúlt* 'outdated'. There are three different edges from this node pointing to three different submeanings: *ósdi* 'shabby', *túlhaladott* 'obsolete', and *anakronisztikus* 'anachronistic'. The figure also shows that the next node after *anakronisztikus* is *történelmietlen* 'ahistorical', which leads to *ál tudományos* 'pseudoscientific' in two steps.

⁸ *divatjamúlt*: 'outdated', *régimódi*: 'old-fashioned', *ódivatú*: 'antiquated', *ósdi*: 'shabby', *meghaladott*: 'outmoded', *túlhaladott*: 'obsolete', *idejétmúlt*: 'outdated', *avított*: 'stale', *avíttság*: 'musty', *anakronisztikus*: 'anachronistic', *történelmietlen*: 'ahistorical', *tudománytalan*: 'unscientific', *ál tudományos*: 'pseudoscientific', *bizonyítatlan*: 'unproven'

Consequently, connected components offer lexicographers a neatly categorized headword list, enabling a more thesaurus-like editing process, as opposed to the traditional alphabetical one, aligning with Stock (1984: 38).

5. Workflow: Unsupervised Extraction of Representations from Corpus Data

The methodology detailed here extends the fairly simple unsupervised graph-based approach described in Héja & Ligeti-Nagy (2022a,b), which was partially inspired by Ah-Pine & Jacquet (2009). Nevertheless, we introduced several significant changes. Firstly, we considered adjectives with lower frequency counts in the HNC to enhance coverage. Secondly, contrary to the previous experiment, we searched the entire HNC for salient noun candidates. Furthermore, our research didn't limit itself to polysemy: in addition to cliques, we generated and explored connected subgraphs from a lexicographic perspective. The key steps of the unsupervised graph induction process are recapped below:

1. Initially, we generated a weighted undirected graph, F , based on adjectival word2vec representations (cf. Subsection 4.2). In this graph, nodes represent adjectives, while edge weights indicate the strength of semantic similarity between every pair of adjectives. The weights were calculated using the standard cosine similarity measure. Importantly, the induced graph's undirectedness is guaranteed by the symmetric nature of cosine similarity.
2. Subsequently, we created an unweighted graph, G , by binarizing F . We used a K cut-off parameter to eliminate edges with low strength. Each edge weight w was set to 1 if $w \geq K$, and w was set to 0, if $w < K$. As a result, the graph G consists only of edges of the same strength ($w = 1$), where edges with $w = 0$ were omitted. During our experiments, K was set to 0.5 or 0.7.

However, in accordance with Zinoviev (2018: 80) we found that determining the optimal value for K presents a challenging task for future research. To illustrate the role of the K cut-off parameter, let us revisit the ego graph shown in Figure 1, where $K = 0.5$ was used. This graph consists of 15 nodes and 27 edges. By contrast, with $K = 0.7$, *érzékeny* becomes an isolated node, i.e., a subgraph containing no edges, since all adjacent nodes are connected with weights where $0.5 < w < 0.7$. Setting $K = 0.65$ results in an ego graph with 6 nodes and 5 edges (cf. Figure 7), indicating that a higher K cut-off value yields a smaller subgraph, both in terms of nodes and edges, likely possessing a less rich microstructure.

Moreover, the manual evaluation of the adjectival graph showed that the edge weights are characteristic of the semantic field to which the investigated adjectives belong. For example, a slicing threshold of $K = 0.9$ results in a graph where the components tend to correspond to referring adjectives with minimal lexical meaning components, such as names of days (cf. 8a), names of months (cf. 8b), or terminological expressions (e.g., *összájú* 'protostome', *újszájú* 'deuterostome').

As expressions with poor lexical meanings are less interesting from a lexicographic perspective, we must reduce the K cut-off value. As implied by Figure 7 and Figure 1, the

⁹ *érzékeny*: 'sensitive', *sérülékeny*: 'vulnerable', *kényes*: 'delicate', *érzéketlen*: 'insensitive', *fogékony*: 'susceptible', *túlérzékeny*: 'hypersensitive'



Figure 7: The ego graph of *érzékeny* 'sensitive' with $K = 0.65$ as cut-off parameter⁹

lower the K cut-off value, the richer the semantic content of the resulting microstructure candidate.

However, a lower K cut-off parameter may lead to more chaotic connected components and cliques, particularly in specific semantic domains. Thus, the precise parameter setting must be guided by meticulous lexicographic inspection, where both the semantic domains and the extent of the coverage need to be considered.

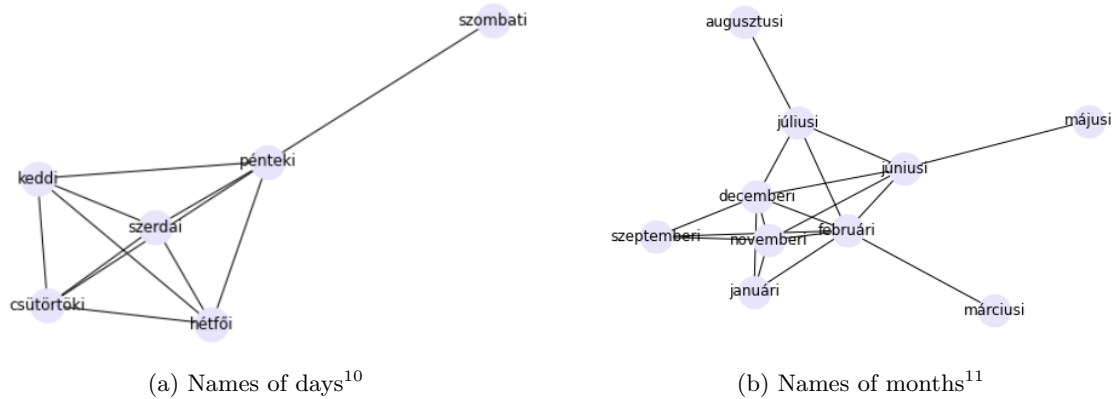


Figure 8: Graphs of referring expressions: $K \geq 0.9$

6. Lexicographic Perspective

In this section, we focus on the potential application of the proposed method for lexicographic purposes, specifically in the compilation of monolingual explanatory dictionaries. To this end, we will compare the automatically induced results with the micro- and

¹⁰ *hétfői*: 'of-Monday', *keddi*: 'of-Tuesday', *szerdai*: 'of-Wednesday', *csütörtöki*: 'of-Thursday', *pénteki*: 'of-Friday', *szombati*: 'of-Saturday'

¹¹ *januári*: 'of-January', *februári*: 'of-February', *márciusi*: 'of-March', *májusi*: 'of-May', *júniusi*: 'of-June', *júliusi*: 'of-July', *augusztusi*: 'of-August', *szeptemberi*: 'of-September', *novemberi*: 'of-November', *decemberi*: 'of-December'

macrostructure of the EDHL. Unfortunately, EDHL does not offer any insight into the selection principles for its adjectival headword list. It merely states that the cataloged headwords, as curated by the editorial board, are “common, widely known, frequently used, and vital in communication and daily interaction in our language” (Bárzsi & Országh, 1959–1962: VII).

From a lexicographic perspective, we tested five hypotheses:

1. The induced cliques can assist lexicographers in constructing the adjectival microstructure.
2. The automatically extracted and clustered nouns, modified by attributive adjectives and represented in the dendrograms, may aid lexicographers in supplementing the data used in EDHL for defining the adjectival microstructure.
3. The clusters of nouns might characterize the adjectival microstructure independently, indicating where distinctions in meaning need to be made, without relying on any pre-existing definitions.
4. We also investigated whether the automatically induced dendrograms can assist lexicographers in identifying inconsistencies in the EDHL, which may arise as a side effect of intuition-based methodologies.
5. The automatically extracted subgraphs, i.e., connected components, may also help in identifying missing headwords, thereby supplementing the macrostructure.

A detailed analysis of the ego graphs for 20 frequent adjectives, cut at a $K = 0.7$ threshold, revealed that in 8 instances, corresponding cliques included relevant adjectives not found in the EDHL. For example, the headword *bárgyú* ‘silly’ does not include the subsense *bugyuta* ‘foolish’. Similarly, the headword *bizarr* ‘bizarre’ lacks *morbid* ‘morbid’ and *szürreális* ‘surreal’ (refer to Figure 9a), while the headword *megdöbbentő* ‘shocking’ does not comprise the subsense *mellbevágó* ‘gut-wrenching’ (see Figure 9b). When we lower the threshold to $K = 0.5$, the cliques become more granular, highlighting additional missing subsenses in the microstructure. For instance, the adjective *érzékeny* (refer to Figure 1) lacks references to subsenses *sebezhető* ‘vulnerable’ and *problematikus* ‘problematic’ in the EDHL.

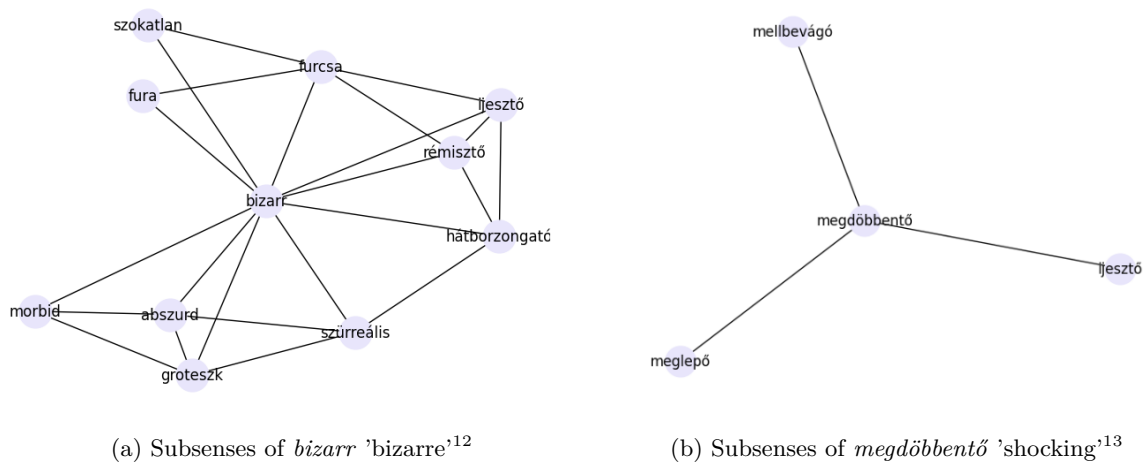


Figure 9: Ego graphs compared to the microstructure of EDHL; $K \geq 0.7$

The second hypothesis proved to be completely correct based on the assessment of randomly selected dendrograms. This outcome isn't surprising, particularly considering that nodes near the terminals in the dendrogram align with cohesive, semantically related noun classes. For example, *fontos* 'important' can co-occur with military events such as *csata* 'battle', *hadművelet* 'military operation', and *küldetés* 'mission', or with various legal acts like *rendelet* 'order', *törvénytervezet* 'legislative proposal', *egyezmény* 'convention', and *szerződés* 'contract'. Similarly, *alacsony* 'low' often modifies financial terms related to money such as *áfakulcs* 'VAT rate', *alapanyagár* 'raw material price', *áramár* 'power tariff', and *adósságállomány* 'debt portfolio'.

The third hypothesis was partially validated. It was discovered that only nodes close to the terminals in the dendrogram—those with low cosine distances—indicate accurate meaning distinctions. For instance, as the red branch in Figure 10 demonstrates, military-related light weapons, including *kard* 'sword', *szablya* 'saber', *cirkáló* 'cruiser', *puska* 'rifle', and *ágyú* 'cannon', share the definition of 'a <smaller-sized weapon> that does not require much effort to carry, transport, and handle' in EDHL and group together convincingly. The data pertaining to the definition of 'a <military unit> equipped with such weapons' (*gyalogság* 'infantry', *tüzérség* 'artillery') is also well-differentiated.

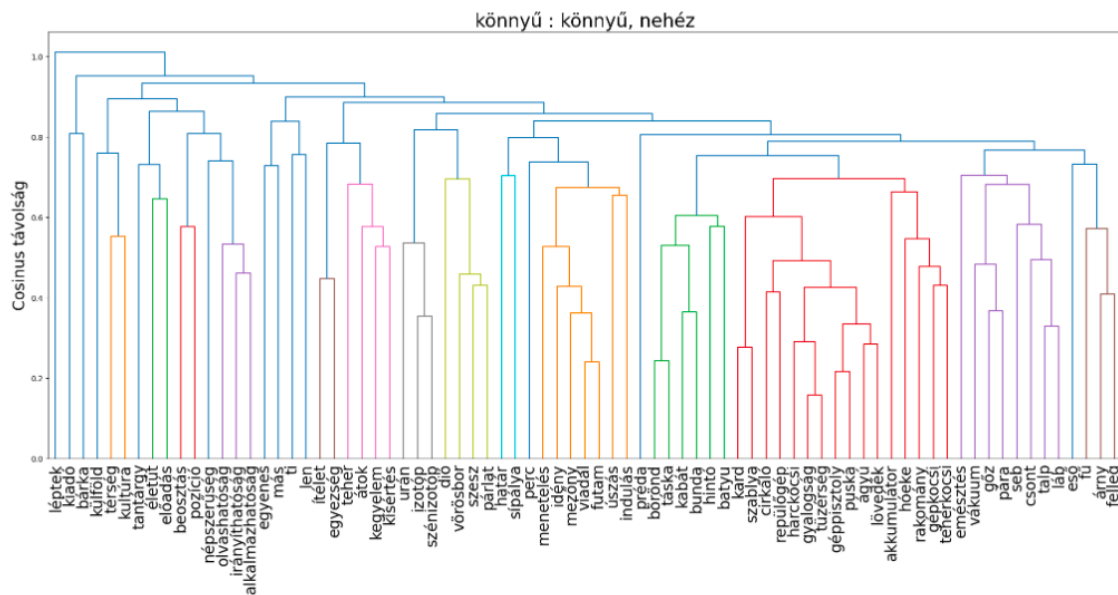


Figure 10: The dendrogram of the adjective *easy* with its antonym *difficult/heavy* and their salient nominal contexts¹⁴

¹² *bizarr*: 'bizarre', *szokatlan*: 'unusual', *fura*: 'strange', *furcsa*: 'peculiar', *ijesztő*: 'scary', *rémisztő*: 'frightening', *háttorzongató*: 'spine-chilling', *szürreális*: 'surreal', *abszurd*: 'absurd', *groteszk*: 'grotesque', *morbíd*: 'morbid'

¹³ *megdöbbentő*: 'shocking', *mellbevágó*: 'striking', *ijesztő*: 'scary', *meglepő*: 'surprising'

¹⁴ *könnyű*: 'light', *nehéz*: 'heavy', *léptek*: 'steps', *kiadó*: 'publisher', *bárka*: 'boat', *külföld*: 'foreign country', *tértség*: 'region', *kultúra*: 'culture', *tantárgy*: 'subject', *életút*: 'life path', *előadás*: 'lecture', *beosztás*: 'schedule', *pozíció*: 'position', *népszerűség*: 'popularity', *olvashatóság*: 'readability', *irányíthatóság*: 'controllability', *alkalmazhatóság*: 'applicability', *egyenes*: 'straight', *más*: 'other', *ti*: 'you', *len*: 'be', *ítélet*: 'judgment', *egyezség*: 'agreement', *teher*: 'load', *átok*: 'curse', *kegyelem*: 'grace', *kísértés*: 'temptation', *urán*: 'uranium', *izotóp*: 'isotope', *szénizotóp*: 'carbon isotope', *dió*: 'walnut', *vörösbor*: 'red wine', *szesz*: 'liquor', *párlat*: 'spirit', *határ*: 'border', *sípálya*: 'ski slope'

Comparing manually the automatically induced results with the microstructures in EDHL revealed that without adequate context, it can often be challenging to determine the appropriate placement of the adjective-noun construction within the microstructure. While this doesn't necessarily imply overlapping sense distinctions in EDHL's microstructure, this potentiality should be considered in future evaluations.

In line with this, we encountered several issues during the disambiguation of the attributive adjectives *fontos* 'important' and *jelentős* 'significant' in EDHL due to strongly overlapping definitions in the microstructures. The correct interpretation of *fontos* 'important' was particularly problematic when the modified nouns were one of the following: *munkatárs* 'colleague', *tisztség* 'position', *bizottság* 'committee', *jogintézmény* 'legal institution', *rendelet* 'order', *törvénytervezet* 'legislative proposal', *egyezmény* 'convention', *szerződés* 'contract', etc. Indeed, the following two senses of *fontos* appear to overlap:

1. <jelentőségénél fogva különös gondot, figyelmet érdemlő, jelentős, lényeges>¹⁵
2. <vmely cél elérésében, ill. a gyakorlati élet vmely területén jelentős szerepet betöltő, alig nélkülözhető>¹⁶

Determining whether *jelentős diadal* 'significant triumph', *jelentős térnyerés* 'significant expansion', *jelentős fölény* 'significant advantage', *jelentős emberveszteség* 'significant loss of life', *jelentős jövedelemforrás* 'significant source of income', *jelentős kiegészítés* 'significant supplement', *jelentős ismeret* 'significant knowledge' can be subsumed under multiple senses in EDHL, such as <'very important, of great significance'>, <'above average, considerable, significant, noteworthy'>, or <'playing an important role; significant, influential as a result of its effects or outcomes'> is also challenging. The overlapping senses suggest that providing more textual context would probably be insufficient to enable the lexicographer to find the correct meaning in this case.

Regarding the fifth hypothesis, the comparison of EDHL and the automatically retrieved semantically related adjectives, extracted via the connected graph components, was rather telling. For example, the graph-based algorithm cataloged 90 adjectives referring to quantities from the training corpus, of which only 8 are listed in EDHL (*ujjnyi* 'one/two inch' or 'a finger-sized', *arasznyi* '5-6 inches', *körömnnyi* 'nail-sized', *késhegyny* 'knife edge-sized', *tenyérynnyi* 'palm-sized', *mázsás* 'two hundred pounds heavy', *mérföldes* 'mile-long', *púpozott* 'rounded' as in a 'rounded tablespoon of sg.'). Regrettably, important adjectives are missing from the headword list: the corpus data clearly indicate that *gyűszűnyi* 'thimble-sized' and *ökölnyi* 'fist-sized' should form headwords on their own, but they are only included in the microstructure of the corresponding nominal headwords (e.g., *gyűszű* 'thimble' and *ököl* 'fist').

, *perc*: 'minute', *menetelés*: 'marching', *idény*: 'season', *mezőny*: 'field', *viadal*: 'tournament', *futam*: 'race', *úszás*: 'swimming', *indulás*: 'departure', *préda*: 'prey', *bőrönd*: 'suitcase', *táska*: 'bag', *kabát*: 'coat', *bunda*: 'fur coat', *hintó*: 'carriage', *batyu*: 'sack', *kard*: 'sword', *szablya*: 'saber', *cirkáló*: 'cruiser', *repülőgép*: 'airplane', *harcokosi*: 'tank', *gyalogság*: 'infantry', *tüzérség*: 'artillery', *géppisztoly*: 'machine gun', *puska*: 'rifle', *ágyú*: 'cannon', *lövedék*: 'bullet', *akkumulátor*: 'battery', *hóeke*: 'snowplow', *rakomány*: 'cargo', *gépkocsi*: 'car', *teherkocsi*: 'truck', *emésztés*: 'digestion', *vákuum*: 'vacuum', *gőz*: 'steam', *pára*: 'vapor', *seb*: 'wound', *csont*: 'bone', *talp*: 'sole', *láb*: 'foot', *eső*: 'rain', *fű*: 'grass', *árny*: 'shadow', *felle*: 'cloud'.

¹⁵ 'By virtue of its significance, it deserves special care, attention, and is significant and essential.'

¹⁶ 'Playing a significant role in achieving a particular goal or in a certain area of practical life; being scarcely dispensable.'

Apart from the insufficient coverage of certain semantic fields, additional inconsistencies emerged during the random testing of certain headwords. For instance, both *kisbirtokos* 'smallholder' and *középbirtokos* 'medium-sized landowner' (lit. mediumholder) appeared in the headword list in their adjectival forms. However, the adjectival form *nagybirtokos* 'large landowner' (lit. largeholder) was absent: only the nominal form was cataloged as a headword: '<Feudális v. kapitalista rendszerben> nagybirtokkal rendelkező, nagybirtokán mezőgazdasági (és állattenyésztési) munkát végeztető és dolgozóit kizsákmányoló személy.|| a. jelzői használat(ban) Ilyen személyekből álló <csoport>. Nagybirtokos arisztokrácia, család, kaszt.'¹⁷. Another inconsistency is the absence of the adjective *kisméretű* 'small-sized', which should be a headword given that *nagyméretű* 'great-sized' is part of the headword list, and that *kisméretű* is used rather frequently in the definitions of other headwords.

7. Future work

An unsupervised graph-based methodology is described in this paper. The aim is to support the work of expert lexicographers in compiling the macro- and microstructure of a monolingual explanatory dictionary for Hungarian. Although the proposed framework seems promising, there are multiple issues that need to be addressed to fully realize the method's potential.

Most importantly, the optimal value of the slicing parameter K should be set so that the automatically obtained results best suit the specific objectives of the lexicographers. Determining the optimal parameter setting requires robust collaboration between lexicographers and computational linguists for several reasons.

First, the selection principles of the adjectives are significantly determined by the purpose and target audience of the dictionary. Secondly, further lexicographic inspection is needed to set the K cut-off parameter, which depends not only on the network topology and weight distribution but also on the specific semantic classes of the adjectives.

Thirdly, the editing principles of the planned dictionary should be explicitly stated: those morphologically or semantically productive cases that, due to their productivity, should not form part of the dictionary, should be cataloged. As the randomly sampled lexicographic observations indicated, the described algorithm seems to be useful for these purposes as well. Various types of subgraphs may yield information both on the morphology-semantics interface and on the systematic subcategorization patterns of adjectives. Again, careful lexicographic work is indispensable to compile a comprehensive list of these attributes.

Finally, the prototype algorithm should be implemented as a software tool to enhance the efficiency of lexicographers' work.

8. References

Adamska-Sałaciak, A. (2006). *Meaning and the Bilingual Dictionary. The Case of English and Polish. (Polish Studies in English Language and Literature 18)*. Frankfurt am Main:

¹⁷ '<In the era of a feudal or a capitalist system> a person who owns a large estate, employs agricultural (and livestock) workers, and exploits them.|| a. (attributive use) A <group> formed by such persons. Large landowner aristocracy, family, caste.'

- Peter Lang.
- Ah-Pine, J. & Jacquet, G. (2009). Clique-Based Clustering for Improving Named Entity Recognition Systems. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*. Athens, Greece: Association for Computational Linguistics, pp. 51–59. URL <https://aclanthology.org/E09-1007>.
- Atkins, B.T.S. & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford University Press.
- Burnard, L. (ed.) (2007). *Reference Guide for the British National Corpus (XML Edition)*. URL: <http://www.natcorp.ox.ac.uk/XMLedition/URG/>.
- Bárczi, G. & Ország, L. (eds.) (1959–1962). EDHL = *A magyar nyelv értelmező szótára I–VII. [The Explanatory Dictionary of the Hungarian Language]*. Budapest: Akadémiai Kiadó.
- Camacho-Collados, J. & Pilehvar, M.T. (2018). From Word to Sense Embeddings: A Survey on Vector Representations of Meaning. URL <https://arxiv.org/abs/1805.04032>.
- Frege, G. (1892). Uber Sinn und Bedeutung. In M. Textor (ed.) *Funktion - Begriff - Bedeutung*, volume 4 of *Sammlung Philosophie*. Göttingen: Vandenhoeck & Ruprecht.
- Geeraerts, D. (2010). *Theories of lexical semantics*. Oxford University Press.
- Hanks, P. (2010). Compiling a Monolingual Dictionary for Native Speakers. *Lexikos*, 20, pp. 580–598.
- Hanks, P. (2012). The Corpus Revolution in Lexicography. *International Journal of Lexicography*, 25, pp. 398–436.
- Héja, E. & Ligeti-Nagy, N. (2022a). A Clique-based Graphical Approach to Detect Interpretable Adjectival Senses in Hungarian. In *Proceedings of TextGraphs-16: Graph-based Methods for Natural Language Processing*. Gyeongju, Republic of Korea: Association for Computational Linguistics, pp. 35–43. URL <https://aclanthology.org/2022.textgraphs-14>.
- Héja, E. & Ligeti-Nagy, N. (2022b). A proof-of-concept meaning discrimination experiment to compile a word-in-context dataset for adjectives – A graph-based distributional approach. *Acta Linguistica Academica*, 69(4), pp. 521 – 548. URL <https://akjournals.com/view/journals/2062/69/4/article-p521.xml>.
- Ittész, N., et al. (ed.) (2006–2021). *CDH = A magyar nyelv nagyszótára I–VIII. [Comprehensive Dictionary of Hungarian]*. Budapest: Nyelvtudományi Kutatóközpont.
- Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P. & Suchomel, V. (2013). The TenTen corpus family. In *7th international corpus linguistics conference CL*. Lancaster University, pp. 125–127.
- Juhász, J., Szőke, I., O. Nagy, G. & Kovalovszky, M. (eds.) (1972). ÉKSz¹ = *Magyar értelmező kéziszótár. [Concise Hungarian Explanatory Dictionary]*. Budapest: Akadémiai Kiadó.
- Kiefer, F. (2003). How much information do adjectives need in the lexicon? In *Igék, főnevek, melléknevek [Verbs, nouns, adjectives]*. Tinta Könyvkiadó, pp. 32–43.
- Kiefer, F. (2008). A melléknevek szótári ábrázolásáról. In *Strukturális magyar nyelvtan 4. A szótár szerkezete*. Akadémiai Kiadó, pp. 505–538.
- Kipper-Schuler, K. (2005). *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. Ph.D. thesis, University of Pennsylvania, Philadelphia.
- Kupietz, M., Belica, C., Keibel, H. & Witt, A. (2010). The German Reference Corpus DeReKo: A primordial sample for linguistic research. In N. Calzolari & et al. (eds.) *Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC 2010)*. Valletta, Malta: European Language Resources Association (ELRA), pp. 1848–1854.

- Levin, B. (1993). *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press.
- Lipp, V. & Simon, L. (2021). Towards a new monolingual Hungarian explanatory dictionary: overview of the Hungarian explanatory dictionaries. *Studia Lexicographica*, 15, pp. 83–96.
- Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013a). Efficient Estimation of Word Representations in Vector Space. URL <https://arxiv.org/abs/1301.3781>.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. & Dean, J. (2013b). Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems*, 26.
- Moon, R. (1987). The Analysis of Meaning. In J. Sinclair (ed.) *Looking Up*. pp. 86–103.
- Nemeskey, D.M. (2020). *Natural Language Processing Methods for Language Modeling*. Ph.D. thesis, Eötvös Loránd University.
- Oravecz, C., Váradi, T. & Sass, B. (2014). The Hungarian Gigaword Corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), pp. 1719–1723. URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/681_Paper.pdf.
- Országh, L. (1953). A magyar nyelv új szótáráról. [On the new dictionary of the Hungarian Language]. *Magyar Nyelvőr*, 77(5–6), pp. 387–407.
- Ploux, S. & Victorri, B. (1998). Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes. *Traitement automatique des langues*, 1(39), pp. 146–162.
- Pustejovsky, J. (1995). *The Generative Lexicon*. Cambridge, MA: MIT Press.
- Pusztai, F. & Csábi, S. (eds.) (2003). ÉKSz² = *Magyar értelmező kéziszótár*. [Concise Hungarian Explanatory Dictionary]. Budapest: Akadémiai Kiadó.
- Rehurek, R. & Sojka, P. (2011). Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2).
- Sass, B., Váradi, T., Pajzs, J. & Kiss, M. (2010). *Magyar igei szerkezetek. A leggyakoribb vonzatok és szókapcsolatok szótára*. Budapest: Tinta Könyvkiadó.
- Stock, P.F. (1984). Polysemy. In R.R.K. Hartman (ed.) *LEXeter '83: Proceedings. Papers from the International Conference on Lexicography at Exeter, 9–12 September 1983*. Tübingen: Max Niemeyer, pp. 131–140.
- Svensén, B. (2009). *A Handbook of Lexicography: The Theory and Practice of Dictionary-Making*. Cambridge: Cambridge University Press.
- Váradi, T. (2002). The Hungarian National Corpus. In *Proceedings of the Second International Conference on Language Resources and Evaluation, Las Palmas*. pp. 385–389.
- Véronis, J. (2003). Sense tagging: does it make sense? In A. Wilson, P. Rayson & T. McEnery (eds.) *Corpus Linguistics by the Lune: a festschrift for Geoffrey Leech*. Frankfurt: Peter Lang.
- Zinoviev, D. (2018). *Complex Network Analysis in Python: Recognize - Construct - Visualize - Analyze - Interpret*. The Pragmatic Bookshelf.

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

Magdalena Gapsa¹, Špela Arhar Holdt^{1,2}

¹ University of Ljubljana, Faculty of Arts, Aškerčeva cesta 2, 1000 Ljubljana, Slovenia

² University of Ljubljana, Faculty of Computer and Information Science, Večna pot 113, 1000 Ljubljana, Slovenia

E-mail: Magdalena.Gapsa@ff.uni-lj.si, Spela.ArharHoldt@fri.uni-lj.si

Abstract

User involvement can be a valuable asset in expediting the process of language resource development, given that a thoughtful methodology is implemented. A successful example is the Thesaurus of Modern Slovene, which incorporates user participation to improve its automatically generated content. To shed light on the otherwise invisible lexicographic decision-making processes and to develop editorial protocols based on the needs of dictionary users, we investigated how differently lexicographers evaluate user-suggested synonyms compared to other user groups. We conducted an evaluation of nearly 1,000 user-suggested synonyms, assessed by a total of 42 evaluators from 7 user groups, and tested four hypotheses about lexicographers as evaluators. After evaluation, the Inter-Annotator Agreement (IAA) in all groups was calculated using Krippendorff's alpha and entropy, the evaluators' comments were classified into bottom-up categories, and the data were statistically analysed. In accordance with our assumptions, the lexicographers provided the most detailed arguments and identified the highest number of potential shortcomings of the suggested synonyms. However, they also scored the second lowest IAA among all groups and were more opposed to discarding user suggestions. We discuss the possible reasons for these results and emphasise their value for the further development of responsive dictionaries.

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

1. Introduction

The Thesaurus of Modern Slovene is a state-of-the-art example of a digitally-born dictionary created automatically from pre-existing openly available language resources (Krek et al., 2017).¹ It was prepared to address the lack of openly available synonym data for modern Slovene, and it serves as a benchmark for data reusability and user involvement for other languages facing similar issues. The development of the Thesaurus is based on a responsive dictionary model (Arhar Holdt et al., 2018), where the initial version of the resource is generated automatically and made available to the public under an open licence as soon as it is deemed useful. The data is then gradually revised, with the help of users, to ensure ongoing improvement. This iterative process is vital due to the presence of noise and the absence of certain types of essential lexical

¹ Thesaurus of Modern Slovene is available in the interface at <https://viri.cjvt.si/sopomenke/eng/> and as a database at <http://hdl.handle.net/11356/1166>.

information in the automatically generated database.²

In the Thesaurus, the users are allowed to suggest new synonym candidates and evaluate existing ones. The possibilities for user participation, as well as many other novelties introduced by the responsive dictionary model, were positively rated and well accepted by the user community (Arhar Holdt, 2020). In practice, allowing the option of suggesting new synonyms has proven especially fruitful, as the number of collected synonym candidates is high: 60,976 at the time of writing. To ease participation, user suggestions are displayed in the dictionary interface immediately and without editorial intervention. However, a lexicographic review and approval process is required before suggestions can be included in the openly accessible dictionary database.

Although a preliminary study by Arhar Holdt and Čibej (2020) suggested that a very limited number of user inputs were malicious, there is currently no large-scale study on the content and relevance of user-suggested data. Conducting such a study would enable an assessment of the quality of user contributions and identification of potential problems that could be addressed to enhance user participation. To address this gap, we carried out an evaluation campaign utilising almost 1,000 user-suggested synonyms from the Thesaurus of Modern Slovene. A total of 42 evaluators, chosen based on their profession or interests, participated in the study.³ In Gapsa (2023), a summative analysis of the results was presented, while this paper focuses specifically on how lexicographers evaluated user-suggested synonyms in comparison to other user groups, such as language editors, translators, and teachers.

2. Related work

The present study belongs to the field of lexicographic user research and builds upon established methodological frameworks (a comprehensive overview of existing methodologies is provided in Welker, 2013a, 2013b). Lexicographic user research emphasises the importance of user-centred design in the development and evaluation of lexicographic products. It has a tradition reaching back to the 1960s (e.g. Barnhart, 1962; Householder, 1967), but the research area was firmly established later in the 1980s and 1990s (e.g. Tomaszczyk, 1979; Hartman, 1987; Atkins, 1998; Nesi, 2000). The emergence of the digital medium in the 2000s offered a vast array of new methodological possibilities (e.g. Bergenholtz and Johnsen, 2013; Müller-Spitzer, 2014; Lew and De Schryver, 2014). In the last decade, existing approaches were also critically evaluated and surpassed (Bogaards, 2003; Tarp, 2009; Lew, 2015; Kosem et al., 2018):

² The data published in Thesaurus 1.0 was not lexicographically post-processed. The entries and synonym candidates were presented in a form of lemmata (without part-of-speech or other metadata that would help disambiguate between forms), semantic descriptions were replaced by automatically obtained semantic clusters, and the data also lacked dictionary labels, apart from domain ones. Version 2.0, currently undergoing testing, aims to address some of these issues, as outlined by Arhar Holdt et al. (In press).

³ The gathered data are available in the Repository of the University of Ljubljana: <http://hdl.handle.net/20.500.12556/RUL-144064>

older studies have most often been criticised for having too few participants or for being too homogeneous (students were the most likely group to participate, as they are the easiest for researchers to access).

In our case, the participants in the study represent dictionary users, while at the same time serve as evaluators of user-suggested synonyms. Previous studies, mainly from the field of NLP, have shown that non-experts are capable of successfully performing tasks of assessing synonymy or word similarity. Crowdsourced evaluations of synonyms have been applied in various contexts, such as evaluating the degree of similarity between words (Schnabel et al., 2015) and creating gold standards for evaluation and training tasks (e.g. Hill et al., 2015; Schneidermann et al., 2020). Human annotations of similarity have been used as evaluation methods in Word-in-Context and SemEval tasks (e.g. Pilehvar and Camacho-Collados, 2019; Breit et al., 2021; Armendariz et al., 2020), and crowdsourcing-oriented tools have been developed for different wordnets to detect and correct errors (e.g. Braslavski et al., 2014; Fišer et al., 2014; Rambousek et al., 2018).

3. Methodology

3.1 Preparation of the dataset

Similar to intrinsic evaluations in NLP tasks (see e.g. Schnabel et al., 2015 and Schneidermann et al., 2020), where pre-selected inventories of word pairs are used, we used a list of 546 Slovene nouns occurring as headwords (or headword-like units) in various openly available language resources for modern Slovene: the Thesaurus of Modern Slovene 1.0 database (Krek et al., 2018), the sloWNet 3.1 database (Fišer, 2015), the Lexical Database for Slovene (Gantar et al., 2013), the Comprehensive Slovenian-Hungarian Dictionary (Kosem et al., 2021), and the database of nouns labelled with semantic types (Kosem and Pori, 2021).⁴ We then extracted user-suggested synonyms for these nouns from the Thesaurus of Modern Slovene 1.0 interface using a custom made script, prepared specifically to track user contributions. The number of suggestions varied for each noun, and not all nouns had suggestions. In total, we extracted 972 synonyms for 307 nouns.

3.2 Selection of user groups

We selected the desired user groups based on the typology of potential dictionary users by Arhar Holdt et al. (2016, pp. 181-184) and the results of a study on user attitudes towards the lexicographic novelties introduced by the Thesaurus (Arhar Holdt, 2020, p. 477). On the one hand, the typology provided a theoretical overview of the user

⁴ This work is part of a larger study in a PhD research project aiming to improve the connectivity and reusability of Slovene synonym data in the digital environment. Certain decisions, e.g. the selection of headwords for the evaluation data, were made with other research objectives in mind.

groups according to the main situations of dictionary use (in the educational process, for professional purposes or for leisure activities). On the other hand, the user study indicated which user groups were most represented among the participating active users of the Thesaurus.

Combining both pieces of information as well as our research questions, we have selected 7 user groups, as presented in Figure 1: Lexicographers (L), Translators (T), Language Editors (LEd), Marketers (M), Teachers of Slovene (ToS), Language Enthusiasts (LEn), and Students (S) of linguistic studies. Our aim was to cover all three scenarios of dictionary usage. We included lexicographers in the study due to their critical role in the editorial process of evaluating synonyms. In addition to representing the educational aspect of the study, we also included students to pilot the research before its wider implementation (Gapsa, 2022).

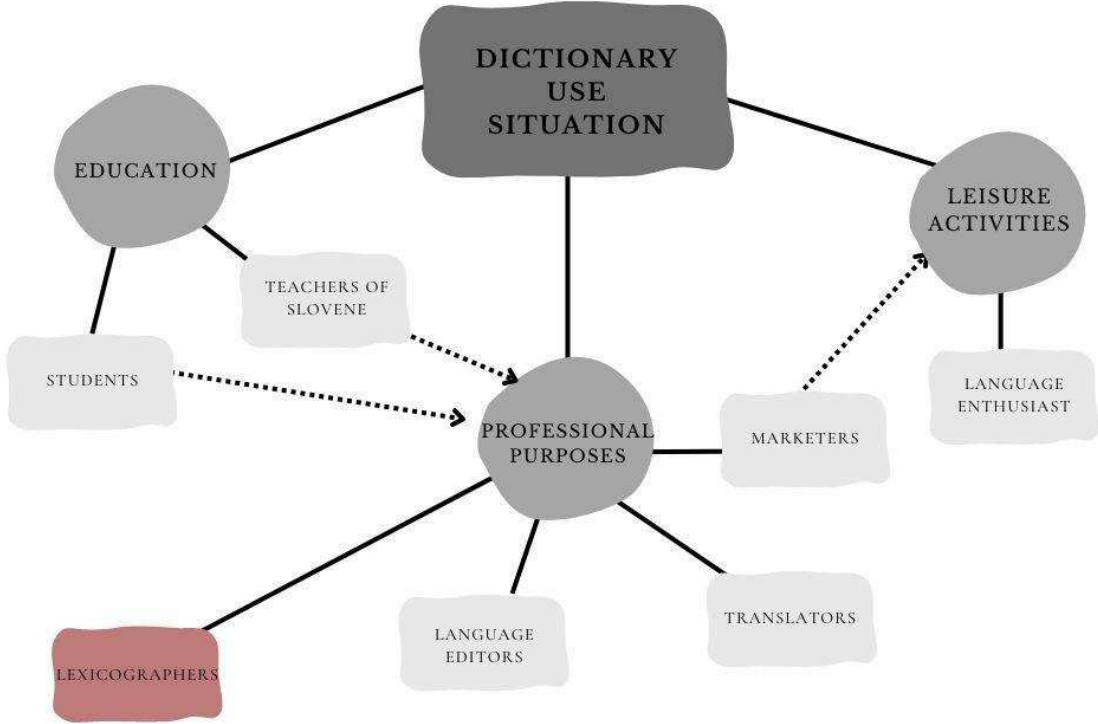


Figure 1: Overview of the selected user groups based on three main dictionary use situations

3.3 Recruiting participants

Considering the cautionary notes against qualitative user studies with a too limited number of participants (Tarp, 2009, 290), and taking into account the resources available for our study, we opted to include six evaluators per group, for a total of 42

evaluators.

The first groups recruited were Students, who had at the time participated in the development of the Thesaurus from 1.0 to 2.0.⁵ They already knew the Thesaurus and had experience in analysing linguistic data and could help test the evaluation process, tools and guidelines, as well as estimate the time needed for the task and set a financial compensation for the participants. Secondly, the group of Lexicographers was assembled under the umbrella of the same project. Recruitment of representatives from other user groups took place in several rounds. The call for applications for Teachers of Slovene, Translators and Language Editors was published, first via the CJVT newsletter and then via the CJVT Facebook profile. A call for applications for Language Enthusiasts, which was also answered by Marketers, was posted in two Facebook groups, which serve as a forum for asking and answering language-related questions: ‘For at least approximately correct use of the Slovene language’ and ‘Association of Amateur Orthographers AND Grammarians’.⁶ The call briefly presented the task and the conditions of participation, including the payment.

3.4 Data evaluation

The participant data was prepared in separate Google Sheets spreadsheets,⁷ where we listed all 972 user-suggested synonyms and their corresponding headwords. Each participant was asked to evaluate whether the words in each pair were synonyms or not by answering the question, “Are the words in the pair synonyms?” for all 972 pairs. Table 1 presents the four possible answers and their suggested uses. In cases where participants answered “CONDITIONAL YES,” it was mandatory for them to explain the specific issues they identified. While comments were encouraged for the other three answer options, they were not mandatory.

Answer	When to use
YES	If you believe that the words in the pair are synonyms.
NO	If you believe that the words in the pair are not synonyms or in the case of obvious errors, typos, etc.

⁵ Project Synonyms and Collocations 2.0 – SoKol, Upgrading fundamental dictionary resources and databases of CJVT UL.

⁶ In Slovene: Za vsaj približno pravilno rabo slovenščine and Društvo ljubiteljskih pravopisarjev IN slovníčarjev.

⁷ Google Sheets was used due to its accessibility, popularity, cost-effectiveness and option for continuous editing and saving of the answers.

CONDITIONAL YES	If you believe that the words in the pair can be synonyms, but at the same time you see limitations or have doubts, e.g. because the words are synonyms only in a certain meaning or context, one or both words are marked, etc.
--------------------	--

NOT SURE/DON'T KNOW	If you are not sure whether the words are synonyms, you do not know one or both of the words in the pair or the meaning of one or both of the words in the pair, or you have difficulty deciding.
---------------------------	---

Table 1: Overview of possible answers in the guidelines for evaluators

The objective was to test the evaluators' understanding of relevant synonymous data. The guidelines provided to participants were intentionally general, without defining synonymy or providing examples of potential synonym pairs (as opposed to e.g. Hill et al., 2015, where a brief definition of similarity was provided together with examples of similar word pairs to better illustrate the difference between similarity and association or relatedness) or suggesting where borderline cases should be classified to avoid influencing the participants' answers. Similarly to Hill et al., 2015, we wanted the participants to rely on their language intuition (thus we discouraged them from consulting other language resources like dictionaries, corpora, etc.) and presented them with context-free word pairs (which is also an experience users get when browsing Thesaurus 1.0, as synonym candidates are listed without sense disambiguation or examples of use).

To ensure quality control of the evaluation process, participants also completed a brief questionnaire using the online survey tool 1ka.⁸ The questionnaire was designed to collect background information about the evaluators and confirm their placement in the designated user groups. It also enabled participants to provide feedback about potential problems with the evaluation process.

3.5 Research Hypotheses

For this study, we tested 4 hypotheses about Lexicographers as an evaluator group:⁹

⁸ Online survey tool 1ka: <https://www.1ka.si/d/en>

⁹ The formulation of the four hypotheses was driven by the aim of ensuring quality control in the user participation aspect of the dictionary-making process. In our workflows, lexicographers, who serve as the editors of the dictionary and possess first-hand experience in organising synonyms in the Thesaurus, undertake the evaluation of user contributions on behalf of the participating community. For this study, it is crucial to establish the lexicographers' evaluations as a gold standard and explore the divergence of their decisions from those made by other participating groups. Consequently, we are also testing hypotheses that may appear obvious or counterintuitive from this particular standpoint.

- H1: Lexicographers’ evaluation would be more consistent and their Inter-Annotator Agreement would be higher than in other groups.
- H2: Lexicographers would argue their decisions in more detail than other groups.
- H3: Lexicographers would make statistically different decisions about (un)acceptability of user-suggested synonyms and identify more potential problems with user-suggested synonyms than other groups.
- H4: Lexicographers would be more reserved to include user-suggested synonyms than other groups.

3.6 Data Analysis

To address the hypotheses, different approaches were used.

Firstly, Inter-Annotator Agreement (IAA) between the evaluators was calculated using Krippendorff’s alpha (Krippendorff, 1970).¹⁰ Calculations were made for each of the synonym pairs within each user group to facilitate clustering of IAA levels (as opposed to manually identifying all possible IAA levels) and to make the data more comparable between groups. The total number of answers received was 40,801, as a total of 23 answers were missing. Since the possible answers were nominal categories and not a scale, entropy¹¹ was calculated to determine the distribution of possible answers.

Secondly, evaluators’ comments were manually categorised according to their content. The categories were created bottom-up, based on the material analysed. The final list of categories comprised 11 possible categories, some of which allowed for further sub-categories, notably the category Other. Multi-layered categorisation was used because some of the comments, although coming from a single commentator, contained multiple pieces of information that could be classified into different categories, e. g. “dialectal and calque”¹², “a stylistic label would be needed, in one of the meanings”, etc. The categories and their definitions, as well as selected examples of categorised comments, can be found in Table 2.

¹⁰ The IAA is usually calculated using Fleiss’ Kappa (see M. Vila et al. 2015, p. 85), however, Krippendorff’s alpha (Krippendorff, 1970) was used here because of rare cases of missing answers.

¹¹ Both calculations are very sensitive to the subtlest differences in answers, therefore both were used as a filtering tool to facilitate the analysis and comparison of the results.

¹² Translations are approximate and may not cover all specifics. Slovene headwords and suggestions are provided with English translations. Evaluators’ comments are presented in English. Translations aim to aid understanding and fluency of reading.

Category name	Definition	Example of evaluators' comments	Synonym pair
limited context or certain sense(s) of the word(s)	context or certain sense; limited usage; other senses or a need for sense disambiguation	Synonyms only in one meaning.	žoga – podaja ('a ball' - 'a pass')
insufficient sense	additional qualifiers seem to be a necessary component of the meaning	A piece of fabric intended for cleaning can be a cloth, let's say.	blago – krpa ('a fabric' - 'a cloth')
semantic discrepancy	semantically related but not necessarily always interchangeable words ; related but different concepts	The customer is not necessarily the subscriber. It can be a random customer or just a visitor to the shop/store etc.	stranka – naročnik ('a client' - 'a subscriber')
alternative semantic relation	other semantic relationship (e.g. hyper-/hyponymy, meronymy/holonymy)	The suggested synonym is a hypernym of the headword.	hotel – prenočišče ('a hotel' - 'an accommodation')
unknown word or sense	unfamiliarity with word or suggested sense	I do not know the second word.	izseljenec – ezul ('an emigrant' - 'an exile')
definition	explanation, definition or description	The suggested synonym sounds more like a definition to me.	anatomija – veda o telesni zgradbi ('an anatomy' - 'science of body structure')
incomplete word units	multi-word expressions suggested as single words	In the form of <i>imeti pogum</i> - <i>imeti jajca</i> .	pogum – jajca ('a courage' - 'balls')
opinionizing	evaluators opinion on the	Perhaps a little	elita – creme de la crème

	suggested synonym	bit too French.	('an elite' - 'crème de la crème')
foreign words	loanword, foreignism, calque or non-standard loanword	Merely as a literal translation of a foreign word from Latin.	aplikacija – namestitev ('an application' - 'an installation')
marked	marked word or a qualifier or tag needed, sometimes very specific, e.g. dialectal, pejorative	Colloquially.	cigareta – dim ('a cigarette' - 'a smoke')
other	remarks that do not fall into the above categories	Consider singular-plural.	pošta – maili ('a post' - 'mails')

Table 2: Comment categories with definitions and examples of use

It was also possible to identify certain problems that occurred with the user-suggested synonyms, but which were not frequent enough to be included in a separate category. Such comments were subcategorised within main categories. This was particularly the case with e.g. phraseological units or metaphorical senses, which created subcategories within main category *Limited context or certain sense(s) of the word(s)*, cases of meronymy, which were put under main category *Alternative semantic relation* or specific semantic labels that were mentioned with comments regarding a headword or user-suggested synonym being *Marked*.

Thirdly, to determine possible dependencies between the user groups and their most frequent comments, statistical tests were carried out, i.e. contingency tables were prepared and a chi-square test was run, followed by calculations of Pearson residuals to determine whether there were statistically significant differences between the groups. Pearson residuals below -1.92 or above 1.92 indicate a statistically significant difference. In the following chapter, we present the results of the study.

4. Results

4.1 Consistency and Inter-Annotator Agreement

Our first hypothesis was that Lexicographers would be the group with the highest IAA of all groups, which would indicate that their answers are more inherently consistent than those of the other groups. The hypothesis is based on the presumption that lexicographers evaluate user-suggested synonyms on the basis of common and

comparable expert knowledge and experience, which would facilitate higher consistency.

To test the hypothesis, we compared: “full IAA”, where all evaluators within the group chose the same answer; “very high IAA”, where 5 out of 6 evaluators chose the same answer; “high IAA”, where 4 out of 6 evaluators chose the same answer; and “moderate IAA”, where 3 out of 6 evaluators chose the same answer. Here, we distinguished “tied answers”: the instances where 3 evaluators agreed on one answer and the remaining 3 evaluators agreed on another answer. Figures in Table 3 show that, on average, evaluators scored at least *high IAA* on almost 60% of the whole evaluation set and *moderate IAA* on 33% of the set.

User group	Full IAA	Very high IAA	High IAA	TOTAL at least high IAA	Moderate IAA	Tied answers
Lexicographers	28 (3 %)	136 (14 %)	341 (35 %)	505 (52 %)	395 (41 %)	130 (13 %)
Language Editors	139 (14 %)	222 (23 %)	286 (29 %)	647 (67 %)	271 (28 %)	58 (6 %)
Language Enthusiasts	52 (5 %)	149 (15 %)	336 (35 %)	537 (55 %)	359 (37 %)	109 (11 %)
Marketers	188 (19 %)	256 (26 %)	272 (28 %)	716 (74 %)	219 (23 %)	59 (6 %)
Translators	46 (5 %)	195 (20 %)	300 (31 %)	541 (56 %)	349 (36 %)	32 (3 %)
Students	34 (3 %)	140 (14 %)	263 (27 %)	437 (45 %)	396 (41 %)	72 (7 %)
Teachers of Slovene	165 (17 %)	209 (22 %)	285 (29 %)	658 (68 %)	255 (26 %)	55 (6 %)

AVERAGE	93 (10 %)	187 (19 %)	298 (31 %)	577 (59 %)	321 (33 %)	74 (8 %)
----------------	---------------------	----------------------	----------------------	----------------------	----------------------	--------------------

Table 3: Distribution of number of pairs with at least high IAA between groups

Lexicographers achieved the second lowest *at least high IAA* among all groups (the only group that scored lower were Students, see Discussion). Their *full* and *very high IAA* was the lowest of all the evaluator groups, at only 3% and 14% respectively (again, a similar percentage was achieved by the Student group). On the other hand, their *high IAA* (35%) was the highest of all groups, followed by Language Enthusiasts. Lexicographers also scored the second highest number of pairs with *moderate IAA*, closely after Students. Finally, they scored the highest number of pairs with tied answers. These results reject the first hypothesis: data shows that Lexicographers were below average in terms of IAA, their answers within the group were less consistent and most often tied in comparison to other groups.

4.2 Detailed argumentation of the decisions

The second hypothesis assumed that the Lexicographers would give a more detailed argumentation of their decisions indicating that they were better informed about the potential problems of the data than other evaluator groups. To test this assumption, we compared the number of comments made and categorised between the different groups and the number of categorised comments for each category within the groups. The numbers are shown in Table 4.

User group	L	LEd	LEn	M	T	S	ToS	TOTAL	AVG.
Comments made	<u>2,717</u>	363	783	640	1,234	2,593	252	8,582	1,226
Comments categorised¹³	1,802	388	708	609	1,249	<u>1,845</u>	246	6,846	978

¹³ Repeating comments were deduplicated – if multiple evaluators in the same group made comments that fell into the same category, it was only counted once.

limited context or certain sense(s) of the word(s)	<u>625</u>	51	121	65	166	435	18	1,481	212
insufficient sense	5	28	40	31	<u>89</u>	60	35	288	41
semantic discrepancy	36	56	57	92	<u>200</u>	188	35	664	95
alternative semantic relation	75	44	35	28	80	<u>190</u>	19	471	67
unknown word or sense	247	53	115	194	166	<u>276</u>	83	1,134	162
definition	<u>93</u>	0	17	1	22	65	0	198	28
incomplete word units	23	1	9	9	<u>58</u>	17	5	122	17
opinionizing	6	17	9	11	11	<u>22</u>	1	77	11
foreign words	0	19	15	<u>43</u>	36	22	0	135	19
marked	425	92	247	84	279	<u>426</u>	27	1580	226
other	<u>267</u>	27	43	51	142	144	23	697	100

Table 4: Number of comments made and categorised per user group and the distribution of the comment categories among the user groups. The abbreviations are: L – Lexicographers, LEd – Language Editors, LEn – Language Enthusiasts, M – Marketers, T – Translators, S – Students, ToS – Teachers of Slovene, AVG. – average

As the figures in Table 4 show, the Lexicographers indeed made the highest number of comments of all the evaluator groups. When comparing the number of categorised comments, Lexicographers scored second highest. The group that behaved most similarly to Lexicographers were again Students.

As mentioned in Section 3.6, some categories were further divided, particularly the category *Other*. Not only did Lexicographers contribute the most comments to this category, their comments also generated most subcategories: about 30 subcategories

compared to 10-15 subcategories¹⁴ in the other evaluator groups. The subcategories most frequently observed in the Lexicographers group were:

- coined synonyms - the comments indicated that this vocabulary is probably characteristic of the suggester's idiolect, and therefore hardly understood or used by the wider community, e.g. *klitoris* 'a clitoris' – *gumbek* 'a button', *menstruacija* 'a menstruation' – *rdeča armada* 'red army',
- terminological correctness - the comments indicated that it needed to be checked whether the suggested synonym can be used in a terminological sense of the headword, e.g. *epidemija* 'an epidemic' – *pandemija* 'a pandemic', *mandarina* 'a mandarine' – *klementina* 'a clementine',
- collocations - the comments indicated that the suggested synonym might be collocative or part of a collocation, e.g. *avtoriteta* 'an authority' – *spoštovan strokovnjak* 'a respected professional', *babica* 'a granny' – *starejša gospa* 'an elderly lady',
- alternative spellings - the comments indicated that a word has no standard written form or that different spellings are possible, e.g. *bonbon* – *bombon* 'a candy', *parfum* – *parfem* 'a perfume',
- doubts on actual use - the comments indicated that it needed to be checked whether the user-suggested synonym is confirmed in modern language, e.g. *alkohol* 'alcohol' – *veselje* 'a joy', *ogrlica* 'a necklace' – *kolje* 'a necklace', *smrad* 'a stench' – *zaudarek* 'a reek',
- doubts on the frequency of use - the comments indicated that it needed to be checked whether the user-suggested synonym is frequent enough in the modern language, e.g. *avtoriteta* 'an authority' – *veščak* 'an expert', *izseljenec* 'an emigrant' – *ezul* 'an exile'.

Overall, Lexicographers made more comments in total and those categorised as *Other* than other groups. Moreover, their comments revealed more subcategories, especially within the category *Other*. These subcategories reflect issues identified and commented on more often or typically by Lexicographers. Both facts support the hypothesis that Lexicographers would give more detailed and informed argumentation of their answers and decisions.

4.3 Focus on different problems

The third hypothesis assumed that Lexicographers' decisions about (un)acceptability of users suggestions would be statistically different from decisions of other groups, as

¹⁴ Except for Students, whose comments contained ample explanations that could be sorted into nearly 30 subcategories.

lexicographers are likely to identify different potential problems than other evaluator groups. To test this assumption, contingency tables were prepared and a chi-square test of independence was performed to finally calculate the Pearson’s residuals. The calculations of the Pearson’s residuals are shown in Table 5.

Category	L	LEd	LEn	M	T	S	ToS
limited context or certain sense(s) of the word(s)	11,915	-3,594	-2,597	-5,814	-6,337	1,798	-4,827
insufficient sense	-8,132	2,891	1,873	1,064	5,031	-1,998	7,664
semantic discrepancy	-10,496	2,995	-1,407	4,286	7,167	0,679	2,282
alternative semantic relation	-4,397	3,351	-1,964	-2,146	-0,638	5,600	0,505
unknown word or sense	-2,978	-1,405	-0,209	9,274	-2,841	-1,692	6,620
definition	5,664	-3,350	-0,768	-3,958	-2,349	1,594	-2,667
incomplete word units	-1,607	-2,249	-1,018	-0,562	7,577	-2,769	0,295
opinionizing	-3,169	6,050	0,368	1,586	-0,813	0,275	-1,062
foreign words	-5,961	4,104	0,279	8,944	2,292	-2,384	-2,202
marked	0,450	0,261	6,542	-4,769	-0,543	0,012	-3,951
other	6,170	-1,988	-3,424	-1,396	1,318	-3,197	-0,408

Table 5: Pearson residuals of the distribution of the comment categories among the user groups. The abbreviations are: L – Lexicographers, LEd – Language Editors, LEn – Language Enthusiasts, M – Marketers, T – Translators, S – Students, ToS – Teachers of Slovene

The group of Lexicographers was the one that most frequently commented on the need for sense disambiguation, while other groups were less concerned about it. Secondly, different evaluator groups frequently commented that the suggestion lacked an essential sense component to be considered synonymous while Lexicographers rarely made such comments. Thirdly, Lexicographers rarely commented on semantic discrepancies between the headword and the user-suggested synonyms, while other groups reported such cases quite frequently. Furthermore, they also reported cases of alternative semantic relations less frequently than other groups. The data also show that Lexicographers were less likely to report cases of unknown word(s) or meaning(s). On the other hand, they were more likely than other groups to comment that the suggestion is a “definition” or “description” rather than a synonym. There were no significant differences between Lexicographers and other evaluators in reporting cases of incomplete word units.

The data presented in Table 5 also clearly show that Lexicographers were less inclined to comment on the foreign origin of word(s), while other groups (with the exception of the Teachers of Slovene) emphasised this relatively frequently. They were also somewhat less likely than other groups to provide comments that had no other value but to express opinions. Marked vocabulary was commented on by the Lexicographers at approximately the same rate as within other groups. They did, as already mentioned, contribute more comments that were categorised as *Other* than the remaining groups.

If we summarise the above results and the data from the previous section, we can conclude that the third hypothesis is true. Lexicographers did indeed focus on other issues. Possible explanations for these findings are addressed in the Discussion.

4.4 Rigour and reserve in incorporating user suggestions

The fourth hypothesis assumed that Lexicographers are more rigorous in their decisions and more reserved to accept user suggestions and consequentially include them in the Thesaurus database. To test this assumption, we compared the total number of NO and CONDITIONAL YES answers within each evaluator group and the distribution of answers chosen by the evaluators in the *full*, *very high* and *high IAA* cases. Table 6 shows the total number of answers given by each group. The highest values for each answer are underlined and in bold.

User group	TOTAL given answers ¹⁵	YES	NO	CONDITIONAL YES	NOT SURE/DON'T KNOW
Lexicographers	5,829	2,720	492	<u>1,956</u>	661
Language Editors	5,823	3,009	1,908	467	439
Language Enthusiasts	5,828	2,916	<u>1,924</u>	611	377
Marketers	<u>5,832</u>	<u>3,590</u>	1,404	300	538
Translators	5,831	2,614	1,687	742	788
Students	5,831	1,797	1,187	1,940	<u>907</u>
Teachers of Slovene	5,827	3,383	1,556	407	481
AVERAGE	5,829	2,861	1,451	918	599

Table 6: Total number of answers given per evaluators group.

As the figures in Table 6 show, Lexicographers gave the answer CONDITIONAL YES more frequently than other evaluators groups. Students achieved an almost identical total number of CONDITIONAL YES answers, while other evaluators gave this answer much less frequently. The total number of CONDITIONAL YES answers supports the assumption that Lexicographers would be more cautious and reserved to include user-suggested synonyms as they were suggested. However, the total number of NO answers proves that the assumption that Lexicographers would reject more data was wrong, as

¹⁵ Occasional missing answers were noted, therefore the numbers given in column 2 vary between groups and rarely equals the total number of possible answers in a group (6 evaluators x 972 pairs = 5,832 possible answers).

Lexicographers gave the NO answer significantly less often than other groups.

Similar results can be observed when looking at the distribution of answers in pairs with *at least high IAA*, which is shown by Figure 2. It shows the summarised number of pairs with each of the possible answers per evaluator group and the average distribution of answers in the case of *full, very high and high IAA*.

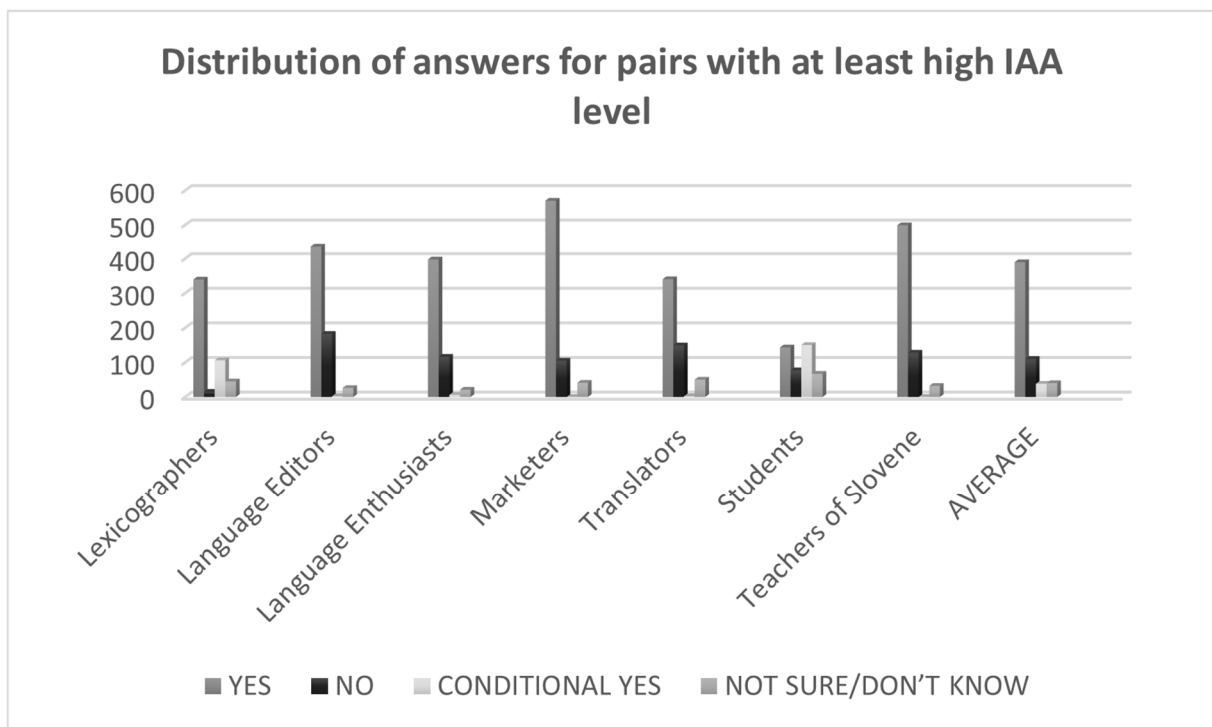


Figure 2: Distribution of number of answers per IAA level.

As the data in Figure 2 show, the two groups that chose the answer **CONDITIONAL YES** more often than other groups and at the same time achieved at least a high IAA were **Students** and **Lexicographers**, suggesting that they made more comments explaining their scruples about the synonym pair, but were also less decisive than other groups who tended to answer **YES** or **NO**. The strictest group that rated most pairs as unsuitable were the **Language Editors**, while **Lexicographers** turned out to be the least strict and rigorous group of all evaluators.

5. Discussion

The results yielded valuable information about **Lexicographers** as evaluators. Out of four hypotheses concerning **Lexicographers** and their decisions when evaluating synonymy, only two were corroborated. The data revealed that **Lexicographers** were the least consistent group, with the second lowest overall Inter-Annotator Agreement

(IAA) score (factoring in *full*, *very high*, and *high IAA* cases) and the highest number of tied responses. Furthermore, they were the least rigorous, deeming only a small proportion of data unsuitable for the Thesaurus. However, Lexicographers demonstrated a broader perspective than other groups, frequently selecting **CONDITIONAL YES** as their answer and offering insights into issues and problems that other evaluators addressed less frequently. The results also indicate that Lexicographers prioritised different issues than other evaluator groups.

Initially, the Lexicographers' answers were meant to serve as a benchmark for evaluation. It was assumed that the lexicographic team's expertise would uniformly reflect the main problems and needs of Thesaurus users and that other evaluator groups would validate this. However, the presented analysis of the Lexicographers' answers revealed that this would not be possible. While the low inter-annotator agreement (IAA) among evaluators was partially due to the four possible decisions allowed, it was surprising that the Lexicographers scored below average on IAA and were more indecisive than other groups. The only group with a lower *at least high IAA* was the Students, however, their performance may have been influenced by imperfect guidelines and a poorer understanding of the task since they were simultaneously evaluating the data and testing the evaluation design (see Gapsa, 2022).

We had expected the Lexicographers to identify both more and different issues with the user-suggested content, while also covering the most common problems and limitations of the Thesaurus and its data. We were surprised to find that they placed disproportionate emphasis on certain issues, which highlights the fact that not all evaluator groups have a universal opinion of the Thesaurus's limitations. Lexicographers focused more on the lack of sense disambiguation and cases of definition instead of actual synonymy, while semantic discrepancies, insufficient senses, or foreign origin of vocabulary were issues raised more frequently by other evaluators. It is possible that the Lexicographers were biased by previous attempts to identify user needs and develop updating solutions, leading them to identify such cases more frequently than other groups. They also operated with more precise terminology, which can explain some of the differences.¹⁶

The design of the research itself may have influenced the Lexicographers' responses. The evaluators were not limited to binary YES-NO choices, but could also select a **NOT SURE/DON'T KNOW** response or a **CONDITIONAL YES** response. Lexicographers, in particular, were more likely to choose the latter option than other evaluator groups (with the exception of Students). From a lexicographic perspective, the difference between YES and **CONDITIONAL YES** responses, especially when combined with comments, is significant. It indicates that either the suggestion or the

¹⁶ Lexicographers' familiarity with "dictionary definitions" facilitated their recognition, but some of the user-suggested synonyms identified by Lexicographers as definitions were actually between descriptions or hypernyms, which other evaluators considered as alternative semantic relations.

headword requires further review and editing, which should be prioritised due to the inadequacy of the current data. Interestingly, Lexicographers were less likely to give a NO response than other groups, perhaps due to their desire to preserve as many synonym candidates as possible and thus provide Thesaurus users with multiple options to choose from. To assist users in making their choice, Lexicographers wanted to ensure that the suggested synonyms were accompanied by semantic information, labels, usage examples, and so on, rather than simply discarding imperfect data. Additionally, Lexicographers did not hesitate to acknowledge that they were unfamiliar with some of the vocabulary. However, the total number of such responses in the Lexicographers group was only slightly higher than average.

The Students group and the Lexicographers shared some interesting similarities. The Inter-Annotator Agreement and number of comments made were almost identical in both groups. Notably, the Students also provided detailed comments, particularly those that were further subcategorized under the “Other” category. They also emphasised alternative spellings, terminological correctness, and issues related to the frequency of use or actual usage of vocabulary. Both groups displayed a greater awareness of the Thesaurus' limitations and had a better understanding of how to name and address them. They were also involved in the updating process and understood the tools and technologies available to facilitate lexicographic review processes, such as verifying data with corpora. Additionally, both groups appeared to take the task more seriously than the other groups, as evidenced by the considerable number of comments as well as the lack of humorous remarks. This could potentially explain the other similarities observed between them.

6. Future work

In this paper, we aimed to explore the differences in how synonymy is perceived and evaluated by Lexicographers, who are experts in the field and typically viewed as the primary evaluators of user-suggested data, and six other groups representing a broader community of dictionary users with diverse professions and interests in language data. The results of the evaluation campaign not only provide a basis for future studies but also have practical implications. They will serve as a guide for drafting editorial protocols, prioritising tasks, and improving the Thesaurus of Modern Slovene. The findings clearly indicate the need for detailed lexicographic guidelines that define appropriate data and the types of additional information pertaining to user suggestions. The guidelines should be based on the priorities identified in the study and supported by empirical data from corpora, as evidenced by the Lexicographers' comments in the "Other" category. The comments highlighted issues such as alternative spellings, frequency of use, and evidence of use in specific meanings, which must be considered in the editorial protocols for future Thesaurus updates. An application-oriented approach would be to add new types of information to the Thesaurus, such as semantic disambiguation, labels, and metadata. Some of these solutions have already been incorporated in the updated version of Thesaurus 2.0.

This paper provides insights into the development of similar online language resources for other languages, based on the involvement of users as collaborators. The study shows that users can offer relevant and useful synonym candidates, but it is also important to involve them as evaluators. The significant differences in the evaluation of synonymy between Lexicographers and other evaluator groups highlight the ongoing need to monitor community priorities and needs and to address them to ensure the actual responsiveness of the responsive lexical resources.

7. Acknowledgements

The authors acknowledge that the project Empirical foundations for digitally-supported development of writing skills (J7-3159) and the programme Language Resources and Technologies for Slovene (P6-0411) were financially supported by the Slovenian Research Agency. The Agency funds the first Author PhD research fund within the programme P6-0411, from which the majority of the evaluators were compensated. Additional funding source for evaluators was the project Upgrading fundamental dictionary resources and databases of CJVT UL, funded by the Ministry of Culture of the Republic of Slovenia in the period 2021–2022.

8. References

- Arhar Holdt, Š. (2020). How Users Responded to a Responsive Dictionary: the Case of the Thesaurus of Modern Slovene. *Rasprave Instituta za hrvatski jezik i jezikoslovlje*, 46(2), pp. 465–482. <https://doi:10.31724/rihjj.46.2.1>
- Arhar Holdt, Š., & Čibej, J. (2020). Rezultati projekta “Slovar sopomenk sodobne slovenščine: Od skupnosti za skupnost“. In D. Fišer, & T. Erjavec (eds.) *Zbornik konference Jezikovne tehnologije in digitalna humanistika, 24. – 25. september 2020, Ljubljana, Slovenija*, pp. 3–9. Available at: http://nl.ijs.si/jtdh20/pdf/JT-DH_2020_Arhar-Holdt-et-al_Rezultati-projekta_Slovar-sopomenk-sodobne-slovenscine.pdf
- Arhar Holdt, Š., Čibej, J., Dobrovoljc, K., Gantar, P., Gorjanc, V., Klemenc, B., . . . 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, 17-21 July 2018, Ljubljana*, pp. 401–410). Available at: <https://euralex.org/wp-content/themes/euralex/proceedings/Euralex%202018/118-4-2991-1-10-20180820.pdf>
- Arhar Holdt, Š., Gantar, P., Kosem, I., Pori, E., Robnik Šikonja, M. & Krek, S. (In press). Thesaurus of Modern Slovene 2.0. *Proceedings of the Eighth Biennial Conference eLex 2023, Brno, Czech Republic, 27–29 June 2023*.
- Arhar Holdt, Š., Kosem, I., & Gantar, P. (2016). Dictionary User Typology: The Slovenian Case. In T. Margalitadze, & G. Meladze (eds.) *Proceedings of the XVII EURALEX International Congress: Lexicography and Linguistic Diversity*, pp. 179–187. Available at https://euralex.org/wp-content/themes/euralex/proceedings/Euralex%202016/euralex_2016_015_p179

.pdf

- Armendariz, C.S., Purver, M., Pollak, S., Ljubešić, N., Ulčar, M., Vulić, I., & Pilehvar, M.T. (2020). SemEval-2020 Task 3: Graded Word Similarity in Context. In A. Herbelot, X. Zhu, A. Palmer, N. Schneider, J. May, & E. Shutova (eds.) *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pp. 36–49. <https://doi:10.18653/v1/2020.semeval-1.3>
- Atkins, S.B.T. (ed.). (1998). *Using Dictionaries: Studies of Dictionary Use by Language Learners and Translators*. Tübingen: Max Niemeyer Verlag.
- Barnhart, C.L. (1962). Problems in editing commercial monolingual dictionaries. *International Journal of American Linguistics*, 28(2), pp. 161–181.
- Bergenholtz, H. & Johnsen, M. (2005). Log files as a tool for improving Internet dictionaries. *Hermes. Journal of Linguistics*, 34, pp. 117–141.
- Bogaards, P. (2003). Uses and users of dictionaries. In P. Van Sterkenburg (ed.) *A Practical Guide to Lexicography*. Amsterdam and Philadelphia: John Benjamins, pp. 26–33.
- Braslavski, P., Ustalov, D., & Mukhin, M. (2014). A Spinning Wheel for YARN: User Interface for a Crowdsourced Thesaurus. In S. Wintner, M. Tadić, & B. Babych (eds.) *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 101–104. <https://doi:10.3115/v1/E14-2026>
- Breit, A., Revenko, A., Rezaee, K., Pilehvar, M.T., & Camacho-Collados, J. (2021). WiC-TSV: An Evaluation Benchmark for Target Sense Verification of Words in Context. In P. Merlo, J. Tiedemann, & R. Tsarfaty (eds.) *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 1635–1645. Available at: <https://aclanthology.org/2021.eacl-main.140/>
- Fišer, D., Tavčar, A., & Erjavec, T. (2014). sloWCrowd: A crowdsourcing tool for lexicographic tasks. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, . . . S. Piperidis (eds.) *Proceedings of the Ninth International Conference on Language Resources and Evaluation. LREC'14, May 26-31, 2014, Reykjavik, Iceland*, pp. 3471–75. Available at: http://www.lrec-conf.org/proceedings/lrec2014/pdf/1106_Paper.pdf
- Gapsa, M. (2022). Ocenjevanje uporabniško dodanih sopomenk v Slovarju sopomenk sodobne slovenščine – pilotna študija. In D. Fišer, & T. Erjavec (eds.) *Proceedings of the Conference on Language Technologies and Digital Humanities, September 15th - 16th 2022, Ljubljana, Slovenia*, pp. 308–316. Available at: https://nl.ijs.si/jtdh22/pdf/JTDH2022_Gapsa_Ocenjevanje-uporabnisko-dodanih-sopomenk-v-Slovarju-sopomenk-sodobne-slovenscine.pdf
- Gapsa, M. (2023). “But why??” *Evaluation of user-suggested synonyms in the Thesaurus of Modern Slovene*. Research Square. <https://doi.org/10.21203/rs.3.rs-2775161/v1>
- Hartmann, R.R.K. (1987). Four perspectives on dictionary use: a critical review of research methods. In A.P. Cowie (ed.) *The Dictionary and the Language Learner*.

- Tübingen: Niemeyer, pp. 11–28.
- Hill, F., Reichart, R., & Korhonen, A. (2015). SimLex-999: Evaluating Semantic Models With (Genuine) Similarity Estimation. *Computational Linguistics*, 41(1), pp. 665–695. https://doi:10.1162/COLI_a_00237
- Householder, F.W. (1967). Summary report. In F.W. Householder and S. Saporta (eds.) *Problems in lexicography*. Bloomington: Indiana University Publications, pp. 279–282.
- Kosem, I., & Pori, E. (2021). Slovenske ontologije semantičnih tipov: samostalniki. In I. Kosem (ed.) *Kolokacije v slovenščini*, pp. 159–202. <https://doi:10.4312/9789610605379>
- Kosem, I., Krek, S., Gantar, P., Arhar Holdt, Š., Čibej, J., Laskowski, C. (2018). Collocations dictionary of modern Slovene. In J. Čibej, V. Gorjanc, I. Kosem, & S. Krek (eds.) *Proceedings of the XVIII EURALEX International Congress, Lexicography in Global Contexts, 17-21 July 2018, Ljubljana*, pp. 989–997. Available at: <https://euralex.org/publications/collocations-dictionary-of-modern-slovene>
- Krek, S., Laskowski, C. & Robnik-Šikonja, M. (2017). From translation equivalents to synonyms: creation of a Slovene thesaurus using word co-occurrence network analysis. In I. Kosem, C. Tiberius, M. Jakubíček, J. Kallas, S. Krek & V. Baisa (eds.) *Proceedings of eLex 2017: Lexicography from Scratch, 19-21 September 2017, Leiden, Netherlands*, pp. 93-109. Available at: <https://elex.link/elex2017/wp-content/uploads/2017/09/paper05.pdf>
- Krippendorff, K. (1970). Estimating the reliability, systematic error, and random error. *Educational and Psychological Measurement*, 30(1), pp. 61–70.
- Lew, R. & de Schryver, G.-M. (2014). Dictionary users in the digital revolution. *International Journal of Lexicography*, 27(4), pp. 341–359.
- Lew, R. (2015) Opportunities and limitations of user studies. In C. Tiberius & C. Müller-Spitzer (eds.) *Research into dictionary use. Wörterbuchbenutzungsfor-schung. 5. Arbeitsbericht des wissenschaftlichen Netzwerks „Internetlexikografie“*. Mannheim: Institut für deutsche Sprache, pp. 6–16.
- Müller-Spitzer, C. (2014). *Using Online Dictionaries*. Berlin – Boston: De Gruyter Mouton.
- Nesi, H. (2000). *The Use and Abuse of EFL Dictionaries. How learners of English as a foreign language read and interpret dictionary entries*. Tübingen: Max Niemeyer Verlag.
- Pilehvar, M.T., & Camacho-Collados, J. (2019). WiC: the Word-in-Context Dataset for Evaluating Context-Sensitive Meaning Representations. 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, Volume 1 (Long and Short Papers)*, pp. 1267–1273. <https://doi:10.18653/v1/N19-1128>
- Rambousek, A., Horák, A., & Pala, K. (2018). Sustainable long-term WordNet development and maintenance: Case study of the Czech WordNet. *Cognitive*

- Studies/Études cognitives*, 18. <https://doi:10.11649/cs.1715>
- Schnabel, T., Labutov, I., Mimno, D., & Joachims, T. (2015). Evaluation methods for unsupervised word embeddings. In L. Màrquez, C. Callison-Burch, & J. Su (eds.) *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 298–307. <https://doi:10.18653/v1/D15-1036>
- Schneidermann, N., Hvingelby, R., & Pedersen, B. (2020). Towards a Gold Standard for Evaluating Danish Word Embeddings. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, . . . S. Piperidis (eds.) *Proceedings of the 12th Language Resources and Evaluation Conference (LREC) 2020, Marseille, France, 13th-15th May 2020*, pp. 4754–4763.
- Tarp, S. (2009). Reflections on lexicographical user research. *Lexikos*, 19(1), pp. 275–296. <https://doi:10.5788/19-0-440>
- Tomaszczyk, J. (1979). Dictionaries: users and uses. *Glottodidactica* 12, pp. 103–119.
- Vila, M., Bertran, M., Martí, M. A., & Rodríguez, H. (2015). Corpus Annotation with Paraphrase Types: New Annotation Scheme and Inter-annotator Agreement Measures. *Language Resources and Evaluation*, 49, pp. 77–105. <https://doi.org/10.1007/s10579-014-9272-5>
- Welker, H.A. (2013a). Methods in Research of Dictionary Use. In R.H. Gouws, U. Heid, W. Schweickard & H.E. Wiegand (eds.) *Dictionaries. An International Encyclopedia of Lexicography: Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography*. Berlin, New York: Walter de Gruyter, pp. 540–547.
- Welker, H.A. (2013b). Empirical Research into Dictionary Use since 1990. In R.H. Gouws, U. Heid, W. Schweickard & H. E. Wiegand (eds.) *Dictionaries. An International Encyclopedia of Lexicography: Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography*. Berlin, New York: Walter de Gruyter, pp. 531–540.

Databases:

- Fišer, D. (2015). *Semantic lexicon of Slovene sloWNet 3.1*. Slovenian language resource repository CLARIN.SI. Available at: <http://hdl.handle.net/11356/1026>
- Gantar, P., Krek, S., Kosem, I., Šorli, M., Kocjančič, P., Grabnar, K. ... Nina Drstvenšek, N. (2013). *Leksikalna baza za slovenščino 1.0*. Slovenian language resource repository CLARIN.SI. Available at: <http://hdl.handle.net/11356/1030>
- Kosem, I., Bálint Čeh, J., Ponikvar, P., Zaranšek, P., Kamenšek, U., Koša, P. ... Gorjanc, V. (2021). *Comprehensive Slovenian-Hungarian Dictionary 1.0*. Slovenian language resource repository CLARIN.SI. Available at: <http://hdl.handle.net/11356/1453>.
- Krek, S., Laskowski, C., Robnik-Šikonja, M., Kosem, I., Arhar Holdt, Š., Gantar, P. ... Dobrovoljc, K. (2018). *Thesaurus of Modern Slovene 1.0*. Slovenian language resource repository CLARIN.SI. Available at: <http://hdl.handle.net/11356/1166>

Word-sense Induction on a Corpus of Buddhist Sanskrit Literature

Matej Martinc¹, Andraž Pelicon¹, Senja Pollak¹, Ligeia Lugli²

¹Jožef Stefan Institute, Ljubljana, Slovenia

²Mangalam Research Center for Buddhist Languages, Berkeley, CA, USA

E-mail: matej.martinc@ijs.si, andraz.pelicon@ijs.si, senja.pollak@ijs.si,
ligeia.lugli@kcl.ac.uk

Abstract

We report on a series of word sense induction (WSI) experiments conducted on a corpus of Buddhist Sanskrit literature with an objective to introduce a degree of automation in the labour-intensive lexicographic task of matching citations for a lemma to the corresponding sense of the lemma. For this purpose, we construct a Buddhist Sanskrit WSI dataset consisting of 3,108 sentences with manually labeled sense annotations for 39 distinct lemmas. The dataset is used for training and evaluation of three transformer-based language models fine-tuned on the task of identifying intended meaning of lemmas in different contexts. The predictions produced by the models are used for clustering of lemma sentence examples into distinct lemma senses using a novel graph-based clustering solution. We evaluate how well do the obtained clusters represent the true sense distribution of new unseen lemmas not used for model training and report the best Adjusted Rand Index (ARI) score of 0.208, and how well do the clusters represent the true lemma sense distribution when the classifier is tested on new unseen sentence examples of lemmas used for model training and report the best ARI score of 0.3. In both scenarios, we outperform the baseline by a large margin.

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

1. Introduction

Buddhist Sanskrit literature constitutes the textual foundation of Mahāyāna, one of the main branches of Buddhism, which flourished in India from around the first couple of centuries BCE to the XII century CE. The experiments reported in this paper stem from a long-standing lexicographic project aimed at creating a first corpus-based dictionary of Buddhist Sanskrit vocabulary (Lugli, 2019, 2021a). Relatively little is known about the semantic permutations that this vocabulary undergoes in different periods and text types, a corpus of relevant sources having become available only recently ((Lugli et al., 2022)). Hence, mapping word senses across various subcorpora of Buddhist literature is a priority in our dictionary and, more generally, in the field of South Asian Buddhist studies. Alas, such mapping is extremely laborious. It requires close reading large quantities of citations for a given lemma. Many of these citations are extracted from highly specialised philosophical discourse and are often challenging to interpret. It took us the most part of a decade to semantically categorize a sample of just over four thousand citations that instantiate word-senses for about 130 lemmas in different genres and periods of Buddhist Sanskrit literature.

Accelerating the process of semantic categorization is clearly the key to scaling up our lexicographic endeavor and achieve a good coverage of the Buddhist Sanskrit lexicon. In this paper we report on a series of word sense induction experiments that we attempted in an effort to integrate a degree of automation in our semantic categorization workflow.

A word sense is a discrete representation of one aspect of the meaning and is context dependent. Dictionaries and lexical databases, such as WordNet (Miller, 1992), organise the entries according to different word meanings. Word Sense Disambiguation (WSD) and Word Sense Induction (WSI) are two fundamental tasks in Natural Language Processing, i.e., those of, respectively, automatically assigning meaning to words in context from a predefined sense inventory and discovering senses from text for a given input word (Navigli, 2012). Both tasks are most frequently applied to open-class words, as those are carrying most of a sentence’s meaning and contain higher level of ambiguity. While for WSD the task consists of associating a word in context with its most appropriate sense from a predefined sense inventory, WSI refers to automatically identifying and grouping different senses of meanings of a word in a given textual context, without exploiting any manually sense-tagged corpus to provide a sense choice for a word in context. The output of WSI is a set of different occurrence clusters, which represent different meanings of a word. When dealing with languages with available large sense inventories, usually WSD methods are being used. On the other side, in less-resourced settings, such as in our case of Buddhist Sanskrit literature, large sense repositories are not available and therefore WSI methods are of core interest.

Therefore, the main aim of this paper is to introduce novel resources for Buddhist Sanskrit related to WSI¹, including:

- a novel word sense induction dataset for Buddhist Sanskrit containing 3108 sentences with manually labeled sense annotations (see Section 3);
- a novel graph-based WSI solution that leverages predictions produced by the transformer-based (Vaswani et al., 2017) language models fine-tuned on the binary classification task of predicting whether the target lemma in two concatenated sentences containing the lemma has the same sense or not;
- an extensive experimental evaluation of three distinct language models and two clustering algorithms, one of them being the widely used Louvain algorithm (Que et al., 2015).

The paper is structured as follows. After related work described in Section 2, we describe the data used in Section 3. Section 4 covers the training of transformer models and the novel clustering solution. Section 5 provides details on the evaluation scenarios, the baselines used and the evaluation measures. While in Section 6 we present the results of our experiments, the paper concludes with final remarks in Section 7.

2. Related work

Word sense induction and disambiguation tasks gained traction more than a decade ago, when several shared tasks on the topic were organized, the most influential being the

¹ The code for experiments is publicly available under the MIT license at <https://gitlab.com/matej.mar-tinc/buddhist-sanskrit-sense-induction>.

Semeval-2010 task 14: Word sense induction and disambiguation (Manandhar et al., 2010) and the SemEval-2013 Task 13: Word Sense Induction for Graded and Non-Graded Senses (Jurgens & Klapaftis, 2013). In these challenges, one of the most common approaches was to build a word co-occurrence graph and use the relations in the graph to obtain word communities, which distinguish senses (Jurgens, 2011).

More recent approaches employ contextual embeddings (Devlin et al., 2019) for the WSI task. For example, the approach by Amrami & Goldberg (2018) is based on the intuition that occurrences of a word that share a sense, also share in-context substitutes. They use a masked language model to derive nearest word substitutes for each word and then cluster the obtained substitute vectors to derive word senses. This substitute-based approach was improved on in the study by Eyal et al. (2022). They show that the approach by Amrami & Goldberg (2018) can be adapted to efficiently and cheaply annotate all the corpus occurrences of all the words in a large vocabulary. They induce senses to a word using contextual representations from a language model and subsequently cluster them into sense clusters. More specifically, for each sentence in which the word appears, they generate k substitute tokens for the target word using a language model. Finally, they cluster all the substitutes into sense clusters. We employ their approach as one of the baselines in our work.

Another WSI method based on contextual embeddings is called PolyLM and was proposed in Ansell et al. (2021). This method combines the task of learning sense embeddings by jointly performing language modeling and word sense induction. This allows the model to utilize the advantages of contextualization at the same training step as modelling senses. PolyLM is based on two underlying assumptions about word senses: firstly, that the probability of a word occurring in a given context is equal to the sum of the probabilities of its individual senses occurring; and secondly, that for a given occurrence of a word, one of its senses tends to be much more plausible in the context than the others. Similar to the other language models, PolyLM is trained in an unsupervised manner on large corpora of unlabeled data and at inference time performs word sense induction without supervision.

Another way to tag words senses is to employ Word Sense Disambiguation (WSD), if a predefined sense inventory is available. These approaches can be roughly divided into supervised WSD and knowledge-based WSD (see Bevilacqua et al. (2021) for a recent survey). Knowledge-based approaches leverage lexical resources, including databases, such as WordNet (Miller, 1992). One of the most popular knowledge-base WSD approaches is the Lesk dictionary-based algorithm (Lesk, 1986), which also inspired one of the baseline approaches in our work. More recent vector-based approaches leverage contextualized word representations and sense embeddings to perform disambiguation (Wang & Wang, 2020). Other popular approaches leverage graph structure of knowledge graphs. A variety of graph algorithms have been employed, including random walks (Agirre et al., 2014) and Personalized PageRank (Scozzafava et al., 2020). While knowledge-base WSD (Pasini & Navigli, 2017) does not require large annotated word-to-sense corpora, they on the other hand do require a language-specific sense inventory, such as for example WordNet.

For the supervised WSD, an adequate amount of annotated data for training is required. One of the first approaches was proposed by Zhong & Ng (2010), who decided to tackle the task with an SVM-based approach. More recent studies on the other hand opted to include neural representations into the workflow. For example, several contextual embeddings based WSD approaches were tested in the scope of the SemDeep-5’s Word-in-Context task

(Pilehvar & Camacho-Collados, 2019). During the task, several sense embedding systems were tested on a binary classification task of determining whether a certain “focus” word has or does not have the same sense in two concatenated sentences containing the word. The approach employing BERT performed the best.

Following Bevilacqua et al. (2021), recent supervised WSD approaches can be grouped into 1-nn vector-based ones (e.g., Wang & Wang (2020)), token tagger-based-ones (e.g., Bevilacqua & Navigli (2020) or sequence tagging-based ones (e.g., Huang et al. (2019)).

As far as we are aware, as of yet no WSI or WSD approaches have been employed for Buddhist Sanskrit.

3. Dataset

The dataset used for our experiments is derived from the data we annotated for our dictionary of Buddhist Sanskrit², with some notable modification. First, for this study we have considered only words for which more than 20 sentences have been manually annotated for sense. Second, we simplified our lexicographic dataset to include a single level of semantic annotation, out of three. We only use here annotations for word sense, leaving aside the more fine-grained categorization into subsenses, as well as the more general categorization into semantic fields—both of which are less closely linked to lexical context and therefore less amenable to automation than word sense. Subsenses especially have proven too complex to model due to their high number, with several words being associated to more than eight of them. Finally, in a few cases we have altered the hierarchy between senses and subsenses for this study, so that, whenever possible, senses are clearly connected to a specific lexical context. In our original lexicographic data, our priority was to convey the continuity between different senses of a word, especially between specialised and general-language uses (Lugli, 2021b). So, in our dictionary data we typically subsume terminological applications under the general-language sense from which they stem, even when the lexical contexts in which the specialized uses occur are markedly different from the general-language ones (see e.g. our dictionary sub voce “vitarka”). Given the importance of lexical context for automated word-sense-induction, we revised our dataset so that terminological uses that occur in specific contexts correspond to senses, rather than subsenses, and are therefore considered as distinct semantic categories in this study. The sense labels used in the dataset are the fruit of our lexicographic work and have been crafted to serve as English paraphrases of the senses expressed by a Sanskrit lemma.

The Buddhist Sanskrit word sense induction (WSI) dataset we used here consists of 3,108 sentences with manually labeled sense annotations for 39 distinct lemmas. The dataset statistics are presented in table 1. 26 of these lemmas have more than one sense (on average 3.3 distinct senses), while 13 are monosemous, and are only used in some of the experiments (see Section 5 for details).

The WSI dataset is used for fine-tuning and evaluation of three distinct transformer-based language models (Devlin et al., 2019), pretrained on a corpus of Buddhist Sanskrit literature.

² <https://zenodo.org/record/7972951>

	Num. lemmas	Num. sent.	Num. tokens	Average num. of senses
Monosemic	13	862	14,471	1
Polysemic	26	2,246	42,059	3.31
All	39	3,108	56,530	2.54

Table 1: The word sense induction dataset statistics.

4. Methodology

4.1 Transformer model training

In our experiments we test three distinct transformer-based language models trained on the Buddhist Sanskrit corpus described in Lugli et al. (2022). Namely, we trained two versions of the BERT model (Devlin et al., 2019), i.e. a “BERT base” model with 12 encoder layers, a hidden size of 768 and 12 self-attention heads, and a “BERT small” model with 8 encoder layers, a hidden size of 768 and 8 self-attention heads. Additionally, we also trained a smaller version of the GPT-2 model (Radford et al., 2019) with 8 encoder layers, a hidden size of 256 and 8 self-attention heads³.

The main reason for testing of smaller models with less parameters are the overfitting issues reported in the studies by Sandhan et al. (2021) and Lugli et al. (2022), when large language models are pretrained on corpora that are magnitudes smaller than the e.g., English corpora on which these models were trained originally. In the study by Sandhan et al. (2021), where they trained a general Sanskrit model, they decided to tackle the overfitting issue by training a lighter version of BERT (a so-called ALBERT model (Lan et al., 2019)), which is a strategy that we also employ in this work in order to assess if possible improvements in performance can be obtained by employing a smaller model.

In our previous study (Lugli et al., 2022), where we tested several contextual embeddings models on the Buddhist Sanskrit corpus, we reported serious overfitting issues with a GPT-2 model⁴, a model almost 10 times larger than the base version of BERT in terms of number of parameters, which resulted in very low embedding quality. For this reason, in this study we do not conduct experiments with a GPT-2 model of original size, but rather just test a much smaller version, which did not overfit on the small pretraining corpus.⁵

For language model pretraining (employing the masked language modeling objective for BERT models and autoregressive language modeling for GPT-2), we follow the regime proposed in Lugli et al. (2022). We pretrain both contextual models on the general Sanskrit corpus described in Lugli et al. (2022) for up to 200 epochs, and then on the Buddhist corpus, again for up to 200 epochs. The reason for pretraining on the general Sanskrit corpus is a considerable overlap in the vocabulary and grammar of general and Buddhist Sanskrit, which we believe the models might be able to leverage and learn some useful lexical, semantic, and grammatical information, and therefore compensate for the relatively small size of the Buddhist corpus. Same as in Lugli et al. (2022), we preprocess the corpus with the compound splitter proposed in Hellwig & Nehrlich (2018) to obtain word tokens.

³ All these models are monolingual and were trained only on Sanskrit data.

⁴ https://huggingface.co/docs/transformers/model_doc/gpt2

⁵ The final model’s size was determined by gradually reducing the number of encoder layers, attention heads and the embedding size until the overfitting problem has been overcome, i.e. until the perplexities the models have achieved on the train and test set were comparable.

The pretrained models are fine-tuned on a binary classification task of predicting whether the same lemma in two distinct sentences has the same sense. More specifically, for each lemma in the WSI dataset presented in Section 3, we define a set of its example sentences as L_i and build a binary classification dataset consisting of lemma sentence pairs that we obtain as a Cartesian product of L_i with itself. Note that we remove sentence pairs in which the first sentence is the same as the second sentence. We define the final dataset D as a union of sentence pairs L_i consisting of sentences s_1 and s_2 containing the same target lemma. More formally, D is defined with the following equation:

$$D = \bigcup_{i=1}^n (L_i \times L_i | (s_1 \in L_i) \neq (s_2 \in L_i))$$

For each sentence pair in the dataset D , we label whether the lemmas in it have the same sense or not. This dataset is used for fine-tuning and evaluation of language models.

4.2 Clustering examples into senses

The binary predictions produced by the models are used for clustering of lemma sentence examples into distinct lemma senses. We build one graph $G = (V, E)$ per lemma, comprised of a set of vertices V representing lemma sentence examples, and a set of edges $E \subseteq V \times V$, which are ordered pairs, representing connections between vertices. Vertices in the graph are connected if they contain lemma with the same sense. This allows us to build a (0,1)-adjacency matrix for each lemma, in which ones indicate whether pairs of vertices (in our case sentences) are adjacent (i.e., contain lemmas with the same sense) in the graph.

The resulting adjacency matrix is used for clustering of vertices (i.e. sentence examples containing the same target lemma) into sense clusters using a novel clustering solution, in which the rows of the matrix are used for construction of initial clusters. More specifically, in the first step, we create a different cluster containing the target vertice and its adjacent sentences for each example, resulting in n initial clusters, where n is a number of vertices in the graph. To obtain the final clusters, these initial clusters are merged by recursively combining the clusters with the largest intersection up to a predefined threshold of minimum intersection or maximum number of clusters. The threshold for minimum intersection was experimentally set to 0.8 and maximum number of clusters was set to 10, i.e., the merging of clusters with the largest intersection continues until at most 10 distinct clusters remain. The threshold of 10 was set due to the observation that very few lemmas in Buddhist Sanskrit have more than 10 distinct main senses. Note that due to a large variability in cluster sizes, the merging of clusters is based on normalized intersection that also takes into the account the number of vertices in the two clusters we potentially wish to merge. More specifically, the intersection I between two sets (clusters) of vertices S_i and S_k is normalized by dividing it with the size of the smaller cluster:

$$I = S_i \cap S_k / \min(|S_i|, |S_k|) \tag{1}$$

The final step in the proposed clustering solution is to remove duplicate vertices, which appear in more than one cluster. Here we opted for a simple solution, which proved experimentally effective, and remove all duplicates but the one in the largest cluster. The

logic behind this strategy relies on a simple probability estimate that these outlier vertices, which do not fit neatly in a single cluster, have the greatest probability of belonging to the biggest cluster in a clustering.

5. Experimental setup

5.1 Evaluation scenarios

The obtained clusters, representing sense distributions for each lemma, are evaluated in two 5-fold cross-validation (CV) scenarios. All pretrained models are fine-tuned for 5 epochs on the binary classification task described in Section 4 for each fold in the cross-validation evaluation. We evaluate the performance of the models on the binary classification task using two measures, accuracy and macro-averaged F1-score. The latter was chosen in addition to accuracy due to unbalance between the two classes in the language model’s test set.

In the first scenario, we test how well do the obtained clusters represent the true sense distribution of new unseen (polysemous and monosemous) lemmas not used for model training. In this scenario, we maintain a strict division between lemmas in the models’ train set and lemmas in models’ test set. We do not remove monosemous lemmas from the test set, in order to simulate a real life scenario of employing the model on new lemmas with unknown number of senses. We call this the “lemma division” setting. In the second scenario, we test how well do the clusters represent the true lemma sense distribution when classifier is tested on new unseen sentence examples for polysemous lemmas used for model training. Here, there is no division between lemmas in the train and test set, just a division between train and test set lemma sentence examples, since we wish to simulate a real life scenario of employing the model on new unlabeled occurrences of lemmas on which the model was trained, with known number of senses. In this scenario, we remove the monosemous lemmas from the test set, since sense induction on these lemmas is trivial for the models. We call this the “no lemma division” setting.

Note that in both scenarios the obtained train sets in the 5-fold CV evaluation are balanced, i.e., the number of sentence pairs with the same target lemma sense and the number of sentence pairs with the different target lemma sense are balanced by downsampling the majority class for each lemma. This also means that in both scenarios the models are only trained on the polysemous lemmas. On the other hand, we do not balance the test sets.

5.2 Baselines

The proposed approach is compared to three distinct baselines. To compare the novel clustering solution to a more commonly used graph-based clustering algorithm, we once again use binary predictions produced by the transformer models to build a graph $G = (V, E)$ for each lemma, comprised of a set of vertices V representing lemma sentence examples, and a set of edges $E \subseteq V \times V$. Two vertices (i.e. sentences) in the graph are again connected if they contain lemmas with the same sense. We fed the constructed graph to the popular Louvain clustering algorithm (Que et al., 2015) to obtain the final sense clusters.

The second baseline we apply only in the “no lemma division” scenario was inspired by the Lesk dictionary-based algorithm for word sense disambiguation (Lesk, 1986). More

specifically, a sentence containing a target lemma for which we wish to determine a sense, is considered as a bag of words (BOW). We calculate normalized intersection (see equation 1) between the set of words in the new sentence in the test set and all the sentences containing the same target lemma in the train set. The lemma in the test set sentence is assigned the sense of the target lemma in the train set sentence with the largest intersection. Note that this approach is only feasible in the “no lemma division scenario” and can only be employed for disambiguation of lemmas for which a set of labeled sentences already exists. We call this the “BOW intersection” approach.

The third baseline is an approach for large-scale word-sense induction by Eyal et al. (2022) described in Section 2. We re-implemented the approach from the original work but omitted the building of the inverted word index which was used to conserve space. Since in our experiment the dataset is several orders smaller in size, this step was unnecessary for our purpose. In our case, we generate the substitutes with a pretrained Buddhist Sanskrit “BERT base” language model. In each sentence, we mask the target word w and we generate the probability distribution across all the tokens in the vocabulary with the language model. We then take the k most probable tokens and treat them as the substitutes for the word w . This way we leverage the context in trying to induce senses for the target word. In our experiment we set the k to 20 experimentally.

For forming sense clusters, we first build a graph with substitutes as nodes where two nodes are connected if they represent substitutes that were being generated for the same word. We then cluster this graph using the Louvain clustering algorithm. The resulting clusters represent sense clusters. Using the Louvain algorithm allows us to not set the number of clusters prior to clustering but induce the number of clusters automatically from the data. This makes this method completely unsupervised as no sense labels nor the number of clusters for target words are needed to be known in advance. For this reason, we use it as a baseline in the “lemma division” setting.

5.3 Evaluation measures

We employ two distinct measures for evaluation of the clustering algorithm, Adjusted Rand Index (ARI) score (Hubert & Arabie, 1985) and an F1-score (Manandhar et al., 2010).

The F1-score measure for evaluation of word sense induction was first proposed in the Semeval-2010 task 14: Word sense induction and disambiguation (Manandhar et al., 2010) and was motivated by a similar evaluation measure used for information retrieval. The F-Score of a gold standard sense gs_i (denoted as $F(gs_i)$ in the equation below), is the maximum $F(gs_i, c_j)$ value attained at any cluster, where the F1-score of gs_i with respect to c_j , $F(gs_i, c_j)$, is defined as the harmonic mean of precision of class gs_i with respect to cluster c_j and recall of class gs_i with respect to cluster c_j . The F1-score of the entire clustering solution is finally defined as the weighted average of the F1-scores of each gold standard sense, where q is the number of gold standard senses and N is the total number of sentence examples for a specific lemma. More formally, the score is defined with the following equation:

$$F1 - score = \sum_{i=1}^q \frac{|gs_i|}{N} F(gs_i)$$

The main advantage of the F1-score evaluation is that it penalises systems that produce higher number of clusters (low recall) or lower number of clusters (low precision) than the gold standard number of senses. On the other hand, F1-score suffers from the matching problem, which results in the score not being able to evaluate the entire membership of a cluster, or by not evaluating every cluster (Rosenberg & Hirschberg, 2007), especially when gold standard distribution is very unbalanced. In this case, the F1-score tends to not consider the make-up of the clusters beyond the majority class.

For this reason, we employ an additional evaluation measure, ARI, which does not suffer from the matching problem, is equal to zero in the cases of trivial clustering, such as random clustering, or when the model produces a separate cluster for each context or a single cluster for all contexts, even in the case of uneven gold standard distribution. The measure was used for evaluation of WSI in several shared tasks (Navigli & Vannella, 2013; Panchenko et al., 2018). We adopt the ARI implementation from the scikit-learn library⁶, which produces scores between 1 (when the clusterings are identical) and -0.5 (for especially discordant clusterings). ARI is based on the Rand Index (RI), which calculates a similarity score between two clusterings by looking at all pairs of samples and then counting pairs that are assigned in the same or different clusters in the predicted and gold standard clusterings.

ARI is calculated by adjusting the Rand Index for chance using the following equation:

$$ARI = \frac{RI - Expected_RI}{max(RI) - Expected_RI}$$

Both measures, F1-score and ARI, are calculated for each lemma. We obtain an overall score for a cross-validation fold by averaging the lemma scores. Finally, we average the scores across five cross-validation folds to obtain the overall cross validation scores. We also report the standard deviation of fold scores for both measures.

6. Results

The results the different language models achieve on the binary classification task of predicting whether the two lemmas have the same sense or not are presented in Table 2. According to both evaluation criteria, macro-averaged F1-score and accuracy, the best performing model in the “no lemma division” setting is GPT-2 small, achieving an F1-score of 69.33% and an accuracy of 72.09%. In the “lemma division” scenario, the best performing model is BERT base with an F1-score of 57.21% and an accuracy of 60.44%. The performances of all models in both scenarios are nevertheless comparable and standard deviation intervals intersect.

The results of different clustering solutions are presented in Table 3. In the “no lemma division” scenario, the best solution in terms of ARI is employing the novel clustering solution (in the Table 3 labeled as “custom clustering”) on binary predictions generated by the BERT base model, with an ARI score of 0.3. While using the combination of the GPT-2 small model (which achieved the best macro-averaged F1-score and accuracy in the binary classification task) and the novel clustering solution also produces competitive

⁶ https://scikit-learn.org/stable/modules/generated/sklearn.metrics.adjusted_rand_score.html

Model	F1 Macro	F1 Macro STD	Accuracy	Accuracy STD
No lemma division				
BERT base	67.94	2.97	69.23	3.37
BERT small	67.25	1.81	69.06	2.19
GPT-2 small	69.33	1.02	72.09	1.19
Lemma division				
BERT base	57.21	5.55	60.44	6.21
BERT small	56.47	3.44	60.09	4.93
GPT-2 small	55.20	3.17	59.71	5.85

Table 2: The results of different language models on the binary classification task.

Approach	ARI	ARI STD	F1	F1 STD
No lemma division				
BERT base + Custom clustering	0.300	0.034	76.10	0.58
BERT small + Custom clustering	0.217	0.041	73.96	1.58
GPT-2 small + Custom clustering	0.286	0.035	75.74	0.65
BERT base + Louvain	0.271	0.039	73.60	1.24
BERT small + Louvain	0.205	0.048	70.84	2.25
GPT-2 small + Louvain	0.258	0.043	74.28	2.13
BOW intersection	0.254	0.042	76.78	1.68
Lemma division				
BERT base + Custom clustering	0.099	0.026	79.78	4.56
BERT small + Custom clustering	0.116	0.029	79.94	4.38
GPT-2 small + Custom clustering	0.208	0.159	80.36	4.03
BERT base + Louvain	0.041	0.026	61.58	2.37
BERT small + Louvain	0.055	0.022	65.04	4.66
GPT-2 small + Louvain	0.023	0.022	69.14	2.56
Eyal et al. (2022)	0.024	0.010	35.50	1.10

Table 3: The results of different clustering solutions.

results in terms of ARI, employing the novel clustering solution on binary predictions produced by the BERT small model surprisingly leads to a much worse performance in terms of ARI. This finding is interesting since all the models achieved comparable performance on the binary classification task, therefore we expected that the clustering results would also be competitive. Using the Louvain clustering, we achieve lower ARI and F1-scores than using the custom clustering no matter the model we use for binary predictions. Again, employing the Louvain algorithm on binary predictions produced by the BERT small model leads to much worse results in terms of ARI than if two other models are used.

In terms of the F1-score, the non-neural BOW intersection baseline achieves the best performance of 76,78%. Using the combination of BERT base or GPT-2 small and custom clustering is also a competitive strategy, leading to F1-scores around 76%. We believe that the best performance of the BOW intersection baseline in terms of F1-score is to some extent caused by the unbalanced distribution of senses in the gold standard distribution and the fact that the F1-score tends to not consider the make-up of the clusters beyond the majority class due to the matching problem. Nevertheless, since the BOW intersection

Lemma	Key Word in Context (KWIC)	Translation	Sense	Assigned cluster
nāman	niṣikte nāma rūpe tu ṣaḍ āyatana sambhavaḥ / ṣaḍ āyatanam āgamyā saṃsparśaḥ saṃpravartate // [...] yad idam a vidyā	When name and form develop, the six senses emerge. In dependence upon the six senses, impact actually occurs. [Batchelor]	nāma-rūpa (one of the twelve nidānas)	1
nāman	pratyaḥ saṃskārāḥ saṃskāra pratyaḥ vijñānam vijñāna pratyaḥ nāma rūpaṃ nāma rūpa pratyaḥ [...]	translation not available	nāma-rūpa (one of the twelve nidānas)	2
nāman	tadyathāpi nāma subhūte ratn ārthikaḥ puruṣo mahāsamudraṃ dṛṣṭvā n āvagāheta /	Just as if a person who desires jewels would not look for them in the great ocean, [...]. [Conze 235]	namely	3
nāman	bhoḥ puruṣa kas tav āsyām upary anunayo yan nāma madiyām ājñām vilānghya n echaṣy enām praghatayituṃ [...]	Man! What regard do you have for her that, violating my order, you do not wish to kill her? [Rajapatirana 19]	namely	3
nāman	tadyathāpi nāma ānanda rāja cakravartīṃ prāsādāt prāsādaṃ saṃkrāmet /	A universal monarch can pass from palace to palace, [...]. [Conze 366]	namely	3
nāman	tasya parama siddha yātravāt supāraga ity eva nāma babhūva /	His voyages proved so extremely successful that he came to be called Supāraga. [Khoroché 96]	name/word	4
nāman	asyām ānanda mathurāyām mama varṣa śata parinirvṛtasya gupto nāma gāndhiko bhaviṣyati /	Ananda, right here in Mathurā, one hundred years after my parinirvāṇa, there will be a perfumer named Gupta. [Strong 174]	name/word	4
nāman	tasya vistareṇa jātimahaṃ kṛtvā pṛcchati kiṃ kumārasya bhavatu nāma /	When the prince’s full birth festival was being celebrated, she was asked what his name should be. [Strong 205]	name/word	4
nāman	paśy ājīta aīka sattvaṃ api nām otsahayitv eyat puṇyaṃ prasavati /	Mark, Agita, how much good is produced by one’s inciting were it but a single creature; [Kern 333]	indeed/really/actually	4
nāman	arthibhiḥ pṛīta hṛdayaiḥ kīrtyamānam itas tataḥ / tyāga saury onnatam nāma tasya vyāpa dīṣo daśa //	[1] His petitioners were well-contented and praised him far and wide, so that the name he earned for his largesse spread to every corner of the earth. [Khoroché 22]	name/word	4

Table 4: Word sense induction examples in the “no lemma division” setting for lemma *nāman* with four distinct labeled senses, when BERT base and custom clustering is employed. In KWIC examples, and tags are used for denoting the target lemma and / (daṇḍa) for punctuation.

baseline also offers solid performance in terms of ARI (0.254), and since it does not require any additional cluster mapping⁷, this approach seems like a viable option, especially since it is extremely fast and requires very few computational resources.

In the “lemma division” setting, the usage of custom clustering tends to outperform all the baseline approaches by a large margin according to both evaluation criteria. By far the best ARI score of 0.208 is achieved if we use custom clustering on the binary predictions produced by the GPT-2 small model. In this setting, the standard deviation between folds in the 5 fold CV setting is nevertheless very large, 0.159. In fact, the ARI score across folds varied between 0.477 and 0.029, which means that the score very much depends on which lemmas are in the train set, when GPT-2 small model is used for production of binary predictions. This indicates that the model might have issues finding general rules that can be applied for sense disambiguation on different lemmas and rather relies on a set of features that only work for some lemmas.

⁷ While the BOW intersection baseline works as a word sense disambiguation approach by assigning target lemmas in new sentences predefined senses, the other approaches work as word sense induction strategies, producing clustering distributions without labeled clusters. While the latter approaches are useful if all word senses for a specific target lemma are not known in advance, an additional cluster mapping step, in which the produced unlabeled clusters are mapped to the actual lemma senses is nevertheless required in order to obtain actual senses.

The usage of BERT base or BERT small models leads to more consistent ARI scores across different folds of around 0.1. This means that there is a substantial drop in terms of ARI, if we compare the “lemma division” approach to the “no lemma division” approach, which suggests that all transformer models (not just the GPT-2 small model) have issues in finding general rules that can be applied for sense disambiguation on different lemmas. Most likely this is due to the limited size of the fine-tuning dataset, which only contains 39 different lemmas.

In terms of the F1-score, all approaches based on custom clustering achieve comparable and very competitive scores around 80%. Again, we believe that this is partially caused by the matching problem of the evaluation score and unbalanced distribution of senses in the gold standard distribution.

Examples of word sense induction for lemma *nāman* in the “no lemma division” setting when BERT base and custom clustering is employed are presented in Table 4. Note how the sentence examples containing lemmas with majority senses (“namely” and “name/word”) tend to be clustered correctly, while the clustering perform worse for examples containing lemmas with minority senses (“indeed/really/actually” and “nāma-rūpa (one of the twelve nidānas)”).

7. Conclusion

In the paper, we released the first word sense induction dataset and proposed the first WSI approach employed for Buddhist Sanskrit, with an intention to automate the time and labor intensive lexicographic task of assigning senses to target lemmas in sentences. The approach relies on pretrained transformer language models fine-tuned on a binary classification task of predicting whether two identical target lemmas in two sentences have the same sense or not. The produced predictions are then used in a novel graph-based clustering solution.

While the proposed approach outperforms several WSI baselines in terms of ARI, we do observe several potential problems with the method, which will need to be thoroughly addressed before it can be fully integrated in a lexicographic pipeline for Buddhist Sanskrit. First, the large difference in performance between the two tested approaches, the “lemma division” approach and the “no lemma division” approach, indicates that transformer models tend to rely on lemma specific features during binary classification and fail to find general contextual features to distinguish between senses. Another indication of that is the standard deviation between folds in the 5 fold CV setting in the “lemma division” setting, when the best performing GPT-2 small model is used. The latter suggests that the selection of lemmas, on which the model is trained, is important. We believe that both of these problems could be resolved by a larger training dataset in terms of both sentence examples for a specific lemma and number of different lemmas in the dataset. The construction of such bigger training dataset will be the object of future work, but it seems likely that only the number of different lemmas included in the data will increase substantially, as lexicographers will in any case progressively annotate sentences for more lemma as they expand the dictionary. By contrast, expanding the number of sentences annotated for each lemma may prove difficult to align with lexicographic goals, since manual annotation is extremely laborious and WSI is needed to reduce the amount of manual annotation required for dictionary development.

When it comes to the evaluation scores, we believe that the F1-score is not appropriate for evaluation in our setting, because the unbalanced classes resulting from the above-mentioned matching problem interfere with the score’s ability to evaluate entire membership of the cluster, especially in scenarios where a prevailing gold standard majority cluster is accompanied by several smaller clusters. Since the score is calculated as the weighted average of the F1-scores of each gold standard cluster, in such scenarios the memberships of smaller clusters are neglected due to relatively small weights. In our case, this leads to a relatively small differences between different approaches in terms of F1-score (this was especially the case in the “no lemma division” scenario), since all approaches were able to assign membership to a majority cluster to a reasonably good degree, since this is the easiest part of the task. On the other hand, there were significant differences between different approaches when it comes to successfully assigning membership to minority clusters, and these were not captured by the F1-score. While the ARI score tends to do better in this respect, we will nevertheless explore other evaluation scores in future work, in order to try to improve our evaluation scenario even further.

8. Acknowledgements

This work was funded by a NEH Digital Advancement Grant level 2 (HAA-277246-21), while the creation of the Buddhist Sanskrit Corpus was partly funded by the British Academy (NF161436) and the Khyentse Foundation. We also acknowledge the Slovenian Research Agency core programme Knowledge technologies P2-0103. Finally, we would like to thank Luis Quiñones for his contribution to the creation of the evaluation dataset.

9. References

- Agirre, E., de Lacalle, O.L. & Soroa, A. (2014). Random Walks for Knowledge-Based Word Sense Disambiguation. *Computational Linguistics*, 40(1), pp. 57–84. URL <https://aclanthology.org/J14-1003>.
- Amrami, A. & Goldberg, Y. (2018). Word Sense Induction with Neural biLM and Symmetric Patterns. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 4860–4867. URL <https://aclanthology.org/D18-1523>.
- Ansell, A., Bravo-Marquez, F. & Pfahringer, B. (2021). PolyLM: Learning about Polysemy through Language Modeling. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, pp. 563–574. URL <https://aclanthology.org/2021.eacl-main.45>.
- 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 *International Joint Conference on Artificial Intelligence*.
- 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 Linguis-*

- tics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186.
- Eyal, M., Sadde, S., Taub-Tabib, H. & Goldberg, Y. (2022). Large Scale Substitution-based Word Sense Induction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 4738–4752. URL <https://aclanthology.org/2022.acl-long.325>.
- Hellwig, O. & Nehrdich, S. (2018). Sanskrit word segmentation using character-level recurrent and convolutional neural networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. pp. 2754–2763.
- Huang, L., Sun, C., Qiu, X. & Huang, X. (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)*. Hong Kong, China: Association for Computational Linguistics, pp. 3509–3514.
- Hubert, L. & Arabie, P. (1985). Comparing partitions. *Journal of classification*, 2, pp. 193–218.
- Jurgens, D. (2011). Word sense induction by community detection. In *Proceedings of TextGraphs-6: Graph-based Methods for Natural Language Processing*. pp. 24–28.
- Jurgens, D. & Klapaftis, I. (2013). SemEval-2013 Task 13: Word Sense Induction for Graded and Non-Graded Senses. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. Atlanta, Georgia, USA: Association for Computational Linguistics, pp. 290–299. URL <https://aclanthology.org/S13-2049>.
- Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P. & Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.
- Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*. pp. 24–26.
- Lugli, L. (2019). Smart lexicography for under-resourced languages: lessons learned from Sanskrit and Tibetan. In *Smart Lexicography: eLex 2019*. pp. 198–212.
- Lugli, L. (2021a). Dictionaries as collections of data stories: an alternative post-editing model for historical corpus lexicography. In *Post-Editing Lexicography: eLex 2021*. pp. 216–231.
- Lugli, L. (2021b). Words or terms? Models of terminology and the translation of Buddhist Sanskrit vocabulary. In A. Collett (ed.) *Buddhism and Translation: Historical and Contextual Perspectives*. pp. 149–172.
- Lugli, L., Martinc, M., Pelicon, A. & Pollak, S. (2022). Embeddings models for Buddhist Sanskrit. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. pp. 3861–3871.
- Manandhar, S., Klapaftis, I., Dligach, D. & Pradhan, S. (2010). Semeval-2010 task 14: Word sense induction & disambiguation. In *Proceedings of the 5th international workshop on semantic evaluation*. pp. 63–68.
- Miller, G.A. (1992). WordNet: A Lexical Database for English. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*. URL <https://aclanthology.org/H92-1116>.
- Navigli, R. (2012). A Quick Tour of Word Sense Disambiguation, Induction and Related Approaches. In M. Bieliková, G. Friedrich, G. Gottlob, S. Katzenbeisser & G. Turán

- (eds.) *SOFSEM 2012: Theory and Practice of Computer Science*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 115–129.
- Navigli, R. & Vannella, D. (2013). Semeval-2013 task 11: Word sense induction and disambiguation within an end-user application. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. pp. 193–201.
- Panchenko, A., Lopukhina, A., Ustalov, D., Lopukhin, K., Arefyev, N., Leontyev, A. & Loukachevitch, N. (2018). RUSSE’2018: a shared task on word sense induction for the Russian Language. *arXiv preprint arXiv:1803.05795*.
- 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*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 78–88. URL <https://aclanthology.org/D17-1008>.
- Pilehvar, M.T. & Camacho-Collados, J. (2019). WiC: the Word-in-Context Dataset for Evaluating Context-Sensitive Meaning Representations. 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. 1267–1273. URL <https://aclanthology.org/N19-1128>.
- Que, X., Checconi, F., Petrini, F. & Gunnels, J.A. (2015). Scalable community detection with the louvain algorithm. In *2015 IEEE International Parallel and Distributed Processing Symposium*. IEEE, pp. 28–37.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).
- Rosenberg, A. & Hirschberg, J. (2007). V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*. pp. 410–420.
- Sandhan, J., Adideva, O., Komal, D., Behera, L. & Goyal, P. (2021). Evaluating Neural Word Embeddings for Sanskrit. *arXiv preprint arXiv:2104.00270*.
- Scozzafava, F., Maru, M., Brignone, F., Torrissi, G. & Navigli, R. (2020). Personalized PageRank with Syntagmatic Information for Multilingual Word Sense Disambiguation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Online: Association for Computational Linguistics, pp. 37–46. URL <https://aclanthology.org/2020.acl-demos.6>.
- 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.
- Wang, M. & Wang, Y. (2020). A Synset Relation-enhanced Framework with a Try-again Mechanism for Word Sense Disambiguation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 6229–6240. URL <https://aclanthology.org/2020.emnlp-main.504>.
- Zhong, Z. & Ng, H.T. (2010). It Makes Sense: A Wide-Coverage Word Sense Disambiguation System for Free Text. In *Proceedings of the ACL 2010 System Demonstrations*. Uppsala, Sweden: Association for Computational Linguistics, pp. 78–83. URL <https://aclanthology.org/P10-4014>.

Word sense induction for (French) verb valency discovery

Naïma Hassert, François Lareau

OLST, Université de Montréal
C.P. 6128, succ. Centre-Ville, Montréal QC, H3C 3J7, Canada
E-mail: first.lastname@umontreal.ca

Abstract

We explore the use of Transformers in word sense induction for the automatic construction of a valency dictionary of French verbs. To account for the way the arguments of a verb change depending on its sense, this type of dictionary must distinguish at least the main senses of a lemma. However, constructing such a resource manually is very costly and requires highly trained staff. That is why one important subtask in the construction of this resource is to automatically identify the polysemy of the verbs. For each of the 2,000 most frequent French verbs, we extract the word embeddings of 20,000 of their occurrences in context found with Sketch Engine, and we cluster those embeddings to find the different senses of each verb. In order to identify the language model and clustering algorithm most suited to our task, we extract the word embeddings of the sentences in the FrenchSemEval evaluation dataset with one language-specific model, CamemBERT, and two multilingual models, XLM-RoBERTa and T5. These vectors are then clustered with three different algorithms that do not require a predetermined number of clusters: Affinity Propagation, Agglomerative Clustering and HDBSCAN. Our experiments confirm the potential of unsupervised methods to identify verb senses, and indicate that monolingual language models are better than multilingual ones for word sense induction tasks involving a single language.

1. Introduction

Valency dictionaries such as DEM (Dubois & Dubois-Charlier, 2010), Dicovalence (van den Eynde et al., 2017), *Lefff* (Sagot, 2010), LVF (Hadouche & Lapalme, 2010), VerbNet (Kipper et al., 2006) and Verbønet (Danlos et al., 2016) are useful in many natural language processing applications, in particular for rule-based natural language generation. This type of dictionary indicates precisely how a predicate expresses its arguments in syntax, including information on selected part-of-speech, preposition or case. However, the way a word expresses its arguments can change significantly depending on its sense. For example, the verb *change* requires a direct object when it means ‘modify’, as in *The discussion has changed my thinking about the issue*, but with the sense ‘become different’, as in *She changed completely as she grew older*, then there is no object at all (examples taken from WordNet¹; Fellbaum 1998). Therefore, a valency dictionary must distinguish at least the main senses of a lemma. Constructing this kind of resource manually, however, is very costly and requires highly trained staff.

Our goal is thus to automate the construction of valency dictionaries. In this paper we focus on how we tackled an important subtask: automatically identifying the polysemy of verbs. Our data is drawn from French, but the method we present here is language-independent.

¹ <https://wordnet.princeton.edu/>

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

We tackled this WSI task in three main steps: first, we extracted contextualized vectors of the sentences in the FrenchSemEval evaluation dataset (Segonne et al., 2019) with one language-specific model, CamemBERT (Martin et al., 2020), and two multilingual models, XLM-RoBERTa (Conneau et al., 2020) and T5 (Raffel et al., 2020). This dataset is comprised of 66 French verbs in context, each having around 50 sense-annotated examples. Then, we tested three unsupervised clustering algorithms that don't require knowing the number of clusters beforehand: Affinity Propagation (Frey & Dueck, 2007), Agglomerative Clustering (Szekely & Rizzo, 2005) and HDBSCAN (McInnes et al., 2017). The best results were achieved with CamemBERT vectors clustered with Agglomerative Clustering, obtaining a BCubed F_1 score (Bagga & Baldwin, 1998) of 65.20%. As a comparison, the FlauBERT team (Le et al., 2020), also using CamemBERT vectors, obtained an F_1 score of 50.02% on the same dataset, although they used a supervised method and measured their results with the traditional F_1 score, which could not be used in our case since we used an unsupervised method. Finally, for each verb present in the evaluation dataset, we add the word embeddings of 20,000 instances of this verb in context extracted via Sketch Engine² (Kilgarriff et al., 2014). We then cluster each group of approximately 20,050 verbs separately (the 20,000 verbs in context previously extracted, plus the 50 examples from the evaluation data), and evaluate the performance of the clustering on the evaluation data. Our experiments allow us to pinpoint the best combination of language model, clustering algorithm and parameter to identify the senses of a verb from raw data.

This paper begins with a brief summary of previous work in the WSI field in §2. Follows a presentation of the language models (§3.1) and the clustering algorithms §3.2 on which we experimented. Then, we describe in §4 how we evaluated the combinations of those algorithms. Finally, we present in §5 an analysis of our results, and conclude in §6.

2. Automatic identification of the polysemy

2.1 Word sense disambiguation (WSD)

The automatic identification of the sense of an ambiguous word in context has been a research topic for decades and is still an unresolved task. Yet, it is crucial in many applications, such as:

- automatic translation, where a word in a language can have many different translations in another;
- information retrieval, where search queries often contain ambiguous words;
- information extraction, where we want to automatically retrieve specific information related to a specific topic;

² <https://www.sketchengine.eu>

- lexicography, where we often want to obtain lexical information specific to a given word sense.

A common way of tackling this task is by using a knowledge-based method or a supervised one. Knowledge-based methods rely heavily on existing resources like WordNet (Fellbaum, 1998), BabelNet (Navigli & Ponzetto, 2012), FrameNet (Baker, 2014) or other dictionaries, and use the content of those resources to compare with the data on hand and deduce the word sense. Supervised methods rely instead on sense-annotated data, which is then used to annotate raw data. Most state-of-the-art methods are hybrid, i.e., combine features of knowledge-based and supervised methods (Bevilacqua et al., 2021).

In the context of dictionary creation, however, knowledge-based or supervised methods are not necessarily the most appropriate way to identify the sense of an ambiguous word, for the following reasons:

1. The senses listed in major lexical resources are often too fine-grained.

A popular lexical resource in the field of natural language processing (NLP) is WordNet, an electronic dictionary of English based on *synsets*, i.e., sets of synonymous lexemes. If one looks up a word in WordNet, one ends up with all the synsets that contain it. In the case of *change*, for instance, there are 10 synsets related to the noun *change*, and 10 synsets related to the verb *change*. Its multilingual counterpart, BabelNet (Navigli & Ponzetto, 2012), is a result of the merging of WordNet and Wikipedia, where the synsets are provided in part by the human-generated translations provided by Wikipedia and in part by a machine translation system. It has been pointed out, however, that WordNet’s senses are very fine-grained, to the point where inter-annotator agreement when using the WordNet inventory is around 70 % (Navigli, 2006), only 5 % more than the most frequent sense (MFS) baseline, which consists of annotating each word with its most frequent sense (Raganato et al., 2017).

2. Most resources are based on English.

WSD systems rely heavily on sense-annotated data. This type of data exists in English, thanks mainly to SemCor (Miller et al., 1993), which is sense-annotated based on WordNet. However, since manual semantic annotation is very costly, this data is scarce or non-existent for languages other than English. As a result, most of the lexical resources in other languages are derived from the English ones, like Europarl (Koehn, 2005), a corpus annotated with the senses of BabelNet (which itself is derived in part from WordNet). Relying on those resources can thus be misleading if we want to do WSD for, say, French.

3. Relying on external resources prevents the discovery of new senses.

As mentioned earlier, lexical resources have the inconvenience that they are costly to create and update. However, new words and senses are created continually. Thus, hand-curated lexical resources can easily become outdated.

2.2 Word sense induction (WSI)

When WSD is performed without the help of an external resource, it is called word sense *induction* (or *discrimination*). WSI methods can be a solution to the knowledge acquisition

bottleneck, since they only rely on raw, non annotated, data. This method does not assign a sense to a word *per se*: instead, it aims to detect *how many* senses there are, based on the assumption that two occurrences of a word have the same sense if they occur in similar contexts. Common approaches in the field are:

- **Context Clustering**

This algorithm, developed by Schütze (1998), interprets senses as groups, or clusters, of similar contexts of an ambiguous word. More specifically, each word is represented as a vector whose components are counts of the number of times another word appears in its context (the context can be a sentence, a paragraph, or any other length of text). The original algorithm dealt with vectors built from second-order co-occurrences, i.e., where vectors of the words in the context of the ambiguous word are themselves built from their own context. These context vectors can then be clustered into groups based on their similarity. Each group is represented by the mean of all the vectors of this group, namely the *centroid*. This is the method closest to the one we decided to adopt in this paper.

- **Word Clustering**

This algorithm has been developed by Lin (1998). It identifies words that are similar to a target word based on their syntactic dependencies. The context is parsed syntactically and represented as triples, each of them consisting of the target word, a syntactic dependent and the syntactic relationship between them. Common information between two words are the triples that appear in the description of both of the words. One can then use this information to calculate the similarity between two words. Finally, a tree is created with the help of a clustering algorithm. The nodes directly under the main node are considered as the different senses of the word.

- **Co-occurrence graph**

Véronis (2004) presented HyperLex, arguing that the problem with clustering vectors is that it can exclude less frequent word senses, which will tend to be considered as noise by the algorithm even if those senses are not rare ones for an average speaker. In an attempt to solve this problem, a graph is built where the nodes are words and they are connected according to their co-occurrences in a given context size. One ends up with *small worlds*, i.e., highly connected groups that are said to correspond to a sense and that are all linked in some way.

- **Recurrent neural networks**

Recurrent neural networks (RNNs) are a class of artificial neural networks that recursively define the output at any stage as a function of the previous output. They have been useful for several NLP tasks, but suffer from the vanishing gradient problem, which makes them only possible to use in short sequences. Long short-term memory (LSTM) (Hochreiter & Schmidhuber, 1997) is a RNN variation that avoids the vanishing gradient problem to a certain extent and allows recurrent networks to learn over many more steps. However, they require a lot of resources and time to train, and still do not have a huge memory.

- **Transformers**

Transformers were first introduced by Vaswani et al. (2017) and have revolutionized

the field. Their attention mechanism allows them to process the entire input at once, reducing training times drastically and achieving state-of-the-art results in NLP. Transformer models pretrained on huge datasets can be easily downloaded from Hugging Face³ and further trained.

3. Method

3.1 Word embeddings

We used three different language models for our experiments, all downloaded from Hugging Face. We used one monolingual model for French, CamemBERT, and two multilingual models, XLM-RoBERTa and T5. Monolingual models have been shown to yield better results than multilingual language models such as mBERT (Martin et al., 2020). However, in 2021, XLM-RoBERTa showed impressive results on the SemEval-2021 Task 2: Word in Context Disambiguation (Martelli et al., 2021), including for French, so we included it in our experiments. We also experimented on another multilingual language, T5, released by Google. We used the large version of each model, and got our contextualized vectors by calculating the mean of all hidden layers.

CamemBERT (Martin et al., 2020) is a monolingual model constructed especially for French. Its architecture is based on RoBERTa’s, a method that builds on BERT’s language masking strategy while modifying key hyperparameters and training with much larger mini-batches and learning rates. RoBERTa reportedly have better downstream task performance than BERT (Liu et al., 2019). CamemBERT is trained on 138 GB of raw data. On its release in 2020, it has improved the state of the art for part-of-speech tagging, dependency parsing, named entity recognition and natural language inference tasks for French.

XLM-RoBERTa (Conneau et al., 2020) is a multilingual language model pretrained on 2.5 TB of data from 100 languages. At its release, it showed a very significant improvement over the multilingual models mBERT and XLM-100, and obtained competitive results over state-of-the-art monolingual models, including RoBERTa, in English. They demonstrated that multilingual models can improve their monolingual BERT counterpart. To the best of our knowledge, however, CamemBERT and XLM-RoBERTa have not been compared for WSI, so we have yet to verify whether XLM-RoBERTa can improve on its monolingual counterpart for French.

T5 (Raffel et al., 2020) is an alternative to BERT. Instead of having class label or a span of the input as outputs, as with BERT-style models, T5 has text string only as input as well as output. T5-large, the checkpoint that we used, has 770 million parameters. T5 was trained on a dataset containing 4 languages: English, French, German and Romanian.

3.2 Clustering

Clustering methods aim at finding structure in a set of unlabeled data. Several clustering algorithms need a number of cluster beforehand; however, since our goal is to eventually be able to find senses that have not been listed in a lexical resource, we tested three clustering algorithms that can choose the optimal number of clusters without being explicitly told.

³ <https://huggingface.co>

3.2.1 Affinity propagation

Affinity propagation (Frey & Dueck, 2007) is a clustering algorithm that exchanges messages between data points until members of the input that are representative of clusters, “exemplars”, are obtained. There are two parameters that can be tuned: damping and preference. Damping affects the convergence of the algorithm. Preference adds noise to the similarity matrix, and thus affects the number of clusters.

3.2.2 Agglomerative clustering

Agglomerative clustering (Szekely & Rizzo, 2005) is a type of hierarchical cluster analysis that uses a bottom-up approach. It begins by considering every element in the data as its own cluster and successively agglomerates similar clusters until all clusters have been merged into a single one that contains all the data. In scikit-learn (Pedregosa et al., 2011), instead of specifying the number of clusters, one can simply specify the distance threshold, i.e., the linkage distance threshold above which clusters will not be merged. Basically, it indicates the limit at which to cut the dendrogram tree. We used the default linkage parameter, namely “ward”, and tested distance thresholds ranging from 10 to 300,000.

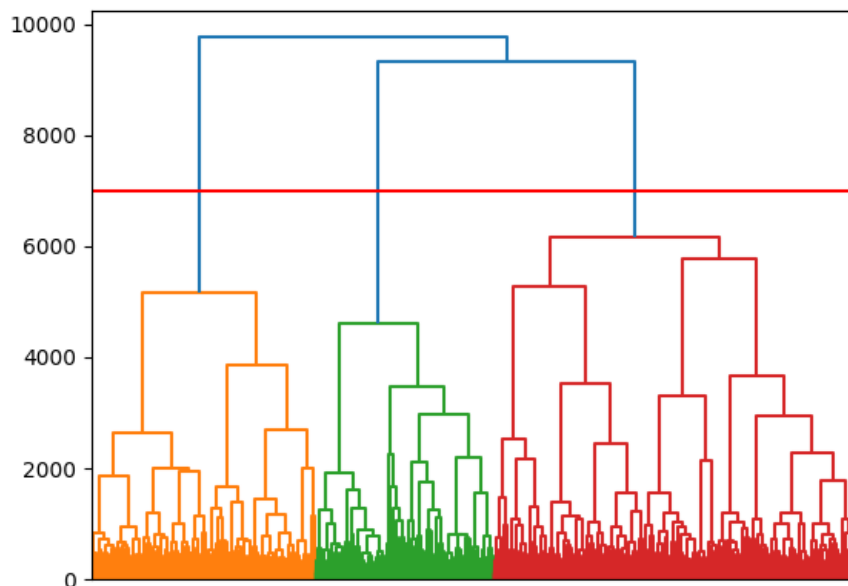


Figure 1: Dendrogram representing the clustering of 20,000 contextual embeddings of the verb *adopter* (‘adopt’) with agglomerative clustering. The horizontal red line represents the final number of clusters (3) obtained with a distance threshold of 7000.

3.2.3 HDBSCAN

HDBSCAN (Campello et al., 2013) is a clustering method that extends DBSCAN by converting it into a hierarchical clustering algorithm. It assumes there is a significant number of noise points, among which one can find islands of higher density. Density methods have the advantage of being efficient even when the data is not clean and the clusters are weirdly shaped. It begins by identifying the densest parts of the data space,

and deciding if those densest parts should be merged or kept separate. The algorithm produces a probability score for each data point of belonging to their cluster, and a cluster quality score.

Three main parameters can be fined-tuned in HDBSCAN: the minimum cluster size (the smallest size grouping that is to be considered a cluster), the minimum number of samples (the higher the value, the more conservative the algorithm will be and the more data will be considered as noise) and the clustering selection method (by default “eom”—for *excess of mass*—and can be changed to “leaf”, which tends to produce more fine-grained clustering).

4. Evaluation

4.1 FrenchSemEval

FrenchSemEval (Segonne et al., 2019) is an evaluation dataset constructed specifically for the WSD of French verbs. It was built after the authors of this dataset inspected Eurosense (Delli Bovi et al., 2017), a multilingual corpus extracted from Europarl (Koehn, 2005) and automatically sense-annotated using the BabelNet multilingual sense inventory. This resource presented good results in terms of inter-annotator agreement, and for English, the high-precision Eurosense annotations cover 75 % of the content words and have a precision score of 81.5 %. As can be expected, though, the French results are lower: coverage is 71.8 % and precision is 63.5 %. Furthermore, the situation gets worse with verbs, which can be expected since the disambiguation of verbs is known to be more difficult (Raganato et al., 2017). When the authors examined the verbs that had been automatically annotated in Eurosense, they realized that the proportion they judged correct was only 44 %. They also confirmed that BabelNet had a very high number of senses per verb; indeed, on a sample of 150 sentences, they found that the average number of BabelNet senses per verb type occurring in these sentences was 15.5, and that the difference between the senses was sometimes difficult to perceive. In short, like most of the available resources, Eurosense is a resource based on English and thus of a lesser quality for French, and in which senses are too fine-grained.

In contrast, Segonne et al. (2019) observed that in Wiktionary,⁴ the granularity level of the senses was usually quite natural and that the sense distinctions were easy to grasp. They thus decided to use the Wiktionary senses as a basis for manual annotation. FrenchSemEval is the result of this effort. It consists of 3,121 sense-annotated sentences, with 66 different verb forms, each having an average of 3.83 senses. All of those verbs were present in the 2,000 most frequent verbs we had identified via Sketch Engine. As indicated in the paper, the MFS baseline for this is 30 % in accuracy.

4.2 MCL-WiC

The Multilingual and Cross-Lingual Word-in-Context Disambiguation task (Martelli et al., 2021) is the first SemEval task to examine the aptitude of systems to discriminate between word senses without any requirement of a fixed sense inventory. The multilingual sub-task is binary: the system must determine if two target words in two different contexts in the

⁴ <https://www.wiktionary.org>

same language has the same meaning or not. The verbs were selected according to their number of senses (it had to have at least three senses in BabelNet) and the sentence pairs were extracted from either the United Nations Parallel Corpus (Ziemski et al., 2016) or Wikipedia. The sentences selected contained sufficient semantic context to determine with certainty the meaning of the target words.

Gupta et al. (2021) got the best result for the Fr-Fr task, attaining 87.5% accuracy. They obtained fine-tuned contextualized embeddings of the target words from XLM-RoBERTa and passed them to a logistic regression unit. It must be noted that even though we tested XLM-RoBERTa too, our results cannot be directly compared, since we evaluated our results on verbs only (and not on all part-of-speech tags as they did).

4.3 Score measure

In this paper, we use the BCubed F_1 scores. This is because the standard F_1 is designed to compare data that is clustered using the same cluster labels, which is useful if the clusters in question have a specific meaning, but not otherwise. Let us take for example the verb *change* mentioned in the introduction. If we want to find all the tokens that have the sense ‘modify’ in cluster A and all the tokens that have the sense ‘become different’ in cluster B, then the cluster labels are important. If all the words put in cluster B should have been instead in cluster A and vice-versa, then the standard F_1 score will be very low.

In our case, though, the cluster labels have no significance: all that matters is to group all the tokens that have similar senses. That is when the BCubed F_1 comes in handy. Instead of calculating the precision and recall based on the number of true and false positives and negatives in all the examples, these scores are calculated for each element individually. The numbers computed for each example in the document are then averaged to produce the recall and precision scores for the entire dataset. The formulas to compute the final BCubed recall and precision are the following:

$$\text{Precision} = \sum_{i=1}^N w_i \times \text{Precision}_i$$

$$\text{Recall} = \sum_{i=1}^N w_i \times \text{Recall}_i$$

The formula for the BCubed F_1 score does not differ from the standard one:

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

In other words, standard F_1 is perfect to evaluate the performance in a WSD task, where the classes are already determined. However, for WSI, we want to evaluate the performance of an algorithm that creates clusters from scratch against an evaluation dataset that will necessarily have its own cluster labels. In this case, BCubed F_1 must be used.

5. Results

5.1 FrenchSemEval

5.1.1 Clustering the evaluation dataset

We clustered the 3,121 sentences of the test set of FrenchSemEval, and calculated BCubed F_1 . We report in table 1 the best results for each language model-clustering algorithm combination.

Clustering algorithm	T5	CamemBERT	XLM-RoBERTa
Affinity Propagation	14.86	14.87	14.86
Agglomerative Clustering	46.02	65.39	56.06
HDBSCAN	30.41	33.76	35.30

Table 1: Best BCubed F_1 scores on the FrenchSemEval dataset

As we can see, the combination of Agglomerative Clustering and CamemBERT is by far the best one for our task, yielding impressive results for an unsupervised method. Indeed, the FlauBERT team (Le et al., 2020), using a combination of CamemBERT and a supervised method, attained an F_1 score of 50.02%. For this clustering method, the distance threshold parameter that allowed each language model to attain the best score varies: in the case of CamemBERT, it was of 650; for T5, it was 100,000; and for XLM-RoBERTa, it was 725.

The worst results we obtained were with Affinity Propagation. Even by doing a grid search with various values of damping and preference, we were not able to achieve more than 14.87% BCubed F_1 . The algorithm achieved a good precision, but a really poor recall in every parameter combination, which indicates that the algorithm was not able to generalize, assigning instead approximately one sense per sentence.

HDBSCAN had the opposite effect: recall was generally much higher than precision, which indicates that it tends to assign only one sense to the entirety of the dataset. The best result with this algorithm was obtained by XLM-RoBERTa, with a minimum cluster size of 10, a minimum number of samples of 2 and the “leaf” cluster selection method. We can also note that an enormous amount of data is considered as noise by the algorithm, and that in almost every parameter configuration, the “leaf” cluster selection method yields much better results than “eom”.

5.1.2 Clustering each verb individually

For each verb in the FrenchSemEval dataset, we clustered the 20,000 instances previously collected (cf. §1), to which we had added the 50 or so sentences of the evaluation dataset, and evaluated the performance of our clustering on the evaluation sentences.

It turns out that BCubed F_1 is maybe not the best indicator of the quality of the clustering for our purpose. Indeed, when we increase the distance threshold, recall approaches 100%, which boosts the score. But it only means that the clustering is more and more severe, so that there is only one cluster or two remaining. For example, if we set the distance

threshold to 19,000, it gets to 67.68%, which is better than our results on the entire dataset. But if we look at the mean number of clusters, we realize that it is not a good clustering: on average, each verb has only one cluster (which is likely not better than the MFS baseline).

The goal could then be to get a number of clusters that is approximately the same as the mean number of clusters for the FrenchSemEval dataset, which is 3.83. We achieve a mean number of clusters of 3.89 with a distance threshold of 6000, with a BCubed F_1 score of 59.93%, which is still satisfactory. The mean number of clusters goes down as the distance threshold goes up, while, on the contrary, BCubed F_1 goes up (as shown in figure 2); at 7000, the mean number of clusters is 3.13, with a score of 62.75%.

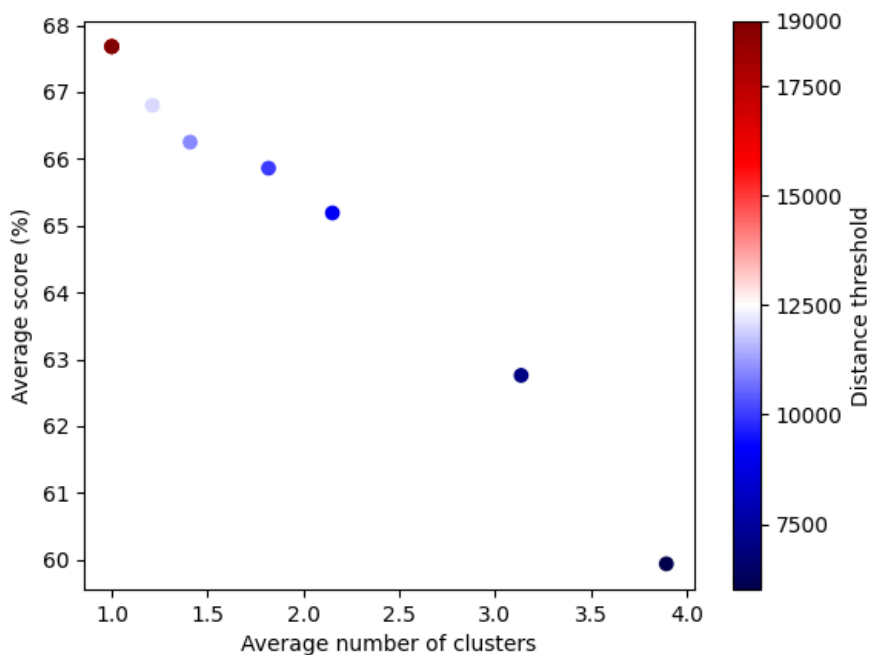


Figure 2: Relation between the score, the number of clusters and the distance threshold when clustering each verb individually with the agglomerative clustering algorithm. The score is expressed in BCubed F_1 and calculated with the FrenchSemEval dataset as a gold standard.

5.2 WiC

After our tests on the FrenchSemEval dataset, we ended up with some uncertainty on the best parameters, knowing we had to strike a balance between the mean number of clusters and BCubed F_1 . For this reason, we decided to test our clustering on another evaluation dataset: the Word-in-Context dataset (Martelli et al., 2021). We proceeded the same way as before: for each verb of the dataset, we first extracted the contextualized embeddings of the test sentences with CamemBERT, then merged them with our own CamemBERT embeddings. We then clustered all the embeddings of each verb with the Agglomerative Clustering algorithm, comparing the three possible parameter values: 6000, 6500, and

7000. The results are in table 2. One can observe that the accuracy score is the highest with a distance threshold value of 6000 on this dataset.

Distance threshold	WiC (accuracy)	FrenchSemEval (BCubed F_1)
6000	61.83 %	59.93 %
6500	59.92 %	61.30 %
7000	59.54 %	62.75 %

Table 2: Performance on the Word-in-Context and FrenchSemEval evaluation datasets according to the distance threshold value of the agglomerative clustering algorithm parameter.

6. Conclusion

In this paper, we have explored how the clustering of contextual embeddings could help discover the senses of French verbs in context. The best results were achieved with a combination of CamemBERT embeddings and the agglomerative clustering algorithm. We noticed that when we augmented the main parameter of the agglomerative clustering algorithm, the distance threshold, the mean number of senses per verb went down while the BCubed F_1 score went up when we evaluated ourselves against the FrenchSemEval dataset. Comparing these results with those obtained for the WiC dataset did not really help us to make a wise decision concerning the distance threshold to use on our data, since the tendency was the opposite in this case (in the WiC dataset, the accuracy went down while the distance threshold went up). Since the FrenchSemEval dataset is bigger and has more similarities with the task we want to achieve, we decided to select the distance threshold of 7000, which gives satisfactory results on the FrenchSemEval dataset (62.75 %) while yielding an adequate number of senses per verb. Now that we have identified the best combination of language model, clustering algorithm and parameter for our task, the clustering for the 2000 most frequent verbs can be done.

7. References

- Bagga, A. & Baldwin, B. (1998). Entity-Based Cross-Document Coreferencing Using the Vector Space Model. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, volume 1*. USA: ACL, pp. 79–85. URL <https://doi.org/10.3115/980845.980859>.
- Baker, C.F. (2014). FrameNet: A Knowledge Base for Natural Language Processing. In *Proceedings of Frame Semantics in NLP: A Workshop in Honor of Chuck Fillmore (1929-2014)*. Baltimore, MD, USA: ACL, pp. 1–5. URL <https://aclanthology.org/W14-3001>.
- Bevilacqua, M., Pasini, T., Raganato, A. & Navigli, R. (2021). Recent Trends in Word Sense Disambiguation: A Survey. In Z.H. Zhou (ed.) *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. IJCAI Organization, pp. 4330–4338. URL <https://doi.org/10.24963/ijcai.2021/593>.
- Campello, R.J., Moulavi, D. & Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining:*

- 17th Pacific-Asia Conference, PAKDD 2013, Gold Coast, Australia, April 14-17, 2013, Proceedings, Part II 17*. Springer, pp. 160–172.
- Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., 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*. Online: ACL, pp. 8440–8451. URL <https://aclanthology.org/2020.acl-main.747>.
- Danlos, L., Pradet, Q., Barque, L., Nakamura, T. & Constant, M. (2016). Un Verbnet du français. *TAL*, 57(1), pp. 33–58. URL <https://hal.inria.fr/hal-01392817>.
- Delli Bovi, C., Camacho-Collados, J., Raganato, A. & Navigli, R. (2017). EuroSense: Automatic Harvesting of Multilingual Sense Annotations from Parallel Text. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, volume 2 (short papers)*. Vancouver, Canada: ACL, pp. 594–600. URL <https://aclanthology.org/P17-2094>.
- Dubois, J. & Dubois-Charlier, F. (2010). La combinatoire lexico-syntaxique dans le Dictionnaire électronique des mots. Les termes du domaine de la musique à titre d’illustration. *Langages*, 179-180(3–4), pp. 31–56.
- Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. The MIT Press. URL <https://doi.org/10.7551/mitpress/7287.001.0001>.
- Frey, B.J. & Dueck, D. (2007). Clustering by passing messages between data points. *science*, 315(5814), pp. 972–976.
- Gupta, R., Mundra, J., Mahajan, D. & Modi, A. (2021). MCL@IITK at SemEval-2021 Task 2: Multilingual and Cross-lingual Word-in-Context Disambiguation using Augmented Data, Signals, and Transformers. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*. ACL, pp. 511–520. URL <https://aclanthology.org/2021.semeval-1.62>.
- Hadouche, F. & Lapalme, G. (2010). Une version électronique du LVF comparée avec d’autres ressources lexicales. *Langages*, 179-180(3–4), pp. 193–220.
- Hochreiter, S. & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8), pp. 1735–1780.
- 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(1), pp. 7–36.
- Kipper, K., Korhonen, A., Ryant, N. & Palmer, M. (2006). Extending VerbNet with Novel Verb Classes. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC’06)*. Genoa, Italy: ELRA, pp. 1027–1032. URL http://www.lrec-conf.org/proceedings/lrec2006/pdf/468_pdf.pdf.
- Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of Machine Translation Summit X: Papers*. Phuket, Thailand, pp. 79–86. URL <https://aclanthology.org/2005.mtsummit-papers.11>.
- Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L. & Schwab, D. (2020). FlauBERT: Unsupervised Language Model Pre-training for French. In *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC’20)*. Marseille, France: ELRA, pp. 2479–2490. URL <https://aclanthology.org/2020.lrec-1.302>.
- Lin, D. (1998). Automatic Retrieval and Clustering of Similar Words. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, volume 2*. USA: ACL, pp. 768–774. URL <https://doi.org/10.3115/980691.980696>.

- 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. arXiv:1907.11692.
- Martelli, F., Kalach, N., Tola, G. & Navigli, R. (2021). SemEval-2021 Task 2: Multilingual and Cross-lingual Word-in-Context Disambiguation (MCL-WiC). In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*. ACL, pp. 24–36. URL <https://aclanthology.org/2021.semeval-1.3>.
- Martin, L., Muller, B., Ortiz Suárez, P.J., Dupont, Y., Romary, L., de la Clergerie, É., Seddah, D. & Sagot, B. (2020). CamemBERT: a Tasty French Language Model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: ACL, pp. 7203–7219. URL <https://aclanthology.org/2020.acl-main.645>.
- McInnes, L., Healy, J. & Astels, S. (2017). hdbscan: Hierarchical density based clustering. *JOSS*, 2(11), p. 205.
- Miller, G.A., Leacock, C., Teng, R. & Bunker, R.T. (1993). A Semantic Concordance. In *Human Language Technology: Proceedings*. pp. 303–308.
- Navigli, R. (2006). Meaningful Clustering of Senses Helps Boost Word Sense Disambiguation Performance. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*. Sydney, Australia: ACL, pp. 105–112. URL <https://aclanthology.org/P06-1014>.
- 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. URL <https://www.sciencedirect.com/science/article/pii/S0004370212000793>.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, pp. 2825–2830.
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. & Liu, P.J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. 1910.10683.
- Raganato, A., Camacho-Collados, J. & Navigli, R. (2017). Word Sense Disambiguation: A Unified Evaluation Framework and Empirical Comparison. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, volume 1 (long papers)*. Valencia, Spain: ACL, pp. 99–110. URL <https://aclanthology.org/E17-1010>.
- Sagot, B. (2010). The Lefff, a Freely Available and Large-coverage Morphological and Syntactic Lexicon for French. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valletta, Malta: ELRA, pp. 2744–2751. URL http://www.lrec-conf.org/proceedings/lrec2010/pdf/701_Paper.pdf.
- Schütze, H. (1998). Automatic Word Sense Discrimination. *Computational Linguistics*, 24(1), pp. 97–123.
- Segonne, V., Candito, M. & Crabbé, B. (2019). Using Wiktionary as a resource for WSD: the case of French verbs. In *Proceedings of the 13th International Conference on Computational Semantics — Long Papers*. Gothenburg, Sweden: ACL, pp. 259–270. URL <https://aclanthology.org/W19-0422>.
- Szekely, G. & Rizzo, M. (2005). Hierarchical Clustering via Joint Between-Within Distances: Extending Ward's Minimum Variance Method. *Journal of Classification*, 22, pp. 151–183.
- van den Eynde, K., Mertens, P. & Eggermont, C. (2017). Dicovallence. URL <https://hdl.handle.net/11403/dicovallence/v1>. ORTOLANG.

- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. & Polosukhin, I. (2017). Attention is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*. Red Hook, NY, USA: Curran Associates Inc., pp. 6000–6010.
- Véronis, J. (2004). HyperLex: Lexical cartography for information retrieval. *Computer Speech & Language*, 18, pp. 223–252.
- Ziemski, M., Junczys-Dowmunt, M. & Pouliquen, B. (2016). The United Nations Parallel Corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia: ELRA, pp. 3530–3534.

Towards a Comprehensive Dictionary of Middle Persian

Francisco Mondaca¹, Kianoosh Rezania²,
Slavomír Čéplö³, Claes Neufeind⁴

¹ ⁴Cologne Center for eHumanities, University of Cologne
² ³Center for Religious Studies, Ruhr University Bochum
E-mail: f.mondaca@uni-koeln.de, kianoosh.rezania@rub.de,
slavomir.ceploe@rub.de, c.neufeind@uni-koeln.de

Abstract

This paper discusses the process of developing a flexible and comprehensive model for a bilingual corpus dictionary with a dead language, in this case, Zoroastrian Middle Persian, as the source language, and a particular focus on accommodating *termini technici* and multi-word expressions. Advanced search capabilities are achieved through the integration of state-of-the-art technologies, with plans to further enhance the system by implementing advanced natural language processing techniques. The project offers two distinct API solutions to cater to diverse user needs and ensure efficient access to lexical data. One of these is a dedicated API designed specifically for the web application. The other is a REST API, which simplifies data access and promotes scalability. The project acknowledges the potential for future integration with large language models, underlining the prospect for future enhancements. This approach encourages collaboration and innovation in historical linguistics, highlighting the crucial role of adaptable and cutting-edge technologies in developing a robust lexicon for historical languages.

Keywords: corpus-based dictionary; middle persian; api; rest; graphql

1. Middle Persian Language and Texts

1.1 Middle Persian

Middle Persian was spoken in the province Persis (Fārs) in the first millennium CE. It served as the official language of the Sasanian Empire (224 - 651 CE), a dominant power beside the Roman Empire during late antiquity. Middle Persian derives from Old Persian, the language spoken in the same area until the third century BCE. In the last centuries of the first millennium CE, Middle Persian has developed into New Persian, the major language of people in today's Iran, Afghanistan, and Tajikistan. With an estimated 800,000 words, the Middle Persian corpus is undoubtedly the most comprehensive among Old and Middle Iranian text corpora.

In the vast territory of the multi-lingual and multi-national Sasanian Empire, Middle Persian served as a sort of *lingua franca*. As a vital linguistic bridge, it significantly contributed to the transcultural communication between diverse regions and civilizations during the first millennium CE. In this role, it enabled the exchange of ideas, knowledge, and cultural practices and fostered a rich and interconnected environment. Middle Persian's prominence within the Sasanian empire highlights its importance in both the administrative and cultural spheres, facilitating effective governance and fostering cultural communication across the empire's vast territories.

One of the cultural fields in which Middle Persian played a prominent role was religion. It was the language of choice for the documentation of numerous religious traditions, including Zoroastrianism, Manichaeism, and Christianity. Whereas Middle Persian texts are transmitted fragmentary to us, religious texts build the largest Middle Persian corpora. Zoroastrian Middle Persian (ZMP) texts account for more than 85 percent of the whole corpus, the Manichean ones to circa five percent. Administrative texts and Sasanian and post-Sasanian inscriptions constitute the rest of the Middle Persian corpus.

Different scripts have been used to write down Middle Persian texts. Sasanian inscriptions are written in the inscriptional Pahlavi script, Manichaean texts in the Manichaean, and most Zoroastrian and administrative texts in the Pahlavi cursive script. In the post-Sasanian period, Zoroastrians also used other scripts, such as Avestan and Perso-Arabic, to document their texts.

Middle Persian texts have been written on a wide range of materials. Inscriptions are engraved in stone and metal coins and seals, administrative texts are transmitted via papyri, parchment, and ostraca. Most extant Zoroastrian and Manichean texts are written on paper whereas the Zoroastrian texts written in the Pahlavi script are transmitted in codices. Some of these texts can be dated only roughly to the last centuries of the Sasanian period, more than half of them have been authored in the early Islamic period, specifically in the ninth and tenth centuries CE. The oldest extant Zoroastrian codices date back to the 14th century (Macuch, 2009; Rezaia, 2023).

1.2 Zoroastrian Middle Persian

ZMP texts are of paramount importance in late antiquity, in terms of linguistic history, history of religions, cultural development, and scientific thought. These texts not only transmit and interpret ancient Iranian religious ideas but also showcase intellectual connections with other religions and cultures of late antiquity. However, the current state of scholarly engagement with these texts is uneven. While many unsophisticated texts have been repeatedly edited, some complex texts remained relatively neglected.

The ZMP texts transmitted in codices can be categorized into the following genres:

- Zand texts, i.e. Middle Persian translations of and commentaries on the Avestan texts
- Zand-related literature, i.e. texts based on Zand texts
- Theological works, mostly from the early Islamic period
- Juridical literature
- Moral-didactic literature
- Narrative-historical or mythological works
- Ritual literature, texts dealing with the performance of Zoroastrian rituals

In essence, ZMP literature comprises two related text layers: Zand, which is primarily a legacy of the Sasanian period (with some later additions), and ninth-century theological literature. An intermediate textual layer, the Zand-related literature is closely dependent on Zand. Besides this multi-layered religious textual material, several other significant texts have been preserved. Although these texts belong to the Zoroastrian literary corpus,

they are not strictly religious in the narrow sense. Among them are epic texts with religious motifs that are not part of the Zoroastrian doctrinal literature. Some of these texts were translated into Arabic and New Persian and are frequently cited in Muslim compositions.

2. Zoroastrian Middle Persian: Digital Corpus and Dictionary (MPCD)

A comprehensive digital corpus of Middle Persian literature is a desideratum. There are two attempts to this end: the TITUS project¹ and the Pārsīg Database². Both corpora are neither based on manuscripts, nor provide an extensive annotation, nor a dictionary. The MPCD project aims to address this gap.

To provide a platform for the study of Zoroastrian Middle Persian texts, the MPCD project³ aims at creating an annotated corpus of ZMP texts written in Pahlavi script and transmitted via codices. It will also provide a comprehensive corpus-based dictionary covering the whole lemmata in the corpus. Our corpus annotation consists of four layers:

- orthographical (transliteration) and phonographical (transcription)
- grammatical (morpho-syntactic)
- semantic
- intertextual (linking Zand texts with their Avestan original)

Within the overall structure of the project, the corpus and dictionary function as closely interrelated analytical tools. For this purpose, the project employs a web-based working environment that facilitates collaborative work on the corpus and dictionary. For the sake of verifiability, all texts in our corpus are linked to the images of the corresponding folios in the codices. In doing so, users have the possibility to switch from the dictionary via the corpus to the folio images and vice versa.

The project intends to create a platform to also be used in the future for the remaining Middle Persian sub-corpora, as well as for the corpora of other Middle Iranian languages. It is thus conducted with a view to the eventual creation of a complete dictionary of Middle Persian (incorporating all of its sub-corpora) as well as to an expansion of the corpus into the domains of other Middle Iranian languages.

3. Middle Persian Lexicography

3.1 Traditional Lexicography

Compared with ZMP, the lexicographical analysis of Manichaean Middle Persian and Middle Persian inscriptions has progressed more significantly. Durkin-Meisterernst (2004) nearly covers the whole Manichaean Middle Persian vocabulary, offering English equivalents, full attestations for each lemma, and bibliographies for scholarly discussions of individual words. This renders older lists obsolete. The vocabulary of Sasanian royal inscriptions is documented in Back (1978), while Gignoux (1972) and Humbach & Skjærvø (1980, 1983)

¹ <https://titus.uni-frankfurt.de/indexd.htm>

² <https://www.parsigdatabase.com/>

³ <https://www.mpcorpus.org>

remain valuable resources. However, revisions and updates are needed to incorporate newly discovered rock inscriptions, seals, and numismatic materials. The status of ZMP literature, the most extensive corpus, remains challenging. Three dictionaries were created in the 20th century: Nyberg (1928), MacKenzie (1971), and Nyberg (1974). These relatively brief works, based on a few non-Zand Pahlavi texts, contain approximately 3,100 lemmata in MacKenzie’s dictionary and ca. 3,000 lemmata in Nyberg’s *Manual of Pahlavi*. Although MacKenzie’s work is more popular due to its adoption of the currently preferred transcription system, Nyberg’s entries provide references and textual quotations. Both works contain some etymological information. For Zand literature whose vocabulary is not included in these dictionaries, researchers must rely on lexicographical works such as Dhabhar (1949) and Kapadia Kapadia (1953). For theological works from the early Islamic period, the largest part of the ZMP corpus, researchers should be satisfied with glossaries accompanying text editions. They are all valuable means, with their limitations, however.

3.2 Digital Lexicography

The advent of digital media in lexicography has transformed the concept of a corpus and its importance for lexicographical work (Granger et al., 2012). A digital corpus is characterized by its interconnection with the lexicon, enabling accurate representation of various levels of primary data, metadata, and structural and linguistic insights that can be utilized in lexicographic analysis. By creating a mesostructure (i.e., establishing internal connections within the lexicon) and assigning (semasiologically organized) lemmata to taxonomical concepts, it becomes possible to implement onomasiological access to the dictionary. The MPCD dictionary is the first attempt in Old and Middle Iranian Linguistics to fulfill this wish.

4. Modeling a Middle Persian-English Dictionary

4.1 A Pragmatic Approach to Dictionary Compilation

To fully grasp the compilation process, it is essential to understand the joint efforts and shared roles between philologists and computational linguists on this project. On the one hand, philologists begin their annotation tasks utilizing computationally pre-annotated texts as a base. On the other hand, computational linguists concentrate on creating a platform that enables the interconnection of corpora and dictionaries. In this process, computational linguists modify a React application developed for Kosh (Mondaca et al., 2019), a framework designed for developing and maintaining APIs for dictionaries. The glossaries and dictionaries added to Kosh are employed by philologists to search for and enrich their annotations and the dictionary. To facilitate collaboration between the two fields and to address potential issues, we jointly developed our data model using RelaxNG Compact⁴, a syntax similar to EBNF (Extended Backus-Naur Form). This cooperative methodology allows both teams to effectively communicate their ideas and needs. At the same time, this model aids the computational linguists in the project in devising a Django⁵-based model that relies on PostgreSQL, a relational database management system, as its foundation. It is crucial to account for the scope of the dictionary and

⁴ RELAX NG’s Compact Syntax

⁵ <https://www.djangoproject.com>

its possible complexities from the beginning, as modeling a dictionary for a relational database requires the establishment of a flexible model. This approach helps prevent intricate refactoring processes during the process of compiling the dictionary, which could be difficult and time-consuming to resolve.

4.2 Dictionary Compilation as Part of the Annotation Process

In the interim period, i.e. before a common interface that integrates both the corpus and dictionary is available to the philologists, they annotate and refine the corpus texts on spreadsheets. A consensus has been reached between computational linguists and philologists to employ an extended version of the CoNLL-U format ⁶. The rationale behind this extension is multifaceted, encompassing the provision of lexicographic data as well as details concerning the physical location of tokens within the context of a manuscript. Furthermore, the extension facilitates the conveyance of pertinent information regarding a text’s metastructure, including elements such as chapters and sections, thus contributing to a more comprehensive understanding of the text’s organization and layout. While modeling the data in RelaxNG provides a conceptual foundation and a depiction of the desired dictionary structure, the actual outcomes in projects are often significantly impacted by the accessible data. As a result, our principal emphasis is to employ a pragmatic strategy. This involves examining the existing data in the CoNLL format and to take into account the philologists’ requirements delineated in the RelaxNG schema in order to create a straightforward and potentially flexible data model using the Django framework.

4.3 Extending CoNLL-U for Lexicographic Purposes

In order to address specific lexicographical requirements, two additional columns have been introduced in addition to the standard CoNLL-U “lemma” field: “meaning” (see Figure 1) and “term_tech” (not in the Figure). The “lemma” and “meaning” columns are intended to provide the lemma and meaning associated with each token, while the “term_tech” column contains a selection from a finite set of values used to identify the category of the terminus technicus if the lemma in question is such one, such as “judicial”, “religious”, and so forth. These technical terms are translated and explained by philologists in a separate shared spreadsheet. This serves as a foundation for the explanation of Zoroastrian technical terms to be applied later in the dictionary. It is crucial to acknowledge that a lemma and its corresponding meaning may not always align perfectly in a one-to-one relationship for a given token. In other words, a single lemma could be associated with multiple meanings. This distinction is significant and should be considered when developing the dictionary model within the Django framework. A notable challenge encountered by the philologists during annotation pertained to the handling of multi-word expressions (MWEs) (Měchura, 2016, 2018). MWEs are not only a crucial aspect of lexicography but also have practical implications in our project. As computational linguists, our responsibility extends beyond merely modeling MWEs; we must also parse this information consistently. To address this issue, we collaborated with the philologists to denote each MWE in the extended CoNLL-U file by marking the corresponding transcription and transliteration fields with an underscore (Figure 2). This marker indicates the presence of an MWE. While processing each sentence, we then search the list of parsed lemmata for the lemmata associated with the MWE and link them accordingly.

⁶ <https://universaldependencies.org/format.html>

LEMMA	MEANING	UPOS	FEATS
pad	in	ADP	AdpType=Prep
nām	name	NOUN	Animacy=Inan
ud	and	CCONJ	_
šnāyišn	honour	NOUN	Animacy=Inan
ī	ezafe	DET	_
wisp-sūd	all-beneficial	ADJ	_
dādār	creator	NOUN	Animacy=Anim
ohrmazd	Ohrmazd	PROPN	NameType=Giv Animacy=Anim Transc=Yes

Figure 1: Snippet of the extended CoNLL-U file

huruāxm.	urwāhm	joy
ī.	ī	ezafe
və/hā.	weh	good person
vašōvəṭ.	wišuftan	destroy
/	\$	\$
u.	ud	and
āštī.	āštīh	peace
bē.	bē	away
barəṭ.	burdan	carry
_	bē burdan	drive away
u.	ud	and
anāštī.	anāštīh	discord
aṇ/dar.	andar	in
āβarəṭ.	āwardan	bring
_	anāštīh andar āwardan	breed discord

Figure 2: Handling Multi-Word Expressions

5. Modeling Lexical Data with RelaxNG Compact Syntax

5.1 Using RelaxNG Compact Syntax for Effective Scholarly Modeling and Collaboration

The RelaxNG Compact syntax is highly valued in scholarly applications for its concise nature, expressive power, and validation capabilities. This syntax provides a clear and succinct representation of schema specifications, facilitating easy comprehension and modification during development. Its expressiveness allows for the creation of complex models while still maintaining readability. The syntax’s adaptability caters to a wide range of modeling requirements and accommodates evolving project needs. Moreover, its compatibility fosters interdisciplinary collaboration by offering a shared language for schema development, enhancing communication between philologists and computational linguists. Lastly, the syntax supports XML document validation against a schema, ensuring data integrity and early identification of potential issues. These benefits make the RelaxNG Compact syntax suitable for scholarly projects requiring efficient collaboration and the creation of intricate, adaptable models.

5.2 Preliminary RelaxNG Model: An Overview and Its Role in the Project

It is key to understand that the existing RelaxNG model primarily serves as a tool for reflection, rather than a definitive schema for our project. The primary purpose of the RelaxNG model is to foster communication between the philologists and the computational linguists, allowing the philologists to convey their vision of the dictionary

```

start = dictionary # dictionary metadata in TEI header

dictionary = element dictionary {taxonomy, entry+} # reference to taxonomy and entries

taxonomy = attribute taxonomy {string} # the taxonomy is defined in taxonomy.rng

entry = element entry {
  attribute entryId {xsd:ID}, # unique ID for the entry
  element lemma {xsd:string}, # the normalized form
  headword?, # If headword is NULL, lemma (and POSs) will be showed as the headword of the entry
  | # We will use headword for e.g. verbs with different infinitive forms and present stems
  language,
  # We use crossRef for linking different lemmata to one entry,
  # for example different forms/readings of one lemma
  attribute crossRef {xsd:IDREF}, # if the entry is a cross reference the other elements will be empty

  # TODO: to generally follow TEI, we can put the following information into this structure:
  # form (including orth, forms), gramGrp (pos), def (hierarchizedMeanings = senses)
  morphology, # different components shown in different locations, see infra
  timeline, # shows the frequency of the occurrences of the lemma in different cent.
  element relativeFrequency {xsd:int}, # from corpus, per 100,000

  types, # attested types in the corpus linked to the lemma
  orthographicVariants, # union of transliterations
  occurrences,
  hierarchizedSenses,
  internalReferences?,
  equivalentents?,
  etymology?,
  biblioEntry?,
  comment?,

  attribute stage {"inprogress" | "finished" | "published"},
  attribute DOI {xsd:anyURI}
}

```

Figure 3: Entry element in RelaxNG - Current status

and its relationship to the corpus. Although the RelaxNG model is crucial to the overall process, the practical implementation and effectiveness of the schema are achieved at the Django level due to its practical implications. Figure 3 and 4 offer an overview of the “entry” model, specifically focusing on the elements “lemma” and “hierarchizedSenses”.⁷ A considerable challenge lies in identifying the most appropriate taxonomy to employ, as a consensus has not yet been reached. The continuing discussions and adjustments will aid in the development of a comprehensive and effective model, ensuring a solid foundation for the integration of the dictionary and corpus.

6. Modeling Lexical Data in Django

The MPCD project has opted to employ the Django web framework due to its nature as a proven and mature technology with a large user base and support network. This, as evidenced by its use in a number of major open-source⁸ and commercial⁹ applications, ensures that the technical foundation of the project can remain supported beyond the life span of the project itself. The framework’s built-in tools and ability to accommodate varying levels of complexity contribute to streamlined project advancement and expansion.

⁷ The RelaxNG data model for our corpus and dictionary is available on GitHub: https://github.com/middlepersian/relaxng_model

⁸ <https://github.com/mozilla/pontoon>

⁹ <https://www.instagram.com/>

Django’s features along with its modular architecture, contribute to the stability and maintainability of the MPCD project. The availability of a supportive community and extensive documentation enhances the integration of various libraries and tools. Furthermore, the implementation of a distinct app for the dictionary has the potential to streamline development and debugging processes, due to improved code organization, maintainability, and reusability. This methodological approach should facilitate subsequent project integrations and augmentations, contributing to the project’s long-term viability.

6.1 Base Models

Upon examining the data in the CoNLL-U extended files and the RelaxNG model, it becomes evident that the project must address lemmata, meanings/senses, taxonomic references, and MWEs. Simultaneously, it is essential to develop a flexible model to handle and organize semantic information. By considering these aspects, the resulting model will be better equipped to manage the complexity and nuances inherent in lexical data. It is worth noting that, as of the time of writing, the final model for the dictionary has not been determined. Internal discussions are ongoing to establish the most suitable model. Additionally, we are incorporating insights from our discussions with other scholars in the field during the international workshop, “Towards a Comprehensive Middle Persian Dictionary,” held in April 2023 at the University of Cologne. This collaborative approach is set to significantly bolster the development of a robust and comprehensive model for the project.

6.2 Lemma Model

At the heart of the dictionary application lie the “lemma” and “meaning” models. Although the RelaxNG model presents a more intricate structure outlining the appearance of an entry, the “lemma” and “meaning” models in Django are specifically designed for maximum flexibility. These models not only mirror the columns in the CoNLL-U extended files but also represent a graph, facilitating an adaptable and robust foundation for handling linguistic data.

Měchura (2016, 2018) identifies the challenges associated with managing MWEs in tree structures for lexical data and proposes addressing them as “shareable entries”. He delves into the implications of using tree structures or graphs, especially considering the human perspective. As humans may struggle with the sometimes intricate nature of graphs, XML trees serve as a compromise between human comprehension and computational processing, despite the enhanced flexibility and ease of handling offered by graphs. A primary concern with MWEs in tree structures lies in their inability to possess multiple parent nodes. Měchura introduces “graph-augmented trees” as a solution, enabling the sharing of MWEs among entries and their subsequent serialization in XML, in compliance with the XLink standard. This approach fosters more efficient and adaptable handling of MWEs in lexicographic data. Měchura’s method aligns with a feature of the Lexical Markup Framework (LMF)¹⁰, where multi-word entries can exist independently and be linked to specific senses of other entries through their ID. Our project also adopts this strategy, facilitating the seamless integration and management of MWEs in the lexicographic data

¹⁰ <https://www.lexicalmarkupframework.org>

```

stem = element stem {
  attribute type {"past" | "present"},
  xsd:string
}

# In which centuries a lemma is attested can be retrieved from the metadata of the texts,
# in which the texts are authored.
timeline = element timeline {
  attribute century {xsd:int},
  attribute freq {xsd:decimal} # frequency of the tokens linked to the lemma in the century
}

hierarchizedSenses = element hierarchizedSenses { # one per entry
  semantic+
}

semantic = element semantic {
  attribute serialId {xsd:int}, # internal numbering

  # reference to the parent node to produce a flexible hierarchy
  # In MPCD, we will often use only two levels for grouping the senses.
  # Nevertheless, we would like to be able to produce a hierarchy with
  # more levels in case of more complex lemmata.
  node?,
  semanticCore,
  element morphologicalForms {morphologicalForm+},
  element mwes {mwe*} # alphabetic sorted
}

semanticCore = element semanticCore {
  element sense {xsd:string}, # in the case of term. tech. = its definition

  # explanation may include detailed description of the use of the lexeme,
  # reasons for transcription, and prototypical examples
  element explanation {xsd:string}?,

  # grammar may include some grammatical explanations about the lemma
  # including valences
  element grammar {xsd:string}?,

  # reference to the taxonomy
  # We will take the liberty to eventually link a sense to more than one concept
  element semanticDomain {concept+}?,
  termTech?,

```

Figure 4: Selected elements from an entry in RelaxNG - Current status

structure. Currently, our software functions as a relatively basic yet efficient dictionary writing system, and we by now have no immediate or long-term plans to offer XML-based views to users. Nevertheless, to comprehend the process of serializing data in XML format is essential for improving the system's capabilities and to manage the lexicographic data generated by our project effectively. Acquiring this knowledge will facilitate the continuous refinement and expansion of the software, ensuring that it remains a valuable and adaptable resource for both users and researchers.

During the development phase, we have intentionally not implemented an entry model, as its final structure is still evolving. However, based on our current understanding, we have identified the core components that are likely to be included in the entry. In addition to standard elements such as word(form) and language, a lemma may possess

```

class Lemma(models.Model):
    id = models.UUIDField(primary_key=True, default=uuid_lib.uuid4, editable=False)
    word = models.CharField(max_length=100)
    language = models.CharField(max_length=3, null=True, blank=True)
    categories = ArrayField(
        models.CharField(max_length=50, blank=True, null=True), null=True, blank=True
    )
    multiword_expression = models.BooleanField(default=False)
    related_lemmas = models.ManyToManyField(
        "self", blank=True, related_name="lemma_related_lemmas", through="LemmaRelation"
    )
    related_meanings = models.ManyToManyField(
        "Meaning",
        blank=True,
        related_name="lemma_related_meanings",
        through="LemmaMeaning",
    )
    created_at = models.DateTimeField(auto_now_add=True)

    history = HistoricalRecords()

class Meta:
    constraints = [
        models.UniqueConstraint(
            fields=["word", "language"], name="word_language_lemma"
        )
    ]
    indexes = [
        models.Index(fields=["word", "language"]),
    ]
    ordering = ["word"]

```

Figure 5: Lemma Model in Django - Current status

related objects of the same type that can be categorized into a specific classification (e.g., `term_tech`) or classified as a MWE. Furthermore, it may have associated meanings. This preliminary understanding informs our current approach, allowing for greater flexibility and adaptability as the project evolves. The inherent simplicity of our system enables a lemma to be associated with multiple lemmata and meanings, even across different languages. Although the primary objective of this project is to develop a Middle Persian-English dictionary, we have designed the model with the potential to create multiple dictionaries with the existing lemmata, rather than exclusively focusing on a single language pair. This flexibility allows for broader applications and adaptability in various linguistic contexts.

6.3 Meaning Model

The meaning model (Figure 6) is simple, yet effective. It comprises the meaning itself, an associated language, and a boolean flag indicating whether the meaning is related to a lemma. This design choice stems from the existence of another model, section, which can have a meaning. Sections are especially useful for sentences, as we consider them as ranges of tokens. By incorporating this boolean field, we can more efficiently filter meanings without the need to create additional models for sentence translations or other potential objects that may require translations in the future. This streamlined approach promotes versatility and adaptability for a variety of translation requirements.

```

class Meaning(models.Model):
    id = models.UUIDField(primary_key=True, default=uuid_lib.uuid4, editable=False)
    meaning = models.TextField(null=True, blank=True, db_index=True)
    lemma_related = models.BooleanField(default=True)
    language = models.CharField(max_length=10, blank=True, null=True, db_index=True)
    related_meanings = models.ManyToManyField('self', blank=True, related_name='meaning_related_meanings')
    created_at = models.DateTimeField(auto_now_add=True)

    def related_lemmas(self):
        return self.lemma_related_meanings.all()

    history = HistoricalRecords()
    class Meta:
        constraints = [
            models.UniqueConstraint(
                fields=['meaning', 'language'], name='meaning_language_meaning'
            )]
        ordering = ['meaning']
        indexes = [
            models.Index(fields=['meaning', 'language']),

```

Figure 6: Meaning Model in Django - Current status

6.4 Token Model

At the heart of the platform lies the token model, which has been designed to connect with one or more lemmata and meanings. This design allows lemmata and meanings to relate to each other, with the relationship originating from the lemma. Although the relationship is symmetrical, enabling access to the lemma from the meaning, it must be initiated from the lemma. By parsing data from the extended CoNLL-U files, we can create lemmata along with their associated meanings and independently link lemmata and meanings to a token. This approach enables us to access tokens within meanings, providing us with increased flexibility and adaptability in managing the corpus data.

7. Exploring Middle Persian Lexical Resources

7.1 APIs for Efficient and Collaborative Historical Linguistics Research

APIs function as vital intermediaries that enable communication and data exchange between disparate software components, playing a critical role in modern software development. As highlighted by Amundsen (2020), they offer several key advantages, including reduced computational time and cost, facilitated ease of computations, and the ability to tackle previously unresolved issues. Furthermore, APIs contribute to standardization by providing a consistent and structured method for software components to interact, simplifying the development process. They also promote modularity, allowing developers to create and modify individual components without disrupting the entire system, thus enhancing maintainability and flexibility. APIs facilitate extensibility, as they enable software to be easily expanded or integrated with new features and services. Additionally, they improve security by enabling controlled access to specific functionalities, ensuring that sensitive data remains protected. APIs hold significant relevance for lexical data in historical languages such as Middle Persian, primarily due to their capacity to facilitate access, promote interoperability, and support data enrichment. By offering standardized methods for interaction between software applications, APIs enable researchers and developers to access linguistic information without in-depth knowledge of the underlying data structure. This accessibility encourages collaboration and allows for integration of lexical data from


```

class Token(models.Model):
    id = models.UUIDField(primary_key=True, default=uuid_lib.uuid4, editable=False)
    number = models.FloatField(null=True, blank=True)
    number_in_sentence = models.FloatField(blank=True, null=True)

    root = models.BooleanField(default=False)
    word_token = models.BooleanField(default=True)
    visible = models.BooleanField(default=True)

    text = models.ForeignKey('Text', on_delete=models.CASCADE, null=True, blank=True, related_name='token_text')
    image = models.ForeignKey('images.Image', on_delete=models.CASCADE, null=True, blank=True, related_name='token_image')

    language = models.CharField(max_length=3, null=True, blank=True)
    transcription = models.CharField(max_length=50)
    transliteration = models.CharField(max_length=50, blank=True)
    lemmas = models.ManyToManyField('dict.Lemma', blank=True, through='TokenLemma', related_name='token_lemmas')
    meanings = models.ManyToManyField('dict.Meaning', blank=True, through='TokenMeaning', related_name='token_meanings')

    avestan = models.TextField(null=True, blank=True)

    previous = models.OneToOneField('self',
                                    related_name='next',
                                    blank=True,
                                    null=True,
                                    on_delete=models.SET_NULL, db_index=True)

    gloss = models.TextField(blank=True, null=True)

    multiword_token = models.BooleanField(default=False)
    multiword_token_number = ArrayField(models.FloatField(blank=True, null=True), null=True, blank=True)
    related_tokens = models.ManyToManyField('self', blank=True)
    created_at = models.DateTimeField(auto_now_add=True)

    history = HistoricalRecords()

```

Figure 7: Token Model in Django - Current status

various sources, fostering innovative research approaches and insights. Furthermore, APIs enable efficient updating and customization, ensuring that users have access to the most current and accurate information tailored to their specific research needs. In the context of the project, two distinct APIs have been developed: a REST (Representational State Transfer) API (Fielding, 2000) and a GraphQL API (GraphQL, 2021). Each offers unique advantages that cater to different user requirements and preferences, enhancing efficiency, flexibility, and collaboration across various platforms and systems. The MPCD project has been designed with an API-centric approach. Through its web application, it offers access to APIs for both the dictionary and corpus apps, which are supported by Django and specific API libraries. Furthermore, the web application displays the most recent publications related to the project through the Zotero API. A team of philologists diligently curates a Zotero catalog featuring a comprehensive bibliography of Middle Persian literature and publications associated with the project.

7.2 REST API

The REST API follows a well-defined architectural style, which simplifies the development process due to its standardized and straightforward nature. It is designed to support scalability, thereby ensuring seamless access to data and services as the user base grows. Moreover, the REST API enables caching of responses, reducing the load on servers and enhancing performance for frequently accessed data. To further facilitate the development process and improve accessibility for developers, we have integrated a comprehensive and interactive documentation interface for the API. This integration allows developers to explore the API's endpoints, understand the data structures, and test API requests and responses directly within the documentation. By providing this interface, we aim to foster

a user-friendly environment for developers to interact with and utilize our lexical data in a more efficient and effective manner.

7.2.1 Integration with Large Language Models

The adoption of a REST API offers significant advantages, such as the potential for interoperability with large language models. This approach enables seamless integration of our lexical data with advanced natural language processing systems, supporting the development of plugins to enhance user experience and expand the applications of historical language data. The flexibility of this project promotes collaboration and innovation within the field of historical linguistics, aligning it with the rapidly evolving sphere of artificial intelligence and language modeling research. Using APIs has been proven to be effective when applied to models with fewer parameters than large-scale language models, indicating promising future developments in this area (Schick et al., 2023).

7.3 GraphQL API

The GraphQL API offers flexible querying, allowing clients to request precisely the data they need, which results in enhanced efficiency by reducing the amount of unnecessary information transferred. The GraphQL API uses a single endpoint for all data requests, simplifying the management of multiple resources and streamlining the development process. One of the reasons GraphQL has become so popular among frontend developers is its ability to enhance productivity and improve the overall development experience. With its strong typing system, developers can easily discover the available data and understand the structure of the API, which leads to fewer errors and better maintainability.

7.4 React Web Application

In our web application, which is built using the React JavaScript library and the Relay data-fetching framework, we have opted to employ the GraphQL API. This choice offers several advantages in the context of a React-based application. GraphQL's flexibility in querying allows the React components to request precisely the data they need, ensuring an efficient data transfer and minimizing over-fetching. This feature is particularly beneficial in the context of a dynamic web application, where various components might have specific data requirements. Furthermore, the combination of GraphQL with Relay allows for seamless integration with the React library, enabling efficient and scalable data-fetching operations. Relay manages the GraphQL requests and optimizes data fetching, reducing the complexity of managing data within the application and promoting a more maintainable codebase.

8. Search

One of the most pertinent features a dictionary should possess is a well-organized macrostructure to facilitate effective information retrieval. Digital dictionaries hold a distinct advantage over their printed counterparts in terms of accessing efficient search modalities, as opposed to the wordlists commonly found in printed dictionaries. Although

traditional databases, including both NoSQL and SQL types, offer full-text search capabilities, they are not specifically designed for search purposes like dedicated search engines. To address this need, we employ Elasticsearch¹¹ as a search engine, enhancing the search experience and improving the overall usability of the digital dictionary.

8.1 Main Search Modi

Prefix, regular expression (regex), match, and wildcard searches are advantageous for retrieving lexical data due to their ability to accommodate diverse query patterns and efficiently filter relevant information from extensive datasets. These search techniques enable users to explore linguistic data with greater precision and flexibility. Prefix searches allow users to identify words or phrases beginning with a specified sequence of characters, which is particularly helpful when investigating lexical data for morphological patterns or etymological connections. By focusing on the initial characters, prefix searches facilitate the discovery of related terms and enable researchers to gain insights into language structure and development. Regular expression searches, or regex searches, offer a powerful method for retrieving lexical data based on complex patterns. By employing a combination of symbols and characters, users can create highly specific search criteria that match a wide variety of linguistic patterns. This flexibility allows researchers to uncover intricate relationships between words or phrases and investigate language phenomena that may otherwise be difficult to discern. Match searches provide a more straightforward approach to lexical data retrieval, locating exact or partial matches within the dataset. This method proves advantageous in cases where users are searching for specific terms or phrases, as it ensures that only the most relevant results are returned. Match searches contribute to the efficiency and accuracy of information retrieval in linguistic research. Wildcard searches introduce an additional layer of flexibility by allowing users to substitute one or more characters within a search query with a wildcard symbol. This functionality enables researchers to locate words or phrases with varying character combinations, accommodating uncertain or incomplete search criteria.

8.2 Implementing Semantic Search

Elasticsearch enables the incorporation of vector embeddings into the search process, an approach commonly referred to as semantic search. The advantage of this method lies in its ability to retrieve documents, lemmata, or meanings that are related to the search term or the terms found, based on their appearance in a corpus or their relationships within a language model. While numerous embedding models are available for English, resources for Middle Persian are lacking. Consequently, the initial implementation of vector search will focus on representing Middle Persian meanings in English. For this purpose we will develop a word-embedding model for Middle Persian. This approach aims to facilitate the discovery of meaningful connections and insights in Middle Persian linguistic data, thereby enhancing the overall search experience and supporting more comprehensive research.

9. Future Enhancements

Integrating a dictionary with a corpus requires technical development and effective communication between philologists and computational linguists. As we continue to refine

¹¹ <https://www.elastic.co/>

and expand this integration, we are actively seeking areas for improvement and further development.

In an effort to facilitate access to our data for as many researchers as possible, we will provide our corpus data in both TEI and CoNLL formats. Moreover, we might be able to make our lexical data available in both TEI-Lex0 (Tasovac et al., 2018) and Ontolex (McCrae et al., 2017) formats. Additionally, we aim to integrate our data with large language models, which are undergoing rapid development at the time of writing.

10. References

- Amundsen, M. (2020). *Design and build great web APIs: robust, reliable, and resilient*. The pragmatic programmers. Raleigh, North Carolina: The Pragmatic Bookshelf.
- Back, M. (1978). *Die sassanidischen Staatsinschriften. Studien zur Orthographie und Phonologie des Mittelpersischen der Inschriften zusammen mit einem etymologischen Index des mittelpersischen Wortgutes und einem Textcorpus der behandelten Inschriften*. Number 18 in Acta Iranica. Leiden: Brill.
- Dhabhar, B.N. (1949). *Pahlavi Yasna and Visperad*. Bombay: The Trustees of the Parsee Panchayet Funds and Properties.
- Durkin-Meisterernst, D. (2004). *Dictionary of Manichean Middle Persian and Parthian (Dictionary of Manichaean Texts. Vol. III Texts from Central Asia and China, Part 1)*. Turnhout: Brepols.
- Fielding, R.T. (2000). *Architectural Styles and the Design of Network-based Software Architectures*. Ph.D. thesis, University of California, Irvine. URL https://www.ics.uci.edu/~fielding/pubs/dissertation/fielding_dissertation.pdf.
- Gignoux, P. (1972). *Glossaire des inscriptions pehlevies et parthes*. Number 1 in Corpus Inscriptionum Iranicarum, Supplementary series. London: Lund Humphries.
- Granger, S., Paquot, M., Granger, S. & Paquot, M. (eds.) (2012). *Electronic Lexicography*. Oxford, New York: Oxford University Press.
- GraphQL (2021). GraphQL Specification. <https://spec.graphql.org>. Accessed: April 18, 2023.
- Humbach, H. & Skjærvø, P.O. (1980). *The Sassanian inscription of Paikuli. Part 2: Synoptic Tables*. Wiesbaden: Dr. Ludwig Reichert.
- Humbach, H. & Skjærvø, P.O. (1983). *The Sassanian inscription of Paikuli. Part 3.1: Restored text and translation. Part 3.2: Commentary*. Wiesbaden: Dr. Ludwig Reichert.
- Kapadia, D.D. (1953). *Glossary of Pahlavi Vendidad*. Bombay.
- MacKenzie, D.N. (1971). *A concise Pahlavi dictionary*. London: Oxford University Press.
- Macuch, M. (2009). Pahlavi literature. In R.E. Emmerick & M. Macuch (eds.) *The literature of pre-Islamic Iran. Companion volume I to 'A history of Persian literature'*, number 17 in A history of Persian literature. London: I.B. Tauris, pp. 116–196.
- McCrae, J.P., Gil, J., Gràcia, J., Bitelaar, P. & Cimiano, P. (2017). The OntoLex-Lemon Model: Development and Applications. In *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference*. URL <https://elex.link/elex2017/wp-content/uploads/2017/09/paper36.pdf>.
- Mondaca, F., Schildkamp, P. & Rau, F. (2019). Introducing Kosh, a Framework for Creating and Maintaining APIs for Lexical Data. In *Electronic Lexicography in the 21st Century. Proceedings of the eLex 2019 Conference, Sintra, Portugal*. Brno: Lexical Computing CZ, pp. 907–21. URL https://elex.link/elex2019/wp-content/uploads/2019/09/eLex_2019_51.pdf.

- Měchura, M. (2016). Data Structures in Lexicography: from Trees to Graphs. In *The 10th Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2016, Karlova Studanka, Czech Republic, December 2-4, 2016*. pp. 97–104. URL <http://nlp.fi.muni.cz/raslan/2016/paper04-Mechura.pdf>.
- Měchura, M. (2018). Shareable Subentries in Lexonomy as a Solution to the Problem of Multiword Item Placement. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (eds.) *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*. Ljubljana, Slovenia: Ljubljana University Press, Faculty of Arts, pp. 223–232.
- Nyberg, H. (1928). *Hilfsbuch des Pehlevi*. Uppsala: Almqvist & Wiksell. {issued:1928/1931}.
- Nyberg, H. (1974). *A manual of Pahlavi. Part II: Glossary*. Wiesbaden: Harrassowitz.
- Rezania, K. (2023). Zoroastrianism in Early Islamic Period: its participation in °Abbāsīd theological-philosophical discourse and its absence in the transmission of Sasanian culture. In S. Heidemann & K. Mewes (eds.) *The reach of empire – the Early Islamic Empire at work*, volume 2 of *Studies in the History and Culture of the Middle East*. Berlin: De Gruyter.
- Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., Cancedda, N. & Scialom, T. (2023). Toolformer: Language Models Can Teach Themselves to Use Tools. URL <http://arxiv.org/abs/2302.04761>. ArXiv:2302.04761 [cs].
- Tasovac, T., Romary, L., Banski, P., Bowers, J., de Does, J., Depuydt, K., Erjavec, T., Geyken, A., Herold, A., Hildenbrandt, V., Khemakhem, M., Lehečka, B., Petrović, S., Salgado, A. & Witt, A. (2018). TEI Lex-0: A baseline encoding for lexicographic data. <https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html>.

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

Francisco Mondaca¹, Philip Schildkamp¹, Felix Rau², Luke Günther¹

¹ Cologne Center for eHumanities, University of Cologne

² Data Center for the Humanities, University of Cologne

E-mail: f.mondaca@uni-koeln.de, philip.schildkamp@uni-koeln.de, f.rau@uni-koeln.de, luke.guenther@uni-koeln.de

Abstract

This paper presents the Kosh Suite, an API-centric framework designed to efficiently manage and access lexical data. The Kosh Suite aims to address the challenges in working with XML and lexical data, providing a flexible and customizable solution. The Kosh Suite architecture features a backend powered by Elasticsearch, which forms the foundation for efficient data management and retrieval. This backend offers two APIs per dataset for accessing the lexical data - a REST API and a GraphQL API per dataset. In addition, the Kosh Suite includes a frontend implemented in form of a React-based user interface, ensuring a user-friendly experience and adaptability to various use cases. Deployment specifications are described for the backend, with reference implementations for FreeDict and Cologne Sanskrit Dictionaries (CSDS). Future enhancements include asynchronous request handling using FastAPI, integration with CSV files, and leveraging advancements in large language models (LLMs). These improvements have the potential to significantly enhance the system's performance and accessibility, promoting the integration of underrepresented languages into mainstream LLMs.

Keywords: api;rest;graphql;xml;elastic

1. Introduction

1.1 An API-Centric Framework for Lexical Data

Application Programming Interfaces (APIs) play a crucial role in enabling efficient data exchange and remote functionality invocation among distributed applications. APIs, through their well-documented, stable, and user-friendly services, present a sustainable alternative to both monolithic web applications, which frequently pose maintenance challenges, and data repositories, which necessitate computational processing for effective use. According to Amundsen (2020) APIs reduce computational time and cost, facilitate easier computations, and tackle previously unresolved issues. Importantly, APIs are intended not only for application integration but also for human interaction. As such, recognizing the target consumer and use case is of utmost importance. APIs accommodate a wide array of devices and software stacks, including Java-enabled smartphone applications and Python-based desktop programs. As a result, development efforts concentrate on the data output and functionality of the API. This emphasis indirectly boosts the sustainability of the underlying system supporting the API, as the computations take place within this system rather than prioritizing data presentation. In case of system errors, all consumers are affected, highlighting the necessity for a robust and dependable API infrastructure.

As of Kosh’s¹ initial development (Mondaca et al., 2019b), and persisting until now, no frameworks have been exclusively dedicated to APIs for lexical data. Lexical data, with its inherent adaptability, can be applied in single-page dictionary applications, incorporated into corpora, and utilized in diverse NLP tasks. While alternative tools, such as Lexonomy (Měchura, 2017), permit users to create and publish dictionaries that can be integrated into the CLARIN network or have their data downloaded in XML format, these tools lack web API offerings. Despite the growing importance and application of APIs in the industry (Medjaoui et al., 2021), such focus has not yet been reflected in the academic advancements within the field of lexicography.

1.2 Background on XML and Lexical Data

XML (eXtensible Markup Language) is a popular serialization format for lexical data. The hierarchical and structured format of XML allows lexicographers to represent complex linguistic data in a clear and organized way, thus facilitating easier processing, query, and share between different systems and applications. In addition, the standardization of XML as a widely adopted format for data serialization lead to many tools and libraries being available for parsing and processing XML data, making it a reliable and well-supported choice for lexicographers. The extensibility of XML also makes it a beneficial choice for digital lexicography, as it enables lexicographers to customize lexical data structures to meet the specific needs of a given project. This allows for the creation of specialized lexicons and dictionaries that cater to specific domains or user groups, and can be used in a variety of contexts, such as natural language processing and machine translation. The development of the TEI (Text Encoding Initiative) has been a relevant endeavor in the digital humanities, as it provides a flexible model that can cover multiple needs of different communities working with digital data. However, the flexibility of the TEI Guidelines (TEI Consortium, 2023) for encoding dictionaries often results in inconsistent encoding, which hinders the processing, analysis, and sharing of lexical data across systems and applications. To address this challenge, TEI Lex-0 (Tasovac et al., 2018) was created to establish a standardized and interoperable format for encoding lexical data within a community of practice. TEI Lex-0 offers a baseline encoding and target format that ensures the interoperability of heterogeneously encoded lexical resources, providing a lightweight and standardized format for creating structured and machine-readable lexical resources. This makes it easy for lexicographers to use and adopt while still adhering to a consistent model for searching, visualizing, or enriching multiple lexical resources. The use of TEI Lex-0 enables the building and management of large-scale lexical infrastructures by facilitating the creation of high-quality and interoperable lexical resources that can be easily processed, analyzed, and shared across different systems and applications. XML-encoded lexical datasets are commonly used due to their flexibility, hierarchical structure, cross-platform compatibility, standardization, and extensibility. While TEI-encoded dictionaries are a relevant part of the digital lexicography scholarly landscape, there are also many other types of XML-encoded dictionaries used in academia and industrial contexts. These dictionaries can have a wide range of structures, and searchable fields can vary greatly between them. By initially focusing solely on XML as an input format, we could concentrate on designing a straightforward and efficient framework. Although data in other formats must be converted into XML to be used by Kosh, multiple tools are available for accomplishing this transformation.

¹ <https://kosh.uni-koeln.de>

1.3 Motivation for the Kosh Suite

Since its inception in 2019, Kosh has been employed in various research and community projects, including the Cologne Digital Sanskrit Dictionaries (2019), VedaWeb (Kiss et al., 2019), Zoroastrian Middle Persian Corpus and Dictionary (MPCD) (Mondaca et al., 2022), and FreeDict². The VedaWeb project was the first to employ Kosh, linking every token of the RigVeda to a lemma of Grassman’s dictionary (Grassmann, 1873) through a Kosh API and its corresponding information (Mondaca et al., 2019a). The primary impetus behind Kosh’s development was to foster the decentralization and increased utilization of lexical data pertaining to datasets that are typically underrepresented in the digital realm for various reasons. Although APIs enable developers and researchers to craft bespoke applications that harness these APIs, it has been observed that numerous scholars possessing XML-encoded lexical data often lack the necessary resources to devise user-centric client applications tailored for data exploration and retrieval. To address this issue, we have developed a client application as part of the Kosh Suite, which consumes and visualizes data provided by the Kosh APIs, thus enabling users to search through the data. By providing a user-friendly interface for accessing data, we aim to make consuming and serving XML-encoded lexical data more accessible to researchers and scholars who may not have the resources or expertise to develop their own applications for using the Kosh APIs.

2. Architecture

2.1 Overview

The Kosh Suite is a comprehensive software framework designed to manage and access lexical data in XML format. The structured nature of XML, facilitated by the use of tags, allows for easy identification and navigation of different elements of the data. The framework relies on Elasticsearch, a search engine that can be used to index lexical data, making it easily searchable. Kosh provides two APIs per dataset: a REST (Representational State Transfer) API (Fielding, 2000) and a GraphQL API (GraphQL, 2021). The REST API allows for read operations on the indexed data and is easy to use with a wide range of programming languages and frameworks. The GraphQL API allows clients to request exactly the data they need and retrieve multiple resources in a single request. Using these APIs, the Kosh frontend offers a user-friendly interface for searching and filtering the indexed data. The Kosh Suite’s frontend is developed using React³ and Tailwind CSS⁴, offering a web-based user interface to search through the lexical data provided by the Kosh APIs. Users have the ability to perform complex searches, filter results based on various criteria, and view detailed information about the lexical entries. Additionally, users have the option to configure the fields they wish to search on through a JSON file, ensuring that the indexed data remains easily searchable.

² <https://freedict.org/>

³ <https://react.dev>

⁴ <https://tailwindcss.com>

2.2 Backend

2.2.1 Elasticsearch

Elasticsearch⁵ is an open-source distributed search engine that is specifically designed to handle large datasets with high performance and scalability. It offers advanced features such as full-text search, faceting, and geospatial search, which make it a popular choice for indexing and searching large datasets. Elasticsearch is built on top of Apache Lucene⁶, a powerful and widely used search library that provides advanced search capabilities. The Elasticsearch platform is broadly used in various applications, including digital lexicography, due to its ability to handle large datasets in near real-time. The full-text search capabilities of Elasticsearch enable users to search for relevant data using natural language queries, making it an ideal choice for indexing and searching textual data. In the context of the Kosh Suite, Elasticsearch serves as the backend for indexing and searching lexical data in XML format. The system's ability to index large amounts of data quickly is crucial for digital lexicography, which often deals with massive datasets. Elasticsearch's full-text search capabilities and real-time search enable complex searches and filtering based on multiple criteria. Additionally, Elasticsearch's distributed architecture facilitates easy scaling of the system, enabling high availability and fault tolerance. The system can be deployed across multiple nodes, providing redundancy and load balancing, making it suitable for large-scale projects. Elasticsearch also offers APIs, making it easy to integrate with other systems and applications, enhancing its flexibility and versatility.

2.2.2 REST API

A REST API is an interfacing standard for creating web services that enables communication between different systems. It follows a client-server architecture and allows for read and write operations on the indexed data, making it easy to use with a wide range of programming languages and frameworks. In the context of the Kosh Suite, the REST API enables read-only access to the indexed lexical data in XML format. Kosh only accepts HTTP GET requests on the REST API, as it is used for reading the indexed data. Users can perform complex searches and filter results based on various criteria using the REST API. It is designed to be easily integrated with other systems and applications, further enhancing its versatility. OpenAPI⁷ provides a user-friendly interface for developers to interact with the Kosh APIs. It offers a visual representation of the API's endpoints and parameters, making it easier to understand how to use the API. Additionally, OpenAPI offers interactive documentation, allowing developers to test the API's endpoints and view the results in real-time. This feature saves time and improves efficiency as it eliminates the need to manually test each endpoint using external tools.

2.2.3 GraphQL API

GraphQL is a query language that provides a more flexible approach to retrieving data compared to traditional REST APIs. One of the key benefits of GraphQL for managing lexical data in Kosh is its ability to allow clients to request only the data they need. This

⁵ <https://www.elastic.co/elasticsearch/>

⁶ <https://lucene.apache.org>

⁷ <https://www.openapis.org>

means that clients can retrieve precisely the data they require without being limited by the constraints of a fixed data structure. As a result, GraphQL provides greater flexibility and reduces the amount of data that needs to be transferred over the network, which can improve the efficiency of data retrieval. In the context of digital lexicography, this flexibility is particularly valuable, as it allows lexicographers to create more complex data structures without worrying about the impact on data retrieval performance. This can enable the development of more advanced and customizable search interfaces for lexical data. Additionally, the ability to perform nested queries enables clients to retrieve related data in a single request, reducing the number of requests required to access all the necessary data. The use of GraphQL in Kosh provides a powerful and flexible tool for managing and retrieving lexical data.

2.3 Frontend

2.3.1 React-based User Interface

React is a widely-used and popular JavaScript library for building user interfaces (UIs). Its component-based architecture enables high modularity and reusability in building UIs, while its virtual DOM feature provides performance advantages by selectively updating only the modified parts of the UI. Moreover, React boasts a vast ecosystem of libraries and tools that facilitates rapid development of complex applications. When combined with a GraphQL API, React offers various advantages that can enhance the performance of web applications that manage large datasets, such as lexical data. By allowing clients to request only the necessary data, GraphQL minimizes the volume of redundant data transferred over the network, leading to faster data retrieval and improved application performance. GraphQL's capability to execute nested queries simplifies data fetching, reducing the number of requests necessary to retrieve data. These benefits are especially relevant when the network data is limited, as GraphQL streamlines data retrieval, resulting in more efficient resource utilization, faster data loading times, and improved application responsiveness. React's component-based structure also promotes code efficiency and reusability, supporting the development of scalable and efficient applications.

2.3.2 Tailwind CSS for UI-Customization

Tailwind CSS⁸ is a utility-first framework that offers more customization and flexibility than alternatives like Bootstrap⁹ or Foundation¹⁰. Its low-level utility classes can be composed to create unique designs, and it offers a vast collection of predefined classes for easy modification. It also provides responsive design classes for optimization across screen sizes and devices. In the Kosh Suite, Tailwind is used for frontend development, offering advantages such as easy customization, and responsive design options. This ensures a user-friendly interface that remains accessible and usable across devices and platforms while giving developers greater control over design and layout.

⁸ <https://tailwindcss.com>

⁹ <https://getbootstrap.com>

¹⁰ <https://get.foundation>

3. Deployment

3.1 Backend

Kosh is designed to offer minimal prerequisites and an uncomplicated setup. It can be deployed on Linux systems or through Docker. For deployment, Kosh requires a Kosh dotfile, a JSON file with mappings, and it processes files in XML format.

3.1.1 Kosh Dotfile

This file provides details on: (i) the name of the dataset's index; (ii) the location of XML files containing lexical information; (iii) the location of the configuration JSON file, utilized for parsing and configuring Elasticsearch; (iv) the title of the dataset; (v) any other, additional metadata, that should be made available through the Kosh API.

3.1.2 JSON Mappings

The JSON file referenced from within a Kosh dotfile contains information about XML nodes and their subnodes, specified in XPath 1.0 notation, which is used for indexing. It also includes details about handling different data types, such as “keyword” for unprocessed strings and “text” for preprocessed strings analyzed by Elasticsearch. Additionally, the file provides instructions on handling arrays of elements and automatically generating entry IDs if not present in the dictionary. Lastly, it outlines the default indexing behavior, which involves indexing the entire entry without analyzing XML tags.

3.1.3 Endpoints

Kosh offers a REST and a GraphQL API for each dataset indexed, and it also indicates in JSON which datasets are accessible for each Kosh instance. Each dataset comprises: (i) information about the queryable fields, such as “id”, “lemma”, and “sense”; (ii) available Elasticsearch query types, like “wildcard” and “prefix”; (iii) the number of entries available. This endpoint information holds computational significance, as it can be utilized by other applications, including the Kosh Suite frontend. For instance, detailed information about each dataset is accessible at <https://kosh.uni-koeln.de/api> for all datasets deployed under a specific Kosh instance. Similarly, individual datasets also contain this information, as seen in https://kosh.uni-koeln.de/api/de_alcedo.

3.2 Frontend

Analogous to the backend, the frontend is deployed with a Docker container. As illustrated in Section 3.1.3 Endpoints, the frontend capitalizes on the information furnished by Kosh's backend to dynamically generate user interface components for the purpose of querying and exhibiting data provided by the backend. This distinctive attribute permits Kosh to accommodate a diverse range of datasets while simultaneously affording users the flexibility to establish naming conventions, as they possess the freedom to determine field names in the JSON file delivered to the backend.

4. Reference Implementations

4.1 FreeDict

4.1.1 About FreeDict

The FreeDict project aims to serve as the preeminent repository for free bilingual dictionaries. These resources not only come at no cost but also confer the rights to examine, modify, and adapt them, provided that users extend these liberties to others. Established in 2000, FreeDict currently offers a compendium of more than 200 multilingual dictionaries, spanning approximately 45 languages, with its continuous growth thanks to the contributions of its members.

4.1.2 FreeDict Implementation

Freedict hosts its hand-written dictionaries in TEI format on GitHub, while also provides a comprehensive list of all its dictionaries in JSON format. There is a repository on GitHub that generates the necessary data for Kosh, including Kosh dotfiles and JSON files, to enable deployment¹¹. The implementation of this repository facilitates continuous monitoring of updates to the database, thereby ensuring synchronization between both Kosh and the FreeDicts. Kosh APIs generated for FreeDict are available at: <https://kosh.uni-koeln.de/freedict>

4.2 Cologne Digital Sanskrit Dictionaries (CDS D)

4.2.1 About the CDS D

The Cologne Digital Sanskrit Dictionaries (CDS D) project began with the efforts of Thomas Malten from the University of Cologne in 1994, initially focusing on digitizing the Monier-Williams Sanskrit-English Dictionary (Monier-Williams, 1899). As of now, the project boasts a collection of 38 dictionaries. The CDS D portal relies on data hosted on GitHub¹², where a diverse team of scholars and users from around the world work together to maintain and improve the available information.

4.2.2 CDS D Implementation

The CDS D project initially utilized a unique markup language for encoding. Presently, the dictionaries on the CDS D are derived from XML files that have been transformed using various scripts, available on a GitHub repository¹³. We employ these repositories in conjunction with a third one¹⁴ to generate the required JSON and Kosh dotfiles essential for deploying these dictionaries with Kosh. Access to the Kosh CDS D APIs is available at: <https://kosh.uni-koeln.de/cdsd>

¹¹ <https://github.com/freedict/fd-kosh>

¹² <https://github.com/sanskrit-lexicon/csl-orig/tree/master/v02>

¹³ <https://github.com/sanskrit-lexicon/csl-pywork>

¹⁴ <https://github.com/cceh/csl-kosh>

5. Exploring Search Queries

In this section, we provide a variety of search examples that can be used with Kosh, illustrating both GraphQL and REST queries. For GraphQL, the necessary parameters to be inputted by the user for effective interaction with the GraphQL interface are explicitly provided.

1. Term: The term query finds documents that contain the exact term specified in the field specified. Examples:

- RESTful: https://kosh.uni-koeln.de/api/de_alcedo/restful/entries?field=lemma&query=santiago&query_type=term&size=20
- GraphQL: https://kosh.uni-koeln.de/api/de_alcedo/graphql

```
{
  entries(queryType: term,
    query: "santiago",
    field: lemma,
    size: 20) {
    lemma
    sense
  }
}
```

2. Fuzzy: The fuzzy query generates all possible matching terms that are within a certain maximum edit distance, allowing for variations in the terms. Examples:

- RESTful: https://kosh.uni-koeln.de/api/de_alcedo/restful/entries?field=lemma&query=ica&query_type=fuzzy&size=50
- GraphQL: https://kosh.uni-koeln.de/api/de_alcedo/graphql

```
{
  entries(queryType: fuzzy,
    query: "ica",
    field: lemma,
    size: 50) {
    lemma
    sense
  }
}
```

3. Match: The match query is a standard query that is useful for single word and phrase queries. Examples:

- RESTful: https://kosh.uni-koeln.de/api/ducange/restful/entries?field=xml&query=viispublicis&query_type=match&size=30
- GraphQL: <https://kosh.uni-koeln.de/api/ducange/graphql>

```
{
  entries(queryType: match,
    query: "viis publicis",
    field: xml,
    size: 30) {
    lemma
  }
}
```

```

        xml
      }
    }
  }

```

4. Match Phrase: The match phrase query is like match, but it only returns documents where the matched words are in the order specified in the query. Examples:

- RESTful: https://kosh.uni-koeln.de/api/de_alcedo/restful/entries?field=sense&query=reynodechile&query_type=match_phrase&size=300

- GraphQL: https://kosh.uni-koeln.de/api/de_alcedo/graphql

```

{
  entries(queryType: match_phrase,
    query: "reyno de chile",
    field: sense,
    size: 300) {
    lemma
    sense
  }
}

```

5. Prefix: The prefix query matches documents where the value of the specified field begins with that prefix. Examples:

- RESTful: https://kosh.uni-koeln.de/api/hoenig/restful/entries?field=lemma_ksh&query=Hau&query_type=prefix&size=20

- GraphQL: <https://kosh.uni-koeln.de/api/hoenig/graphql>

```

{
  entries(queryType: prefix,
    query: "Hau",
    field: lemma_ksh,
    size: 20) {
    lemmaKsh
    translationDeu
  }
}

```

6. Wildcard: The wildcard query matches documents where the specified field matches a wildcard expression. Examples:

- RESTful: https://kosh.uni-koeln.de/api/ducange/restful/entries?field=lemma&query=e*en&query_type=wildcard&size=50

- GraphQL: <https://kosh.uni-koeln.de/api/ducange/graphql>

```

{
  entries(queryType: wildcard,
    query: "e*en"
    field: lemma,
    size: 50) {
    lemma
  }
}

```

7. Regexp: The regexp query matches documents where the specified field matches a regular expression. Examples:

- RESTful: https://kosh.uni-koeln.de/api/tunico/restful/entries?field=lemma&query=d.*m&query_type=regexp&size=50
- GraphQL: <https://kosh.uni-koeln.de/api/tunico/graphql>

```
{
  entries(queryType: regexp,
    query: "d.*m",
    field: lemma,
    size: 50) {
    lemma
    transEn
  }
}
```

6. Future Directions and Enhancements

6.1 Backend

6.1.1 Asynchronous Request Handling

Our objective is to improve Kosh by integrating asynchronous capabilities. To accomplish this, we will migrate from the existing web framework, Flask¹⁵, to FastAPI¹⁶. FastAPI is an asynchronous web framework that delivers a notable performance boost. It inherently supports REST and accommodates GraphQL through the Strawberry¹⁷ library. Moreover, we intend to introduce asynchronous queries in Elasticsearch and implement nested fields. The nested field type, a specialized version of the object data type, enables the indexing of object arrays in a manner that allows them to be queried separately from one another. These changes are anticipated to significantly enhance the system's overall performance and usability.

6.1.2 Integration with CSV Files

Drawing upon our expertise in the Zoroastrian Middle Persian Corpus and Dictionary (MPCD) project, as well as collaborations with other scholars, it has been observed that numerous researchers maintain their lexical data utilizing CSV files. While it is feasible to convert this data into XML format, this task imposes additional workload. Consequently, we endeavor to develop a methodology for directly incorporating data from spreadsheets into Kosh.

6.1.3 Integration with Large Language Models

Recent advancements in the field of large language models (LLMs) present a promising landscape for APIs handling natural language processing data in the upcoming future. The

¹⁵ <https://flask.palletsprojects.com/en/2.2.x>

¹⁶ <https://fastapi.tiangolo.com>

¹⁷ <https://strawberry.rocks/docs/integrations/fastapi>

progress in this domain is unparalleled, given the rapidity and depth of the transformations witnessed in recent months. Toolformer (Schick et al., 2023), a model designed to determine which APIs to invoke, when to initiate them, which arguments to transmit, and how to optimally integrate the outcomes into subsequent token predictions, operates in a self-supervised manner, necessitating only a few demonstrations for each API. The researchers trained a GPT-J model akin to GPT-3, albeit with significantly fewer parameters—6B compared to 175B—and achieved comparable results to GPT-3 across various benchmarks. Furthermore, OpenAI is incorporating external APIs into Chat-GPT, adhering to the methodology delineated in Toolformer, and accessing external APIs through plugins. Although plugin access remains in limited beta at the time of writing, the initial draft outlining the creation of a Chat-GPT plugin has been released. The prevailing specification employs Open-API, or Swagger, which is also utilized by Kosh for their REST APIs. While this specification may evolve and encompass GraphQL in the future, it is likely to remain the standard employed by Chat-GPT and other LLMs. This development is highly propitious for the evolution of Kosh and the integration of knowledge from underrepresented languages into mainstream LLMs.

7. References

- Amundsen, M. (2020). *Design and build great web APIs: robust, reliable, and resilient*. The pragmatic programmers. Raleigh, North Carolina: The Pragmatic Bookshelf.
- Cologne Digital Sanskrit Dictionaries (2019). Version 2.4.79. Cologne University. Accessed on April 13, 2023. <https://www.sanskrit-lexicon.uni-koeln.de>.
- Fielding, R.T. (2000). *Architectural Styles and the Design of Network-based Software Architectures*. Ph.D. thesis, University of California, Irvine. URL https://www.ics.uci.edu/~fielding/pubs/dissertation/fielding_dissertation.pdf.
- GraphQL (2021). GraphQL Specification. <https://spec.graphql.org>. Accessed: April 18, 2023.
- Grassmann, H.G. (1873). *Worterbuch zum Rig-veda*. Wiesbaden: O. Harrassowitz. OCLC: 184798352.
- Kiss, B., Kölligan, D., Mondaca, F., Neufeind, C., Reinöhl, U. & Sahle, P. (2019). It Takes a Village: Co-developing VedaWeb, a Digital Research Platform for Old Indo-Aryan Texts. In S. Krauwer & D. Fišer (eds.) *TwinTalks at DHN 2019 – Understanding Collaboration in Digital Humanities*, volume 2365 of *CEUR Workshop Proceedings*. URL http://ceur-ws.org/Vol-2365/05-TwinTalks-DHN2019_paper_5.pdf.
- Medjaoui, M., Wilde, E., Mitra, R. & Amundsen, M. (2021). *Continuous API Management*. O’Reilly Media, Inc., 2 edition. ISBN: 9781098103521.
- Mondaca, F., Esser, M., Neufeind, C. & Eide, Ø. (2022). MPCD: An API-based Research Environment. URL <https://doi.org/10.5281/zenodo.7839927>.
- Mondaca, F., Rau, F., Neufeind, C., Kiss, B., Kölligan, D., Reinöhl, U. & Sahle, P. (2019a). C-SALT APIs – Connecting and Exposing Heterogeneous Language Resources. URL <https://doi.org/10.5281/zenodo.3265782>.
- Mondaca, F., Schildkamp, P. & Rau, F. (2019b). Introducing Kosh, a Framework for Creating and Maintaining APIs for Lexical Data. In *Electronic Lexicography in the 21st Century. Proceedings of the eLex 2019 Conference, Sintra, Portugal*. Brno: Lexical Computing CZ, pp. 907–21. URL https://elex.link/elex2019/wp-content/uploads/2019/09/eLex_2019_51.pdf.

- Monier-Williams, M. (1899). *A Sanskrit-English dictionary: Etymologically and philologically arranged with special reference to Cognate indo-european languages*. Oxford: The Clarendon Press.
- Měchura, M. (2017). Introducing Lexonomy: an open-source dictionary writing and publishing system. In *Electronic Lexicography in the 21st Century. Proceedings of the eLex 2017 Conference, Leiden, Netherlands*. Leiden: Lexical Computing CZ, p. 18. URL <https://elex.link/elex2017/wp-content/uploads/2017/09/paper41.pdf>.
- Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., Cancedda, N. & Scialom, T. (2023). Toolformer: Language Models Can Teach Themselves to Use Tools. URL <http://arxiv.org/abs/2302.04761>. ArXiv:2302.04761 [cs].
- Tasovac, T., Romary, L., Banski, P., Bowers, J., de Does, J., Depuydt, K., Erjavec, T., Geyken, A., Herold, A., Hildenbrandt, V., Khemakhem, M., Lehečka, B., Petrović, S., Salgado, A. & Witt, A. (2018). TEI Lex-0: A baseline encoding for lexicographic data. <https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html>.
- TEI Consortium (2023). TEI P5: Guidelines for Electronic Text Encoding and Interchange. <http://www.tei-c.org/Guidelines/P5/>. Accessed on 18.04.2023.

Humanitarian reports on ReliefWeb as a domain-specific corpus

Loryn Isaacs

Department of Translation and Interpreting, EDHCSJ,
University of Granada, Buensuceso, 11, 18002 Granada (Spain)

E-mail: lisaacs@ugr.es

<https://orcid.org/0000-0003-0267-4853>

Abstract

This paper presents an assessment of the content available on ReliefWeb’s API for its suitability as a domain-specific corpus. ReliefWeb’s position as a primary information resource for humanitarian response, boasting a database of nearly a million reports, lends it considerable value for the corpus-based study of humanitarian discourse. However, the service’s content is under-explored in this regard. To this end, a Python package is introduced to manage the creation of ReliefWeb corpora. The composition of ReliefWeb’s HTML reports in English is examined and compared with a corpus from the Humanitarian Encyclopedia. The comparison includes a keyness analysis of the Encyclopedia’s 129 concepts and an assessment of diachronic trends for six concepts (HUMANITARIAN REFORM, SUSTAINABILITY, RESILIENCE, GENDER-BASED VIOLENCE, SETTLEMENT, and SOVEREIGNTY), as well as an analysis of hypernymic and definitional knowledge-rich contexts. Results indicated that ReliefWeb reports, mostly brief news and press release items, have much lower relative frequencies for humanitarian concepts than the reference corpus. Still, the data overlapped considerably and the breadth of the HTML content contributed important thematic diversity for some concepts. The paper concludes with a discussion of how the management of ReliefWeb corpora could be improved in future iterations.

Keywords: humanitarian domain; corpus creation; ReliefWeb; information extraction

1. Introduction

ReliefWeb¹ is a service managed by the United Nations Office for the Coordination of Humanitarian Affairs (OCHA) that aggregates publicly available documents related to current humanitarian issues. In 2023, this 27-year-old database is likely to reach one million reports spanning half a century and representing thousands of diverse actors worldwide. During the first half of 2022 it had 10.8 million users and experienced increases across a variety of usage metrics (OCHA, 2022). In fulfilling its founding objective to “act as the principal information system for prevention, preparedness, and rapid response for the humanitarian community” (Ruso, 1996, p. 18), the service also represents a significant resource in the study of humanitarian communication. The database has been utilized in various fashions, such as in

¹ <https://reliefweb.int/>

ReliefWeb Labs projects,² to discursively track famine (Rubin, 2014), and to extract knowledge via semantic embedding (Shamoug, Cranefield & Dick, 2023). A common goal is to improve humanitarian response by leveraging linguistic data, which can be hampered by the difficulty of synthesizing and transmitting domain knowledge.

This paper approaches the database from a corpus-based linguistics perspective with several aims. One is to provide a thorough analysis of ReliefWeb’s composition, which can be treated superficially despite its relevance in guiding data interpretation. Another is to convert the bulk of ReliefWeb reports into a format readable by language corpus management software and to offer a means to periodically update such corpora. This is with the hope of establishing an accessible and durable means to facilitate research in regards to humanitarian knowledge extraction and representation. Finally, this article assesses the suitability of ReliefWeb’s HTML reports for analyzing domain-specific concepts. While the database’s value is apparent, understanding its limits as a representation of humanitarian discourse is a necessary practice.

The assessment of ReliefWeb data provided here is part of ongoing efforts to expand and refine the corpus-based methods used to generate concept entries for the Humanitarian Encyclopedia platform.³ The Encyclopedia, a project by the Geneva Centre of Humanitarian Studies, offers analyses on 129 humanitarian concepts with a combination of corpus-based linguistic reports and input from domain experts. It documents aspects of humanitarian discourse with a focus on concept variation and multidimensionality (León-Araúz, 2017), and also promotes community discussion of the domain’s lexicon. To develop entries, data are retrieved with semantic and multiword-term querying techniques, and a battery of visualizations are supplied to ground discussion quantitatively (Chambó & León-Araúz, 2021; León-Araúz & San Martín, 2018). Concept analyses for the Encyclopedia have so far been conducted on an internal corpus of 4,824 public humanitarian documents from the last two decades. However, a significant portion of this corpus is likely included in ReliefWeb, which boasts both a mature data management system and continuously updated content. While leveraging the service’s content is a logical progression, this requires converting ReliefWeb’s reports into a tokenized, lemmatized language corpus and studying its shape and limitations.

Section 2 describes an API-based data retrieval method and the conversion of data into a Sketch Engine-compatible format (Kilgarriff et al., 2014). The section summarizes the ReliefWeb corpus’s composition and describes the methodology used to compare it with the Humanitarian Encyclopedia’s corpus. Section 3 reports results for a keyness analysis of 129 humanitarian concepts, as well as results regarding diachronic trends for six concepts and the density of their hypernymic and

² <https://labs.reliefweb.int/>

³ <https://humanitarianencyclopedia.org/>

definitional contexts. Section 4 discusses how the results pertain to the use of ReliefWeb as a domain-specific corpus.

2. The ReliefWeb English corpus

2.1 Data collection, structure and limits

ReliefWeb reports are one of several content types available via API request to the service. The category contains the bulk of the site’s primary content: an empty query returned over 988,000 results as of April 2023, including documents, maps, and other digital formats aggregated from internet sources. The corpus described in the following sections focuses on reports in English with an original publication date from 2000 to 2022, excluding several text-poor formats, namely maps, interactives, and infographics, as well as the heterogeneous “other” category. This returned 662,473 API responses (67% of reports at the time).

The above figures require some contextualization. Importantly, many of the report metadata fields (a total of 98 were detected) allow multiple values. Language is one of these, meaning that a request for English content includes any report with at least a tag for English. Reports of this type may have multiple texts with different languages, as in the uncommon case where several PDF translations are available; just under 2% of the collected reports contained other language tags in addition to English. Tagging errors can also introduce some non-English texts in the data recovered (e.g., as of publication, report no. 21366).⁴ While these sources of noise appear minimal, employing language detection algorithms would likely be necessary to establish more exact figures.

Defining what “report” means in ReliefWeb’s database is also a prerequisite. Each text available through the API⁵ is given a unique identifier. Each identifier refers to at least one form of content, but a report may be a collection of related materials. The content visible on ReliefWeb’s website is all of or a portion of what is understood as the report’s primary document. For shorter texts, like press releases, most or all body text (e.g., excluding footers) may be displayed as HTML content. For longer texts, however, only a portion is displayed, such as a document’s introduction, executive summary, or first page.

An example report is given below for a publication from Humanitarian Practice Network (web page content is in Figure 1 and the source PDF in Figure 2).⁶ In this case, the HTML text on ReliefWeb contains 366 words: this is a portion of the PDF’s

⁴ <https://reliefweb.int/node/21366>

⁵ <https://api.reliefweb.int/v1/reports>

⁶ <https://reliefweb.int/node/23456>

first page, with one of the middle paragraphs removed (starting with “The articles in the special feature”). In other words, a report may consist of a portion of altered text that meets ReliefWeb’s editorial constraints for size and content. In this example the full PDF is 52 pages, while the searchable text is the report’s HTML body.

Humanitarian Exchange Magazine No. 29 - Good Humanitarian Donorship

📄 Analysis • Source: [ODI - HPN](#) • Posted: 1 Mar 2005 • Originally published: 1 Mar 2005 •

Origin: [View original](#) ↗

<p>Donors and agencies alike have long sought means of improving the performance, accountability and transparency of humanitarian action. Whilst a proliferation of NGO and agency</p>	<div style="border: 1px solid gray; padding: 5px; display: inline-block;"> Download Report (PDF 535.88 KB) </div>	<p>Primary country: World</p> <p>Source: ODI - Humanitarian</p>
--	--	---

Figure 1: Report no. 23456 HTML content

Humanitarian Practice Network

HPN

Managed by

Humanitarian Policy Group

Number 29
March 2005

Humanitarian Exchange

Commissioned and published by the Humanitarian Practice Network at ODI

In this issue

Good Humanitarian Donorship

- 2** Welcome to the club
- 4** Good Donorship: how serious are the donors?
- 8** Too good to be true? US engagement in the GHD initiative
- 10** The EU: Good Humanitarian Donorship and the world's biggest humanitarian donor

Donors and agencies alike have long sought means of improving the performance, accountability and transparency of humanitarian action. Whilst a proliferation of NGO and agency initiatives followed the



Figure 2: Report no. 23456 original PDF

Report, then, is used here to refer to each unique item in the database and, more specifically, the HTML content for these items that users view when browsing (the *body-html* API field). Since a corpus built from this data excludes full-text PDF content, which exists for nearly a third of the downloaded reports, full-text analysis is

not possible. Conversely, since two thirds of reports have no PDF data, HTML-only content may be more complete, albeit for genres with shorter average lengths.

In addition to the aforementioned architectural limitations, authors have noted concerns familiar to corpus linguists. As one of the service’s founders states, “Information is not neutral. The user must judge the reliability of content and the biases in reporting” (Ruso, 1996, p. 120). While the methods and principles guiding the service have developed over the course of decades, some factors regarding neutrality or bias may still be relevant. The first is perhaps the primacy of English language texts, a challenge recognized in past recommendations (Naidoo, 2007, p. 57). The future also holds new issues for the online data aggregation service, including disinformation (Wackernagel & Footner, 2021).

Aspects of data collection aside, authors have pointed to several considerations for interpreting ReliefWeb’s linguistic data. One is that content published during and immediately after emergencies may suffer in quality, originality, and substance (von Schreeb et al., 2013). In other words, while quality issues with crisis reporting may be corrected in successive documents, initial errors may remain in the corpus as artifacts that could later skew results. Other fundamental concerns for approaching the domain’s discourse include humanitarian concepts lacking standardization; poor contextualization of term frequencies; politically and institutionally motivated uses and omissions; changes in the distribution and representation of organizations; under- and over-reporting of geographic areas due to accessibility; and data reported on national levels obscuring local trends (Rubin, 2014; von Schreeb et al., 2013).

2.2 Corpus compilation

After JSON API response data was flattened and stored in an SQLite table, *html-body* text was processed with a Stanza NLP pipeline utilizing the default Universal Dependencies English Web Treebank (EWT) model (Qi et al., 2020; Silveira et al., 2014). Output was reshaped into a vertical format, with an XML string containing metadata inserted into each text. Given that 98 metadata tags were detected in the data set, only those judged most valuable for corpus queries were included, 22 in total. Discarded tags include country coordinates, URLs to associated images, and redundant categories (*country.iso3* being preferred over *country.id*).

Since many ReliefWeb metadata fields allow for multiple values, fields with lists of values were concatenated into strings with a pipe separator. For example, a report with multiple values in *source.name* appears as “World Health Organization|Government of Nigeria”, and the corresponding values in related tags maintain the same list order, with *source.type.name* being “International Organization|Government”). The original data structure is maintained in this way,

although it can cause confusion if tags are viewed in isolation, as some include over thirty values, e.g., report no. 630723.⁷

A tagset file was produced by detecting unique XPOS values (52) generated from the Stanza pipeline. This tagset was compared with Sketch Engine’s default tagger,⁸ a modified version of TreeTagger (Marcus et al., 1993), to identify dissimilarities that end users should take into account. Most of the EWT and modified TreeTagger tags were functionally equivalent, with the exception of verb tags, given that Sketch Engine has a separate set of tags exclusively for the verb *be*. When designing queries, Stanza’s more atomic tokenization for hyphenation should be taken into account: for example, “gender-based” gets split into three tokens rather than one.

Compressed archives of vertical text were then exported and fed to Sketch Engine’s corpus compilation tool. Specifically, this took the form of a local NoSketch Engine server run as a Docker container (Kilgarriff et al., 2014; National Laboratory for Digital Heritage, Eötvös Loránd University Department of Digital Humanities, 2023; Rychlý, 2007). The Python functions to replicate data retrieval and corpus creation have been made available in a GitHub repository (Corpusama, v0.1.1).⁹ The workflow was configured to update the corpus by last date of modification, meaning that prior versions of content could be overwritten without being tracked if a report is updated post-publication.

2.3 Corpus composition

In total, 662,473 API responses produced 431,170,905 tokens, 366,049,459 words, 16,809,660 sentences, and 657,098 documents, with the last figure being slightly lower than the total number of API responses due to some reports lacking any text content (i.e., only containing other HTML elements). The average document length amounted to 557 words. Table 1 summarizes corpus attributes, ordered by the number of unique items and combination of items (“structure frequency” in Sketch Engine). For example, among the 47,152 values for *country.shortname* are “Israel”, “oPt”, “Israel|oPt”, “oPt|Israel”, and other combinations with over a dozen countries. For this attribute a total of 248 countries are represented, meaning that most of its values are combinations of several countries. This multivalued format preserves ReliefWeb’s data structure, but it also inflates tallies for some attributes: “Israel|oPt” and “oPt|Israel” may have no meaningful difference but are nonetheless counted separately.

⁷ <https://reliefweb.int/node/630723>

⁸ <https://www.sketchengine.eu/english-treetagger-pipeline-2>

⁹ <https://github.com/engisalor/corpusama>

Attribute	Items ^a	Unique ^b	NA% ^c	Example ^d
id	657,098	657,098	0	100001
url	657,098	657,098	0	https://reliefweb.int/node/100001
title	652,358	652,358	0	Food Security Outlook Update May2011
origin	381,773	381,773	38	https://www.ifrc.org/appeals
country.iso3	47,152	248	0	afg
country.shortname	47,152	248	0	Afghanistan
theme.name	14,492	21	12	Protection and Human Rights
disaster.name	13,041	2,621	62	Haiti: Earthquakes - Jan 2010
disaster.glide	12,888	2,508	61	OT-2011-000205-NER
source.name	11,155	2,708	< 1	The New Humanitarian
source.shortname	11,133	2,684	< 1	TNH
source.homepage	10,884	2,485	< 1	http://www.unhcr.org/
date.original	8,395	8,395	0	2017-06-30T00:00:00+00:00
source.spanish_name	3,848	197	48	Gobierno de Filipinas
disaster_type.name	2,436	22	45	Tropical Cyclone
source.type.name	1,122	8	< 1	International Organization
primary_country.iso3	235	235	0	wld
primary_country.shortname	235	235	0	World
date.original.year ^e	23	23	0	2017
language.name	17	5	0	English
ocha_product.name	14	14	97	Flash Update
format.name	9	9	< 1	News and Press Release

^a Includes individual items (*Ethiopia*) and lists (*Ethiopia|Kenya* and *Kenya|Ethiopia* being distinct).

^b Includes individual items only (*Ethiopia*, *Kenya*, *Somalia*).

^c Percentage of NA values in the total frequency: 0 = no missing values; < 1 is a non-zero result.

^d Examples from various reports.

^e Extracted from date.original during compilation.

Table 1: ReliefWeb corpus attributes

Table 2 offers further details on corpus composition, showing the top ten values for several attributes. Frequencies and relative frequencies refer to the total number of instances of an item: “World” occurs 65,395 times in *county.shortname*, whether alone (29,100) or as part of lists, including “Greece|World” (608), “Libya|World” (545), etc. Altogether, “World” appears in 16,209 different lists, almost 98% of which have five or fewer occurrences. This long tail is characteristic of *county.shortname* and similar ReliefWeb attributes.

Attribute	Value	freq	relfreq
country.shortname	World	65,395	151.67
	Sudan	46,418	107.66
	Afghanistan	42,672	98.97
	Syria	37,781	87.62
	DR Congo	34,948	81.05
	Somalia	34,820	80.76
	Iraq	34,497	80.01
	oPt	33,025	76.59
	Pakistan	30,423	70.56
	Ethiopia	27,862	64.62
date.original.year	2020	37,328	86.57
	2015	35,816	83.07

Attribute	Value	freq	relfreq
	2009	34,942	81.04
	2022	34,690	80.46
	2017	34,532	80.09
	2014	33,784	78.35
	2018	33,446	77.57
	2019	32,839	76.16
	2016	32,210	74.7
	2021	31,845	73.86
disaster_type.name	NA	403,656	936.19
	Flood	80,114	185.81
	Epidemic	75,544	175.21
	Drought	48,386	112.22
	Earthquake	40,108	93.02
	Tropical Cyclone	39,872	92.47
	Land Slide	30,945	71.77
	Flash Flood	28,507	66.12
	Other	19,367	44.92
	Drought Other	14,317	33.2
format.name	News and Press Release	456,371	1058.45
	Situation Report	126,377	293.1
	Analysis	37,992	88.11
	Assessment	11,993	27.81
	Appeal	7,055	16.36
	Manual and Guideline	6,823	15.82
	UN Document	5,527	12.82
	Evaluation and Lessons Learned	4,792	11.11
	NA	168	0.39
source.name	The New Humanitarian	30,726	71.26
	UN High Commissioner for Refugees	27,245	63.19
	UN Office for the Coordination of Humanitarian Affairs	25,684	59.57
	World Health Organization	24,030	55.73
	World Food Programme	23,567	54.66
	Reuters - Thomson Reuters Foundation	22,614	52.45
	UN Children's Fund	22,317	51.76
	International Federation of Red Cross And Red Crescent Societies	18,288	42.41
	International Organization for Migration	14,273	33.1
	UN News Service	11,955	27.73
source.type.name	International Organization	277,390	643.34
	Media	115,298	267.41
	Non-governmental Organization	114,691	266
	Government	102,135	236.88
	Red Cross/Red Crescent Movement	34,065	79.01
	Academic and Research Institution	20,744	48.11
	International Organization International Organization	11,902	27.6
	Other	9,200	21.34
	Government International Organization	4,025	9.34
	International Organization Government	2,084	4.83
theme.name	Protection and Human Rights	183,709	426.07
	NA	183,678	426

Attribute	Value	freq	relfreq
Health		172,045	399.02
Food and Nutrition		142,468	330.42
Water Sanitation Hygiene		97,461	226.04
Shelter and Non-Food Items		89,913	208.53
Agriculture		70,714	164
Education		60,230	139.69
Contributions		55,310	128.28
Coordination		53,564	124.23

Table 2: ReliefWeb corpus text type analysis

Frequencies for several key attributes yield some of the corpus's general characteristics:

1. The top countries have relatively comparable frequencies, with the highest counts ranging between 46,418 (3.2%) for Sudan and 16,536 (1.1%) for Haiti.
2. Though an increase in annual document counts is expected, no single year between 2000 and 2022 is a particular outlier.
3. While many disasters are not categorized by type, *flood* and *epidemic* are the most common.
4. Over two thirds of the corpus consist of news and press releases, with another fifth being situation reports.
5. While there are 2,708 contributing organizations, almost 30% of reports originate from the top ten sources, led by The New Humanitarian (4.1%).
6. Most sources fall under international organizations (39%), media (16%), NGOs (16%), and governments (15%).
7. Almost 12% of documents lack a theme. A considerable portion with themes refer to protection and human rights (12%), health (11%), food and nutrition (9%) and water, sanitation and hygiene (6.2%).

In brief, in the first 22 years since the turn of the century, ReliefWeb posted an average of close to 17 million words in English annually. This consists mainly of short news and press release items tagged for natural disasters, with potentially edited content that provides at least a document's summary. The most commonly tagged countries are from the African and Eastern Mediterranean World Health Organization regions¹⁰ and are often grouped into wider affected areas. Among thousands of authors, led by international organizations, a small subset provides a

¹⁰ <https://www.who.int/countries>

substantial amount of content, particularly those affiliated with the United Nations.

2.4 Assessment methodology

After compiling the corpus of ReliefWeb reports, an initial assessment was conducted to compare its characteristics against the corpus developed by the Humanitarian Encyclopedia (the HE corpus). The primary concern was how the low average word count of HTML texts on ReliefWeb (557) could affect the frequencies of the 129 humanitarian concepts studied by the Humanitarian Encyclopedia, given that the HE corpus is made up of PDFs averaging 14,760 words (over 26 times longer). The main objective of the following analysis, then, was to determine whether ReliefWeb's curated HTML content would be suitable for the Encyclopedia's concept analyses in lieu of compiling a complete corpus of the service's PDFs.

In comparison with ReliefWeb's multitudinous reports, the HE corpus is a much smaller set of publicly available humanitarian documents (4,824 texts amounting to 71,201,157 words) published between 2005 and early 2019. As both corpora consist of documents published online by humanitarian organizations, much of their content is expected to overlap, although to what extent is unknown without aligning their metadata. Both corpora have tags for document format, organization type, and geographic region, but these are not necessarily comparable. For example, geographic metadata in ReliefWeb refer to individual affected countries, whereas in the HE corpus they refer to the continent a document originated from. Given limitations of this nature, the current analysis considered only year of publication as a viable text type for comparison.

The analysis was conducted in three steps. First, the frequencies of the Humanitarian Encyclopedia's 129 concepts were collected from each corpus via a Python-based NoSketch Engine API controller (Isaacs, 2022). Corpus Query Language (CQL) rules (Jakubíček et al., 2010) were designed for these queries, most being uncomplicated (`[lemma_lc="knowledge"]` for KNOWLEDGE), while others took into account hyphenation, part-of-speech, or common abbreviations. For instance, the rule for INTERGOVERNMENTAL ORGANIZATION was as follows:

```
( ( [lemma_lc="inter-governmental|intergovernmental"] |  
  [lemma_lc="inter"] [lemma_lc="governmental"] |  
  [lemma_lc="inter"] [lc="-"] [lemma_lc="governmental"]  
  ) [lemma_lc="organisation|organization"]  
  ) | [lc="IGOs?"]
```

To compare the density of concepts across the corpora, normalized frequencies were used to compute an effect-size keyness score (Gabrielatos, 2018; Kilgarriff, 2012), which indicates whether a concept is more common in the focus corpus (ReliefWeb, or simply RW) or reference corpus (Humanitarian Encyclopedia, HE). The

distribution of keyness was analyzed with the assumption that many scores gravitating toward $K=1$ indicates a shared focus for a concept. In contrast, consistently low or high keyness across the 129 concepts could indicate important dissimilarities for the Humanitarian Encyclopedia’s concept analyses.

Second, concepts were assessed by visualizing their frequency over time with the *DATE* text type for HE and *date.original.year* for RW. This utilized the *reltt* measurement in Sketch Engine (relative text type frequency), which is a per million tokens calculation that normalizes frequencies for text type values (in this case, each individual year). However, despite the fact that the corpora both focus on the same time period, they are not fully comparable. RW includes more years (23 versus 14) and HE has some tagging irregularities: documents missing a year (*DATE=0*), multiyear tags (*DATE=2005-2006*), and incomplete data for 2019. With these caveats in mind, comparing the data by year was intended to add perspective to the keyness analysis and identify whether the corpora displayed similar trends for concept frequencies over the first two decades of the century.

Third, six concepts with a range of keyness and which displayed varying diachronic trends (downward, upward, stable) were selected for an analysis of their hypernymic and definitional contexts. Random samples of 1,000 concordances were inspected manually in each corpus to compare the density and diversity of these knowledge-rich contexts, or KRCs (Condamines, 2022; Marshman, 2022; Meyer, 2001). This followed the Humanitarian Encyclopedia’s concept analysis procedure, which is informed by Frame-based Terminology (Faber, 2022) and utilizes KRC-based knowledge extraction techniques (León-Araúz & San Martín, 2018; San Martín et al., 2020).

As part of this analysis, lists of monolexical hypernyms were collected and compared to judge the RW corpus’s potential for knowledge extraction. KRCs with polylexical hypernyms were simplified to facilitate this comparison. For example, in the phrase “Resilience is also a contested term in the literature” in the RW corpus, TERM was extracted as the hypernym, with the (quite valuable) adjective “contested” being left for future discussion. Definitional contexts were identified with a flexible approach to maximize the number of available candidates. This allowed for formal and informal definitions containing genus and differentiae, verbal patterns, or paralinguistic patterns (Meyer, 2001; Sierra et al., 2008).

The number and qualities of the hypernyms and definitional contexts were compared; results were then discussed regarding the content for humanitarian concepts in RW and HE. Results followed the Humanitarian Encyclopedia’s concept entry structure, which manages polysemous terms under a single entry. In other words, although `[lemma_lc="settlement"]` retrieved hypernyms referring to both `SETTLEMENT=COMMUNITY` (a population inhabiting a geographic area) and `SETTLEMENT=BARGAIN` (a mutual agreement), these were grouped together.

3. Results

3.1 Keyness for humanitarian concepts in ReliefWeb

Table 3 and Table 4 display the keyness and normalized frequency in RW by quartile for each of the Humanitarian Encyclopedia’s 129 concepts. Keyness ranged from 0.006, for RIGHT-BASED APPROACH, to 3.850, for SOVEREIGNTY, with $Q1=0.437$, $Q2=0.693$, and $Q3=1.060$. The least frequent concept in absolute terms for RW was also RIGHT-BASED APPROACH, with 39 occurrences, compared to a maximum of 876,392 occurrences for CHILD ($K=0.940$). The large majority of concepts were less common in RW ($K<1$), with 93 concepts or 72%. 38 concepts were at least half as common in RW as HE ($K<0.5$); 10 were at least twice as common in RW ($K>2$).

Q1			Q2		
concept	<i>K</i>	<i>fpm</i>	concept	<i>K</i>	<i>fpm</i>
right-based approach	0.006	0.09	humanitarianism	0.438	1.83
logistic	0.027	0.96	inclusion	0.440	40.93
equity	0.112	12.29	participation	0.446	118.92
remote-sensing	0.122	0.16	private sector	0.451	70.82
humanitarian reform	0.143	1.18	diplomacy	0.480	13.47
advocacy	0.177	56.96	program	0.505	1,208.59
			community-based		
urbanisation	0.229	7.93	approach	0.517	1.96
efficiency	0.245	23.93	cash	0.530	194.94
innovation	0.265	31.54	integrated approach	0.535	6.55
humanitarian action	0.273	32.69	aid dependence	0.536	1.04
sustainability	0.274	29.26	education	0.558	599.00
knowledge	0.279	86.64	humanitarian space	0.558	3.83
empowerment	0.281	34.68	context	0.566	156.18
effectiveness	0.302	39.19	mitigation	0.576	40.04
competition	0.304	17.90	adaptation	0.583	44.74
governance	0.309	95.06	datum	0.583	253.48
policy	0.337	357.08	empathy	0.589	1.99
grand bargain	0.346	3.67	dignity	0.592	52.25
management	0.346	403.67	leadership	0.596	118.09
partnership	0.347	192.63	care	0.598	483.37
capacity-building	0.369	58.73	localisation	0.598	3.88
do no harm	0.380	1.54	resilience	0.603	126.25
acceptance	0.381	12.74	corruption	0.617	46.21
technology	0.382	86.48	ethics	0.630	4.30
quality	0.390	155.98	intervention	0.638	201.79
disaster risk reduction	0.392	52.64	climate change	0.662	180.73
leave no one behind	0.402	4.73	service	0.664	920.50
development	0.402	1,122.52	risk	0.672	602.80
International governmental organisation	0.414	3.04	neutrality	0.672	11.10
poverty	0.418	216.56	prevention	0.676	178.47
accountability	0.429	89.47	implementation	0.685	300.41
culture	0.434	50.45	nutrition	0.693	220.27
local organisation	0.437	10.89			

Note. Smoothing = 0. Reference corpus = Humanitarian Encyclopedia. *fpm* = frequency per million tokens (focus corpus).

Table 3: Keyness of humanitarian concepts in ReliefWeb (lower half)

Q3			Q4		
concept	K	fpm	concept	K	fpm
vulnerability	0.694	107.41	response	1.091	919.16
faith	0.697	27.51	emergency	1.093	957.59
funding	0.706	317.57	coordination	1.097	341.15
humanitarian-development nexus	0.708	0.80	gender-based violence	1.135	80.05
impact	0.735	420.77	recovery	1.165	226.77
communication	0.751	195.08	early warning	1.181	59.35
evidence	0.765	107.06	food security	1.241	231.82
impartiality	0.765	12.82	testimony	1.264	17.50
community	0.795	1,662.05	power	1.293	256.40
justice	0.796	215.03	independence	1.319	78.45
aid	0.802	819.64	security	1.335	1,447.80
community engagement	0.814	11.32	need	1.356	1,183.39
monitoring	0.814	188.13	authority	1.361	629.82
solidarity	0.832	61.41	conflict	1.370	1,043.01
rehabilitation	0.833	178.78	armed actors	1.381	4.86
humanity	0.835	69.21	forced displacement	1.460	15.75
politics	0.859	25.08	needs assessment	1.479	34.94
humanitarian imperative	0.871	1.55	mandate	1.537	167.50
non-governmental organisation	0.871	419.32	epidemic	1.589	86.30
health	0.874	1,810.05	settlement	1.837	207.07
protection	0.884	580.74	crime	1.846	227.17
psychosocial support	0.889	43.00	negotiation	1.921	116.30
contingency planning	0.912	7.97	peace	2.057	844.76
child	0.940	2,032.59	affected population	2.293	65.60
access	0.944	704.60	shelter	2.296	426.33
sanitation	0.945	278.86	famine	2.352	67.74
humanitarian actor	0.947	35.61	terrorism	2.446	79.43
livelihood	0.954	261.32	humanitarian worker	2.708	29.82
transition	1.029	102.09	civilian	3.081	667.50
law	1.044	515.61	responsibility-to-protect	3.109	14.30
crisis	1.051	560.05	evacuation	3.326	89.44
ethnicity	1.060	13.71	sovereignty	3.850	31.76

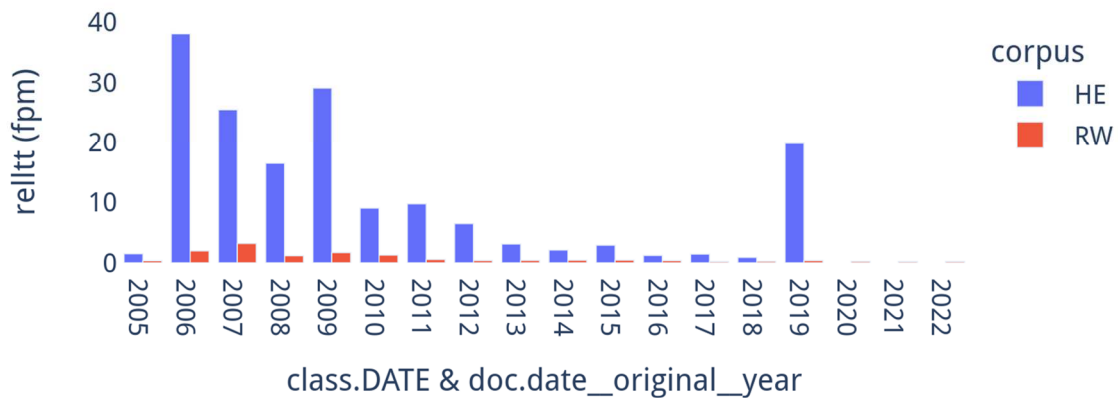
Note. Smoothing = 0. Reference corpus = Humanitarian Encyclopedia. *fpm* = frequency per million tokens (focus corpus).

Table 4: Keyness of humanitarian concepts in ReliefWeb (upper half)

3.2 Diachronic change in humanitarian concept frequencies

As nearly three quarters of the concepts had $K < 1$, most visualizations were similar to the one below for SUSTAINABILITY. These displayed generally flat or upward trending distributions across time for both corpora and higher relative frequencies for HE. An exception was HUMANITARIAN REFORM, one of the few concepts with a marked decline in use. Many of these graphs were punctuated by outlying values for problematic HE date tags (0 for missing values, 2019 being incomplete, and multiyear values like 2004-2005). Whereas typical years in HE have between 109 and 622 documents each, multiyear values each only appear once and hence were excluded below.

humanitarian reform



sustainability

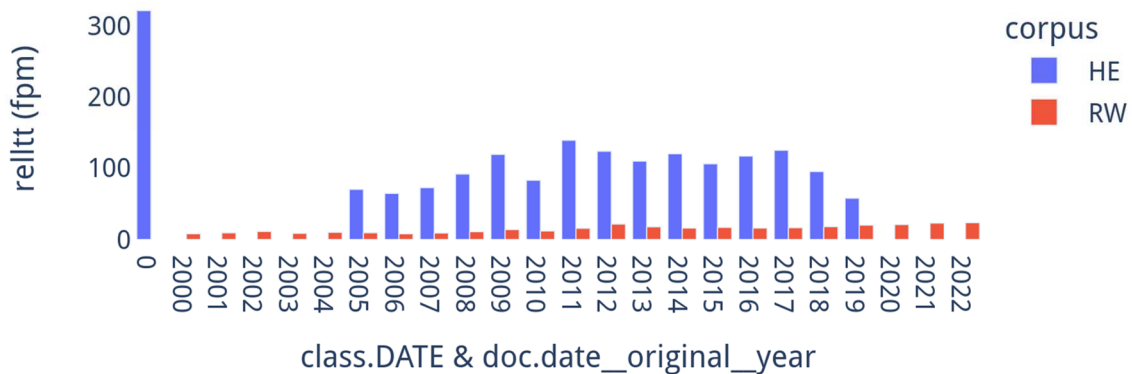
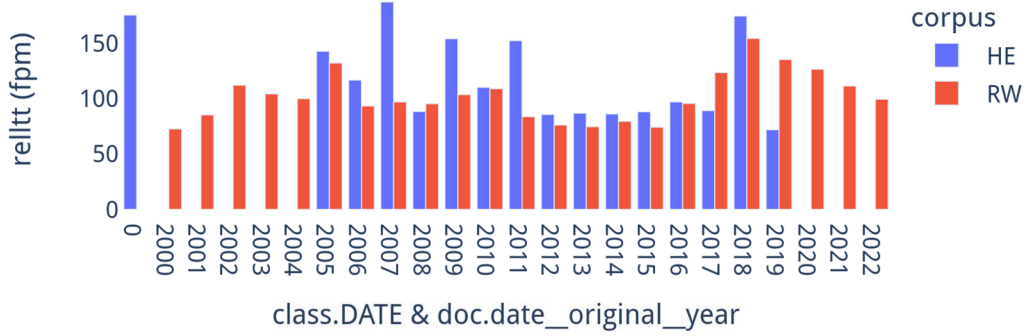


Figure 3: Concepts with $K < Q1$

The additional years covered in RW that HE lacks inflated the corpus-wide keyness for concepts. When keyness was computed for each shared, complete year and then averaged, scores dropped by half: the adjusted quartiles were $Q1=0.232$, $Q2=0.358$, and $Q3=0.552$ (compared to $Q1=0.437$, $Q2=0.693$, and $Q3=1.06$), with a maximum keyness of 2.199 for RESPONSIBILITY-TO-PROTECT. With these data, only ten concepts had $K > 1$ (PEACE, CRIME, SHELTER, HUMANITARIAN WORKER, TERRORISM, AFFECTED POPULATION, CIVILIAN, SOVEREIGNTY, EVACUATION, RESPONSIBILITY-TO-PROTECT), with the final two above $K > 2$. In contrast, 87 concepts had $K < 0.5$, of which 38 had $K < 0.25$. As seen in Figure 4, the annual relative frequencies of common concepts in RW often matched those of HE or were lower than suggested by corpus-wide keyness.

settlement



sovereignty

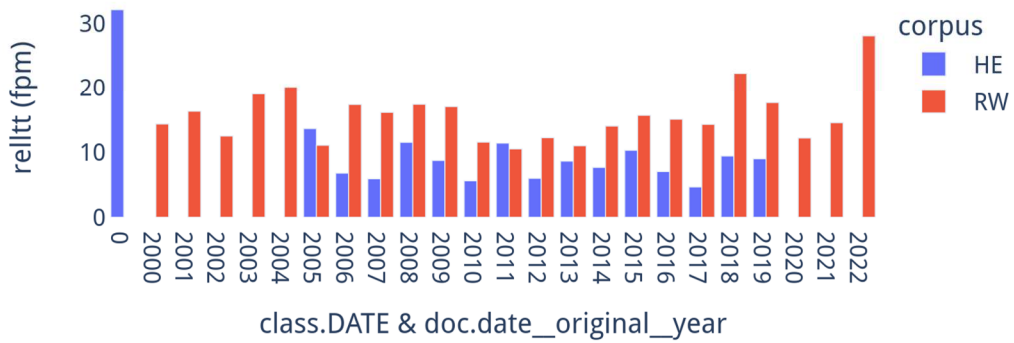
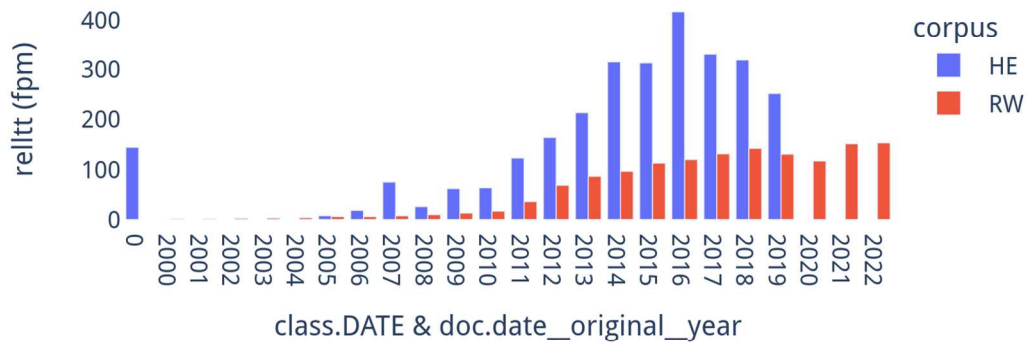


Figure 4: Concepts with $K > Q3$

Several concepts experienced upward trends over the last two decades, such as INNOVATION, EMPOWERMENT, DO NO HARM, INCLUSION, CASH, CONTEXT, PSYCHOSOCIAL SUPPORT, GENDER-BASED VIOLENCE, RESILIENCE, ARMED ACTORS, and FORCED DISPLACEMENT. Figure 5 shows annual relative frequencies for RESILIENCE and GENDER-BASED VIOLENCE beginning near 0 and reaching close to 400 and 200, respectively, as part of generally steady increases.

resilience



gender based violence

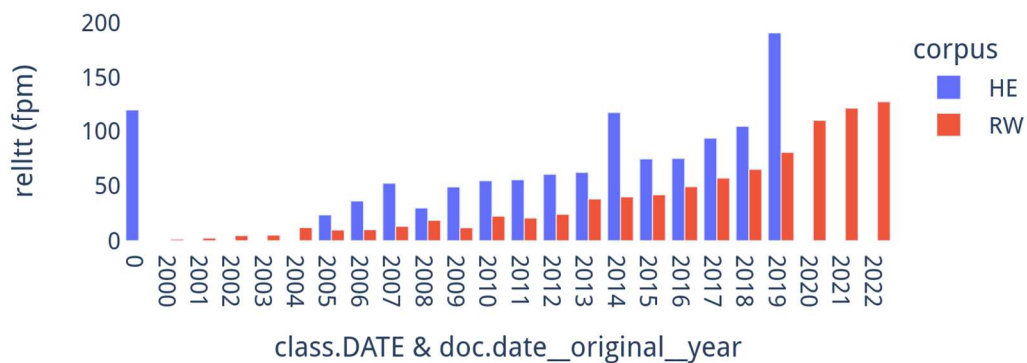


Figure 5: Concepts with shared upward trends

3.3 Hypernym and definitional context comparison

In randomized samples of up to 1,000 contexts (where possible), the average density of hypernyms and definitional contexts fluctuated from 0.10% for SETTLEMENT in RW to 5.40% for GENDER-BASED VIOLENCE in RW. Overall, RW had slightly higher KRC densities for the six concepts, at an average of 2.48% against 2.23% in HE. Two concepts had several-fold differences in density, with KRCs for HUMANITARIAN REFORM being 3.65 times more frequent in RW and KRCs for SETTLEMENT being 12.00 times more frequent in HE. These two concepts happened to be the least and most frequent in absolute terms (509 and 89,283 concordances) in RW.

Concept	K	Concordances		KRCs		Density %	
		HE	RW	HE	RW	HE	RW
humanitarian reform	0.143	699	509	3	8	0.43	1.57
sustainability	0.274	9,060	12,614	20	29	2.00	2.90
resilience	0.603	17,789	54,437	13	12	1.30	1.20
gender-based violence	1.135	5,991	34,516	40	54	4.00	5.40
settlement	1.837	9,572	89,283	12	6	1.20	0.10
sovereignty	3.85	701	13,692	31	32	4.42	3.20
mean	1.307	7,302	34,175	19.8	23.5	2.23	2.48

Note: Sample size = 1,000 random concordances or all if fewer

Table 5: Density of hypernymic and definitional contexts

Among the 260 contexts extracted for the six concepts were 104 monolexical hypernyms (including repeated cases, e.g., with ISSUE appearing separately for three concepts). 25 hypernyms, or 24%, were shared for the same concept across corpora, with HE having 34 additional hypernyms and RW 45. Once again, HUMANITARIAN REFORM and SETTLEMENT stood out for having the fewest shared hypernyms (0 of 9 for the former and 1 of 16 for the latter). In contrast, SOVEREIGNTY had the most homogeneous hypernyms, with 4 of 9 being shared (44%).

Concept	Shared	HE	RW
humanitarian reform	(0/9)	challenge, development [recent change], matter	module, initiative, issue, priority, reform, solution area, catchword,
sustainability	criterion, goal, indicator, issue, principle, theme, topic (7/25)	category, cornerstone, driver, objective	challenge, component, concept, concern, element, journey, measure, pillar, point, priority, problem, struggle
resilience	area, capacity, concept, term (4/13)	ability, notion, objective, priority, theme	accelerator, buzzword, pillar, quality
gender-based violence	abuse, challenge, concern, crime, issue, problem, term, violation, violence (9/32)	act, area, burden, component, crisis, practice, precursor, reaction, topic, weapon	barrier, discrimination, epidemic, exploitation, fact, injustice, phenomenon, plague, risk, scourge, threat, trauma, vulnerability
settlement	area (1/16)	bargain, categorization, concern, crime, need, shelter, slum, town	action, activity, community, facility, measure, village, violation
sovereignty	concept, issue, notion, principle (4/9)	priority, responsibility, right, theme	idea

Table 6: Shared and unique monolexical hypernyms

Among the contexts collected from concordance samples, a small minority were definitional, with only three of the concepts having this type of KRC in both corpora. RW had one context for SUSTAINABILITY, albeit less formal: “sustainability entails "striking a balance between the needs of both human and natural systems"”. RESILIENCE had four definitional contexts in HE, all similar to the example offered in Table 7. In contrast, each of the five contexts for RESILIENCE in RW were complementary but contextualized the concept in distinct settings: road infrastructure and farming, social institutions, livelihood systems, difficult situations, and cities. GENDER-BASED VIOLENCE had five contexts in HE and four in RW, with one subsumed in a definition for VIOLENCE AGAINST WOMEN. SOVEREIGNTY had four contexts in HE, with three being repeats from the same organization (due to the small sample size), and one in RW. For both corpora, these definitional contexts specifically treated FOOD SOVEREIGNTY rather than SOVEREIGNTY generically.

Concept	HE	RW
resilience	GOAL defines resilience as "the ability of communities and households living within complex systems to anticipate and adapt to risks, and to absorb, respond and recover from shocks and stresses in a timely and effective manner without compromising their long term prospects, ultimately improving their well-being.	Resilience refers here to the capacity of these social institutions to absorb and adapt in order to sustain an acceptable level of functioning, structure, and identity under stress.
gender-based violence	This Strategy defines GBV "as violence that is directed at an individual based on his or her biological sex, gender identity, perceived adherence to socially defined norms of masculinity and femininity.	Gender-based violence (GBV) is an umbrella term for any harmful act perpetrated against a person's will based on the socially ascribed (i.e. gender) differences between females and males.
sovereignty	In its own words: "Food sovereignty is the right of peoples to healthy and culturally appropriate food produced through sustainable methods and their right to define their own food and agriculture systems.	Though closely linked to food insecurity, food sovereignty involves the right of a state to be food self-sufficient based on their own democratically-determined policies.

Table 7: Definitional contexts across corpora

4. Discussion

Despite the shared domain of the corpora, the relative frequencies of humanitarian concepts in ReliefWeb's sometimes abbreviated HTML content are very often lower than the Humanitarian Encyclopedia's complete texts. This is especially the case when comparing per million token frequencies in each year shared by the corpora, which offers a more accurate depiction of how common concepts are. Using keyness scores that adjust for shared years, only 1.55% of concepts were at least twice as common in RW, whereas 67.44% were at least twice as common in HE.

That said, the size and scope of the RW corpus offered data that paralleled the HE corpus at each stage in the analysis. Diachronic trends for both stable and unstable concepts often agreed, the density of hypernymic contexts was similar and had

important overlaps, and both the appearance of and content of definitional contexts generally coincided. In other words, despite the large disparity in the average length of texts in each corpus (a 26-fold difference), analysis results for key humanitarian concepts are likely to share many commonalities.

Definitional contexts were found in both corpora precisely for the two concepts that experienced increasing usage in the previous two decades (RESILIENCE and GENDER-BASED VIOLENCE), whereas comparatively stable concepts like SETTLEMENT had no definitions. The one exception was FOOD SOVEREIGNTY, indicating that definitions for important yet less common hyponyms may also be captured to a similar extent. Still, the frequency of hypernyms varied widely, as with the much-reduced frequency of KRCs for SETTLEMENT in RW. The overall number of overlapping hypernyms was also low; along with the varied definitions for RESILIENCE in RW, it is apparent that RW contributes important diversity for some concepts, regardless of keyness.

While an analysis of concepts by organization type and theme would be beneficial in future work, one can still note that the concepts with the highest keyness in RW tend to underscore ReliefWeb’s focus on emergency response. There was a preponderance of EVENT concepts in Q4 that afflict populations (EMERGENCY, GENDER-BASED VIOLENCE, CONFLICT, FORCED DISPLACEMENT, EPIDEMIC, CRIME, FAMINE, TERRORISM, EVACUATION). In contrast, Q1 contained more abstract and process-oriented concepts related to humanitarian action (ADVOCACY, EFFICIENCY, INNOVATION, SUSTAINABILITY, KNOWLEDGE, MANAGEMENT, TECHNOLOGY, DEVELOPMENT, ACCOUNTABILITY). This divergent focus between the corpora may be an important consideration particularly when studying humanitarian development practices with ReliefWeb’s curated HTML reports.

Although this analysis offered perspective on ReliefWeb’s composition, as well as some characteristics relevant to the study of humanitarian concepts, a main limitation was its restriction to HTML content. Including PDF content would provide a more complete vision, likely increasing the relative frequencies of the domain’s core concepts. This task, which is underway, requires a more advanced pipeline with text extraction and language identification. Still, the data collected validate that KRC-based concept analysis can be fruitful with the HTML texts. While a workflow was developed here to build and update ReliefWeb corpora, in English and other languages, optimizing data extraction and its presentation to the humanitarian community is another area to contend with.

5. Conclusion

This paper presented a corpus of two thirds of HTML reports available on the United Nations-managed service ReliefWeb. These were mostly short news articles, press releases, or summaries in English regarding humanitarian response to emergency events. The corpus was compiled and inspected with a mix of open-source software,

including the Stanza NLP package, NoSketch Engine, and a Python package (Corpusama, available on GitHub) that was introduced to manage corpus generation with the service’s API.

A keyness analysis comparing 129 humanitarian concepts in the ReliefWeb corpus with a corpus developed by the Humanitarian Encyclopedia showed that ReliefWeb’s HTML content has consistently low relative frequencies for these concepts. Still, a subsequent knowledge-rich context analysis of six concepts (HUMANITARIAN REFORM, SUSTAINABILITY, RESILIENCE, GENDER-BASED VIOLENCE, SETTLEMENT, SOVEREIGNTY) indicated that the corpus offers both similar and complementary data for hypernym- and definition-centered information extraction. This result contextualizes the database’s potential and limits for humanitarian concept analysis, including the sort conducted by the Humanitarian Encyclopedia.

In the future, the development of a multilingual family of ReliefWeb corpora compatible with popular language corpus management software could be a boon for studying humanitarian discourse across linguistic communities. This is a goal for future versions of the Corpusama package. The current analysis, which focused on the composition of the English corpus and the frequencies of key humanitarian concepts, did not take into account most of the metadata offered via the service’s API. Organization type, document format, and humanitarian theme are prime candidates for more research, in the form of larger, more complete concept analyses and further inspection of the various characteristics of humanitarian reports on ReliefWeb.

6. Acknowledgments

Funding for this work was provided through the Humanitarian Encyclopedia project at the Geneva Centre of Humanitarian Studies and the research project PROYEXCEL_00369 (VariTermiHum), funded by the Regional Government of Andalusia (Spain).

7. References

- Chambó, S., & León-Araúz, P. (2021). Visualising lexical data for a corpus-driven encyclopaedia. In I. Kosem, M. Cukr, J. Miloš, J. Kallas, S. Krek, & C. Tiberius (eds.) *Electronic Lexicography in the 21st Century. Proceedings of the eLex 2021 Conference*. Brno, Czech Republic: Lexical Computing, pp. 29–55.
- Condamines, A. (2022). How the notion of “knowledge rich context” can be characterized today. *Frontiers in Communication*, 7. <https://doi.org/10.3389/fcomm.2022.824711>
- Faber, P. (2022). Frame-based terminology. In P. Faber & M.-C. L’Homme (eds.) *Theoretical Perspectives on Terminology*. Amsterdam: John Benjamins, pp. 353–376.

- Gabrielatos, C. (2018). Keyness analysis: Nature, metrics and techniques. In C. Taylor & A. Marchi (eds.) *Corpus Approaches to Discourse: A Critical Review*. London: Routledge, pp. 225–258.
- Isaacs, L. (2022). Sketch Grammar Explorer (Version 0.5.5) [Computer software]. <https://doi.org/10.5281/zenodo.6812335>
- Jakubíček, M., Kilgarriff, A., McCarthy, D., & Rychlý, P. (2010). Fast syntactic searching in very large corpora for many languages. In R. Ootoguro, K. Ishikawa, H. Umemoto, K. Yoshimoto, & Y. Harada (eds.), *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation (PACLIC 2010)*. Sendai, Japan: Institute of Digital Enhancement of Cognitive Processing, Waseda University, pp. 741–747.
- Kilgarriff, A. (2012). Getting to know your corpus. In P. Sojka, A. Horák, I. Kopeček, & K. Pala (eds.) *Text, Speech and Dialogue 15th International Conference, TSD 2012*. Brno, Czech Republic: Springer, pp. 3–15. https://doi.org/10.1007/978-3-642-32790-2_1
- 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(1), pp. 7–36. <https://doi.org/10.1007/s40607-014-0009-9>
- León-Araúz, P., & San Martín, A. (2018). The EcoLexicon Semantic Sketch Grammar: From knowledge patterns to word sketches. In I. Kerneman & S. Krek (eds.) *Proceedings of the LREC 2018 Workshop “Globalex 2018 – Lexicography & WordNets”*. Miyazaki, Japan: Globalex, pp. 94–99.
- León-Araúz, P. (2017). Term and concept variation in specialized knowledge dynamics. In P. Drouin, A. Francoeur, J. Humbley, & A. Picton (eds.), *Multiple Perspectives on Terminological Variation*. Amsterdam: John Benjamins, pp. 213–258.
- Marcus, M. P., Marcinkiewicz, M. A., & Santorini, B. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2), pp. 313–330.
- Marshman, E. (2022). Knowledge patterns in corpora. In P. Faber & M.-C. L’Homme (eds.) *Theoretical Perspectives on Terminology*. Amsterdam: John Benjamins, pp. 291–310.
- 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 (Vol. 2)*. Amsterdam: John Benjamins, pp. 279–302. <https://doi.org/10.1075/nlp.2.15mey>
- Naidoo, S. (2007). Redesigning the ReliefWeb. *Information Management Journal*, 41(5), pp. 52–58.
- National Laboratory for Digital Heritage, Eötvös Loránd University Department of Digital Humanities. (2023). NoSketch-Engine-Docker (Version 5.0.0) [Computer software]. <https://github.com/ELTE-DH/NoSketch-Engine-Docker>
- Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python

- natural language processing toolkit for many human languages. In A. Celikyilmaz & T.-H. Wen (eds.) *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Online: Association for Computational Linguistics, pp. 101–108.
- Rubin, O. (2014). Diagnosis of famine: A discursive contribution. *Disasters*, 38(1), pp. 1–21. <https://doi.org/10.1111/disa.12030>
- Ruso, S. (1996). ReliefWeb: Mandate and objectives. *Refuge*, 15(4), pp. 18–20. <https://doi.org/10.25071/1920-7336.21881>
- Rychlý, P. (2007). Manatee/Bonito — a modular corpus manager. In P. Sojka & A. Horák (eds.) *First Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2007*. Brno, Czech Republic: Masaryk University, pp. 65–70.
- San Martín, A., Trekker, C., & León-Araúz, P. (2020). Extraction of hyponymic relations in French with knowledge-pattern-based word sketches. In N. Calzolari et al. (eds.) *Proceedings of The 12th Language Resources and Evaluation Conference (LREC-2020)*. Marseille, France: European Language Resources Association, pp. 5953–5961.
- Shamoug, A., Cranefield, S., & Dick, G. (2023). SEmHuS: A semantically embedded humanitarian space. *Journal of International Humanitarian Action*, 8(3). <https://doi.org/10.1186/s41018-023-00135-4>
- Sierra, G., Alarcón, R., Aguilar, C., & Bach, C. (2008). Definitional verbal patterns for semantic relation extraction. *Terminology*, 14(1), pp. 74–98. <https://doi.org/10.1075/term.14.1.05sie>
- 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 N. Calzolari et al. (eds.) *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. Reykjavik, Iceland: European Language Resources Association, pp. 2897–2904.
- United Nations Office for the Coordination of Humanitarian Affairs. (2022). ReliefWeb analytics: 2022 mid-year highlights. <https://reliefweb.int/report/world/reliefweb-highlights-mid-year-2022>
- Von Schreeb, J., Legha, J. K., Karlsson, N., & Garfield, R. (2013). Information for action? Analysis of 2005 South Asian earthquake reports posted on Reliefweb. *Disaster Medicine and Public Health Preparedness*, 7(3), pp. 251–256. <https://doi.org/10.1001/dmp.2010.36>
- Wackernagel, M., & Footner, A. (2021, October 6). Talking Heads: ReliefWeb then and now. *ReliefWeb*. <https://reliefweb.int/blogpost/talking-heads-reliefweb-then-and-now>

A Federated Search and Retrieval Platform

for Lexical Resources in Text+ and CLARIN

Thomas Eckart¹, Axel Herold², Erik Körner¹, Frank Wiegand²

¹ Saxon Academy of Sciences and Humanities in Leipzig, Leipzig, Germany

² Berlin-Brandenburg Academy of Sciences and Humanities, Berlin, Germany

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

Abstract

The landscape of digital lexical resources is often characterized by dedicated local portals and proprietary interfaces as primary access points for scholars and the interested public. In addition, legal and technical restrictions are potential issues that can make it difficult to efficiently query and use these valuable resources. The research data consortium Text+ develops solutions for the storage and provision of digital language resources which are then provided in the context of the unified cross-domain German research data infrastructure NFDI. The specific topic of accessing lexical resources in a diverse and heterogenous setting with a variety of participating institutions and established technical solutions is met with the development of the federated search and query framework LexFCS. The LexFCS extends the established CLARIN Federated Content Search (FCS) that already allows accessing spatially distributed text corpora using a common specification of technical interfaces, data formats, and query languages. This paper describes the current state of development of the LexFCS, gives an insight into its technical details, and provides an outlook on its future development.

Keywords: lexical resources; federated content search; Text+; information retrieval

1. Introduction

The Text+ consortium¹ works on the utilization of text- and language-based research data in a distributed research environment. It is part of Germany's National Research Data Infrastructure (NFDI²) which aims to make research data available for scientific usage, support their interlinkage, and their long-term preservation. Consortia from various research areas are participating in the NFDI and work on establishing an interdisciplinary network of data and services based on common standards and the FAIR principles (findability, accessibility, interoperability, and reusability).

Text+ is organised in three “data domains”. The data domain “lexical resources” deals with all kinds of lexical resources, including dictionaries, encyclopedias, normative data, terminological databases, ontologies etc. Many of the largest German providers of such resources are members of the consortium. The data domain is structured in three thematic clusters with varying focuses (see figure 1).

One salient goal of the data domain is the integration of lexical data in a decentralized dictionary platform. Due to the heterogeneous nature of available resources, formats, levels

¹ <https://www.text-plus.org/en>

² <https://www.nfdi.de/?lang=en>



Figure 1: Data domains and their thematic clusters in the Text+ project

of annotation, and technical architectures in use, the implementation will follow a federated approach based on common protocols and formats. Query and retrieval of lexical data on this platform is based on the protocol of the CLARIN Federated Content Search (FCS³). The CLARIN FCS is an established framework that already allows accessing spatially distributed text corpora using a common specification of technical interfaces, data formats, and query languages. This framework was developed in the European CLARIN⁴ (*Common Language Resources and Technology Infrastructure*) project (Váradi et al., 2008). Its current specification is the basis for adding additional features that support the querying and retrieval of lexical records in the same distributed research environment.

The paper is structured as follows: section 2 gives an overview about related work in the field of providing lexical resources in a distributed research environment. Section 3 describes the state and amount of available lexical resources in the Text+ project. Section 4 outlines the general characteristics of the Federated Content Search, followed by section 5 that describes all extensions made to address specifics of lexical resources. Finally, section 6 presents the conclusion and highlights other current work and plans for further improvements.

2. Related Work

To make different electronic lexical resources available in one place and to allow them to be browsed and queried in a unified way has been a longstanding endeavour for years. Often, such projects were organisationally restricted to single institutions such as is the case with the Trier “Wörterbuchnetz”,⁵ a growing collection of mainly historical and dialectal dictionaries on the German language. Similar projects can be found across the world for different languages. Another early attempt in this regard is the interconnection of wordnets in different languages as pursued by the Global WordNet Association.⁶ However, most of these attempts were focused on lexical resources that are structurally and conceptionally very similar.

With the advent of the creation of common research infrastructures on national and international levels and a strong focus on FAIR data, more general initiatives have tackled

³ <https://www.clarin.eu/content/content-search>

⁴ <https://www.clarin.eu/>

⁵ <https://woerterbuchnetz.de/>

⁶ <http://globalwordnet.org/>

the problem of unifying the ways to access and exploit lexical data, such as the ERICs (*European Research Infrastructure Consortia*) CLARIN⁷ and DARIAH⁸ (*Digital Research Infrastructure for the Arts and Humanities*) together with their national sub-projects as well as the *ELEXIS*⁹ project among a range of smaller initiatives.

The four FAIR principles were not targeted equally for lexical data by the early infrastructures. CLARIN and DARIAH focused their efforts especially on *findability* and *accessibility*, resulting in an elaborate metadata ecosystem (e. g. based on CLARIN’s *component metadata infrastructure, CMDI*¹⁰), and a group of distributed certified data centers to operate repositories that host the actual data. The ELEXIS project on the other hand focused more strongly on *interoperability* and *reusability* in terms of the computational exploitation of lexical data. It had a strong influence on the development of the OntoLex/Lemon¹¹ model for the representation of lexical data (McCrae et al., 2017), and on the ISO 24613 family of standards on the *lexical markup framework*¹² (LMF).

Other initiatives such as the TEI¹³ (Text Encoding Initiative) have also worked on the standardization of dictionary and lexicon mark-up since the 1980s. In this context, the focus is often but not exclusively directed on the faithful representation of digitized print dictionaries. Work on the refinement of the TEI guidelines for the encoding of lexical data has recently been promoted by DARIAH and ELEXIS as well as by individual scholars, most notably in the TEI Lex-0¹⁴ initiative.

Orthogonal to the approaches described above, there are also more communal attempts to the creation of lexical resources which are not exclusively run by academic participants. The most important projects in this regard are Wiktionary,¹⁵ DBpedia,¹⁶ and Wikidata.¹⁷ While Wiktionary is essentially a community-driven multilingual dictionary based on (highly formalized) wiki syntax, DBpedia and Wikidata aim at automatically extracting strictly formalized information (though not restricted to lexical information) from sources like Wiktionary and (foremost) Wikipedia and at providing the extracted knowledge in the linked open data (LOD) paradigm, e. g. in the form of RDF serializations.

3. Lexical Resources in Text+

The diversity of (technical) data representation in lexical resources that was outlined above is also reflected in the actual data the participating institutions contribute to the Text+ project and which they have contributed to earlier projects such as CLARIN. Representation formats range from generic and customised TEI/XML serializations to legacy XML formats to table-like serializations (in the cases e. g. of lemma lists and frequency information) to geographic information captured in images of maps and to many more formats. This

⁷ <https://www.clarin.eu/>

⁸ <https://www.dariah.eu/>

⁹ <https://elex.is/>

¹⁰ <https://www.clarin.eu/content/component-metadata>

¹¹ <https://www.w3.org/2019/09/lexicog/>

¹² <https://www.iso.org/standard/68516.html>

¹³ <https://www.tei-c.org/>

¹⁴ <https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html>

¹⁵ <https://www.wiktionary.org/>

¹⁶ <https://www.dbpedia.org/>

¹⁷ <https://www.wikidata.org/>

heterogeneity makes a unified representation of the data both for retrieval and presentation a challenging task.

The set of data categories for a given lexical resource is typically specific to this resource. It may range from very broad and general categories (e.g. headword, definition) or mostly uncontroversial grammatical features of the headword (e.g. its part-of-speech) to highly specific linguistic properties that only occur in certain types of dictionaries (e.g. cognates, lexical inheritance relations), or to properties that are strongly bound to certain linguistic theories (e.g. different notations of collocation, or the treatment of homonymy/homography).

Moreover, participating institutions typically have operated local systems for serving, updating, and querying their resources and thus created specific environments for the maintenance and exploitation of their resources. These environments need also not necessarily be technically interoperable per se across different institutions. They may rely on different underlying data formats, different query languages, and different protocols for communicating with their server instances.

Given the situation described above, three principal strategies for harmonising the lexical data and the access to the data can be considered:

1. converting all lexical data into a single common format and using generic software to access the data;
2. explicitly annotating all data categories in the different formats and using generic software that operates on the annotations to access the data;
3. not changing the data or the existing infrastructure and using conversion mechanisms to relay standardised queries to the existing infrastructure and possibly also converting the query results to a specific exchange format.

Converting all lexical resources into genuine triple representations (e.g. in the form of an RDF serialization) based on an agreed-upon predicate inventory would be the easiest way to achieve a unification in case one. The generic infrastructural pillar would then be a triple store. This shifts the computational burden to the representation of the results. These will then have to be transformed into a human readable form in the context of a general research infrastructure. In this paper we do not report on preliminary work we have done in this direction.

Following case two, all data categories would have to be marked up in the lexical data. Depending on the granularity of the categories this can lead to tedious manual or automatic annotation work on all lexical data sources. The generic software would still have to be able to work on different serializations (e.g. XML vs. JSON vs. proprietary formats such as relational databases).

The scenario in case three has the advantage that the lexical data does not have to be modified or adapted. This can be essential when the data is also used in other workflows such as is the case for ongoing editions that rely on a software stack of their own. What needs to be implemented, though, is an on-the-fly transformation of incoming standardised queries into the resource specific query language. Note that this transformation might not be guaranteed to be lossless when the source query language has greater expressive power than the target query language.

In the following sections we describe our approach with respect to the third case.

4. Federated Content Search Infrastructure

The CLARIN Federated Content Search (FCS) is an established federated search engine that is specified, developed, and maintained in the context of the European CLARIN project. CLARIN works on an interoperable, integrated research online environment for the support of researchers in the humanities and social sciences. CLARIN is characterised by many participating institutions (so called *CLARIN centres*) that provide linguistic resources for a variety of research communities. These centres agree on and adhere to general requirements on how to provide data, tools, and services and work on an integrated research environment where those resources are linked by and accessed via common data formats and technical interfaces.

In this context, the CLARIN FCS' original focus is to give access to text corpora in this environment. It allows querying distributed corpora by using a standardised RESTful protocol and data formats (Stehouwer et al., 2012) that are based on the *Search/Retrieval via URL* protocol (specified by the Library of Congress, Morgan (2004)) and the *searchRetrieve* protocol (OASIS, 2013), specified by the open standards consortium OASIS (Organization for the Advancement of Structured Information Standards). The protocol allows to query data stored in online available data 'endpoints' via three operations of which the following two are relevant here:

- Operation **Explain** to identify capabilities (like supported query language(s), query vocabulary, and data formats) and available resources that a specific endpoint provides.
- Operation **SearchRetrieve** to query (a subset of) those resources at a specific endpoint.

Based on these operations, a client – including central aggregators or search portals – can query a single endpoint, or multiple endpoints in parallel. Each endpoint functions as a “proxy” for the local technical infrastructure at a specific institution in the FCS infrastructure and is typically hosted by the individual institution itself (see figure 2).

The SRU/*searchRetrieve* protocol includes means to be easily adapted to new requirements. The CLARIN FCS makes use of this mechanism and extends the protocol with a focus on accessing text corpora. These text corpora can be queried based on their fulltext representation or by addressing a variety of linguistic annotation layers, including part-of-speech, word baseforms, or phonetic transcriptions. For this, a dedicated corpus query language FCS-QL (inspired by the popular CQP¹⁸) and data representation formats were defined as key components of the protocol.

The CLARIN FCS acknowledges the problem of heterogeneity in a distributed research environment, where access to data can vary in aspects like the data format used, storage solution, query language and more. By agreeing on a lightweight retrieval protocol and simple default data formats (so-called *DataViews*) those distributed resources can be

¹⁸ https://cwb.sourceforge.io/files/CQP_Manual/

made available to end users via easy-to-use Web interfaces (like the FCS aggregator¹⁹). In this sense, the FCS does not provide a feature-complete replacement of specific search interfaces, but offers a simple way to get an overview of available resources that can also be accessed by the specialised applications at the hosting institution, if needed.

Despite its original focus on text corpora, support for requests and retrieval of lexical entries in the FCS has long been discussed and is currently implemented in an iterative work process coordinated between Text+ and CLARIN’s FCS taskforce. This seemed to be especially reasonable as many German Text+ participants already participate in the current CLARIN FCS infrastructure and therefore have experience in creating and maintaining compatible endpoints.

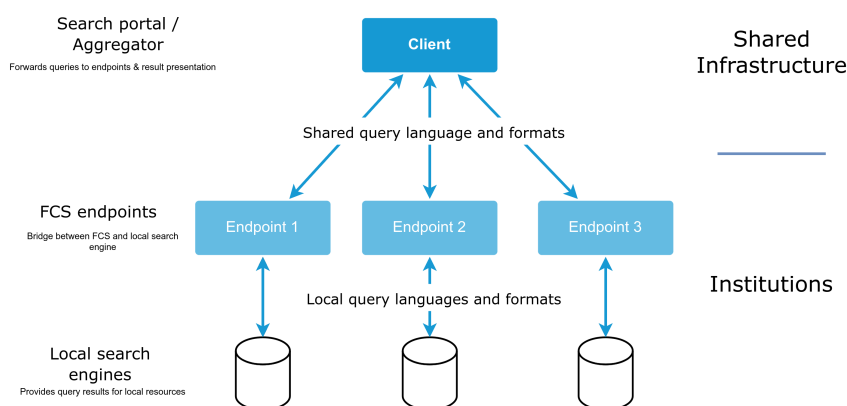


Figure 2: The general FCS architecture

5. FCS Specification Extension for Lexical Resources

The FCS specification (Schonefeld et al., 2014) in its latest version 2.0 describes two search modes:

- **Basic Search** (mandatory) is the minimum requirement to participate in the FCS infrastructure. It specifies a minimal query language for fulltext search and a simple *HITS* (Generic Hits) DataView (basic highlighting of query matches and Keyword-in-Context (KWIC) visualization) for results. This search mode allows to integrate any resource that has some form of textual representation.
- **Advanced Search** (optional) is used for searching in annotated text data with one or more annotation layers. The specification describes six types of layers (text, lemma, part-of-speech, and different forms of normalisation and transcription) for – potentially complex – queries using FCS-QL, a CQP-like query language. The result serialisation is the *ADV* (Advanced) DataView to support structured information in annotation layers. Annotations are (character) streams with offsets which also allows for e.g. audio transcriptions. This search capability is primarily focused on text corpora and similar resources.

We extend the core FCS specification in terms of *announcing*, *querying* and *retrieving* lexical resources while we ensure to seamlessly integrate and remain compatible with the

¹⁹ <https://contentsearch.clarin.eu>

overall FCS architecture. This allows to reuse features such as access control for restricted resources, automatic configuration of clients, and the overall registration of endpoints within the FCS system (see figure 3). We also adapt existing search interfaces to support users in the process of creating lexical queries and dealing with the results offered (see figure 4). The specification extension for lexical resources introduces the **LEX** search capability, and entails:

- Specifying the query language (see section 5.1) which is a “CQL Context Set”²⁰ of the Contextual Query Language²¹ (standardized by the US Library of Congress) dedicated to querying lexical entries. Its specification includes agreements on accessible fields of information (like part-of-speech, definitions, (semantically) related entries etc.) for a lexeme, and how to combine them to complex queries. This is especially challenging due to the inherently hierarchical structure of lexical data.
- Specifying common data formats for a unified result presentation (see section 5.2). On the basic level, this is achieved by a mandatory KWIC representation that allows annotating information types inline and by an advanced tabular-representation of all fields in a key-value-style. It is clearly understood that in most cases these representations can only provide a simplified view on the data. It is therefore endorsed to provide records in their complex native representation as well with examples being different TEI dialects including TEI Lex-0,²² OntoLex/Lemon,²³ and other formats.

For the current draft of the LexFCS proposal for extending the core FCS specification with regards to lexical resources refer to Körner et al. (2023). The document is still under heavy development and subject to change.

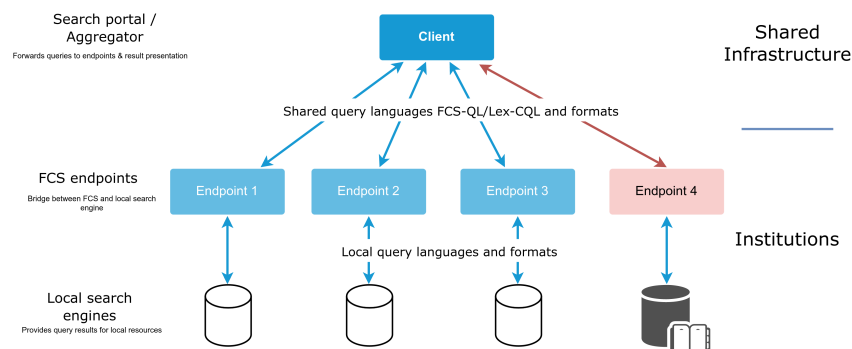


Figure 3: The FCS architecture extended to incorporate endpoints for lexical resources

5.1 Query Language – LexCQL

We propose **LexCQL** as the main query language – a subset of the CQL²⁴ which is customized for searching through fields of lexical resources. In contrast to text corpora that

²⁰ <https://www.loc.gov/standards/sru/cql/contextSets/theCqlContextSet.html>

²¹ <https://www.loc.gov/standards/sru/cql/>

²² <https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html>

²³ <https://www.w3.org/2019/09/lexicog/>

²⁴ Please note that in this paper the term “CQL” is always referring to the *Contextual Query Language* of the Library of Congress and should not be confused with the Sketch Engine’s *Corpus Query Language*.

are subdivided into sentences, paragraphs, documents etc., with their various annotation layers, lexical resources are often organised around single lexical entries with specific information that is frequently represented in the form of property-value pairs. Even though lexical entries can be nested, the LexCQL initially focuses on flat entries only.

A typical minimal set of information available in many Text+ resources contains the following searchable information types (or “indexes” in CQL):

- **lemma**: Lemma or article name,
- **pos**: Part-of-speech; it is encouraged to support – in addition to potentially more specific tagsets – the *Universal POS tags* of the Universal Dependencies project,²⁵
- **def**: Definition or description as fulltext string,
- **xr\$synonymy**, **xr\$hyponymy**, ...: Semantic relations as fulltext strings; analogous to the TEI Lex-0 types,²⁶
- **senseRef** (draft): ID/URI referring to external authority files or lexical databases, like Princeton WordNet, GermaNet, GND, or WikiData.

In the current specification draft, only the relation “=” is defined to separate queried field and value. In general, endpoints should be lenient when processing queries to improve usability and recall of results. This might include to implicitly handle spelling variants, to use normalisation procedures for historic word forms, or to support partial matches for full text fields like definitions. The CQL relation modifier “/exact” should be used and supported when searching for an exact string match.

For more complex queries, Boolean operators²⁷ such as **AND**, **OR** and **NOT** can be used and structured via parentheses if necessary. As specified by CQL, strings containing whitespaces or special characters require quoting using double quotes (”) which are optional otherwise. However, we suggest using quotes for better readability.

Examples:

1. `cat` # searching on default field, e.g. lemma; specified by endpoint
2. `lemma =/exact "läuft"` # exact string match requested
3. `def = "an edible" and pos = "NOUN"` # (implicit) partial match in def
4. `pos = ADJ and xr$synonymy = "tiny"`
5. `senseRef = "https://d-nb.info/gnd/118571249"`

5.2 Data Format for Results

The result formats currently supported by the FCS (HITS and ADV “DataViews”) are insufficient for the structure of lexical resources like dictionaries, encyclopedias, wordnets, or ontologies. The LexFCS specification proposes two additional formats.

The DataView **LexHITS**, an extension of the basic HITS DataView allows endpoints to optionally annotate information (like lemma, part-of-speech, and the record’s definition,

²⁵ <https://universaldependencies.org/u/pos/>

²⁶ https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html#crossref_typology

²⁷ <https://www.loc.gov/standards/sru/cql/contextSets/theCqlContextSet.html#booleans>

explanation or description) in a fulltext representation. This allows endpoints to reuse the mandatory KWIC format of the HITS DataView to present a simple representation of the entry but augments the results with more information. If this is not feasible, endpoints can gracefully fall back to a plain text result. The information types to be annotated are intentionally kept similar to the three most basic LexCQL fields (*lemma*, *pos*, and *def*) to emphasize the relation between queried fields and result presentation. To remain compatible with the HITS DataView the search hits marker (<hits:Hit>) is reused and extended by the XML attribute @kind. For a specific example of its use in a result's presentation, see figure 4.

Example of HITS DataView with Lex annotations extension (highlighted in red):

```
<fcs:DataView type="application/x-textplus-fcs-hits+xml">
  <hits:Result xmlns:hits="http://textplus.org/fcs/dataview/hits">
    <hits:Hit kind="lex-lemma">Apple</hits:Hit>: <hits:Hit
      kind="lex-pos">NOUN</hits:Hit>. <hits:Hit kind="lex-def">An apple
      is an edible fruit produced by an apple tree.</hits:Hit>
    </hits:Result>
  </fcs:DataView>
```

Apple: NOUN. An apple is an edible fruit produced by an apple tree.

For a more structured presentation of results, an optional DataView is currently discussed that allows providing lexical information as key-value pairs. This format aims at an easy conversion of potentially complex formats into a more general – however simplified – flat structure. As it focuses on a shallow representation, nested entries with sub-structures will need to be flattened into their own entries for search and retrieval. Discussions are ongoing to specify a set of recommendations on required and optional information types and a normative list of keys and value formats.

Example of a potential tabular key-value DataView:

```
<fcs:DataView type="application/x-textplus-fcs-lex+xml">
  <Result>
    <Entry>
      <!-- Lexeme entry -->
      <Name type="lemma">Lemma</Name>
      <Value>Lauf</Value>
    </Entry>
    <Entry>
      <!-- Standard POS tag set -->
      <Name type="pos">POS</Name>
      <!-- Multiple values are possible -->
      <Value>NOUN</Value>
      <Value>VERB</Value>
    </Entry>
```



```

<Entry>
  <!-- Custom POS tag set, as additional "pos" entry type -->
  <Name type="pos">STTS</Name>
  <Value>VVIMP</Value>
  <Value>NN</Value>
</Entry>
<!-- ... -->
</Result>
</fcs:DataView>

```

It is also suggested that resources are made available in their native representation, e. g., in various TEI dialects including TEI Lex-0, OntoLex/Lemon, and other formats in custom DataViews. If necessary, stylesheets, e. g. XSL(T), can be used as a generic way to transform TEI-based or XML-serialized RDF formats into a uniform presentation.

5.3 Search Portal / User Interface Prototype

The prototypical LexFCS search portal implementation is already available as a basis for further discussion and refinements.²⁸ It provides access to endpoints maintained by several lexical resource providers of Text+. A first stable version of the specification and an improved user interface implementation is expected until end of 2023. As a means of technical integration, the LexFCS aggregator provides an OpenAPI-compliant specification of its RESTful API.²⁹

5.4 Software and Software Libraries

The source code of all infrastructural components is provided using open-source licenses. This includes the central search portal,³⁰ parsers and validators for LexCQL in various programming languages,³¹ and specification and documentation artifacts.

6. Next Steps and Future Work

All mentioned constituents of the architecture are actively worked on and are incrementally developed. Throughout specification and implementation, feedback is provided by interested parties, particularly from but not limited to the Text+ and CLARIN projects. With a first public release in the coming months – based on the current demonstrator –, the availability and visibility of various lexical resources will be improved, including some that were not easily accessible or even unknown to the general public until now. Future work will therefore be focused on finalising the specification for the lexical search functionality. This includes the broader dissemination of the specification and providing reference implementations by different parties.

One general question that has become salient during the previous work is the problem of accessing restricted resources. Those restrictions – often because of legal obligations with

²⁸ <https://hdl.handle.net/11022/0000-0007-FBF2-D>

²⁹ <https://hdl.handle.net/11022/0000-0007-FBF2-D?urlappend=%2Fopenapi.json>

³⁰ <https://gitlab.gwdg.de/textplus/ag-fcs-lex-fcs-aggregator>

³¹ <https://gitlab.gwdg.de/textplus/>

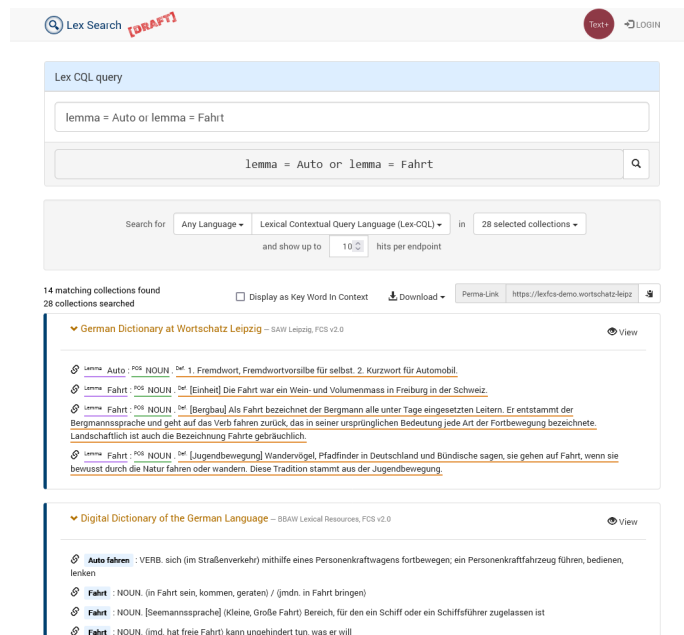


Figure 4: Screenshot of the current frontend demonstrator

publishers that prevent public access – are addressed by extending the current specification on (a) how to notify users about possible restrictions on resources, (b) how to present possibly restricted results to an end-user, and (c) how to formalise the access modalities of those resources. Scenarios might include only authenticated users being able to view results as well as providing meta information about possible hits with users being only able to view actual results at the institute or publisher in question. Using the established *Authentication & Authorization Infrastructure* (AAI) for federated authentication mechanisms with *SAML/Shibboleth* (Needleman, 2004) is currently being worked on and specifications as well as working prototypes are planned during this year in the context of the overall development of the CLARIN FCS.

A major problem of federated search systems is the absence of a global result ranking. Due to the distributed nature of the FCS, each endpoint decides how to rank its results. Those criteria are often not comparable because of differences in local retrieval systems or even the nature of the resources themselves. Results in the aggregator are therefore only grouped by resource and providing endpoint, but not in a joint representation. Using collection or provider based ranking approaches (Shokouhi & Si, 2011), result preference based on specificity of records regarding a concrete query, or other standard information retrieval methods might be a sensible approach.

Records containing lexical information referring to identical lemmas from different providers are also an issue that can significantly reduce the usability for end users. It is planned to evaluate the usage of external references to semantic wordnets – like Princetown WordNet (Fellbaum, 1998) or GermaNet (Hamp & Feldweg, 1997) –, authority files – like the *Integrated Authority File* of the German National Library (DNB, 2023) –, or other knowledge bases – like Wikidata (Vrandečić & Krötzsch, 2014) or Wiktionary – to allow a sense-based combined representation of information from different data providers.

As the software is already publicly available, third parties that wish to make their lexical data accessible over the FCS infrastructure can already set up endpoints for the aggregator.

They can also deploy a self-contained instance of the FCS including their own aggregator. Based on the specification, independent software solutions can also be developed, e. g. based on the TEI publisher.³² However, we are not currently planning to provide software beyond the reference implementations.

7. Acknowledgements

This publication was created in the context of the work of the association German National Research Data Infrastructure (NFDI) e. V. NFDI is financed by the Federal Republic of Germany and the 16 federal states, and the consortium Text+ is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 460033370. The authors would like to thank for the funding and support. Furthermore, the authors would like to thank all members of the Text+ data domain *Lexical resources* for their continuous work.

8. References

- DNB (2023). The Integrated Authority File (GND). https://www.dnb.de/EN/Professionell/Standardisierung/GND/gnd_node.html. Accessed: 2023-04-14.
- Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Bradford Books. URL <https://mitpress.mit.edu/9780262561167/>.
- Hamp, B. & Feldweg, H. (1997). GermaNet - a Lexical-Semantic Net for German. In *Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*. URL <https://aclanthology.org/W97-0802>.
- Körner, E., Eckart, T., Herold, A., Wiegand, F., Michaelis, F., Bremm, M., Cotgrove, L., Trippel, T. & Rau, F. (2023). *Federated Content Search for Lexical Resources (LexFCS): Specification*. URL <https://doi.org/10.5281/zenodo.7849753>.
- 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.
- Morgan, E.L. (2004). An Introduction to the Search/Retrieve URL Service (SRU). *Ariadne*, 40. URL <http://www.ariadne.ac.uk/issue/40/morgan/>.
- Needleman, M. (2004). The Shibboleth Authentication/Authorization System. *Serials Review*, 30(3), pp. 252–253. URL <https://www.sciencedirect.com/science/article/pii/S0098791304000978>.
- OASIS (2013). *searchRetrieve: Part 0*. Organization for the Advancement of Structured Information Standards. URL <http://docs.oasis-open.org/search-ws/searchRetrieve/v1.0/searchRetrieve-v1.0-part0-overview.html>.
- Schonefeld, O., Eckart, T., Kislner, T., Draxler, C., Zimmer, K., Ďurčo, M., Panchenko, Y., Hedeland, H., Blessing, A. & Shkaravska, O. (2014). *CLARIN Federated Content Search (CLARIN-FCS) – Core Specification*. URL <https://www.clarin.eu/content/federated-content-search-core-specification>.
- Shokouhi, M. & Si, L. (2011). *Federated Search*. Federated search edition. URL <https://www.microsoft.com/en-us/research/publication/federated-search/>.
- Stehouwer, H., Durco, M., Auer, E. & Broeder, D. (2012). Federated Search: Towards a Common Search Infrastructure. In *Proceedings of the Eighth International Conference on*

³² <https://teipublisher.com/index.html>

- Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA), pp. 3255–3259. URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/524_Paper.pdf.
- Váradi, T., Krauwer, S., Wittenburg, P., Wynne, M. & Koskenniemi, K. (2008). CLARIN: Common Language Resources and Technology Infrastructure. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco: European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2008/pdf/317_paper.pdf.
- Vrandečić, D. & Krötzsch, M. (2014). Wikidata: A Free Collaborative Knowledgebase. *Commun. ACM*, 57(10), p. 78–85. URL <https://doi.org/10.1145/2629489>.

From Structured Textual Data to Semantic Linked-data

for Georgian Verbal Knowledge

Archil Elizbarashvili¹, Mireille Ducassé², Manana Khachidze¹,
Magda Tsintsadze¹

¹Tbilisi State University, Georgia

² Univ Rennes, INSA Rennes, CNRS, IRISA, France

E-mail: archil.elizbarashvili@tsu.ge, mireille.ducasse@irisa.fr, manana.khachidze@tsu.ge,
magda.tsintsadze@tsu.ge

Abstract

The Georgian language has a difficult verbal system. To help foreigners learn Georgian, a linked-data base of inflected forms of Georgian verbs is being built: KartuVerbs. We use structured textual knowledge developed by Meurer (2007) that has a much broader scope than KartuVerbs. However, accessing its lexicographic data is challenging; the work on its base has stopped; all properties are not systematically present for every verb; some properties, important for us, do not exist. After filtering and reconstructing some properties, KartuVerbs currently contains more than 5 million inflected forms related to more than 16 000 verbs; there are more than 80 million links in the base. Response times are acceptable when running on a private machine, thus validating the feasibility of the linked-data approach. There is still a need to validate, correct and expand data. Considering the mass of data, this requires tools. This paper presents a process to transform textual structured knowledge into semantic linked data, applied to Georgian verbal knowledge. The process successively applies improvement tools. A specific one, using decision tree technique, complement occasional missing values. The scripts produced so far are freely available. They can be adapted to other applications to help transform data produced for given objectives into other data suited for different objectives.

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

1. Introduction

The Georgian language has a difficult grammar. The verbal system, in particular, is challenging. As discussed in more detail in Ducassé & Elizbarashvili (2022), there are numerous irregular verbs. Conjugation can modify both the beginning and the ending of verbs. For example, verb "to work" (mushaoba - მუშაობა), at the first person plural of present tense gives "vmushaobt" (ვმუშაობთ). Note the preradical "v" at the beginning of the verb to mark the first person, and the ending "t" to mark the plural. Some tenses, such as future, often introduce a preverb. For example, for verb "to work", the first person singular future is "vimushaveb" (ვიმუშავებ). An "i" has been inserted after the "v" marker of first person. Note that stem formant "ob" has changed into "eb". See for example Tuite (1998) for an exhaustive description. To help foreigners learn Georgian conjugation, a linked-data base of inflected forms of Georgian verbs is being built: KartuVerbs. It is accessible by a logical information system, Sparklis, see Ferré (2017). Sparklis uses

linked-data and enables powerful access and navigation as demonstrated in Ducassé (2020) and Ducassé & Elizbarashvili (2022).

To build KartuVerbs, we started from a structured textual form of the knowledge developed by Meurer (2007) for the Georgian language within the INESS project, called the Clarino base in the following. INESS is an infrastructure to help linguists explore syntax and semantics. It is multilingual and it has a much broader scope than KartuVerbs. However, accessing its lexicographic data is challenging for our target users who are not necessarily linguists. Furthermore, the work on its base for Georgian has stopped. Integrating its data into KartuVerbs both revives them and allow them to evolve. There are more than 60 possible properties, sometimes attached to inflected forms, sometimes attached to verb paradigms. Some of them are obsolete, kept for historical reasons. There are missing pieces of information. All properties are not systematically present for every verb. Some properties, important for us, do not explicitly exist, for example the ending of a form. The initial data were based on the dictionary of Tschenkéli (1965). The Georgian language has evolved since then.

After filtering and reconstructing some properties, KartuVerbs currently contains more than 5 million inflected forms related to more than 16 000 verbs for 11 tenses; each form can have 14 properties; there are more than 80 million links in the base. Response times are acceptable when running on a private machine, thus validating the feasibility of the semantic linked-data approach. There is still a need to validate, correct and expand data. Considering the mass of data, this requires tools. We are currently building experiments using machine learning algorithms.

Section 2 analyses the Clarino database with respect to our needs and introduces a typology of fields. Section 3 describes the transformation process to go from the structured text to the linked data. The process is in 3 blocks. The first block scraps the web pages into a CSV file. The second block aims at incrementally improving the data. The third block produces RDF data and integrates them into Sparklis. Section 4 describes how the decision tree algorithm can help improve a field that has occasional missing values. The field is the verbal noun, the lemma to represent a Georgian verb; there is no infinitive in Georgian. Verbal noun is crucial for our knowledge base. Section 5 discusses further work and Section 6 concludes the paper.

The main contribution of the described work is that all the scripts of the process are freely available on the web¹. They can be adapted to other applications. Those of the first block could be the base to scrap other textual sources for other languages or applications, not necessarily KartuVerbs. Those of the third block could be used to integrate into KartuVerbs (or another linked-data application) CSV data from other sources than INESS. The scripts to implement the decision tree algorithm dedicated to missing values for verbal nouns could be customized to predict occasional missing values of other fields. Furthermore, the typology of fields described in Section 2 can be used as an analysis grid to help transform data produced for given objectives into another set of data suited for different objectives.

Clarino

...	aorist	vn	...
...	1sg ვაადამიანე, გავაადამიანე	*ადამიანება	...
...

Kartuverbs (CSV)

form	tense	person	number	masdar	...
ვაადამი- ანე	aorist	1	sg	ვაადამიანება	...
გავაადამიანე	aorist	1	sg	ვაადამიანება	...
...

Table 1: First singular aorist tense form of verb ვაადამიანება (gaadamianeba): Clarino’s display and Kartuverbs records

2. The Initial Clarino Base

As already mentioned, the Clarino base is aiming at linguists whereas KartuVerbs is aiming at foreigners learning the Georgian language. Sometimes beginners would have a hard time to interpret Clarino information. For example, we already introduced verbal nouns, the lemmas representatives of verbs. They in general contain a preverb that is important to understand the meaning. As illustrated by Table 1, in Clarino the verbal noun field does not explicitly mention the preverb, because linguists can easily infer the full values of verbal noun with preverbs of the forms. In KartuVerbs, however, we need the full verbal noun, otherwise users will not be able to find the verbs in a dictionary. In the following, we call "masdar" the full version of verbal noun. In the example, Clarino’s verbal noun is *ადამიანება (*adamianeba) whereas the masdar is ვაადამიანება (gaadamianeba), for verb "to humanize somebody". Furthermore, the textual information we have access to is displayed in a condensed way. For example, as also illustrated by Table 1, all the possible inflected forms of a given verb at a given tense and at a given person are all listed in one field. The linked-data approach of KartuVerbs base requires that the relations are not factorized. For example, instead of the list of inflected forms, there should be as many records as there are inflected forms. Our process, thus, parses and interprets records.

In the Clarino base, the verbs are indexed by roots, a given root in general corresponds to several verbs, and conversely a verb can have several roots. Therefore, Clarino chooses one of the possible roots of a verb as an index. It is called the common root. Verbs have inflected forms in 11 tenses, 6 persons. Table 2 shows the Clarino fields for a form of verb "to humanize somebody". "გაადამიანებს" (gaaadamianebs) is the inflected form at 3rd person singular future. The verbal noun is "*ადამიანება". One of the 3 fields related to preverb gives "გა" (ga). The root is "ადამიან" (adamian). The stem formant is "ებ" (eb). There is no causative stem formant. The Tchkhenskeli Class is T1. Morphology Type is

¹ <https://github.com/aelizbarashvili/KartuVerbs>

Form	”გაადამიანებს” (gaaadamianebs)	Causative Stem Formant	”_”
Tense	future	Stem Formant	”ებ” (eb)
Person	3rd	Tchkhenkeli Class	T1
Number	sg	Morphology Type	active
Verbal Noun	”*ადამიანება” (adamianeba)	Verb ID	1
Preverb (3 variants)	”-”, ”გა” (ga), ”-”	Common Root ID	4
Root	”ადამიან” (adamian)		

Table 2: Clarino fields for a form of verb ”to humanize somebody”

active. It is the first verb (Verb ID = 1) of the 4th common root (in the index of Clarino, Common Root ID = 4).

The basic field for us is the inflected form. In addition, we use the following form characteristics: Tense, Person, Number, Verbal noun (that we use to build the Masdar), Preverb (3 variants), Root, Stem Formant, Causative Stem Formant, Tchkhen-keli Class, Morphology Type, verb ID, Common Root ID.

The Clarino fields do not exactly fit our needs. They can be classified as follows. Note also that certain verbs do not have all the forms for all the tenses.

1. Fields that we need, that are systematically present and that seem correct; for example, tense, person, number and some linguistic classification inherited from Tschenkeli’s work.
2. Fields that we need, that are systematically present but with specific encoding that need systematic (easy) decoding. The main example is the root of the form that, in Clarino, can contain Latin characters in the middle of the Georgian characters. They are used to signal alternatives. Another characteristic is that some verbs have different roots at different tenses. As the base is indexed by roots, Clarino decided on a common root and attached all the possible roots to the verb and not to the forms; after extraction, forms have all possible roots of the verb, all but one being incorrect. Note that correcting these two features can be done by simple scripts.
3. Missing fields but there is enough information in the Clarino fields to deduce the information. For example, preverbs can be deduced from 3 different Clarino fields. Masdars can be deduced from preverbs and verbal noun.
4. Fields that we need, that are systematically present but with occasional mistakes; for example, verb ID. This field category is typical of any source of data. It is almost impossible to create a large body of data and make no mistake.
5. Fields that we need, that are not systematically present but for which the absence can be normal; for example, preverb and stem formant. Forms at present tense often do not have any preverb. Forms at aorist tense often do not have any stem formant.
6. Fields that we need that should always be present, but that are occasionally absent; for example, verbal noun.

7. Missing fields and there is no information in Clarino to deduce them; for example, English infinitive.
8. Fields that we do not need (yet).

In the remaining of this article, Section 3.2 briefly presents the processing of fields of categories 1 to 3. Section 4 describes how machine learning is used to address fields of category 6.

3. Transformation Process

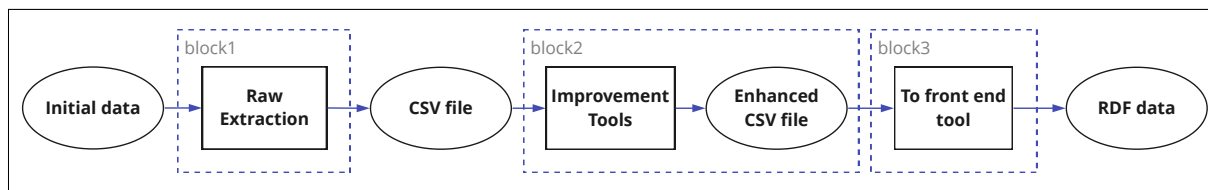


Figure 1: Global structure of the transformation process

As illustrated by Figure 1, the process to transfer data from Clarino to KartuVerbs consists of 3 main blocks. The first one starts from the Clarino web page and generates an intermediate CSV File. The second one consists of several processes to improve the raw data (in relation with issues described in Section 2). The third one transforms the CSV data into RDF data and creates a SPARQL endpoint.

3.1 From Clarino To An Intermediate CSV File

First the Clarino web pages are scraped. The result is a 22 million lines, 625 MB, Json file in pretty format, one line per form with fields. In Clarino, information is hierarchical, whereas our aim is to generate relational data so that information can be accessed from any piece of data (see Ducassé (2020) and Ducassé & Elizbarashvili (2022) for further details). The Clarino structure starts with root, then verb, whereas our key information is inflected form. The process thus flattens the structure and generates tuples whose first field is an inflected form. In Clarino, it is possible to have different values for the same field. In that case, several lines are generated. For example, for a given tense, there may be n possibilities for a given person. In that case, there will be n lines with the same tense and person, and with different forms. Then a Python script converts json to csv. The Clarino properties that are not used for KartuVerbs are filtered out to keep only 14. The result is a 610 MB file.

3.2 Improvement of Intermediate Data

For fields of types 1 to 3, we wrote scripts to improve data.

For example:

- Root: Verb root can contain Latin characters among the Georgian characters. They are used to signal either alternatives (for example, "A" means either "ღ" or nothing

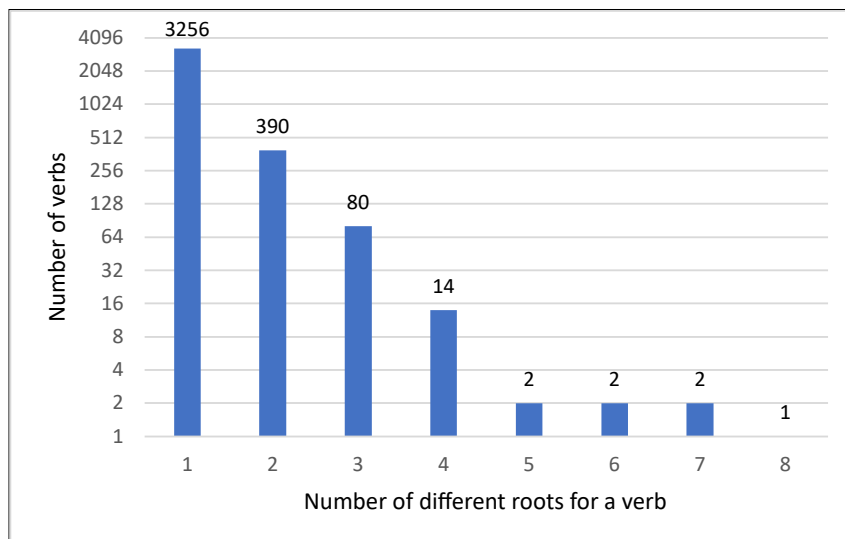


Figure 2: Almost 500 verbs show different roots in their forms

as in "ვად" = "ვად" (vad) OR "ვდ" (vd); or strict absence (for example, "ა" means that the "ა" that may be present in other forms must be absent, as in "თარგმან" = "თარგმნ" (targmn)). The script duplicates alternatives by adding new lines and changes Latin characters into either a Georgian character or no character. In addition, Figure 2 shows that verbs may have different roots in their forms. If the vast majority (3256) of the verbs have only one root throughout all their forms, 390 have 2, 80 have 3, 14 have 4, 2 have 5, 6 or 7, and 1 even has 8 roots in the forms of the verbs built. For example: common root "მბობ" ("mbob") leads to verbs around the meaning of "to say" with 7 different roots: "ამბ" ("amb"), "თქ" ("tk"), "თქვ" ("tkv"), "თხრ" ("tkhr"), "მბობ" ("mbob"), "ტყ" ("t'q"), "უბნ" ("ubn"). Common root "სვლ" ("svl") leads to verbs around the meaning of "to go" and "to come" with 8 different roots: "არ" ("ar"), "დი" ("di"), "ვედ" ("ved"), "ველ" ("ved"), "ვიდ" ("vid"), "ვლ" ("vl"), "ს" ("s"), "სვლ" ("svl"). For a given verb, Clarino gives all the possible roots attached to a common root. However, a given form only contains one of them. The script eliminates all the irrelevant roots from form descriptions.

- Preverb: some forms start with a preverb that gives an indication similar to English postpositions. There are more than 10 possible preverbs. For example, "ა" (a) conveys the same idea as "up". This information is especially crucial to understand Georgian conjugation. It is split into 3 fields in Clarino. If any of the 3 fields is present in a form, the script collects it. If several of the 3 fields are present with different values, the script keeps the value present in the form.
- Verbal noun and Masdar: In section 2 we explained why we must transform Clarino's verbal noun to generate a masdar. When the verbal noun is available, the script merges it with the preverb. For example, for the form "გადა-ვა-კეთ-ეთ" (gada-vaket-et), the preverb is "გადა" (gada), the verbal noun is "*კეთება" (*keteba), the deduced "masdar" is "გადაკეთება" (gadaketeba);

CSV file				
form	tense	person	number	...
ვაადამიანებ	present	1	sg	...
...

RDF triplets		
<ვაადამიანებ>	<tense>	<present> .
<ვაადამიანებ>	<person>	<1> .
<ვაადამიანებ>	<number>	<sg> .
...		

Table 3: Extract from the CSV file and the corresponding RDF triplets

3.3 From CSV To SparkLis

Another Python script converts the CSV format into RDF (Resource Description Framework) Turtle N-triplets to create linked data compatible with Sparklis, the logical information system developed by Ferré (2017) and used in KartuVerbs for navigating in the data. For example, Table 3 shows an extract of the CSV file line and the corresponding Turtle N-triplets entries. Basically, one line of the CSV file with n columns is transformed into $n - 1$ triplets of the form " l_id " " $p_property\ name$ " " $p_property\ l_value$ ". Where " l_id " is the content of the first column of line " l ", " $p_property\ name$ " is the first line of column " p " and " $p_property\ l_value$ " is the content of the slot line " l " column " p ".

In order to support the data into two languages, this script also adds transliteration of Latin characters into Georgian characters. The result is a 3.2 GB turtle file. Considering the huge amount of data, it is crucial that the file is indexed for SPARQL to give answers with acceptable response times. Indexation is done with an open-source packages: *apache-jena* and *apache-jena-fuseki*. The result is a 11 GB file. The final step is to create an endpoint for Sparklis, namely to start a SPARQL database server and load the RDF Turtle N-triplets, a standard procedure for Sparklis applications.

4. Decision Tree to Improve Occasional Missing Fields

This section describes the machine learning experiment we made for the improvement of the verbal noun field that is of type 6. Namely, this field should always be present, but it is occasionally absent. Over 600 000 forms (corresponding to 4640 verbs) do not have a verbal noun. Thus, filling up the blanks can be of significant importance. Furthermore, the needed value might already be present in the other records. Indeed, Figure 3 shows that a common-root can lead to multiple verbs. We can see that only 670 common roots lead to a single verb. On the other end of the range, 1 common root leads to 153 different verbs. A common-root leads in average to 9 different verbs. Even if not all verbs coming from a common root have the same verbal noun, the root is crucial to build a verbal noun. In addition, there is only one common root without any information about verbal noun.

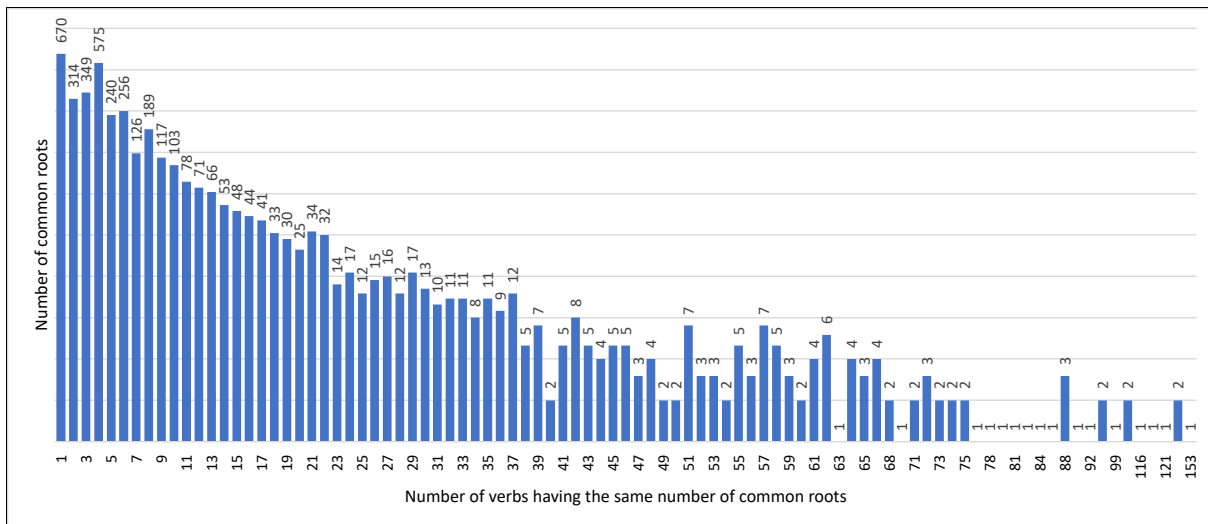


Figure 3: A common root can lead to several verbs - up to 153, 9 in average

The research hypothesis is that there is enough information in the input dataset to predict the missing verbal nouns by machine learning.

The remaining of this section, introduces the experimental setting (input data and training process), presents the results, discusses them and presents some implementation issues.

4.1 Experimental setting

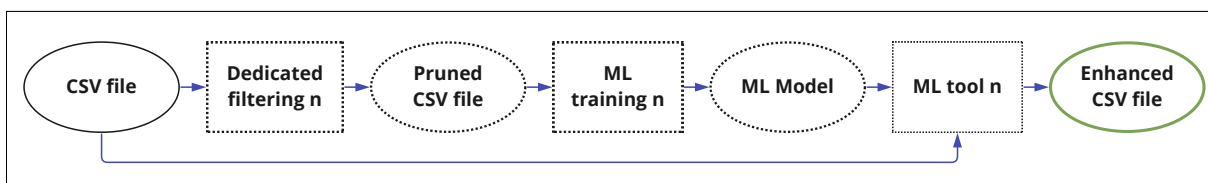


Figure 4: Structure of Machine Learning Tool n

Our improvement process is incremental and, when applying a given tool, not all forms are necessarily reliable. Thus, as illustrated in Figure 4, when using machine learning we first filter out all forms for which a doubt still exists. In particular, for this experiment, all lines without verbal noun have been filtered out before training. After filtering, the input file consists of 3.8 million lines, each one containing a Georgian inflected verb form and 14 of its features. We then train a model, and from this model we generate an enhanced CSV file that can be enhanced further by other tools.

Input Data As discussed in Section 2, fields of type 5 may exhibit an absence of value and it is not necessarily an error. Therefore, forms with such missing values have to be kept in the training data. Missing values, however, have an impact on the chosen machine learning technique (see discussion below). Figure 5 shows the percentage of missing values for these fields. Fields Ending and Preradical have been built from Clarino information by

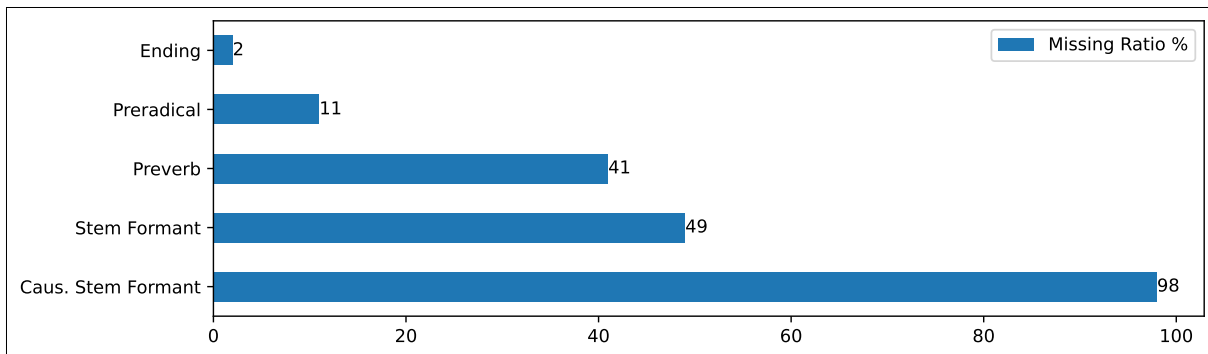


Figure 5: Missing values in input data

scripts not described in this article. Note that Causative Stem Formant is mostly absent. An absence of preverb or stem formant can be perfectly valid for some tenses and verb groups. They are, however, key in the structure of the verbal noun when they exist.

Machine learning algorithms work on numbers. As our fields are mostly symbolic, we had to encode them into numbers. Missing values are encoded by "0". Fields with a finite number of possible values (tenses for example) are simply encoded by constants. String fields require more subtle treatments. For some fields, it is sufficient that the encoding is a function (to one string corresponds a single encoding, the same encoding may correspond to several strings). For other fields, it is crucial to build a bijection between string representations and numerical representations in order to be able to interpret the result properly (in a unique way). Here it is crucial to be able to know what the string is suggested by the ML algorithm for missing verbal nouns. It should be noted that missing verbal nouns most probably already exist in other forms in the database. For the fields where a function is sufficient, to encode Georgian characters we refer to the UTF-8 encoding scheme, where Georgian characters are represented by three-byte sequences. The first two bytes are redundant for the conversion process. Thus, to encode a Georgian word into a numeric representation, we extract the last byte from each character and sum their decimal values.

$$f(string) = \sum_{i=1}^{string_length} Byte_value(Last_Byte(UTF-8(character_i)))$$

where $character_i$ is set of individual characters of a Georgian text $string$.

For example, "გდ" \Rightarrow "გ" + "დ" \Rightarrow $b'\backslash xe1\backslash x83\backslash x92'$ + $b'\backslash xe1\backslash x83\backslash x90'$ \Rightarrow $b'\backslash xe1\backslash x83\backslash x92'$ + $b'\backslash xe1\backslash x83\backslash x90'$ \Rightarrow $value('b\backslash x92')$ + $value('b\backslash x90')$ \Rightarrow $146 + 144 = 290$.

However, for verbal nouns, a one-to-one correspondence (bijection) between text and numeric versions of Georgian Verbal nouns is required. Indeed, Verbal Noun is the target variable for our task prediction. With the previous encoding, 290 can be decoded as "გდ" and "დგ" as well. Therefore, to each verbal noun we assign a different integer in range $[0, 6538]$ and we keep a correspondence table for the 6539 different verbal nouns.

Training process To train our data we use a supervised learning model, Decision Tree, for the following reasons. It is suited to handle multiclass classification tasks (as discussed in

Bansal et al. (2022)). Our task is, indeed, a classification because a predicted best match of verbal noun should be selected from a set of verbal nouns included in the input file. Furthermore, Decision tree model is non-parametric; before training our model we did not have to determine any parameters. Decision tree algorithm possesses very low complexity. This is crucial considering the size of our input data. Last but not least, Decision Tree model is not influenced by missing values. This is also crucial because our original data contain missing values as illustrated in Figure 5. For the experiment, the filtered database is split into 2 parts: 80% for the training subset and 20% for the testing subset, a typical ratio in data science. We split the dataset using either systematic randomization or a different seed number. Both approaches of splitting led to the same evaluation scores across different runs. More than 10 seed numbers were tried and the resulting scores were the same for all the attempts. The actual verbal nouns were removed from the test dataset and kept for later verification.

4.2 Results and Discussion

	Precision	Recall	F1-score	Support
Accuracy			1.00	759663
Macro avg	1.00	1.00	1.00	759663
Weighted avg	1.00	1.00	1.00	759663

Table 4: Classification report for Decision trees with 14 form characteristics

Table 4 shows the classification report, which assesses the prediction performance for a classification model. The report generates three common metrics that we use to access the quality of the model. Precision is the percentage of correct positive predictions relative to total positive predictions. Calculation: Number of True Positives (TP) divided by the Total Number of True Positives (TP) and False Positives (FP).

$$Precision = \frac{TP}{TP + FP}$$

Recall is the percentage of correct positive predictions relative to total actual positives. Calculation: Number of True Positives (TP) divided by the Total Number of True Positives (TP) and False Negatives (FN).

$$Recall = \frac{TP}{TP + FN}$$

F1 score is a weighted harmonic mean of precision and recall.

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

The macro-averaged scores are computed by taking the arithmetic means (unweighted means) of all the per-class scores (in our case of all the VN precision scores, recall scores

and f1 scores). This method treats all classes equally regardless of their support values. The weighted-averaged scores (precision, recall and f1 scores) are calculated by taking the mean of all per-class scores while considering each class’s support. The ‘weight’ essentially refers to the proportion of each class’s support relative to the sum of all support values. On the Table 4, accuracy refers to micro averaging. It computes a global average F1 score by counting the sums of the True Positives (TP), False Negatives (FN), and False Positives (FP).

$$Accuracy\ F1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

In multi-class classification cases where each observation has a single label, the micro-F1, micro-precision, micro-recall, and accuracy share the same value (i.e., 1.00 in our case). For each metric, the closer to 1, the better the model. 1 corresponds to 100% of prediction rate.

	Precision	Recall	F1-score	Support
⋮	⋮	⋮	⋮	⋮
813	1.00	0.99	0.99	96
6507	1.00	0.99	0.99	82
4398	0.98	1.00	0.99	62
4094	0.99	1.00	0.99	77
1021	0.97	1.00	0.99	38
6488	0.92	1.00	0.96	12
4882	1.00	0.89	0.94	9
361	0.80	0.87	0.83	113
6453	0.59	0.47	0.52	47
6448	0.00	0.00	0.00	1

Table 5: Classification report for individual verbal noun prediction scores

In Table 4 all scores (macro, micro, and weighted scores) reflect a 100% prediction rate. However, it is important to note that these scores represent averages. Table 5 gives the individual verbal noun prediction scores that are less than 100%, ranging as low as 83%, 52%, and even 0%. The 0% score is due to the fact that there is only one occurrence of verbal noun *ფშვენა (*pshvena, related to a family of verbs around heavily breathing) in the entire training and test datasets. Therefore, there are no other instances for comparison. The 52% corresponds to ბორძიკ (bordzik’, to stumble); a verb for which in the training dataset, half of the occurrences have verbal noun ბორძიკ (bordzik’) and the other half have verbal noun *ბორძიკება (*bordzik’eba). We applied the trained model to the 600 000 forms with missing verbal noun. We are developing tools to facilitate the validation of the results. We are planning to use a crowd-sourcing platform (see Section 5). We are designing heuristics to reduce the number of results to be manually checked and rank the

results such that experts would be asked to double check the most dubious results first. The heuristics that we have identified so far are,

- A predicted verbal noun is questionable when it does not match the root of the form.
- If forms of the same verb (identified by their Clarino Id) have different verbal nouns these verbal nouns are questionable. It might be the case that all are valid but in that case they should all be attached to all the forms.
- Verbal nouns without a vowel at the end are questionable. Experts can manage with them but not beginners. For example, an expert will understand that verbal noun "ყვილი" (qivil) should be understood as "ყვილი" (qivili, to crow), but a beginner would be lost.
- It is not necessary to check all the forms of a verb. Samples are sufficient, sampling should take into account at least the tense (preferably one that uses a preverb, future for example) and roots. Some verbs exhibit different roots at different tenses or persons.

We have tried the first heuristic combined with the last one, out of the initial the 600 000 forms with missing verbal noun, 100 000 forms have a predicted verbal noun that does not match its root. Taking a sample of these forms resulted in a set of 153 forms that have been checked by hand. Approximately half of them were correct. Although our trained model has achieved a 100% prediction rate, our heuristic observations indicate that the results are not consistently correct. We conjecture that this discrepancy arises from the fact that a limited number of examples in the training data correspond to verbs with a missing verbal noun. Another possibility is that the encoding of Georgian texts utilizes a non-bijective method. Except for verbal nouns, there is no one-to-one correspondence between the Georgian texts and their encoded versions. This unique correspondence presents challenges, as it can result in excessively large and sparse numbers, rendering the machine learning algorithm ineffective or sometimes even impossible to implement.

4.3 Implementation issues

	Virtual Server	Laptop: ROG Zephyrus M16
Model	Intel(R) Xeon(R) Gold 5118 CPU @ 2.30GHz	12th Gen Intel(R) Core(TM) i9-12900H
CPU MHz	2294.612, 32-64 cores	2900.000, 20 cores
Cache size	16896 KB	24576 KB
Memory	64-96 GB	48 GB
Swap Memory	4 GB	400 GB

Table 6: Hardware characteristics

Considering the amount of data of the base (several millions of records), implementation issues are important. Table 6 shows the Hardware characteristics of the experiment. All

the experiments were done in the Linux distributions - Debian 11 (bullseye) and Ubuntu 22.04.2 LTS (jammy). We used free, open-source platform - Python programming language through Jupyter notebook (Anaconda Navigator) and other Unix-tools (awk, sed ...). The Decision tree algorithm is not suitable for variables continuous in nature Bansal et al. (2022). Indeed, using integers instead of floats for verb ID, the F1-score for the predictions went from 77% to 100%. In order to evaluate the performance, we conducted additional tests by training the model on datasets consisting of 10 fields (form, preverb, preadical, root, stem formant, causative stem formant, ending, verb paradigm sub-ID, clarino ID, verbal noun) instead of 15, and 5 fields (form, root, verb paradigm sub-ID, clarino ID, verbal noun). Both datasets yielded similar average results. However, when examining the individual verbal noun prediction rates, the model trained with the larger dataset model outperformed the others. In terms of machine resource consumption and time efficiency, our experiments revealed that there is not a significant disparity between processing 15 fields, 10 fields, and 5 fields. Regardless of the number of fields processed, the model utilized a substantial amount of memory during the prediction phase. Specifically, for our input file, the model required approximately 50 GB of memory, which exceeds the typical memory capacity of machines. To overcome this challenge, we resolved the issue by expanding the SSD-based swap memory. With this configuration in place, our model successfully completed training, testing, and prediction tasks within approximately 2 minutes and 30 seconds for 15 fields, 2 minutes for 10 fields, and 1 minute and 30 seconds for 5 fields input files. Hence, in this particular context, a regular machine or laptop equipped with ample SSD storage can be employed to train extensive datasets using a decision tree algorithm. Although this may lead to a longer processing time, it remains a viable option. We tried another robust model for classification, Support Vector Machine learning model. With only 100,000 rows of input data, and even using maximum cores for parallel computations, it took over 100 times longer than with Decision Tree for the entire data (4 million of lines). It seems impossible to obtain results in a reasonable time for our case.

5. Perspectives

The perspectives are to refine the process and add more improvement tools. We will apply decision tree to occasional incorrect value fields. In Stefanovitch et al. (2022), the authors use machine learning and transformer based models to classify sentiments in Georgian texts. In so doing, they automatically derive all possible morphemes of a verb, based on its root and two additional parameters: a list of potential preverbs, and a dependent noun. We will investigate if their process could be adapted to help improve or validate our morphemes. Another perspective is to investigate how BERT (see Devlin et al. (2019)) could help to add further improvement tools. Language-specific BERT models are not currently available for Georgian. However, there exist multilingual models that include Georgian language, see Wang et al. (2020); Conneau et al. (2019); Pires et al. (2019). Besides conjugation we also plan to use it for different tasks such as morphological tagging and Named Entity Recognition classification, along the lines of the work for Estonian of Kittask et al. (2020). It will enable us to enrich the base with new properties. Of possible interest are also the network structures to learn word embedding, sentence embedding, and sequence generation with transformers like BERT, introduced in Zhou et al. (2020).

We plan to use crowdsourcing in order to give a chance to users and experts to signal mistakes or missing information. The IRISA platform Headwork² will be used. Indeed, the Georgian language contains so many exceptions to the conjugation rules that we do not expect machine learning tools, however efficient, to produce 100% correct information.

6. Conclusion

In this paper, we described a process to transform textual structured knowledge into semantic linked data, applied to Georgian verbal knowledge. The target users and the objectives of the two knowledge bases differ. Hence, initial data have to be reconstructed and interpreted to fit KartuVerbs objectives. The described process aims at applying successively a number of improvement tools. A specific one, using decision tree for machine learning, has been described in detail to complement occasional missing values. The average F1-score for the generated model is 100%. The scripts produced so far are freely available on the net³. They can be adapted to other applications to help transform data produced for given objectives into other data suited for different objectives.

7. Acknowledgements

We are indebted to Paul Meurer who granted us a private access to a web version of the base behind the Georgian functionalities of <https://clarino.uib.no/iness>. We thank Mikheil Sulikashvili for his help to scrap Clarino web pages.

This research PHDF-22-1840 is supported by Shota Rustaveli National Science Foundation of Georgia (SRNSFG) and by ANR Project SmartFCA, ANR-21-CE23-0023.

8. References

- Bansal, M., Goyal, A. & Choudhary, A. (2022). A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning. *Decision Analytics Journal*, 3, p. 100071. URL <https://www.sciencedirect.com/science/article/pii/S2772662222000261>.
- Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzman, F., Grave, E., Ott, M., Zettlemoyer, L. & Stoyanov, V. (2019). Unsupervised Cross-lingual Representation Learning at Scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics*, pp. 8440–8451. URL <https://aclanthology.org/2020.acl-main.747>.
- Devlin, J., Chang, M.W., Lee, K. & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *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://aclanthology.org/N19-1423>.
- Ducassé, M. (2020). Kartu-Verbs: A Semantic Web Base of Inflected Georgian Verb Forms to Bypass Georgian Verb Lemmatization Issues. *Z. Gavriilidou, M. Mitsiaki &*

² <https://druid-garden.irisa.fr/spipollhw/>

³ <https://github.com/aelizbarashvili/KartuVerbs>

- A. Fliatouras (eds.) *Proceedings of XIX EURALEX International Congress, volume 1. SynMorPhoSe Lab, Democritus University of Thrace*, pp. 81–89. URL https://euralex2020.gr/wp-content/uploads/2020/11/EURALEX2020_ProceedingsBook-p081-089.pdf.
- Ducassé, M. & Elizbarashvili, A. (2022). Finding Lemmas in Agglutinative and Inflectional Language Dictionaries with Logical Information Systems: The Case of Georgian verbs. *Proceedings of XX EURALEX International Congress*, p. 6.
- Ferré, S. (2017). Sparklis: An Expressive Query Builder for SPARQL Endpoints with Guidance in Natural Language. *Semantic Web: Interoperability, Usability, Applicability*, pp. 405–418. URL <http://www.semantic-web-journal.net/content/sparklis-expressive-query-builder-sparql-endpoints-guidance-natural-language-1>.
- Kittask, C., Milintsevich, K. & Sirts, K. (2020). Evaluating multilingual BERT for Estonian. A. Utka et al. (ed.) *Human Language Technologies – The Baltic Perspective. IOS Press Online*.
- Meurer, P. (2007). A computational grammar for Georgian. *International Tbilisi Symposium on Logic, Language, and Computation. Springer*, pp. 1–15.
- Pires, T., Schlinger, E. & Garrette, D. (2019). How multilingual is Multilingual BERT? *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4996–5001.
- Stefanovitch, N., Piskorski, J. & Kharazi, S. (2022). Resources and Experiments on Sentiment Classification for Georgian. *International Conference on Language Resources and Evaluation*, pp. 1613–1621.
- Tschenkéli, K. (1965). Georgisch-deutsches Wörterbuch, volume 2. *Amirani-Verlag Zürich*.
- Tuite, K. (1998). Kartvelian morphosyntax: Number agreement and morphosyntactic orientation in the South Caucasian languages. *Lincom Europa Munich*.
- Wang, Z., Mayhew, S., Roth, D. et al. (2020). Extending Multilingual BERT to Low-Resource Languages. *Findings of the Association for Computational Linguistics: EMNLP 2020. Online: Association for Computational Linguistics*, pp. 2649–2656. URL <https://aclanthology.org/2020.findings-emnlp.240>.
- Zhou, M., Duan, N., Liu, S. & Shum, H.Y. (2020). Progress in Neural NLP: Modeling, Learning, and Reasoning. *Engineering*, pp. 275–290. URL <https://www.sciencedirect.com/science/article/pii/S2095809919304928>.

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

Tamila Krashtan

Lviv Polytechnic National University, Lviv, Ukraine

E-mail: tamila.krashtan@gmail.com

Abstract

This paper presents a new search engine developed for the Large Electronic Dictionary of the Ukrainian Language (also known as VESUM)—a project aiming at generating a morphological dictionary for the Ukrainian language, which is also used in a Ukrainian POS-tagger. The aim of the current project is to set up a more user-friendly interface with broader search options, which at the same time provides more information contained in the Dictionary database. The newly developed search functionality for the Ukrainian Dictionary is built upon the search engine created for the Belarusian grammar database and utilizes grammar tags defined in the VESUM database. It enables the usage of wildcards in search queries and allows a user to set up search grammars. The developed system provides more extensive search options and a way of displaying lemma information that is more structured and transparent both for professionals and non-linguists. It is well-suited for the addition of new tags and search parameters (including, but not limited to, conjugation classes and variations in the orthography of certain words) which will be featured in future versions of the software.

Keywords: search engine; online dictionary; Ukrainian; VESUM

1. Introduction

There is a multitude of well-developed NLP tools and databases available for the Ukrainian language that are widely used across various software applications. Nevertheless, the information in such databases is often stored in formats that are not easily usable and are not conveniently consumable by the public.

One such database is the Large Electronic Dictionary of the Ukrainian Language, also known as VESUM, after its Ukrainian acronym (Rysin & Starko, 2005–2023). It is a morphological dictionary that describes the lemmas of the Ukrainian language along with their inflected forms, supplied with grammatical and semantic tags. The data from this dictionary is used in such projects as LanguageTool spellchecker or Wikipedia search (Rysin & Starko, 2020). The creators of the dictionary also provide the data both in the raw format and through a simple search form. However, in order to comprehend the search results, one may need to thoroughly go through pages of documentation, while any search that goes beyond a keyword requires one to create own query scripts to run on the raw data.

The aim of the current project is to build a user-friendly search interface for the VESUM that would leverage most of the data available in the dictionary, including some pieces of information that might be inaccessible through the existing simple search form. This task includes several steps: 1) an analysis of the VESUM structure and the ways to put this structure into queries; 2) an overview of the existing search systems for the dictionaries and grammatical databases of other languages; 3) an implementation of the first versions of the search tool and planning of the future directions of the development. Each of the outlined steps is described in detail in the sections that follow.

2. VESUM

VESUM, the Large Electronic Dictionary of the Ukrainian Language, was created in 2005 as a part-of-speech database. Since then, the dictionary itself or its modifications have been utilized in a number of projects, including search engines of Ukrainian Wikipedia and the General Regionally Annotated Corpus of Ukrainian, or GRAC, as well as the Ukrainian spellcheckers in LibreOffice or LanguageTool (Rysin & Starko, 2020).

Since the dictionary is used in spellchecking software, it contains not only those Ukrainian lemmas that are considered a part of literary language, but also corrupted, colloquial, dialectal, and other non-standard forms, each marked accordingly to indicate its usage mode.

The initial version of the dictionary was based on several printed dictionaries of the Ukrainian language (Krytska et al., 2011; Busel, 2005; Karpilovska, 2013), and the database is constantly being updated with new entries, in particular, the untagged words found in the GRAC (Shvedova et al., 2017–2023). The GitHub page of the dictionary (Project to generate POS tag dictionary for the Ukrainian language) contains the latest available version of the database and enables users to make suggestions on corrections and additions to the dictionary. Those are reviewed by the maintainers.

2.1 Internal Representation

The internal representation of the dictionary does not contain all inflected forms of each lemma, but rather lists lemmas with a group of special tags describing the lemma from grammatical and lexical standpoints. They are then used to automatically generate the visual representation of the dictionary (see section 2.2 below). Examples of lemmas and corresponding forms are shown in the table 1.

Each tag group starts with the base tag showing the part of speech and the inflection class, if applicable. It may be followed by a series of marks showing specifics of the inflected forms' generation for this lemma, including sound alternations or alternative

endings for certain forms. Additional tags placed after them may indicate supplemental grammatical (e.g. perfective vs imperfective aspect for verbs) as well as semantic information (e.g. animate vs inanimate nouns, specific indication of family names or names of cities). Lastly, the lemma can be marked with flags indicating its usage (colloquialisms, vulgarisms, alternative spellings, orthographic variations, etc.)

деренькотання /n2n
деренькотати /v1.cf.advp :imperf
деренькотіння /n2n
деренькотіти /v1.cf.advp :imperf
деренькучий /adj
дерепресія /n10.p1
дерешуватий /adj
держава /n10.p1.i1k1
державдитор /n20.a.p.ke.< :ua_2019
державець /n22.a.p.<

Table 1: An excerpt from the internal representation of VESUM.

For example, the lemma “деренькотати” (“to jar”) shown in the table 1 is tagged as an imperfective (:imperf) verb of the first conjugation group (v1) that may use synthetic future tense forms (cf) and has a corresponding adverbial participle (advp). Similarly, the lemma “державдитор” (“state auditor”) is described as an animate (<) masculine noun of the second declension group (n20), ending in -a in singular genitive form (a), in -e – in singular vocative form (ke) and having no alternations in plural forms (p). Additionally, it’s indicated to follow the spelling norms introduced by the Ukrainian orthography of 2019 (:ua_2019).

2.2 Generated Visual Representation

The internal representation shown in the previous section provides a way for systematic and economic storage of the lemma descriptions. However, for the dictionary to be used in real-life applications, it is preferable to have a list of all the inflected forms shown explicitly. In the VESUM, one of the forms of such a list is called a visual representation. It is generated from the internal representation and has a set of tags of its own: it copies some semantic and lexical information, removes the tags used solely for mechanical forms generation (e.g. the tags “a” or “ke” shown above to represent certain endings), and adds the characteristics of the inflected forms. Table 2 shows examples

of visual representations for several parts of speech (for the sake of brevity, only part of the inflected forms is shown for each lemma).

As can be seen from the table 2, VESUM generates all the inflected forms for a given lemma, including cases when several alternative versions are possible for a certain form. For instance, the masculine (m) locative (v_mis) of the lemma “державдитор” (“state auditor”) can surface both as “державдиторові” and as “державдитору”. Possible differences in usage of the alternative forms are indicated as well: the plural (p) accusative (v_zna) of the adjective “дернуватий” (“soddy”) is “дернуватих” for animates (ranim) and “дернуваті” for inanimates (rinanim).

Verb	Noun	Adjective
деренькотати verb:imperf:inf	державдитор noun:anim:m:v_naz:ua_2019	дернуватий adj:m:v_naz
деренькотать verb:imperf:inf:short	державдитора noun:anim:m:v_rod:ua_2019	дернуватого adj:m:v_rod
деренькочи verb:imperf:impr:s:2	державдиторі noun:anim:m:v_mis:ua_2019	дернуватім adj:m:v_mis
деренькочім verb:imperf:impr:p:1	державдиторові noun:anim:m:v_mis:ua_2019	дернуватому adj:m:v_mis
деренькочем verb:imperf:pres:p:1:subst	державдитору noun:anim:m:v_mis:ua_2019	дернуватуою adj:f:v_oru
деренькотатиму verb:imperf:futr:s:1	державдиторе noun:anim:m:v_kly:ua_2019	дернуватим adj:n:v_oru
деренькотала verb:imperf:past:f	державдитори noun:anim:p:v_naz:ua_2019	дернуватих adj:p:v_zna:ranim
деренькотали verb:imperf:past:p	державдитори noun:anim:p:v_kly:ua_2019	дернуваті adj:p:v_zna:rinanim

Table 2: Examples of visual representation in VESUM.

2.3 Current Search Form

The search form that has been used for VESUM so far (Rysin & Starcko, 2005–2023; see figure 1) provides only the basic functionality: search across the dictionary’s visual representation by lemmas, inflected forms, or their parts with no options to utilize the grammatical information provided by the dictionary. Apart from that, the results are directly replicating the format of the visual representation, i.e., show only the inflected forms with their tags in the machine- rather than human-readable format.



Figure 1: The previous VESUM search interface.

The limited search capabilities of this form are far from providing users with all the data that can be retrieved from the database. A new user-friendly search interface enabling the creation of search queries based on the grammatical features of the lemmas or their forms would make the database more convenient both for linguistic research and for day-to-day usage as a reference dictionary. Apart from that, structured inflection tables instead of the bare lists of forms would constitute a nice addition to the updated interface.

3. Online Dictionaries of Other Languages

The next step in creating the new search interface is an analysis of similar existing tools and surveying the possibility of their adaptation to VESUM. This section provides a short overview of the search engines available for dictionaries and grammatical databases of other languages. It focuses on the Slavic languages since those have similar grammatical categories compared to Ukrainian, and on English as a language with a wide range of lexicographic resources.

3.1 Grammatical Dictionary of Polish

The Grammatical Dictionary of Polish (Kieraś & Woliński, 2017; Grammatical Dictionary of Polish; see figure 2) provides a variety of search options for the lemmas and their forms: lexical classes, frequency, gender, aspect, etc. It provides grammatical information, inflection tables, and—for some of the lemmas—clarifications on the meaning. The downside of this dictionary is its unintuitive interface that makes it hard for the users to do an extensive search without studying the documentation for the dictionary.

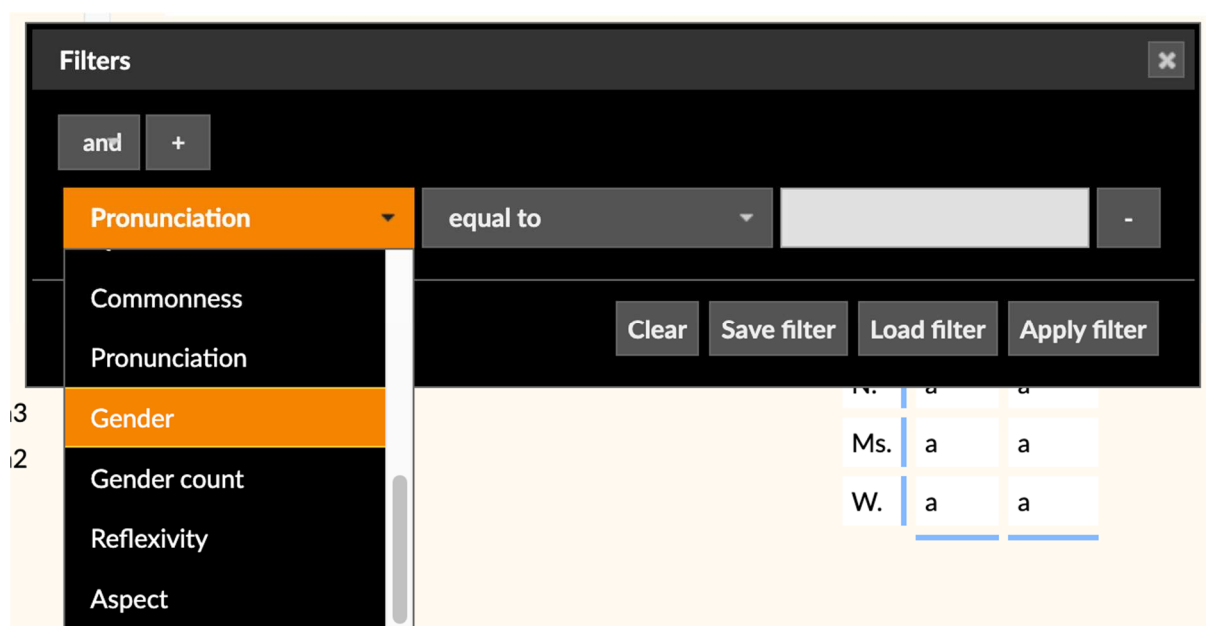


Figure 2: Grammatical Dictionary of Polish.

3.2 Grammar Database of Belarusian

The Grammar Database of Belarusian (Bułojčyk & Koščanka, 2021; Grammar Database) also provides a search interface that uses grammatical and lexical information of the lemmas and provides the inflection tables for each inflected lemma. Compared to the Polish resource, it uses a more visual and structured way of filtering by lemma features, which it presents as “search grammars” (see figure 3). Another advantage of this dictionary is the fact that the source code for the search engine is public (Korpus: Corpus Linguistics Software) and is declared to be adjustable for other languages.

Граматыка

Скасаваць Абраць

дзеяслоў

Пераходнасць

пераходны непераходны пераходны/непераходны

Трыванне

закончанае трыванне незакончанае трыванне

Зваротнасць

зваротны незваротны

Спражэнне

1-е спражэнне 2-е спражэнне рознаспрагалыны


Figure 3: Grammar Database of Belarusian.

3.3 Dictionaries of English

Some examples of well-known online English dictionaries are the Cambridge Dictionary (Cambridge University Press & Assessment, 2023) and the Macmillan Dictionary (Macmillan Education Limited, 2009–2023). Both provide definitions of the words, their pronunciation, basic grammatical information (for instance, part of speech), and usage examples. In addition, the Cambridge Dictionary supplies some articles with pictures, while the Macmillan Dictionary contains inflection tables and may list synonyms and other related words (see figure 4). Nonetheless, they neither provide the possibility to use any word characteristics in the search queries, nor support search by regular expressions.

The RegEx search for English words however is supported by some other web resources, for example, the Word Finder (Word finder 2023). However, this kind of resource typically doesn't contain any grammatical information as it is mostly oriented at crossword puzzle solving, rather than at providing linguistic information.

toe DEFINITIONS AND SYNONYMS ★★

NOUN COUNTABLE UK  /təʊ/

WORD FORMS

singular **toe**
 plural **toes**

DEFINITIONS 2

1 one of the five individual parts at the end of your foot. Your big toe is the largest, and your little toe is the smallest

Vera slipped off her shoes and wiggled her toes.
I stubbed my toe (=hurt it by hitting it) on the step.

on your toes (=with only your toes on the ground): *He stood on his toes to look out of the window.*

Synonyms and related words

General words for limbs and appendages

ankle **appendage**

Figure 4: An entry in the Macmillan Dictionary.

4. Implementation

Based on the overview of the online dictionaries presented in the previous section, it was reasoned that the Grammar Database of Belarusian has the most fitting search interface for online dictionaries of Slavic languages. One of the most important differences between the VESUM and the Belarusian Grammar Database is the format in which the data is stored. In the VESUM the data is stored in two ways: internal (see section 2.1) with several kinds of tags and marks for each lemma and visual representation (see section 2.2) with a narrower set of tags. The Belarusian Grammar Database utilizes a group of XML files describing paradigms, lemmas, and inflected forms (see table 3). It also contains groups of tags describing each of the items. However, the Grammar Database strongly relies on the tag groups having a certain rigid format and order, so the usage of the suggested format required a transformation of the VESUM tag system. The scripts that achieve that are available on GitHub (Dictionary Format Translator, 2023). The base version of the updated VESUM search uses XML files as its internal data representation. At the time of writing efforts are underway to rebuild the search engine so that it uses the SQL-based data source, in order to achieve a more efficient and seamless operation.

<pre> <Paradigm pdgId="1211000" lemma="па-пя+тае" tag="Z"> <Variant id="a" lemma="па-пя+тае" pravapis="A1957,A2008"> <Form tag="" slouniki="krapivabr2012,sbm2012">па-пя+тае</Form> </Variant> </Paradigm> </pre>
<pre> <Paradigm pdgId="1127963" lemma="адзіна+ццацера" tag="MAKS"> <Variant id="a" lemma="адзіна+ццацера" slouniki="piskunou2012:7147" pravapis="A1957,A2008"> <Form tag="PNP" slouniki="прым2009">адзіна+ццацера</Form> <Form tag="PGP" slouniki="прым2009">адзінаццацяры+х</Form> <Form tag="PDP" slouniki="прым2009">адзінаццацяры+м</Form> <Form tag="PAP" slouniki="прым2009" options="inanim">адзіна+ццацера</Form> <Form tag="PAP" slouniki="прым2009" options="anim">адзінаццацяры+х</Form> <Form tag="PIP" type="nonstandard">адзінаццацяры+ма</Form> <Form tag="PIP" slouniki="прым2009">адзінаццацяры+мі</Form> <Form tag="PLP" slouniki="прым2009">адзінаццацяры+х</Form> </Variant> </Paradigm> </pre>

Table 3: Examples of entries in the Belarussian Grammar Database.

Of the two VESUM representation formats, the porting was done for the visual one since it provides the set of data that is closer to the search criteria that the users might be interested in when using the software. Nonetheless, the internal representation does provide some additional data that can be interesting for researchers, so the future development plans include the integration of the internal data into the final search form (see also section 6). The currently supported tags provide information on the part of speech, and POS-specific characteristics for lemmas and their inflected forms: 1) for nouns: animate vs inanimate, common vs proper, abbreviation vs non-abbreviation, gender, number, case, 2) for adjectives: degree of comparison, gender, number, case, usage with animate vs inanimate nouns for certain forms, 3) for verbs: reflexivity, aspect, tense, gender, number, person, 4) for adverbs: degree of comparison, and 5) for conjunctions: coordinating vs subordinating. An example of tag groups describing Ukrainian lemmas is shown in the table 4.

After completing this task, several other items had to be addressed in order to launch the updated VESUM search engine, namely: 1) porting of Korpus search functionality to a more compact framework (from pure Tomcat to Spring Boot), 2) adaptation of search parameters to correspond to the transformed VESUM tag system, 3) adaptation of the inflection tables to correspond to the sets of inflected forms generated by the

VESUM, 4) design adjustments to make the search page more in line with the existing ecosystem of computational linguistic tools for the Ukrainian language. All the listed tasks and future development steps can be followed on the VESUM search GitHub page (2023).

<pre> <Paradigm pdgId="87838" lemma="деренькотати" tag="VMN"> <Variant id="a" lemma="деренькотати"> <Form tag="0">деренькотати</Form> <Form tag="R1S">деренькоч</Form> <Form tag="R2S">деренькочеш</Form> [...] <Form tag="PXP">деренькотали</Form> </Variant> </Paradigm> </pre>
<pre> <Paradigm pdgId="87858" lemma="державдитор" tag="NCANM"> <Variant id="a" lemma="державдитор" orthography="ua_2019"> <Form tag="NS">державдитор</Form> <Form tag="GS">державдитора</Form> <Form tag="DS">державдиторові</Form> <Form tag="DS">державдитору</Form> <Form tag="AS">державдитора</Form> [...] <Form tag="VP">державдитори</Form> </Variant> </Paradigm> </pre>
<pre> <Paradigm pdgId="88489" lemma="дернуватий" tag="AP"> <Variant id="a" lemma="дернуватий"> <Form tag="MNS">дернуватий</Form> <Form tag="MGS">дернуватого</Form> [...] <Form tag="PAP" options="anim">дернуватих</Form> <Form tag="PAP" options="inanim">дернуваті</Form> <Form tag="PIP">дернуватими</Form> <Form tag="PLP">дернуватих</Form> <Form tag="PVP">дернуваті</Form> </Variant> </Paradigm> </pre>

Table 4: Examples of the VESUM entries after the translation of the internal tag system of the VESUM to the one more suitable for the search engine.

5. Results

The set-up search system for the VESUM provides users with functionality to perform a search across lemmas or across all the inflected forms in the dictionary, using both

exact queries and regular expressions (see figure 5). The results are displayed as lemmas with lists of their grammatical features. By clicking on a lemma, the user can view its inflection table (see figure 6).

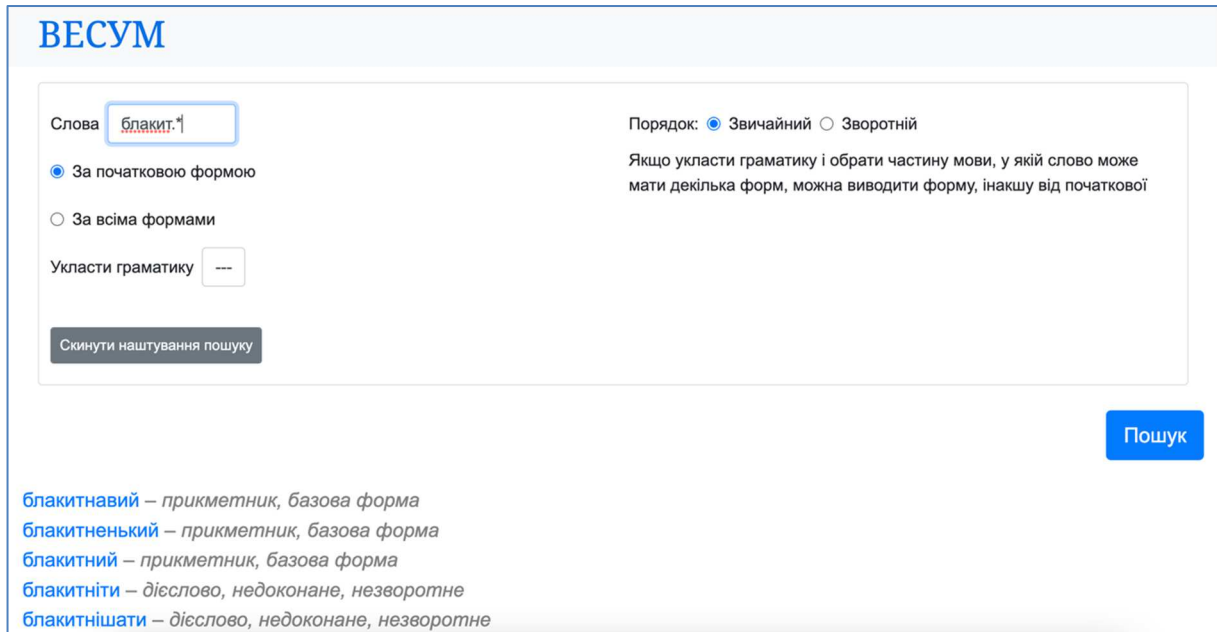


Figure 5: Main page of the developed search interface, search with a regular expression, and a list of displayed results.

одн.	с.	Д.	блакитному
		Зн.	блакитне
		Ор.	блакитним
		М.	блакитнім блакитному
		Кл.	блакитне
	ж.	Н.	блакитна
		Р.	блакитної
		Д.	блакитній
		Зн.	блакитну
		Ор.	блакитною
	М.	блакитній	
	Кл.	блакитна	

Figure 6: A section of an inflection table for the lemma “блакитний” (“light blue”).

The most powerful part of this tool is the search grammars that can be used to filter lemmas by their features. By clicking “Укласти граматику” (“Set up a grammar”) the

user can select a certain part of speech and the POS-specific features of interest. Figure 7 demonstrates a grammar that covers masculine inanimate common nouns.

Граматика Скасувати Обрати

іменник

Власна назва
 загальна назва власна назва

Істота
 істота неістота

Абревіатура
 абревіатура не абревіатура

Рід
 чоловічий рід жіночий рід середній рід

Множинні
 множина

Відмінок
 називний відмінок родовий відмінок давальний відмінок знахідний відмінок
 орудний відмінок місцевий відмінок кличний відмінок

Figure 7: A search grammar that would capture masculine inanimate common nouns.

6. Conclusions and Future Directions

The new search interface that was built for the VESUM is implementing a more user-friendly way of interacting with the database, and by providing advanced search options it allows one to make more use of the information available for each of its items. The use of search grammars (see figure 7) in combination with the regular expressions may be useful for researchers who need to compile lists of words sharing a certain grammatical feature (for example, listing all masculine nouns with typical feminine endings, or verbs that have given prefix, etc.). Apart from that, the neatly structured inflection tables make the dictionary convenient for non-linguists who might be looking for correct spellings of certain words or their forms.

Having said that, there is still a lot of work that can be done on the part of integrating the search form with the VESUM database. The current version implements a search that is based on the tags of the dictionary's visual representation, leaving the internal source files aside. One of the next steps therefore would be to integrate both dictionary representations (as each of them contains some unique information) into the set of search options. That would include grammatical information, such as inflection classes

or alternations, as well as usage mode, such as indications of colloquialisms, vulgarisms, etc.

Other possible directions for improvements include 1) addition of information that might not come directly from the VESUM, e.g., integrating with other online resources for the Ukrainian language, 2) addition of links to the related resources for Ukrainian and other languages, 3) implementing functionality for reporting mistakes and providing suggestions, 4) English localization, and 5) creating the structured public documentation describing all the data available in VESUM and accessible through the search interface.

7. References

- Bułowczyk, A. & Koščanka, U. (2021). Belarusian Language Grammar Database. Minsk: Тэхналогія.
- Busel V. (2005). Великий тлумачний словник сучасної української мови: 250000. Kyiv, Irpin: Perun.
- Cambridge University Press & Assessment. (2023). Cambridge Dictionary. Accessed at: <https://dictionary.cambridge.org/>. (26 March 2023).
- Dictionary Format Translator (2023). Accessed at: <https://github.com/tamila-krashtan/dictionary-format-translator>. (26 March 2023).
- Grammar Database. Accessed at: <https://bnkorporus.info/grammar.en.html>. (26 March 2023).
- Grammatical Dictionary of Polish. Accessed at: <http://sgjp.pl/>. (26 March 2023).
- Karpilovska, Y., Kysliuk, L., Klymenko, N. et al. (2013). Активні ресурси сучасної української номінації: Ідеографічний словник нової лексики. Kyiv: TOV “КММ”.
- Kieraś, W. & Woliński, M. (2017). “Grammatical Dictionary of Polish” – an online version. *Jezyk Polski*, 97(1), pp. 84–93.
- Korpus: Corpus Linguistics Software. Accessed at: <https://github.com/alex73/Software-Korpus>. (26 March 2023).
- Krytska, V., Nedozyrn, T., Orlova, L., Puzdyreva, T., Romaniuk, Y. (2011). Граматичний словник української літературної мови. Словозміна: Близько 140 000 слів. Kyiv: Dmytro Burago Publishing House.
- Macmillan Education Limited. (2009–2023). Macmillan Dictionary. Accessed at: <https://www.macmillandictionary.com/>. (26 March 2023).
- Project to generate POS tag dictionary for Ukrainian language. Accessed at: https://github.com/brown-uk/dict_uk. (26 March 2023).
- Rysin, A. & Starko, V. (2005–2023). Large Electronic Dictionary of Ukrainian (VESUM). Web version 6.0.1. Accessed at: <https://r2u.org.ua/vesum/>. (21 March 2023).
- Rysin, A. & Starko, V. (2020). Великий електронний словник української мови (VESUM) як засіб NLP для української мови. *Галактика Слова*. Галині Макарівні Гнатюк, pp. 135–141.
- Shvedova, M. & von Waldenfels, R., Yaryhin, S., Rysin, A., Starko, V., Nikolajenko, T.

et al. (2017-2023): GRAC: General Regionally Annotated Corpus of Ukrainian. Electronic resource: Kyiv, Lviv, Jena. Available at uacorporus.org.
Word finder 2023. Accessed at: <https://findwords.info/>. (26 March 2023).
VESUM search GitHub page. (2023). Accessed at: <https://github.com/tamilakrashtan/vesum-search>. (26 March 2023).

The Use of Lexicographic Resources in Croatian Primary and Secondary Education

Ana Ostroški Anić, Daria Lazić, Maja Matijević, Martina Pavić

Institute of Croatian Language and Linguistics, Republike Austrije 16, Zagreb

E-mail: aostrosk@ihjj.hr, dlazic@ihjj.hr, mmatijevic@ihjj.hr, mpavic@ihjj.hr

Abstract

This paper presents the findings of an online survey on the use of dictionaries and other lexicographic resources in primary and secondary education in Croatia. Apart from asking teachers to answer questions regarding the frequency of their use of lexicographic resources in classroom and while preparing for classes, the survey also elicited teachers' satisfaction with the dictionaries' content and structure. The survey was conducted at the national level among teachers of different educational backgrounds between 1 February to 15 February 2023. It was completed by 503 respondents. The results provide important statistical data on the usage of dictionaries in Croatian education and the usefulness of lexicographic resources in contemporary teaching practices. Respondents were generally satisfied with the available resources but often combined different resources to find the information they needed. Major shortcomings highlighted in the survey include the lack of resources suitable for students at lower levels of education, outdated or incomplete resources, and substantive and formal deficiencies in the content and form of resources.

Keywords: dictionary usage; education; user needs; survey.

1. Introduction

Dictionary usage has been a well-researched topic in lexicography over the past several decades, with empirical studies gradually shifting their focus towards the analysis of digital dictionary usage, specifically online dictionaries (Müller-Spitzer, 2014). While the general purpose of investigating dictionary usage could be described as determining “in which situations, how and, with which degree of success, etc., lexicographic tools are used” (Müller-Spitzer, Koplénig & Wolfer, 2017: 715), most such studies have primarily concentrated on students and language professionals (Lew, 2015: 234) as the most prominent users of lexicographic resources. Only recently have large-scale surveys on dictionary use among the general public started to emerge, such as the study by Kosem et al. (2019) that included the most participants in an overall study in different European countries.

Research on the use of dictionaries in the context of education has predominantly focused on their role as reference tools and teaching aids in foreign language learning (e.g. Boulton & De Cock, 2017; Tono, 2001), with some exceptions emphasizing their contribution to enhancing literacy and reading skills (e.g. Beech, 2010). Large-scale studies, like the one conducted by Kosem et al. on the use of monolingual dictionaries

(2019), have naturally included teachers as a subgroup of language professionals. However, there have been no specific studies exploring how teachers use dictionaries and other lexicographic resources as teaching aids and during the preparation of teaching materials. Therefore, the research presented in this paper was designed to investigate how dictionaries and other lexicographic and specialized resources, such as encyclopaedias, specialized dictionaries and databases, glossaries, etc., are used by Croatian primary and secondary school teachers.

An online survey was conducted to examine the extent to which teachers of various subjects use dictionaries and other resources when preparing teaching and learning materials for their classes, as well as their use in the classroom. In addition to assessing the frequency of lexicographic resource usage in the classroom and during lesson preparation, the survey also aimed to gauge teachers' satisfaction with the content and structure of dictionaries. The results of the survey offer valuable statistical data regarding dictionary usage in Croatian education and shed light on the usability of lexicographic resources in contemporary teaching practices. However, the qualitative analysis of several open-ended questions proved to be more significant in assessing the actual role of dictionaries in the educational process, and even more so for the discussion of the future development of dictionaries and other educational resources that better meet the needs of teachers and students in the classroom.

The remainder of the paper is structured as follows: previous studies on dictionary usage in education are briefly presented in the next section. Section 3 explains the design and implementation of the current survey in detail. Section 4 presents and analyses the results of the survey. Finally, in section 5, we reflect on the findings and propose potential future studies.¹

2. Background

Pedagogical lexicography has always been a highly active line of lexicography research and practice, which is rather expected given the application of dictionaries as reference tools in education systems, particularly in foreign language teaching. However, dictionaries have gradually lost their prominent position in foreign languages curricula in primary and secondary education, sliding into the background of teaching aids and materials due to a shift away from the translation-oriented methods towards communicative strategies that rely more on language acquisition in modelled surroundings and less on explicit vocabulary instruction. Nevertheless, dictionaries have somewhat regained their role in foreign language teaching with the introduction of the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001), which emphasizes the need for using different strategies to enhance language

¹ The English translation of the survey questions is given in Appendix 1. The entire survey dataset is available on Zenodo, <https://zenodo.org/record/7975264> (Ostroški et al., 2023).

awareness and develop language competence.

Foreign language teaching has contributed to many empirical studies of dictionary usage as well. Nied Curcio (2022: 73) reports on over 250 empirical studies focusing on dictionary usage in the field of foreign language teaching. However, far less studies have been designed to address the use of dictionaries in first language (L1) teaching, particularly in primary education (e.g. Zerdeli, 2021 among the more recent ones), despite the competence in dictionary skills remains to be recognized by many European curricula (Vicente, 2022), in L1 and L2 alike. Children's dictionaries not only contribute to vocabulary learning but also facilitate encyclopaedic and cultural learning, assisting school children in acquiring general knowledge as a foundation for vocabulary development (Tarp & Gows, 2012). Therefore, explicit instruction in dictionary skills proves to be an effective strategy in improving reference skills (Lew & Galas, 2008) needed not only for a variety of language learning and language acquisition tasks, but also for overall cognitive development.

The national curriculum for the Croatian language as L1 (MZO, 2018) includes the active use of a children's dictionary as a learning outcome for students as early as the first grade of primary education: "The student searches for unfamiliar words in a children's dictionary, using the alphabetical order of words, and reads the definition of each word to understand what it means" (MZO, 2018: 12, translation by authors). In the second grade, students should be able to search for explanations of unfamiliar words in a children's dictionary and use them as part of their active vocabulary. In the fourth grade, they need to be able to explain unfamiliar words by using children's dictionaries, as well as distinguish between children's dictionaries, encyclopaedias, and lexicons. In the seventh grade of primary school, students should explain the meaning of unfamiliar words after listening to a text and by using dictionaries. Dictionary skills are explicitly required in the first and final grades of secondary school. Apart from being able to explain unfamiliar words, 15-year-olds should be able to analyse the lexical-semantic relations between words using dictionaries. In addition to being able to explain unfamiliar words, 15-year-olds should be able to analyse the relations between words using dictionaries. They are also explicitly taught lexicography, and therefore, they need to be able to describe the structure of dictionary entries, as well as use dictionaries to develop their vocabulary.

3. The Structure and Implementation of the Survey

3.1 General aims and principles

The general aim of the survey was to investigate the usage of dictionaries by teachers and students in primary and secondary education in Croatia. The survey was designed based on three key research questions:

1. To what extent do teachers incorporate lexicographic resources in their teaching practices, both during class preparation and in the classroom?
2. How do teachers perceive the relevance and accuracy of information provided by lexicographic resources in relation to the curriculum they are teaching? In other words, how satisfied are they with the existing dictionaries and other reference resources they use?
3. How familiar are teachers with specialized dictionaries, databases, and other lexicographic resources?

To ensure maximum participation, various channels were utilized to disseminate the survey. These included mailing lists, social networks (such as teachers' Facebook groups, institutional and personal Facebook, Twitter, LinkedIn, and Instagram profiles), group and individual emails, and personal contacts. Additionally, in order to ensure that the data is representative in terms of participants' age, experience and place of employment, the survey was also distributed to a random selection of primary and secondary schools in Croatia by sending emails to the school principals, whose contact information is publicly available on the website of the Ministry of Science and Education (MZO).

3.2 Structure and implementation

The survey was conducted as an anonymous online questionnaire using Google Forms². It comprised of 24 questions divided into four sections: 1) Personal information, 2) Workplace information, 3) Use of dictionaries when preparing classes, and 4) Use of dictionaries in class. The section on personal information collected data on gender, age, place of birth, and place of employment, while the workplace information section requested participants to indicate their university degree, current occupation (e.g. teacher of Croatian, teacher of biology, librarian, etc.), current place of employment (primary or secondary education institution), years of employment, and overall job satisfaction. In sections 3) and 4), it was explicitly stated that *lexicographic resources* refer to “dictionaries, encyclopaedias, terminological databases, and other specialized resources (in Croatian and foreign languages).” The survey included various question formats, including multiple choice questions as well as short and long open-ended questions. All questions were mandatory.

Prior to the full survey, a pilot survey was conducted with a sample of 20 participants. Based on their feedback, certain questions were modified, and new questions were introduced. The survey was open from 1 February to 17 February 2023. While the time frame may appear short for a nationwide study, previous experiences with similar studies have shown that the majority of responses to online surveys distributed via

² www.google.com/forms/about/ (12 April 2023).

email are typically received in the first few days after the invitations are sent³. We therefore expected the same to hold for invitations posted in closed groups on social networks or sent directly through personal contact, in which cases responses are typically received promptly upon the survey distribution (Saleh & Bista, 2017).

4. Results

4.1 Respondents’ background

The survey was completed by 503 respondents. Among them, 448 respondents (89.1%) identified as female, while 53 (10.5%) identified as male. One participant chose not to answer, and another selected the ‘Other’ option, which was provided for participants not wishing to identify with a binary gender category. The majority of respondents (333 or 66.2%) fell into the age range of 35 to 54 years old. Figure 2 illustrates the distribution of respondents across six age ranges: 5 (1%) participants were under 25, 80 (15.9%) were in the 25–34 group, 183 (36.4%) in the 35–44 group, 150 (29.8%) in 45–54, 83 (16.5%) in 55–64, and 2 (0.4%) participants were over 65 years old.



Figure 1: Years of teaching experience.

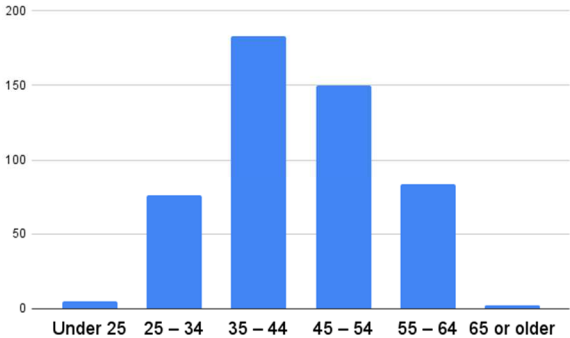


Figure 2: The age of respondents.

The survey successfully reached schools and teachers across all regions of Croatia, as indicated by the diverse range of workplaces reported by the participants. It is worth noting that two responses were received from teachers teaching the Croatian language outside of Croatia, specifically in Pécs, Hungary, and Stuttgart, Germany. Given the last population distribution of Croatia with the majority of its inhabitants living in major cities (Zagreb, Split, Rijeka and Osijek), one would expect that most participants are from Zagreb. However, the capital covers only 28.9% of all participants, while the

³ This is verified in the information available at www.surveymonkey.com/curiosity/time-to-respond (12 April 2023).

remaining participants are evenly distributed among smaller towns (such as Bjelovar, Beli Manastir, Dubrovnik, Đakovo, Imotski, Karlovac, Križevci, Makarska, Našice, Varaždin, Zadar, Županja) as well as villages (such as Biškupci, Ilača, Kraljevec na Sutli, Magadenovac, Retkovci, Velika, Vrginmost). This distribution indicates a good geographical coverage, as depicted in Figure 3.

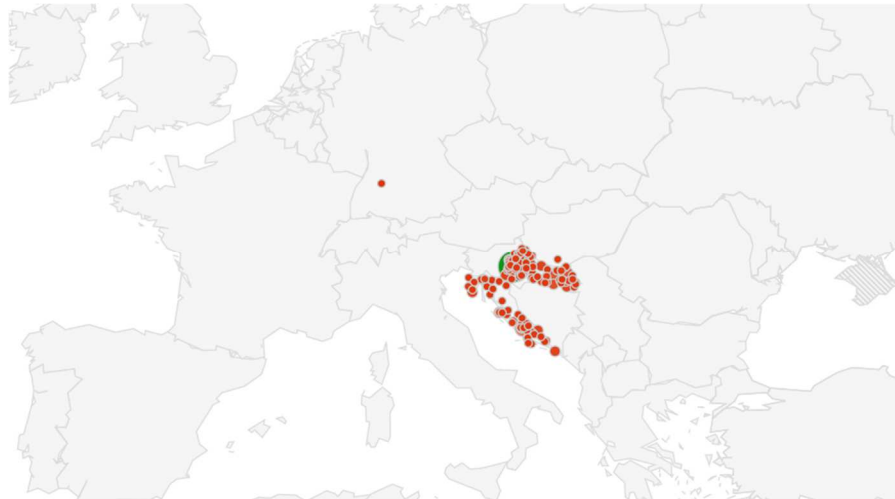


Figure 3: Regional distribution of respondents.

All respondents were equally distributed between primary and secondary schools as their places of occupation: 239 (47.5%) work in a primary school, 253 (50.3%) in a secondary school, while the remaining 11 respondents (2.2%) provided other responses, such as working in both primary and secondary schools, a music school, or a foreign language school. The question about years of experience in education shows a correlation with the respondents' expressed age. As depicted in Figure 1, the largest number of respondents, 92 (18.3%), have 20 to 25 years of experience working in schools. 80 (15.5%) respondents have worked for 10 to 15 years, 76 (15.1%) for 5 to 10 years, 73 (14.5%) for 15 to 20 years, 69 (13.7%) for less than 5 years, 61 (12.1%) for 25 to 30 years, and 54 respondents (10.7%) have worked for over 30 years in education. These figures indicate that the respondents are generally experienced teachers, whose responses should be considered as a result of relevant dictionary use.

The majority of all participants are mostly satisfied (299 or 59.4%) or fully satisfied with their workplace (118 or 23.5%). The remaining respondents indicated varying levels of satisfaction: 68 (13.5%) are satisfied to a certain extent, 13 (2.6%) mostly not satisfied, and only 5 (1%) not satisfied at all. This question was included to ensure that respondents' dissatisfaction with their workplace did not significantly influence their responses concerning the preparation for teaching. As is evident from the responses to this question, this was not the case.

The final information about the respondents' background concerns their teaching

position, and the results were positively surprising. Despite emphasizing in the invitations and survey information that the survey was intended for teachers of all school subjects, we had expected that language experts or teachers of L1 and L2 languages would constitute the predominant groups in the survey. We had also anticipated that primary school teachers, specifically those teaching grades 1 to 4 (usually ages 7 to 11), would form another significant group of respondents. This group is particularly interesting as they teach not only Croatian as L1 but also subjects such as Math, Science, Art, and Music. Figure 4 demonstrates that the survey successfully attracted teachers from all school subjects, including a substantial number of vocational subject teachers.

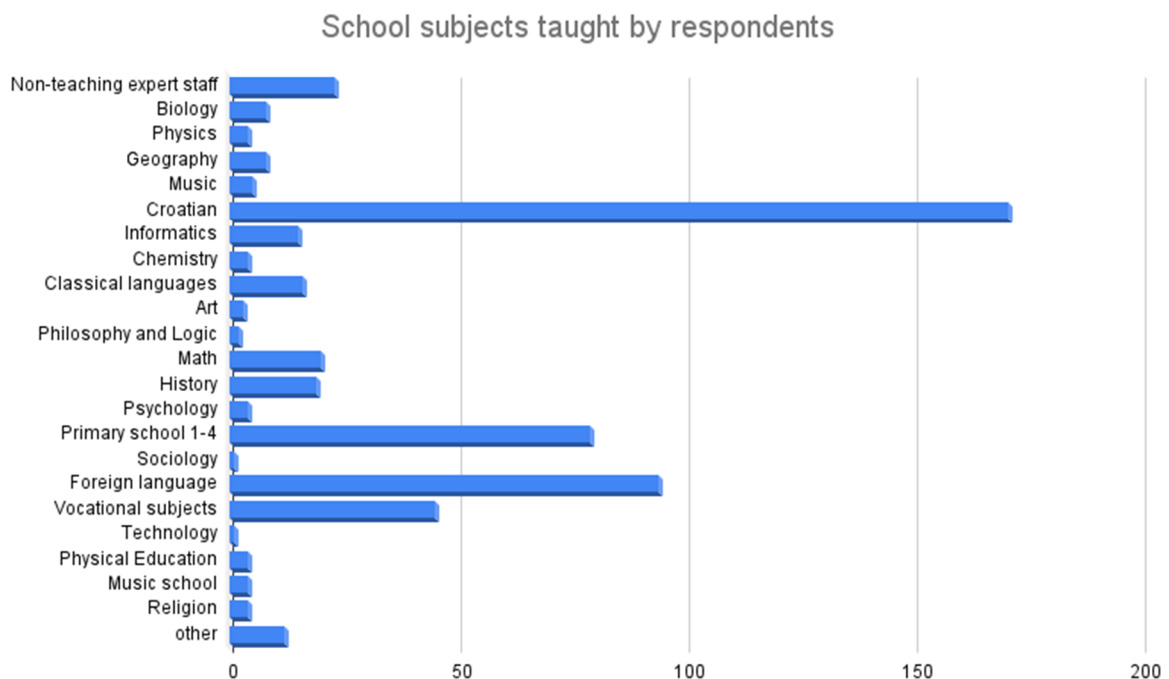


Figure 4: School subjects represented in the survey.

4.2 Using dictionaries when preparing teaching and learning material

Section 3 of the survey gathered questions related to the use of dictionaries when preparing classes. The first five questions focused on the use of general language dictionaries, both in printed and online formats, as well as language portals in Croatian and other languages. The subsequent questions addressed the use of specialized dictionaries, including those specific to particular subject fields, databases, glossaries, and other online resources in Croatian and other languages. Examples of such resources included the *Croatian Encyclopedia*, *Wikipedia*, *Britannica*, *IATE*, *Struna*, *Medical dictionary*, among others. Participants were given the option to provide their own responses if they were not satisfied with the provided options for frequency of usage.

Among the most common responses, 288 respondents (57.3%) indicated that they use

contemporary Croatian monolingual dictionaries once a month or more frequently. 110 respondents (21.9%) stated that they refer to these dictionaries once or twice over the course of several months. Additionally, 51 respondents (10.1%) reported using dictionaries about once a year. On the other hand, 37 participants (7.4%) stated that they never use monolingual Croatian dictionaries. Only 7 individuals reported using them on a daily basis (3), almost every day (2), or once a week (2).

More respondents seem to reach for dictionaries of English and other foreign languages. 126 (25%) respondents use them about once a month, 118 (23.5%) once to two times a month, 133 (26.4%) about once a year, and 115 (22.9%) never use them.

In the case of specialized resources in Croatian, a considerable number of respondents, 281 (55.9%), indicated that they regularly use them, i.e. once a month or more frequently. 136 individuals (27%) reported using specialized reference tools once or twice over the period of several months. Notably, English specialized resources are more frequently consulted than their Croatian counterparts. Specifically, 129 respondents (25.6%) referred to English specialized dictionaries or databases once a month or more often, while 130 (25.8%) did so once or twice over several months. On the other hand, 129 participants (25.6%) reported using English specialized resources about once a year, and 113 (22.5%) stated that they never consult specialized dictionaries or databases in English. The use of specialized resources generally does not come as a surprise if we consider they are more likely to be utilized as reference points for exam preparation, teaching, and learning materials, compared to monolingual Croatian dictionaries.

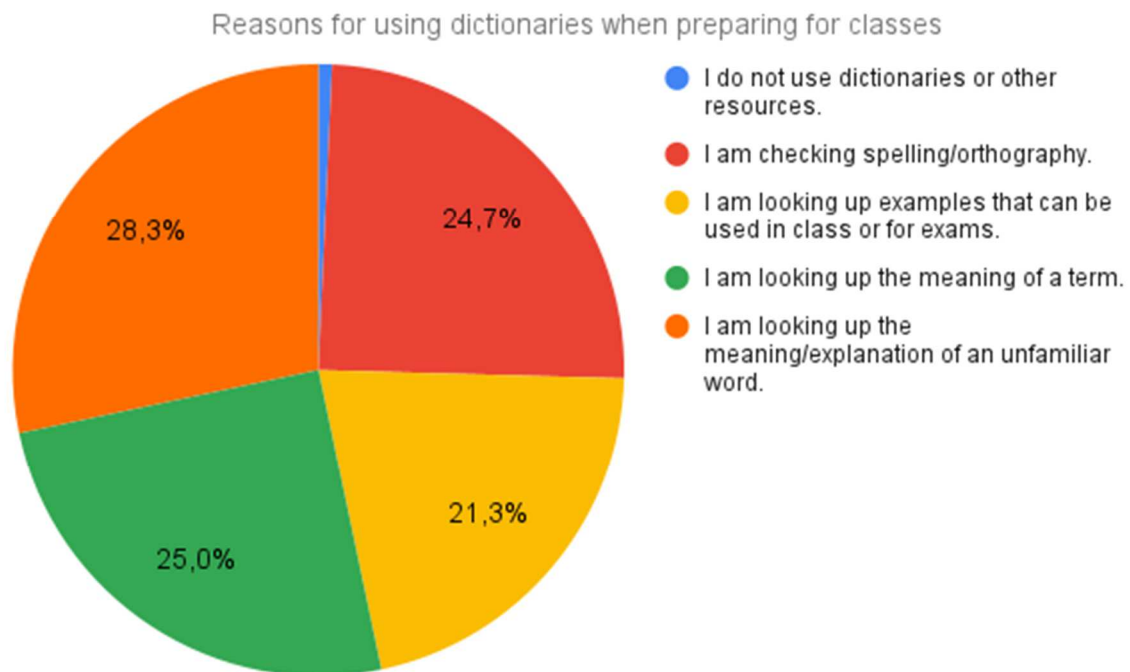


Figure 5: Most common reasons for using dictionaries when preparing for class.

The subsequent question in this section aimed to explore the reasons behind using lexicographic and other resources when preparing materials for class. Participants were provided with five pre-defined answer choices for the question *If you use dictionaries and other above-mentioned resources, what is the reason for this?* and they had the option to select multiple answers. The answer choices included: *I am looking up the meaning/explanation of an unfamiliar word, I am checking spelling/orthography, I am looking up the meaning of a term, I am looking up examples that can be used in class or for exams, and I do not use dictionaries or other resources.*

They could also add their own responses. Additionally, participants were given the opportunity to provide their own responses. Those who wrote their own reasons mentioned using lexicographic resources for checking accents, verifying grammatical features, finding lexemes related in a specific lexical-semantic relation (also using thesauri for this purpose), translation, learning word etymology, finding reliable definitions, examples of word usage in sentences, and preparing lessons where students would need to use dictionaries. Only a small number of respondents (0.7%) stated that they do not use dictionaries or other resources. The percentages of all responses are depicted in Figure 5.

Participants were asked about their opinions regarding the benefits of using dictionaries and similar resources when preparing for class, as well as their satisfaction with the structure and content of the resources they used. The responses received were predominantly positive and encouraging, e.g. *I use dictionaries to get a new idea or find a new example, or to clarify my dilemmas, to express myself more professionally in some situations, to quickly find examples and/or answers I need, to strengthen/expand my own vocabulary; It gives me confidence that I won't make a mistake; It is a reliable source of necessary information, etc.*

4.3 Using dictionaries in class

Section 4 of the survey included questions about the use of dictionaries and other lexicographic resources in class, whether as reference tools or as examples during explicit teaching of lexicography or methods of dictionary use. The first two questions asked respondents about the frequency of using printed and online dictionaries and other resources in class activities, regardless of the language. 68 teachers (13.5%) use printed dictionaries often in class, while 192 (38.2%) use them occasionally. 106 (21.1%) have rarely used them, only a couple of times at all, while 128 (25.4%) have never used printed dictionaries. Although the questions were of the multiple-choice type, participants could add their own responses. Therefore, several responses provided were more comments or descriptions of teachers' teaching habits than they were responses to the question. The numbers are much more in favour of online resources. 162 (32.2%) teachers often use online dictionaries, 213 (42.3%) use them occasionally, while the rest of the respondents are equally distributed into 'rarely' and 'never' options, 62 of them

(12.3%) in each.

The analysis of all responses regarding the grade levels in which dictionaries are used with students reveals an increasing trend in dictionary usage starting from the first grade of primary school. However, there is a slight decline in usage reported for the final grades, specifically the 8th grade of primary school and the 4th grade of secondary or high school, as depicted in Figure 6.

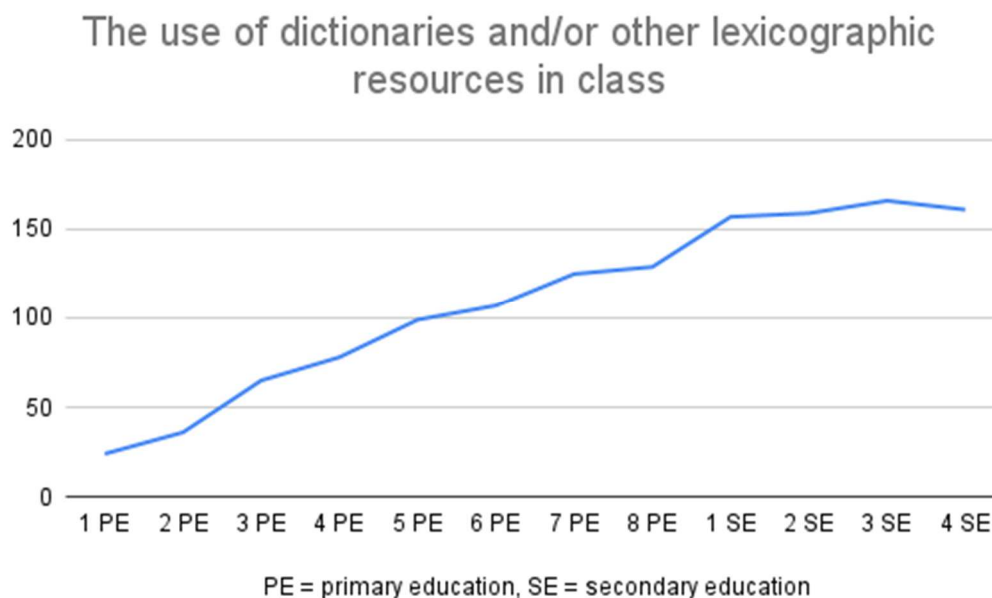


Figure 6: The use of dictionaries during class activities in primary and secondary education.

Since one of our research objectives was to assess teachers' familiarity with different types of lexicographic resources, we included an open-ended question asking them to list dictionaries or other lexicographic resources (in Croatian or other languages) they had used in class. Although this question required manual analysis, it provided valuable insights into the various reference tools employed in teaching activities. To ensure objectivity in the qualitative analysis, two of the authors initially categorized the responses based on their intuition. After thorough discussion and deliberation, they merged similar groups of responses and established a unified classification. Each response was then coded according to this agreed-upon classification. It is important to note that since a single response could encompass multiple activities, each mentioned activity within a response was treated as an individual response during coding.

The majority of respondents mentioned using at least one resource in their classes. Some responses mentioned the general type of resource (e.g. *Latin-Croatian dictionary*, various encyclopaedias, a dictionary of Croatian), while others specifically listed the resources they utilized. Both online and printed resources were well-represented among the mentioned resources. When it comes to Croatian monolingual dictionaries, there

was a preference for online resources. The *Croatian language portal* (*Hrvatski jezični portal*, hjp.znanje.hr) was the most frequently mentioned resource, cited by 160 teachers (31.8%). On the other hand, only three respondents mentioned *A Large Dictionary of the Croatian Standard Language* (*Veliki rječnik hrvatskoga standardnog jezika*), which is the most recent comprehensive general dictionary in Croatian. This dictionary is available in both printed form and electronically as a dictionary app that can be purchased. Respondents also reported using specialized Croatian dictionaries, both online and in printed form. These included dictionaries of foreign words, idioms, synonyms, jargon, personal names, eponyms, dialect dictionaries, and differential dictionaries of Croatian and Serbian.

The list of dictionaries for languages other than Croatian includes various printed and online dictionaries. Among online resources, the respondents most frequently mentioned English online dictionaries (*Oxford, Cambridge, Merriam-Webster, Collins, Macmillan*), as well as online resources, such as thesaurus.com, synonym.com, www.freecollocation.com, acronyms.thefreedictionary.com. Several participants also mentioned the German online dictionary *Duden*. Multilingual lexical databases such as *Wiktionary, Glosbe, the Free Dictionary, Larousse, EUdict, Crodict* (Croatian-German and Croatian-English dictionary), and *DrDicty* (English-Croatian, German-Croatian, and Spanish-Croatian dictionary) were also used. However, many respondents also mentioned using printed dictionaries. Some respondents indicated using dictionaries available in the classroom or resources found in the school library.

Among specialized resources, the *Croatian Encyclopedia* (hr. *Hrvatska enciklopedija*, enciklopedija.hr) was used most frequently, indicated by 65 (12.9%) respondents. It is available both online and in print, and some respondents referred to the online version, while a few mentioned using both versions. Other Croatian online resources were also mentioned, along with numerous resources (encyclopaedias, lexicons, and specialized dictionaries) that exist only in printed form. Resources in other languages included a few online encyclopaedias, primarily *Wikipedia* (including articles in Croatian), and *Encyclopaedia Britannica*.

In addition to dictionaries and encyclopaedic resources, the respondents have used other sources of information about language use and tools that facilitate language production. They primarily consulted online resources published by the Institute of Croatian Language and Linguistics, such as an orthography manual (pravopis.hr), a grammar overview (gramatika.hr), a database of Croatian equivalents to common English neologisms (bolje.hr), a collection of language-related advice (jezicni-savjetnik.hr), and a database of collocations (ihjj.hr/kolokacije). Furthermore, they also consulted various portals and blogs providing answers to language-related questions, and they used online tools for proofreading (Ispravi.me), translating (Google translate, Reverso), and generating lists of synonyms from corpora (Kontekst.io). Some responses indicated that teachers use any available resources and tools they can find on the internet.

Another open-ended question was closely related to the previous one and focused on the active participation of students in class. The responses to this question were also coded using the same procedure as the previous question. Table 1 presents the types of teaching activities used for practicing dictionary skills in the classroom.

Type of activity	Number of responses
looking up unknown words / definitions of words / comparing the meaning of words	158
looking up concepts / definitions of concepts	59
learning about the structure of lexicographical entry; forming a dictionary definition	44
searching for examples of usage	24
looking up grammatical categories, morphological features of words	24
checking spelling/orthography	24
searching for synonyms	17
assistance when writing essays and other assignments	15
searching for definition/explanation of terms	10
looking up foreign words; translation; Croatian word for a foreign equivalent	10
using new words in a text or in speech	9
looking up etymology of words	7
looking up meaning/explanations of phrasemes; collocations	7
searching for antonyms	6
reading for vocabulary enrichment	3
looking up accents and pronunciation	2
defining loanwords	2
looking up abbreviations	1
looking up archaic words	1
learning how to summarize and take notes	1
they did not participate	135
general positive reply	70
other (e.g. practicing language skills)	11
vague reply	11

Table 1: Teaching and learning activities for which students actively used dictionaries.

Responses coded as *general positive reply* (N = 70) refer to responses in which teachers didn't provide specific examples of active participation, e.g. *they participated by searching the dictionaries using their smartphones, we look up into dictionaries together, they look up information, they were active looking up words, I bring dictionaries in class*, etc. Responses coded as *vague reply* (N = 11) included statements such as *I encourage them to use them, they react in a positive manner, positive impressions, students are instructed to use them in homework*, etc., which were insufficient for analysis.

Although respondents had not been previously divided into subgroups according to the subjects they taught, activities mentioned in the responses confirmed our expectations that teachers of Croatian (as L1) and foreign language teachers, e.g. of English, German or Italian, would provide more responses to this question compared to other teachers. The most common activity for which dictionaries are used is looking up the meaning of words, followed by looking up concepts or their definitions. While it may seem reasonable to assume that most teachers do not differentiate between the meaning of words and concepts in the sense that a linguist would, a closer look at teachers who provided specific examples (e.g. *Yes, they used tablets to visit search certain links to search content with the help of given concepts.; They use encyclopaedias when they need the definitions of specific concepts., In groups, they were looking up explanations of certain concepts.; We look up concepts together, definitions they don't know when they come across in teaching material.*) reveals that in most cases, teachers of Croatian and primary school teachers (teaching grades 1 to 4) referred to subject-specific concepts that students need to acquire. Additional examples of subject teachers other than language teachers using dictionaries are provided in the Discussion section of the paper.

The question about students' reactions to using dictionaries in class predominantly yielded positive responses. However, some respondents were confused by the question, mistakenly thinking that they had already provided an answer in the previous question regarding the type of student activity involving dictionaries. As a result, we decided not to conduct a detailed analysis of these responses.

4.4 Self-reflection on lexicographic works meeting the user needs

To gather the participants' opinions on the quality of the content and structure of the dictionaries they use, we asked them whether the structure and content of the dictionaries and other resources they have used corresponded to their needs. We also invited them to share any additions or changes they would suggest. Most respondents indicated that the structure and content of resources corresponded or mostly corresponded to their needs. Some pointed out the lack of certain types of dictionaries for Croatian: etymological dictionary, dictionary of synonyms, dictionary of idioms, frequency dictionary, thesaurus, specialized dictionaries within certain fields, good-quality bilingual dictionaries, and even the lack of a single dictionary which would

comprise various sorts of linguistic information. Although many of these resources exist in printed form, their content is often incomplete or outdated, and they are not available online, which makes them more difficult to access. Participants also highlighted that the content of lexicographic resources is often too complex and extensive for school use (especially for primary school students), and possibly even for a wider circle of non-expert users. Furthermore, some comments were made regarding the microstructure of the articles, particularly in dictionaries (see Figure 7).

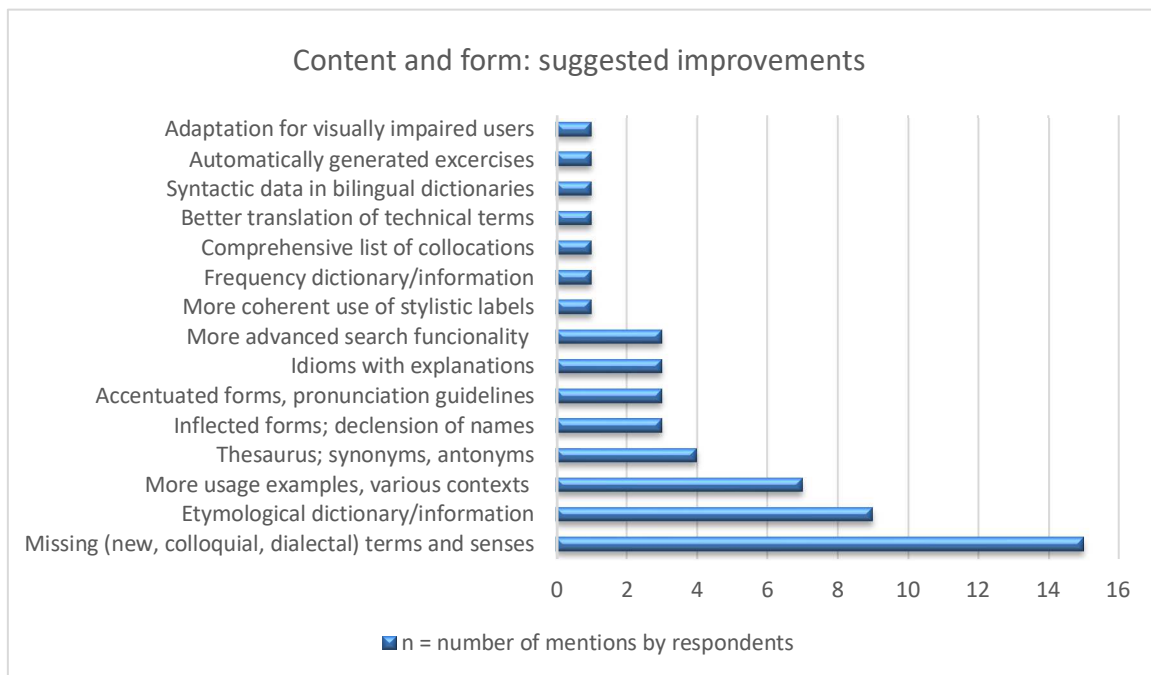


Figure 7: Suggested improvements and additions to the existing lexicographic resources, primarily dictionaries.

The most common concern raised by respondents was the lack of terms and senses in available resources, particularly in relation to newer concepts, technological innovations, and specific language variants (such as jargons, dialects, youth language, and colloquial language use). The need to update and expand existing resources was emphasized. Additionally, some participants expressed the need for more etymological information and a greater variety of usage examples that illustrate different contexts. The latter could be due to the fact that many resources for Croatian are only available in printed form (e.g. dictionaries of foreign words), and online dictionaries are mostly electronic versions of printed ones, where space limitations affect the content. Furthermore, respondents mentioned the absence of complete lists of inflected word forms, including accent marks⁴. The issue of name declension was also raised, as names are typically not

⁴ Croatian is an inflectional language with complex accentuation rules, and accentuation can vary within a single paradigm.

included in dictionaries, making it challenging to find information about their forms.

Other suggestions included improving the coverage of synonyms and antonyms, as well as idioms and their meanings; a need for a frequency dictionary and a further extension of the existing base of collocations was also mentioned. In the realm of bilingual lexicography, participants recommended providing more comprehensive information on syntactic differences between languages and ensuring accurate translations of technical terms in specific fields. Some comments focused on the interface of electronic resources, proposing the incorporation of advanced search options (such as suggested entries based on input, searching by inflected forms or metadata) and the integration of automatically generated exercises for students based on dictionary content. One respondent even suggested the implementation of an audio version of a dictionary to cater to visually impaired users. Finally, a noticeable number of respondents prefer electronic/online resources over printed ones, and many of them mentioned using a combination of different resources to obtain the information they need. However, a few responses expressed scepticism towards online resources and expressed a lack of trust in Croatian resources compared to foreign resources, particularly those in English.

5. Discussion and Conclusions

The survey was intended as a general overview of the use of dictionaries and related reference works at all levels of primary and secondary education. Therefore, no specific subgroups were analysed, but several conclusions can be reached based on the qualitative analysis of open-ended questions. First, teachers of Croatian as L1 make up the largest user group, and their teaching activities involving dictionaries are what one might expect: checking the meaning of unfamiliar words when reading; using dictionaries as reference tools for language production activities; studying morphological and etymological features of words, as well as explicitly teaching lexicology and lexicography in secondary school grades. Teachers of foreign languages are still the predominant users of specialized language reference works such as dictionaries of idioms, collocations and thesauri. However, a more detailed analysis of class activities per particular user subgroup is needed.

Respondents are generally satisfied with the available resources and often find the information they need by combining different resources. The lack of resources suitable for students at lower levels of education, the outdatedness or incompleteness of the existing resources, and some substantive and formal deficiencies in the content and form of the resources were highlighted as major shortcomings. Respondents often mentioned that they lacked certain types of resources (e.g. an etymological dictionary, a dictionary of synonyms), but it can be argued whether they really need a specialized resource or just access such information, e.g. in the form of a comprehensive online dictionary that would contain such information among other things. Namely, some of these resources exist in printed form, but as such are more difficult to access than online resources, and many of them are outdated. Based on the responses, it can be

concluded that users are generally inclined to online resources; some of them stated that they simply Googled things or used whatever resources they could find on the Internet. However, some users expressed doubt about the quality of the data in some online resources. Finally, especially when it comes to L1 Croatian language teaching, teachers need resources that comply with current language norm, e.g. spelling, and resources that contain normative information, normative recommendations, and the like. Such online dictionaries are currently rare.

Among the resources that teachers are using in class, the most frequently mentioned were general dictionaries of Croatian and other languages, as well as dictionaries focusing on a certain aspect of vocabulary, such as dictionaries of idioms, foreign words, synonyms, etc. The variation was also greatest in that group of resources. This is understandable given that the majority of the respondents were teachers of Croatian as L1 and teachers of other L2 languages.

Among the specialized resources, the *Croatian Encyclopaedia*, available both online and in printed form, and *Wikipedia* were more popular. Less frequently, the respondents mentioned other Croatian specialized online and printed resources, as well as foreign (mainly English) online resources such as *Britannica*. Although terminological resources were very rarely mentioned in the responses, several math teachers reported using lexicographic resources as reference tools for definitions of specialized mathematical concepts. This unexpected high awareness of the possibility of using reference works for teaching non-language subjects among this group of teachers should be fostered in future teacher-oriented activities. Surveys like this one are an excellent opportunity to raise awareness about using terminological resources in teaching, particularly for preparing materials for primary education students, who often need help with understanding complex concepts or formulating their own definitions. Understanding the lexicographic needs of young users – though they may not always be recognized as lexicographic – and meeting them in the form of well-developed teaching materials paves the way for a better development of children's categorization skills and their overall cognitive development.

6. Acknowledgements

We thank the anonymous reviewers for their valuable feedback and suggestions that improved the paper. This work has been supported in part by the Croatian Science Foundation under the project UIP-2017-05-7169, and by the Institute of Croatian Language and Linguistics within the research projects *Croatian Web Dictionary – Mrežnik* and *Semantic Frames in Croatian*.

7. References

Beech, J. R. (2010). Using a dictionary: Its influence on a children's reading, spelling, and phonology. *Reading Psychology*, 25(1), pp. 19–36.

- Boulton, A. & De Cock, S. (2017). Dictionaries as aids for language learning. In P. Hanks & G.-M. Schryver (eds.) *International Handbook of Modern Lexis and Lexicography*. Berlin/Heidelberg: Springer.
- CEFR = Council of Europe (2018). *Common Europe Framework of Reference for Languages: learning, teaching, assessment. Companion volume with new descriptors*. Strasbourg. <https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989> (20 April 2023).
- Egido Vicente, M. (2022). Dictionaries in German and Spanish Primary Education Curricula: A Comparative Study. *International Journal of Lexicography*, 35(2), pp. 176–203.
- Kosem, I. et al. (2019). The Image of the Monolingual Dictionary Across Europe. Results of the European Survey of Dictionary use and Culture. *International Journal of Lexicography*, 32(1), pp. 92–114.
- Lew, R. (2015). Research into the Use of Online Dictionaries. *International Journal of Lexicography*, 28.2, pp. 232–253.
- Lew, R. & Galas, K. Can Dictionary Skills Be Taught? The Effectiveness of Lexicographic Training for Primary-School-Level Polish Learners of English. In E. Bernal & J. DeCesaris (eds.) *Proceedings of the XIII EURALEX International Congress*. Barcelona: Universitat Pompeu Fabra, pp. 1273–1285.
- MZO = Ministarstvo znanosti i obrazovanja (2018). *Odluka o donošenju kurikuluma za nastavni predmet Hrvatski jezik za osnovne škole i gimnazije u Republici Hrvatskoj*. narodne-novine.nn.hr/clanci/sluzbeni/2019_01_10_215.html (20 April 2023).
- Müller-Spitzer, C. (ed.). (2014). *Using Online Dictionaries. Lexicographica Series Maior* 145. Berlin: Walter de Gruyter.
- Müller-Spitzer, C., Koplenig, A. & Wolfer, S. (2017). Dictionary Usage Research in the Internet Era. In P. A. Fuertes-Olivera (ed.) *The Routledge Handbook of Lexicography*. London: Routledge, pp. 715–734.
- Nied Curcio, M. (2022). Dictionaries, foreign language learners and teachers. New challenges in the digital era. In A. Klosa-Kückelhaus, S. Engelberg, C. Möhrs & P. Storjohann (eds.) *Dictionaries and Society. Proceedings of the XX EURALEX International Congress*. Mannheim: IDS-Verlag, pp. 71–84.
- Ostroški Anić, A. et al. (2023). The use of lexicographic resources in Croatian primary and secondary education – Survey Data [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.7975264> (26 May 2023).
- Saleh, A. & Bista, K. (2017). Examining Factors Impacting Online Survey Response Rates in Educational Research: Perceptions of Graduate Students. *Journal of MultiDisciplinary Evaluation*, 13(29), pp. 63–74.
- Tarp, S. & Gouws, R. H. (2012). School Dictionaries for First-Language Learners. *Lexikos*, 22, pp. 333–351.
- Tono, Y. (2001). *Research on Dictionary Use in the Context of Foreign Language Learning: Focus on Reading Comprehension*. Berlin/Boston: Max Niemeyer Verlag.

Zerdeli, S. (2021). Teaching Dictionary-Using Strategies in Primary Educational System: An Empirical Research. *Journal of Modern Education Review*, 11(1), pp. 1–8.

Appendix 1.

Survey questions

The use of dictionaries and other lexicographic resources in teaching

Section 1/5

The goal of this research is to examine the extent to which teachers use dictionaries and other lexicographic sources when preparing and carrying out their lessons. The research is completely anonymous, and all collected data will be analyzed on a group level rather than an individual one.

Participation is voluntary, and you can withdraw from answering at any time without any consequences.

Please respond to the questions spontaneously and as honestly as possible. Detailed instructions and the method of responding are provided in the questionnaire, so please read the instructions carefully before providing your answers. As the participants' email addresses are not collected, we are unable to send you a copy of your responses.

By completing the questionnaire, you agree to participate in the research. The questionnaire takes approximately 10 minutes to complete.

We sincerely thank you for your effort and time invested!

Ana Ostroški Anić, Daria Lazić, Maja Matijević and Martina Pavić

Institute of Croatian Language and Linguistics

You will be able to read the research results on the website of the Institute of Croatian Language and Linguistics.

Section 2/5

Personal information

Please provide the following basic demographic information about yourself in this section: gender, age, place of birth, and place of residence.

2.1 What is your gender?

- Male
- Female
- Other
- Prefer not to say.

2.2 How old are you?

- Under 25
- 25–34
- 35–44
- 45–54
- 55–64

- 65 or older

2.3 Where were you born?

Short answer text.

2.4 Where do you work?

Short answer text.

Section 3/5

Workplace related information

Please answer the following questions about your job title and workplace.

3.1 Where do you work?

- In primary school
- In secondary school
- Other

3.2 What is your educational degree?

Short answer text.

3.3 What is your current job position (e.g. primary school teacher, Croatian language teacher, biology teacher, etc.)?

Short answer text.

3.4 How many years of teaching experience do you have (not just in your current position)?

- Less than 5
- 5–10
- 10–15
- 15–20
- 20–25
- 25–30
- More than 30

3.5 Are you satisfied with your job position?

- Completely
- Mostly
- Somewhat
- Mostly not
- Not at all

Section 4/5

Dictionaries and other resources

The following questions pertain to the extent of your use of dictionaries, encyclopedias, terminological databases, and other specialized resources (both in Croatian and foreign languages) in lesson preparation. If none of the provided answers are acceptable, you can enter your response under "Other."

4.1 Do you use contemporary general dictionaries of the **Croatian language** in lesson preparation and

teaching materials (printed Croatian language dictionaries, dictionary portals like the *Croatian Language Portal*, online dictionaries like the *School Dictionary of the Croatian Language*, etc.)?

- Yes, once a month or more
- Occasionally, once or twice every few months
- Rarely, approximately once a year
- No, never
- Other

4.2 Do you use contemporary dictionaries of the **English language** or other **foreign languages** in lesson preparation and teaching materials (printed dictionaries or online dictionaries such as *Merriam-Webster*, *Wiktionary*, *Oxford Learner's Dictionaries*, etc.)?

- Yes, once a month or more
- Occasionally, once or twice every few months
- Rarely, approximately once a year
- No, never
- Other

4.3 Do you use dictionaries of foreign words?

- Yes, once a month or more
- Occasionally, once or twice every few months
- Rarely, approximately once a year
- No, never
- Other

4.4 Do you use specialized and/or technical dictionaries, databases, glossaries, and online resources in the Croatian language (e.g. *Croatian Encyclopedia*, *Wikipedia*, *Struna* database, *Medical Dictionary*, *Chemical Dictionary*)?

- Yes, once a month or more
- Occasionally, once or twice every few months
- Rarely, approximately once a year
- No, never
- Other

4.5 Do you use specialized and/or technical dictionaries, databases, glossaries, and online resources in the English language (e.g. *Wikipedia*, *Britannica*, *IATE* database, *BabelNet*, *thesaurus.com*, etc.)?

- Yes, once a month or more
- Occasionally, once or twice every few months
- Rarely, approximately once a year
- No, never
- Other

4.6 If you use dictionaries and other mentioned resources, what is the reason for doing so? You can select multiple answers or select the provided answers and provide an additional response.

- I search for the meaning/interpretation of unfamiliar words.
- I search for correct spelling of words.
- I search for the meaning of technical terms.

- I search for examples that I can use in teaching or exams.
- I do not use dictionaries or other sources.
- Other

4.7 In your opinion, what are the benefits of using dictionaries and other resources in any form of lesson preparation? Please briefly explain your answer.

Short answer text.

4.8 Do the structure and content of the dictionaries and resources you have used meet your needs? Is there anything you would add or change? Please explain.

Short answer text.

Section 5/5

The use of dictionaries and other lexicographic resources in teaching

The following questions relate to whether you use dictionaries and/or other sources during the actual class. If none of the provided answers are acceptable, you can enter your response under "Other."

5.1 Do you use **printed** Croatian or foreign dictionaries or other lexicographic resources during class?

- Frequently
- Occasionally
- Rarely, a few times throughout your teaching career
- No, never
- Other

5.2 Do you use Croatian or foreign **online** (internet) dictionaries or other lexicographic sources during class?

- Frequently
- Occasionally
- Rarely, a few times throughout your teaching career
- No, never
- Other

5.3 Please list the dictionaries or other lexicographic resources (in Croatian or foreign languages) that you have used during class.

Long answer text.

5.4 If you have used dictionaries or other lexicographic resources, did the students actively participate in that activity? If they did, please briefly describe how they participated.

Long answer text.

5.5 If the students actively participated in the instructional activity that involved using dictionaries or other resources, how did they react to that activity?

Long answer text.

5.6 Do you use dictionaries and/or other lexicographic and specialized resources during class, and in which grades? You can select multiple answers or select the provided answers and provide an additional response.

- 1st grade primary school

- 2nd grade primary school
- 3rd grade primary school
- 4th grade primary school
- 5th grade primary school
- 6th grade primary school
- 7th grade primary school
- 8th grade primary school
- 1st grade secondary school
- 2nd grade secondary school
- 3rd grade secondary school
- 4th grade secondary school
- I haven't used dictionaries and other lexicographic resources in class.
- Other

5.7 What benefits, in your opinion, can students have if they use lexicographic manuals or other sources? Please briefly explain your answer.

Long answer text.

Thank you for taking the time to participate in the survey. If you have any comments or feedback, please feel free to write them.

Long answer text.

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

Chayanon Phoodai¹, Richárd Rikk

¹ European Master in Lexicography (EMLex),
Károli Gáspár University of the Reformed Church in Hungary,
H-1088 Budapest, Reviczky str. 4.

E-mail: chayanon507@gmail.com, rikk.richard@gmail.com

Abstract

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

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

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

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

1. Introduction

Artificial Intelligence (AI) plays a significant role in natural language processing (NLP). Large language models (LLMs) Bahdanau et al. (2014) can provide better solutions than the previous state-of-the-art in areas such as machine translation Brants et al. (2012),

code synthesis Poesia et al. (2022), text summarization Pilault et al. (2020), and more Araci (2019); Dathathri et al. (2019); Kant et al. (2018); Yasunaga et al. (2021).

Despite the success of LLMs, their applicability in lexicography remains mostly understudied. In this paper, we evaluate ChatGPT Brown et al. (2020); OpenAI (2023) in a lexicographical context by comparing it to the 10th Edition of Oxford Advanced Learner’s Dictionary (OALD) Hornby (2019). We do this by using the Wiegand (1989) item classes and similarity scores.

We use Wiegand (1989)’s item classes to determine how answers of ChatGPT to lexicographical questions and the information provided by OALD align with the structural requirements of a dictionary. Item classes provide a comprehensive method to determine which and how lexicographical items should be presented. This allows us to compare OALD and answers of ChatGPT in an objective manner, and gain useful insights of these tools including what information they do and do not provide. In order to use this method, we compile two main datasets containing information from ChatGPT and OALD regarding the lexicographical items of the most frequently used English words according to the British National Corpus (BNC) from Oxford Text Archive (2007). Iterating over our first dataset, we manually check if the given lexicographical item satisfies the criteria given by the item classes. Our findings then collected into a comparative table. We aim to show how effective ChatGPT is in providing lexicographical data for English language learners compared to a conventionally assembled dictionary.

Also, we calculate similarity scores using Bilingual Evaluation Understudy (BLEU) Papineni et al. (2002) and Recall Oriented Understudy for Gisting Evaluation (ROUGE) Lin (2004) methods which are widely used in NLP for determining the syntactic similarities of texts. The scores are calculated programmatically on one of our datasets. This dataset includes only those lexicographical items that can be compared using BLEU and ROUGE, like item giving pronunciation, spelling, part of speech, definition, and etymology. These items have been chosen for their unambiguity, which not only makes it possible to calculate similarity scores on them, but also allows us to consider the information provided by OALD as ground truth. We apply BLEU and ROUGE method to every lemma for every lexicographical item in the dataset, and visualize our results on separate figures. Last, we calculate the average of BLEU and ROUGE scores by lexicographical items only. Since ChatGPT has been trained on large amounts of human generated data gathered from the internet, we aim to show how much ChatGPT deviates from the ground truth. Large deviations have to be examined further as ChatGPT has some tendency to state incorrect information. In these cases, the expected score should be close to zero. Therefore, manual analysis of BLEU and ROUGE scores can allow us to investigate the reliability of ChatGPT as well.

Using Wiegand’s item classes and similarity scores, we provide comparative analyses in a lexicographical context between ChatGPT and OALD. Our research gives insights into the viability of AI-generated dictionary content, and aims to help the adoption of such technologies in language learning and education. Also, it tries to identify some of the limitations and challenges of AI in lexicography to inform the development of models in the field.

In the next sections, we go over our method in detail. First, an overview is provided highlighting all the main parts of our method. Then, the item classes and the comparative table provided by them are discussed in detail. Next, we describe the similarity scores,

results of the calculation, and their meaning. After that, we summarize our results from the two different methods and finish with our conclusions.

2. Related work

Previous studies have explored different aspects of monolingual learner's dictionaries (MLDs), such as their interface, software, structure, and user experience. In this section, we review related work that provides valuable insights and guidelines for conducting comparative studies on MLDs.

Herbst (1996) examines the features of four popular English learners' dictionaries: Oxford Advanced Learner's Dictionary (OALD5), Longman Dictionary of Contemporary English (LDOCE3), Collins COBUILD English Dictionary (COBUILD2), and Cambridge International Dictionary of English (CIDE). The study's methodology involves a detailed analysis and comparison of the dictionaries' features, including their target users, corpus basis, definitions, pronunciation, example policies, valency information, collocations and phrases, labelling system, illustrations, access structure etc. The paper employs a qualitative research approach, relying on the author's expert judgement and critical evaluation of the dictionaries' strengths and weaknesses. The study's findings are based on a thorough and systematic comparison of the dictionaries' features. The author provides clear and detailed explanations of the criteria used for evaluation. Overall, the study's methodology is rigorous and comprehensive, and the findings are based on a thorough analysis of the dictionaries' features and feedback from language experts and users. However, the study does not provide statistical analysis or quantitative data, and the evaluation criteria used by the author are subjective to some extent.

Ivančič & Fabijanić (2017) present an approach for analysing the chronological development of the macro- and microstructure of the OALD. Ten editions were investigated to find out the similarities and differences. This study involves methodology of the analytical standpoint of the authors, because it takes us thoroughly through different lexicographical item within the macro- and the microstructure. The findings are shown comparatively between the ten editions in tabular form. The study shows that the both macro- and microstructure have been expanding increasingly over each edition. Variety of new sections in MLDs has been introduced. This is to encourage the EFL learner's language skills. This study is highly relevant to our research as it focuses on the development of OALD specifically and its treatment of lemmas within the dictionary.

While these two studies provide us with comprehensive framework for conducting detailed manual analysis within the microstructure and offer guidelines for comparing different dictionaries, they lack objective criteria as both studies rely solely on the author's opinion. To address this limitation, we propose the use of reliable criteria for analysis, specifically *Wiegand's item classes* described in the methodology section. By adopting these established criteria, we can ensure a more reliable and unbiased approach to our analysis, moving beyond the subjective viewpoint of the authors alone.

3. Methodology

This research paper is a comparative study that aims to show the capabilities of ChatGPT for lexicographical purposes and compare it with the OALD focusing on the microstructural

elements. To accomplish this, we provide a detailed explanation of the methods employed for this study in the following section. In addition, this section will provide a comprehensive overview of the entire study process (see Figure 1).

3.1 Corpus and lemma selection

In order to notice the differences of microstructural elements we selected the ten most frequently used words from five different parts of speech (POS) including noun, verb, adjective, adverb, and preposition. According to the frequency counts in Davies & Gardner (2013), our chosen five POS belong to the most commonly used functional word classes in English. We choose lemmas from different POS because the lexicographical items in dictionary entries can vary even within the same category. We selected 50 lemmas from the British National Corpus (BNC) Oxford Text Archive (2007). While various corpora may produce slightly different outcomes, our choice of corpus does not significantly affect our study's purpose of showcasing the likeness of the most frequently utilized English words.

3.2 Wiegand's item classes

According to Wiegand (1989), dictionaries have more than 200 classes of functional text segments that serve as structural indicators within the dictionary microstructure. However, for the purpose of our study, we focused only on the lexicographical items suggested by Wiegand for general and learner's dictionaries. Since the OALD falls into this category and our objective is to assess the capabilities of ChatGPT as a learner's dictionary, we have chosen the suggested item classes and lexicographical items as our criteria for evaluation. Engelberg & Lemnitzer (2009) adapted these lexicographical items and simplified their names. To ensure coherence and ease of understanding, we adapted these names of the item classes for our English language research and assigned them acronyms for easy reference in subsequent sections, namely Lemma Sign (LES), Phonetic-Phonological and Orthographic Information (PPOI), Morphological Information (MOI), Syntactic Information (SYI), Syntactic-Semantic Information (SYSI), Semantic Information (SEMI), Pragmatic Information (PRAI), and Other Items (OTI), that contains lexicographical items which do not belong to the other item classes. We use the lexicographical items belonging to the above mentioned classes to test ChatGPT's competence as a learner's dictionary. ChatGPT is assigned to provide information for each item to complete the task.

3.3 Lexicographical prompts

Prompt engineering is becoming an essential skill for effective communication with large language models (LLMs) like ChatGPT. These prompts serve as instructions that enforce rules, automate processes, and ensure specific qualities and quantities of output. They are also a programming tool that can customize the outputs and interactions with the language model White et al. (2023).

In order to make ChatGPT to complete a task effectively, an appropriate prompt needs to be formulated. Different style of prompt may affect the quality of the outputs. Initially, we begin formulating our own prompt giving ChatGPT to *provide dictionary information*

for a given "lemma". With this prompt, the responses generated by ChatGPT can be inconsistent, even with the same prompt used. Each time, ChatGPT produces a different structure. Prompt 1 shows how ChatGPT typically generates the structure. The overall result does not meet our satisfaction due to the lack of data within one dictionary entry. Prompt 1 demonstrates that ChatGPT only provides us data related to five lexicographical items which are the form of the lemma sign, the spelling, part of speech, and polysemy. Only some of the polysemous senses are displayed in Prompt 1, despite the fact that the original response contains approximately 20 senses. Due to the lack of data in an entry we then ask ChatGPT itself for the advice of the possible prompts for our lexicographical tasks. Prompt 2 shows the response of possible prompts for different lexicographical tasks. While the prompts provided by ChatGPT show potential for lexicographical tasks, their effectiveness has not been established in the field of lexicography. As such, we consider them as an initial attempt to assess the response of our lexicographical items within Wiegand's item classes.

PROMPT 1: PROVIDE DICTIONARY INFORMATION FOR "TIME"

Sure, here is the dictionary information for the word "time":

noun

1. as a measurement, "time" refers to a duration or a quantity that is measured in units of seconds, minutes, hours, days, weeks, months, years, or other units depending on the context.
2. the duration of a particular event or situation.

verb

1. to measure or record the speed, duration, or rate of.
2. to adjust (a mechanism) so as to synchronize its operation with a timing device or with a sequence of operations.

PROMPT 2: PROVIDE TEN CONCISE PROMPTS OR TEMPLATES THAT CAN MAKE YOU DO LEXICOGRAPHICAL TASKS

Sure, here are ten concise prompts or templates for lexicographical tasks:

1. Define the word [insert word here].
2. Create a list of synonyms for [insert word here].
3. Provide an antonym for [insert word here].
4. Explain the origin and etymology of [insert word here].
5. Write a sentence using [insert word here] in context.
6. Give an example of a word that is commonly confused with [insert word here].
7. What part of speech does [insert word here] belong to?
8. Identify the syllable stress in [insert word here].
9. Provide a definition for [insert phrase or idiom here].
10. List three words that are derived from [insert word root here].

3.4 Datasets

All the information we gather from the BNC, lexicographical items from Wiegand’s item classes, responses from ChatGPT and OALD are stored in two primary tables. The first table lists 50 selected lemmas from the most frequently used English words across five different parts of speech from BNC. Additionally, the table presents 41 lexicographical items belonging to eight distinct item classes. Each column of the lexicographical items is marked with *present* or *absent* indicating whether ChatGPT and OALD can provide information related to the corresponding lexicographical item for each lemma. The second table lists 50 selected lemmas as like in the first table and five chosen lexicographical items LES, AUSA, RA, WAA, ABED¹, and ETYA. The table also contains the actual answers in textual form that we gather from both sources to be calculated for their similarities with BLEU and ROUGE.

3.5 Manual Analysis

We use the collected dataset to analyze the results for eight item classes: LES, PPOI, MOI, SYI, SYSI, SEMI, PRAI, and OTI. This evaluation helps us assess the capabilities of ChatGPT. For each item class, we examine whether ChatGPT and OALD are capable of providing the corresponding lexicographical items within the microstructure. Additionally, we analyze how they present the corresponding data, if available. The tables display lexicographical items in each class, lemma count², and three different types of symbols: percentages (%), plus signs (+), and minus signs (-). Percentages represent the availability of data provided by both tools for related lexicographical items, while a minus sign indicates unavailability of the data. A plus sign indicates that the related data is available but beyond the scope of our selected 50 lemmas.

3.6 Similarity Scores

In addition we calculate how similar the provided answers from ChatGPT and OALD are by using BLEU Papineni et al. (2002) and ROUGE Lin (2004). It is important to note that these scores do not indicate the quality of the answers, but rather measure the extent to which they align with the human-edited dictionary entries in a learner’s dictionary. Both calculation methods are not simple scoring functions, but robust frameworks aimed at evaluating NLP model outputs using given reference texts. Therefore, we only cover parts of these methods that are relevant for our research purposes.

For clarification, let us describe the most important definitions before we go over our calculations. In the field of NLP, an n -gram is a contiguous sequence of $n \in \mathbb{N}$ tokens from a given sample of text. They are instances of a sequence of characters that are grouped together as a useful semantic unit for processing. Depending on the application in which they are used, tokens can be a simple character, few characters, or even words. This paper considers tokens that represent words. When $n = 1$ the n -gram is called a unigram, $n = 2$ a bigram, and $n = 3$ it is a trigram. In our calculations, we use multiple n values to provide a more complete picture.

¹ This includes polysemous senses of the definition within the entries.

² This indicates the number of lemmas that can undergo certain lexicographical items, as some items are only applicable to certain parts of speech. The provided percentages also correspond to this.

Two other useful definitions are reference and candidate text. The former can be considered as ground truth and it is usually compiled by humans, while the latter is generated by a NLP model. In our case, reference text is information gathered from OALD, while candidate text refers to answers collected from ChatGPT. Comparing reference and candidate texts yields a similarity score $s \in [0, \dots, 1]$. If $s = 0$ the texts are completely different, while $s = 1$ means they are the same according to the used method. However, it is important to highlight that BLEU and ROUGE only considers the syntactics and not the semantics of a text.

3.6.1 Method BLEU

Originally, BLEU is designed for machine translation tasks. However, it is widely used in other areas such as code comparison Rikk et al. (2022) for program synthesis. This section gives an overview of the method and introduces all key concepts of it.

This method calculates the n -gram overlaps between the reference and candidate texts. Usually, we have multiple of the former as there can be multiple correct translation for a given text. However in our case, the reference text is obtained from OALD, because we are only interested in the similarities between it and ChatGPT.

Now, we go over how BLEU is calculated. Let us define the count function which given a text T and a n -gram g returns the number of times g is in T .

$$\text{count}(g, T) = \sum_{\substack{t \in T \\ t = g}} 1 \quad (1)$$

Next, a clipped count count_c function given a list of reference texts \mathcal{R} and candidate text C calculates the maximum number of times a n -gram occurs in any single reference translation. Then clips the total count of each candidate n -grams by its maximum reference count.

$$\text{count}_c(g, \mathcal{R}, C) = \min \left(\text{count}(g, C), \max_{R \in \mathcal{R}} \text{count}(g, R) \right) \quad (2)$$

With Equations (1) and (2), BLEU is calculated as follows. We first compute the n -gram matches sentence by sentence. Next, we add the clipped n -gram counts for all the candidate sentences and divide by the number of candidate n -grams in the test corpus to compute a modified precision score, p_n , for the entire test corpus.

$$p_n = \frac{\sum_{c \in C} \sum_{g \in c} \text{count}_c(g, c)}{\sum_{c' \in C'} \sum_{g' \in c'} \text{count}(g', c)} \quad (3)$$

Then, we take the geometric mean of the test corpus' modified precision scores and then multiply the result by an exponential brevity penalty factor. If k is the length of the candidate translation and r is the effective reference corpus length, then the brevity penalty BP :

$$BP = \begin{cases} 1 & \text{if, } k > r \\ e^{\frac{1-r}{k}} & \text{if, } k \leq r \end{cases} \quad (4)$$

Last, *BLEU* function is calculated as

$$BLEU = BP \cdot \exp \left(\sum_{n=1}^N w_n \log p_n \right) \quad (5)$$

where $w_n \in \mathbb{R}$ is called weight and $\sum_n w_n = 1$. In our calculations, we use a variety of weights to obtain a more robust evaluation. Depending on the n -grams used in Equation (5), it is also referred to as BLEU- n .

3.6.2 Method ROUGE

ROUGE is a set of metrics, rather than just one method. In this section, we cover the main approaches that are used in our tests, starting with ROUGE- N .

Formally, ROUGE- N is an n -gram recall between a candidate summary C and a set of reference summaries \mathcal{R} . ROUGE- N is computed as follows:

$$ROUGE\text{-}N = \frac{\sum_{R \in \mathcal{R}} \sum_{g \in R} count_m(g, C)}{\sum_{R' \in \mathcal{R}} \sum_{g' \in R'} count(g', C)} \quad (6)$$

where g is a n -gram, function $count_m$ is the maximum number of n -grams co-occurring in a candidate summary and a set of reference summaries, while function $count$ is defined as Equation (1). With ROUGE- N , N represents the n -gram that we are using. For ROUGE-1, we would be measuring the match-rate of unigrams between our model output and reference.

ROUGE- N can calculate three different values. These are recall, precision, and F1 score. Recall counts the number of overlapping n -grams found in both the model output and reference, then divides this number by the total number of n -grams in the reference (Equation (6)). This ensures that our model is capturing all of the information contained in the reference, but this is not so great at ensuring our model is not just pushing out a huge number of words to game the recall score. To avoid this, we use the precision metric, which is calculated just as the recall except, we divide by the model n -gram count and not with the reference n -gram count. Last, the F1 score is calculated as

$$F1 \text{ score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \quad (7)$$

ROUGE- L measures the longest common subsequence (LCS) between our model output and reference. We can apply our recall, precision, and F1 calculations just like before, but this time we replace $count_m$ with the LCS count.

4. Manual Analysis

This section presents the findings of our manual analysis, which is organized according to the item classes proposed by Wiegand, each containing relevant lexicographical items.

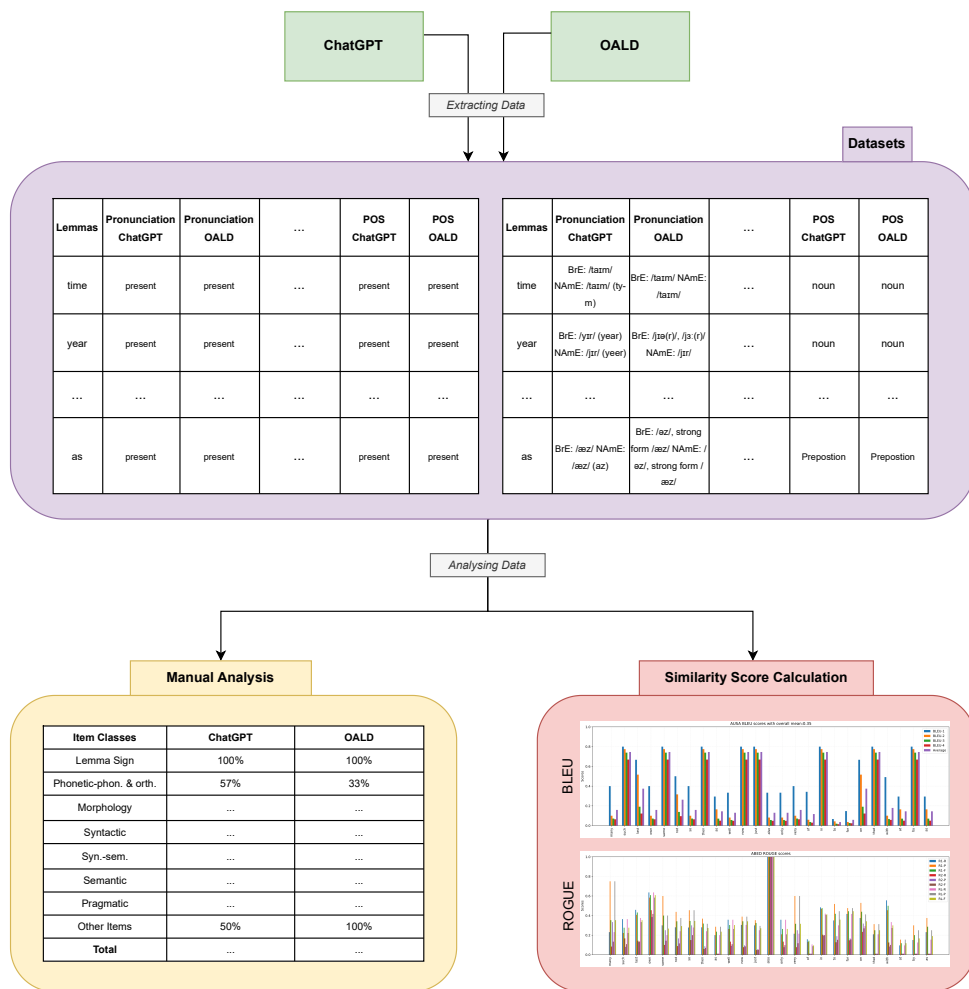


Figure 1: Comparing ChatGPT and OALD. First, we extract the information from both platforms manually. This yields two datasets. The first describes the presence or absence of the lexicographical items, while the second contains the actual answers from both tools. Then, we analyse our datasets using Wiegand’s item classes and similarity scores. Last, the results are presented as tables for the former and as figures for the latter.

4.1 Lexicographical Items Regarding LES

Table 1 shows that both ChatGPT and OALD can provide LES to all of our selected lemmas. When providing dictionary information, ChatGPT displays this item or headword in a plain format without any typographical indicators such as font-style, font-size, or colors that make it more invisible than any other information within the entry. Prompt 1 shows that the headword appears within quotation marks (“..”) in the answer. In contrast, OALD displays the lemma sign in bold and dark blue color at the top of each entry, making it highly visible and distinct from other elements. The font size is adjusted to ensure effective emphasis. Furthermore, the presentation of the headword is not influenced by the different parts of speech.

Lexicographical Item	Lemma Count	ChatGPT	OALD
Item giving the form of LES	50	100%	100%

Table 1: The percentage result of ChatGPT and OALD’s capabilities to provide LES shows that both platforms score 100%. This is essential because the lemma sign is a prerequisite for including an entry in a dictionary.

4.2 Lexicographical Items Regarding PPOI

According to the information presented in Table 2, it is evident that both ChatGPT and OALD provide the pronunciation of all our lemmas using the International Phonetic Alphabet (IPA) transcription system. However, ChatGPT does not include the pronunciation as part of the dictionary entry on the first attempt as shown in Prompt 1. To access the pronunciation, we must specifically ask, but once prompted, ChatGPT delivers just like OALD as demonstrated in Prompt 3. One area where ChatGPT falls short compared to OALD is that OALD provides audio files for users to imitate the sounds. Instead of audio files, ChatGPT compensates by offering detailed instructions on how to pronounce words which related to the lexicographical items giving *accent*, *vowel quantity marker*, *syllable* and *syllable division* which are lacking in OALD. Regarding *item giving the spelling*, the lemma sign serves not only as a means to identify a headword in a dictionary, but it also functions as a reference point for the word’s spelling. As a result, both ChatGPT and OALD are equipped to provide this lexicographical item for 100% of our lemmas.

Lexicographical Item	Lemma Count	ChatGPT	OALD
Pronunciation	50	100%	100%
Accent	50	14%	–
Vowel quantity marker	50	100%	–
Syllable	50	14%	–
Spelling	50	100%	100%
Syllable division	50	14%	–

Table 2: The comparison results for PPOI indicate that ChatGPT matches OALD’s 100% score for pronunciation and spelling items. Additionally, ChatGPT can provide a 100% score for the vowel quantity marker, which is not a feature possessed by OALD. However, for the items giving accent, syllable, and syllable division, ChatGPT only provides 14%, whereas OALD does not include these features.

PROMPT 3: PROVIDE PRONUNCIATION IN AMERICAN AND BRITISH ENGLISH FOR "PEOPLE"

The pronunciation of "people" is generally the same in both American and British English. It is pronounced as 'pi:pl (PEE-pul) with stress on the first syllable in both dialects.

4.3 Lexicographical Items Regarding MOI

Table 3 demonstrates the capabilities of ChatGPT and OALD for providing data on lexicographical items in this class. *Item giving gender*, not applicable in English, is excluded from our criteria. Regarding *inflexion*, ChatGPT and OALD offer inflection for lemmas within certain parts of speech limitations. ChatGPT covers 77.5% of lemmas across multiple parts of speech, while OALD covers 42.5%, primarily for verbs and nouns. For *degrees of comparison*, adjectives and adverbs in English are considered. Among 50 lemmas, we focus on 20 from POS adjective and adverb. ChatGPT provides information for 11 lemmas (55%), mostly adjectives (see example Prompt 4), while OALD only covers four lemmas (20%), three adjectives, and one adverb. All 50 lemmas can have *compounds* and *word families* identified by ChatGPT (see example Prompt 5). Note that some selected lemmas may not form compounds, which ChatGPT can indicate. OALD provides compounds and word families only for noun lemmas (20%), with certain compounds listed separately. Nouns alone offer *item giving number*. ChatGPT provides plural forms for all ten nouns (100%), while OALD only provides only irregular plural forms for two lemmas: *man* and *child* (20%). Regular plural forms ending in *-s* or *-es* are not covered by OALD.

Lexicographical Item	Lemma count	ChatGPT	OALD
Inflection	40	77.5%	42.5%
Degrees of comparison	20	55%	20%
Compound	50	100%	20%
Word family	50	100%	20%
Number	10	20%	4%

Table 3: The comparison of MOI reveals that ChatGPT and OALD can provide all the lexicographical items in this class. However, the ability of ChatGPT and OALD to provide information in this class depends on the POS of the lemmata, which affects the percentage of information provided. In terms of MOI, ChatGPT is capable of providing information for a greater number of lemmata than OALD.

PROMPT 4: PROVIDE DEGREES OF COMPARISON FOR "GOOD"

The degrees of comparison for "good" are:

- Positive degree: good
- Comparative degree: better
- Superlative degree: best

PROMPT 5: PROVIDE COMPOUNDS AND WORD FAMILY FOR VERB "BE"

The verb "be" is an irregular verb and does not typically form compounds. However, here are some related words in its word family:

Word family:

- am (present tense first person singular form)
- is (present tense third person singular form)
- are (present tense plural form and second person singular form)

4.4 Lexicographical Items Regarding SYI

Table 4 indicates that both ChatGPT and OALD can provide all the lexicographical items related to this item class, achieving a 100% coverage. Regarding *part of speech* ChatGPT includes all the relevant POS in a single entry following by the senses as demonstrated in Prompt 1, while OALD tends to present them separately in each individual entry, such as with the lemma "work" that can function as both a noun and verb. Regarding *item giving valency* ChatGPT displays valency using argument structure³ and indicates whether a verb is *transitive* or *intransitive* (see example Prompt 6). OALD provides information on valency for the same verb, but only for the latter. For example, for the sense of *do* as a transitive verb, OALD highlights *do something* in bold to indicate that the verb requires an object or thing to receive the action. *Item giving adjective distribution* is limited to POS adjective. Both ChatGPT and OALD present this data through example sentences and collocations (see Prompt 7). Their approaches are similar.

Lexicographical Item	Lemma count	ChatGPT	OALD
Part of speech	50	100%	100%
Valency	10	100%	100%
Adjective distribution	10	100%	100%

Table 4: The percentage results of the capabilities between ChatGPT and OALD in providing information related to SYI show that both platforms are capable of providing 100% of all related lexicographical items.

PROMPT 6: PROVIDE VALENCY FOR VERB "DO"

The verb "do" is a transitive verb and its valency is 2, which means it requires two arguments to form a complete sentence. The first argument is the subject of the sentence, and the second argument is the object that receives the action of the verb.

[Example sentences]

The verb "do" can also be used as an auxiliary verb to form questions and negative statements, in which case it does not take an object. In these cases, the valency of the verb is considered to be 0.

[Example sentences]

³ This refers to the number of core arguments or complements required by a given verb to form a grammatically correct sentence.

PROMPT 7: PROVIDE ADJECTIVE DISTRIBUTION FOR "NEW"

The adjective "new" can be used to describe a variety of nouns across different categories, including:

1. New objects: new car, new phone, new book
2. New ideas: new concept, new theory, new approach
3. New experiences: new adventure, new opportunity, new relationship
4. New places: new city, new neighborhood, new country
5. New information: new data, new research, new findings

Overall, the distribution of "new" as an adjective is quite wide and versatile.

4.5 Lexicographical Items Regarding SYSI

In terms of the lexicographical items related to this item class, both ChatGPT and OALD demonstrate capability, covering *collocation*, *phraseme*, *proverb*, and *example* (see Table 5). However, there is a distinction between the example sentence types provided. ChatGPT offers *usage examples*, whereas OALD presents *citation texts* without the accompanying *reference*.

ChatGPT provides *collocations* for all 50 selected lemmas (100%) in a list format. Users can choose the lemma's position in the collocations. OALD, however, covers only 18 out of the 50 lemmas (36%), mostly nouns. Its collocations appear in separate information boxes for each definition, sourced from the Oxford Collocations Dictionary.

Regarding *Phraseme* ChatGPT can provide idiomatic expressions for all of the 50 lemmas (100%), although some of the expressions may not include the headword but refer to it by meaning. For POS other than nouns and verbs, ChatGPT may provide some kind of collocations instead of idioms which is not the concept of idiomatic expressions. In contrast, OALD can provide idiomatic expressions for 84% of the lemmas. OALD has a separate section dedicated to idioms located at the end of the dictionary entry. Users can also find a shortcut to this section at the top of the entry below the headword, POS, and pronunciation.

ChatGPT is capable of providing *proverb* for all the lemmas. However, some proverbs may not include the headword, and the accuracy of the provided proverbs is questionable. On the other hand, OALD can only provide this information for 26% of the lemmas, but the proverbs provided are accurate. OALD presents proverbs within the idioms section, indicated by (*saying*).

ChatGPT and OALD are capable of providing *examples* for all of our lemmas (100%). ChatGPT usually offers ten example sentences for each lemma mixed from all the senses. On the other hand, OALD provides sense-specific examples of varying numbers. ChatGPT generates original example sentences using its own language proficiency derived from its training on large amounts of text, thus we consider the examples provided by ChatGPT to be *usage examples*. OALD offers a different type of examples referred to as *citation text* or *corpus examples*. These examples are usually sourced from the dictionary's corpora and other lexicographical sources. However, OALD does not include *item indicating the reference of the citation* within the entry. Since it's apparent that the examples are

extracted from the BNC, this may be the reason why this information is not provided in OALD dictionary entry.

Lexicographical Item	Lemma Count	ChatGPT	OALD
Collocation	50	100%	36%
Phraseme	50	100%	84%
Proverb	50	100%	26%
Example	50	100%	100%
Usage example	50	100%	–
Citation text	50	–	100%
Reference of the citation	50	–	–

Table 5: In the comparison of SYSI capabilities, it was found that ChatGPT can provide collocations, phrasemes, and proverbs for all selected lemmas. In contrast, the percentages of OALD in providing these lexicographical items are consistently lower than those of ChatGPT. While both ChatGPT and OALD can provide example sentences, the approaches used by the two platforms to provide these examples differ.

4.6 Lexicographical Items Regarding SEMI

Regarding the semantic class, ChatGPT and OALD are capable of providing most of the items in this category. However, ChatGPT is unable to provide the *illustration* due to its nature as a text-based LLM. However, as Prompt 8 shows, it can provide detailed and descriptive explanations to help the users understand the concepts and ideas of the lemma. OALD occasionally includes pictures with the definitions in its dictionary entries. However, the entries for our 50 chosen lemmas do not have any illustrations. It is worth noting that OALD has the ability to provide illustrations, but not for the lemmas we selected. We indicate this by using plus symbol (+) in Table 6.

Additionally, neither ChatGPT nor OALD can provide the *item giving an equivalent*, as this belongs to one of the features of bilingual and multilingual dictionaries, that provides translations of the lemmas in the target language(s). However, since OALD is a monolingual learner’s dictionary for English, this feature is not available in OALD. Similarly, we cannot compare this feature in ChatGPT for the same reason.

ChatGPT and OALD are both capable of providing *Item giving meaning* and *polysemy* for all the lemmas. In the cases of polysemy, both platforms enumerate each sense and present them in a list. However, OALD uses a restricted vocabulary, known as The Oxford 3000 Oxford University Press (n.d.), to ensure that definitions are easily understandable for EFL learners.

Both ChatGPT and OALD can provide *synonyms* and *antonyms*. But not all the dictionary entries in OALD possess this feature. Only 26% of our lemmas can provide synonyms and only 2% percent can provide antonyms. Out of the 50 lemmas we examine, only the lemma *new* provides an antonym. In OALD, synonyms are presented in two different ways: a fold-out box and list of synonymous words. On the other hand, ChatGPT presents both synonyms and antonyms in a list format, as demonstrated in Prompt 9 and can provide 100% of the lemmas.

Lexicographical Item	Lemma Count	ChatGPT	OALD
Meaning	50	100%	100%
Synonyms	50	100%	26%
Antonyms	50	100%	2%
Polysemy	50	100%	100%
Illustration	50	–	+
Equivalent	50	–	–

Table 6: Upon comparing the SEMI capabilities, it was discovered that both ChatGPT and OALD can provide complete information for the essential lexicographical items of meaning and polysemy, with a 100% success rate. Both platforms can also provide synonyms and antonyms, with OALD consistently shows lower percentages in providing this information for our selected lemmas. ChatGPT is unable to provide illustrations, while OALD has the capability to do so, although it is not available for our selected lemmas. There is no equivalent feature to be considered in this context.

PROMPT 8: PROVIDE ILLUSTRATION FOR "MAN"

I'm sorry, but I cannot provide illustrations as I am a text-based AI language model. An illustration for the word "man" could be a drawing, painting, or photograph that depicts a male human being. The illustration could show a man in a specific context, such as at work, at home, or in a social setting. It could also show a man in different stages of life, such as a young boy, middle-aged man, or elderly man. The purpose of the illustration is to provide a visual representation of the word "man" to help readers understand and visualize its meaning.

PROMPT 9: PROVIDE SYNONYMS AND ANTONYMS FOR "DAY"

Synonyms for "day":

- daytime
- 24 hours
- sunlight
- ...

Antonyms for "day":

- night
- nighttime
- evening
- ...

4.7 Lexicographical Items Regarding PRAI

OALD typically includes pragmatic information within each individual sense, indicated before the definition in brackets in italic and light grey. An example of *diatechnic labelling* is (*computing*) provided within the entry of *at* for sense 15, referring to the symbol @ used in email addresses. This is the only lemma (2%) out of 50 lemmas that OALD can provide diatechnic label. ChatGPT is capable of providing information on diatechnic

labels, but for certain lemmas, it responds with *No diatechnic labelling* and notes that the words are commonly used in general language rather than specialized terminology. It is worth mentioning that ChatGPT can provide diatechnic labels in other cases, but it is not applicable for our selected lemmas.

OALD can provide *diastratic label* for 12% of our 50 lemmas. This information can be found in the entries of certain lemmas such as *have, do, make, get, know, and well*. Labels such as (*slang*) can be found in these entries. ChatGPT, however, is not capable of providing information related to this label, giving the reason that it requires more context related to the headword.

In terms of *diafrequency labeling*, OALD does not provide this information for our selected lemmas within its dictionary entries. However, it should be noted that the lemmas in OALD are already commonly used and therefore do not require frequency labeling. In some cases, the entries may include a label such as (*rare*), which refers to diafrequency. However, this does not apply to our selected lemmas. In contrast, ChatGPT provides diafrequency information for all lemmas, indicating whether they are *common* or *very common*.

Out of the 50 lemmas we analyzed, OALD provides *diaevaluative labelling* for eight of them, which accounts for 16% of the total. The labels used in OALD for diaevaluative purposes are denoted by phrases such as (*approving*) or (*disapproving*). ChatGPT is also capable of providing this information, although it uses different labels. For our specific list of lemmas, ChatGPT indicates whether a word is *neutral* or *positive* in connotation, since there are no words with negative connotations in our list. However, it's important to note that due to ChatGPT's lack of contextual awareness, caution should be exercised when interpreting these labels.

OALD is capable of providing *diachronic labels* for 11 lemmas (22%) of the lemmas in our sample. These labels, such as (*old used*) and (*old-fashioned*) appear within the dictionary entries. Although ChatGPT is unable to provide diachronic labels for our selected lemmas, it is important to note that this is because the lemmas are still commonly used today. However, it is worth mentioning that ChatGPT has the capability to provide diachronic labels for other entries. When asked if it is possible to provide diachronic labels, ChatGPT indicated that terms such as *historic* or *archaic* are used for some entries.

OALD is capable of providing *diatopic labelling* for 48% of its lemmas, indicating regional varieties of English such as American English, Australian English, British English, Northern English, etc. This labelling is provided within individual senses rather than just for the headword. In contrast, ChatGPT can also provide diatopic labelling, but for our chosen lemmas, it only offers a *neutral* label since the words are universal and not associated with any particular region or culture. We consider this as ChatGPT is capable of providing diatopic label, but just not for our chosen lemmas.

OALD is typically able to provide *Item giving the diainegrative labelling* for loanwords and their original language. However, since our chosen lemmas do not fall under this category, OALD entries do not include this label. Nevertheless, if we were to look up a lemma like "croissant," it would be labeled as (*from French*) for diainegrative purposes. ChatGPT, on the other hand, explains that such labels would fall under etymology and can provide the word's origin instead. We consider ChatGPT unable to provide diainegrative labelling

Regarding *Dianormative labelling*, OALD can provide this information for 10% of the 50 chosen lemmas from different parts of speech. OALD shows this information by presenting typical mistakes made by EFL learners with a crossed-out sentence alongside the correct version. However, ChatGPT is not able to provide this type of information, giving the reason that dianormative labelling is a complex process that requires knowledge of the social, cultural, and historical context of a language and its users. It involves identifying the norms and values associated with the use of certain words and how they may vary across different social groups or contexts. This is a task that requires human expertise and cultural knowledge.

Out of our 50 lemmas, OALD can provide *Item giving the diatextual labelling* for 10 of them (20%). The labels provided in OALD entries include *literally* or *figurative*, which indicate the intended meaning of larger textual units such as phrases, sentences, and definitions. However, ChatGPT cannot offer this type of information as diatextual labelling is not applicable to individual words. It is a labelling system that is used to analyse and describe larger textual units, such as those found in OALD entries.

OALD provides *diamedial labels* such as *spoken* or *written*, but surprisingly, none of our chosen lemmas are labeled as such in the dictionary entries. It is possible that this is because they are commonly used words. However, ChatGPT can provide diamedial labels for all of our chosen lemmas, using terms like *spoken*, *written*, *news*, and *academic* to indicate this information. However, it is important to note that ChatGPT provides all four above mentioned labels to all of our lemmas, which may lead to inaccurate information. Users of ChatGPT should be aware of this potential issue.

According to Wiegand et al. (2010), OALD includes some additional diasystem labels, such as *diaphasic labelling*, which indicates whether a lexeme is considered *formal* or *informal*, and *diaattitudinal labelling*, which includes indications such as *humorous* and *ironic*. However, these labels are not included in item classes or lexicographical items as defined by Wiegand.

Lexicographical Item	Lemma Count	ChatGPT	OALD
Diatechnic labelling	50	+	2%
Diastratic labelling	50	-	12%
Diafrequency labelling	50	100%	+
Diaevaluative labelling	50	100%	16%
Diachronic labelling	50	+	22%
Diatopic labelling	50	+	48%
Diintegrative labelling	50	-	+
Dianormative labelling	50	-	10%
Diatextual labelling	50	-	20%
Diamedial labelling	50	100%	+

Table 7: The pragmatic class percentage outcome indicates that OALD excels in providing all types of labelling, whereas ChatGPT falls short in providing information on this class. This is mainly because the expertise and cultural knowledge of human lexicographers are essential for such labelling, and ChatGPT requires more context in order to provide related information.

4.8 Lexicographical Items Regarding OTI

The two lexicographical items, etymology and cross-reference, do not fall under any of the item classes mentioned previously. As a result, they are categorized separately by Wiegand. Although providing the *etymology* of a word is not a typical feature found in monolingual learner’s dictionaries, most printed dictionaries do not provide this information for every lemma in their dictionaries due to space constraints. However, both ChatGPT and OALD are capable of providing etymology for all 50 lemmas with a 100% success rate. In OALD, the etymological information is available in a separate fold-out section titled *Word Origin*. The manner in which ChatGPT provides etymological information is illustrated in Prompt 10.

Regarding *Item giving a cross-reference*, ChatGPT generates responses one by one and does not have the capability to refer to external sources of information within a single answer. On the other hand, OALD has a cross-reference feature marked in blue for all of the lemmas, and by hovering over it with the cursor, the linked source is immediately displayed. Additionally, users can double-click on any word in the dictionary entry, which will lead them to the entry for that word. This feature is particularly useful when encountering unfamiliar words in a definition.

Lexicographical Item	Lemma Count	ChatGPT	OALD
Etymology	50	100%	100%
Cross-reference	50	–	100%

Table 8: The percentage result of the OTI comparison shows that OALD is capable of providing information for both lexicographical items in this class, achieving a score of 100%. In contrast, ChatGPT can only provide information related to one item, demonstrating a limitation in its capabilities. This is due to the fact that ChatGPT does not possess the feature to refer to external sources.

PROMPT 10: PROVIDE WORD ORIGIN FOR "CHILD"

The word "child" comes from the Old English word "cild", which meant "a fetus, an infant, or a young person". The word is thought to be derived from the Proto-Germanic word "*kiltham", which meant "womb" or "uterus." This word is related to the Old Norse word "kildr" and the German word "Kinder".

4.9 Additional Lexicographical Features

In this section, we have compiled extra lexicographical features from OALD that were not classified by Wiegand but can benefit EFL learners, as illustrated in Table 9. We compare their availability in ChatGPT and note that while both dictionaries offer these features, ChatGPT requires specific prompts for providing certain information.

5. Similarity Scores

In this section, we present our key findings obtained through similarity scores. We first discuss the BLEU results, followed by the ROUGE results. Additionally, we provide

Lexicographical Item	Lemma Count	ChatGPT	OALD
CEFR level	50	100%	100%
Topic	50	+	100%
Abbreviation	50	10%	10%
Cultural Information	50	+	2%
Political statement	50	+	2%
Notes on usage	50	+	28%

Table 9: The percentage comparison results for the compiled additional lexicographical features in OALD demonstrate that ChatGPT is capable of providing information related to those lexicographical items as well.

interpretations for each of the findings. We calculate the similarity scores on the dataset containing the responses from ChatGPT and data from OALD. This contains the following lexicographical items: LES, AUSA, RA, WAA, ABED, and ETYA. Not counting the lemma sign, all lexicographical items define a category. For each category, we collect the answers of ChatGPT and OALD. Since we have 5 categories, this yields $1 + 5 \cdot 2$ many columns (features) with 50 rows for our dataset.

The similarity scores calculated by iterating over all categories row by row. For each row, the calculations return a vector $v \in \mathbb{R}^{1 \times l}$, where l is determined by the method used. When using BLEU, the last element of v is the average of the previous elements. The overall mean is calculated by taking the average of the last element in every v in a given category. For each category, our results are described by a matrix $X \in \mathbb{R}^{50 \times l}$. These matrices are visualized in the next sections.

5.1 BLEU Scores

This section contains our most important BLEU results and interpretations of these. In the calculations, we have used different n -grams with $n = 1, 2, 3, 4$. Also, the averages of all the n -grams are provided.

The used weights for BLEU-1 to BLEU-4 in order are $w_1 = [1]$, $w_2 = [0.5, 0.5]$, $w_3 = [0.33, 0.33, 0.33]$, and $w_4 = [0.25, 0.25, 0.25, 0.25]$. Additionally, we use a smoothing function Chen & Cherry (2014). This is needed because if there is no n -gram overlap for any order of n -grams, BLEU returns 0. Due to the precision for the order of n -grams without overlap is 0, and the geometric mean in the final BLEU score computation multiplies the 0 with the precision of other n -grams. This results in 0 independently of the precision of the other n -gram orders. Specifically, we use ϵ -smoothing which adds a small ϵ value to the numerator when it is 0 in Equation (3). In our case, $\epsilon = 0.1$.

BLEU scores consistently show that lexicographical items containing more n -grams receive lower scores, indicating that ChatGPT’s responses match better with single words (unigrams) than with phrases (multigrams). This trend is observed across all evaluated lemmas and the five chosen lexicographical items. Figure 2 highlights that ABED’s complex text elements result in lower scores, compared to single-word representations like RA or WAA. The bar charts clearly demonstrate that shorter candidate texts, like those in AUSA, RA, or WAA, receive higher scores, while longer ones, like ABED and ETYA,

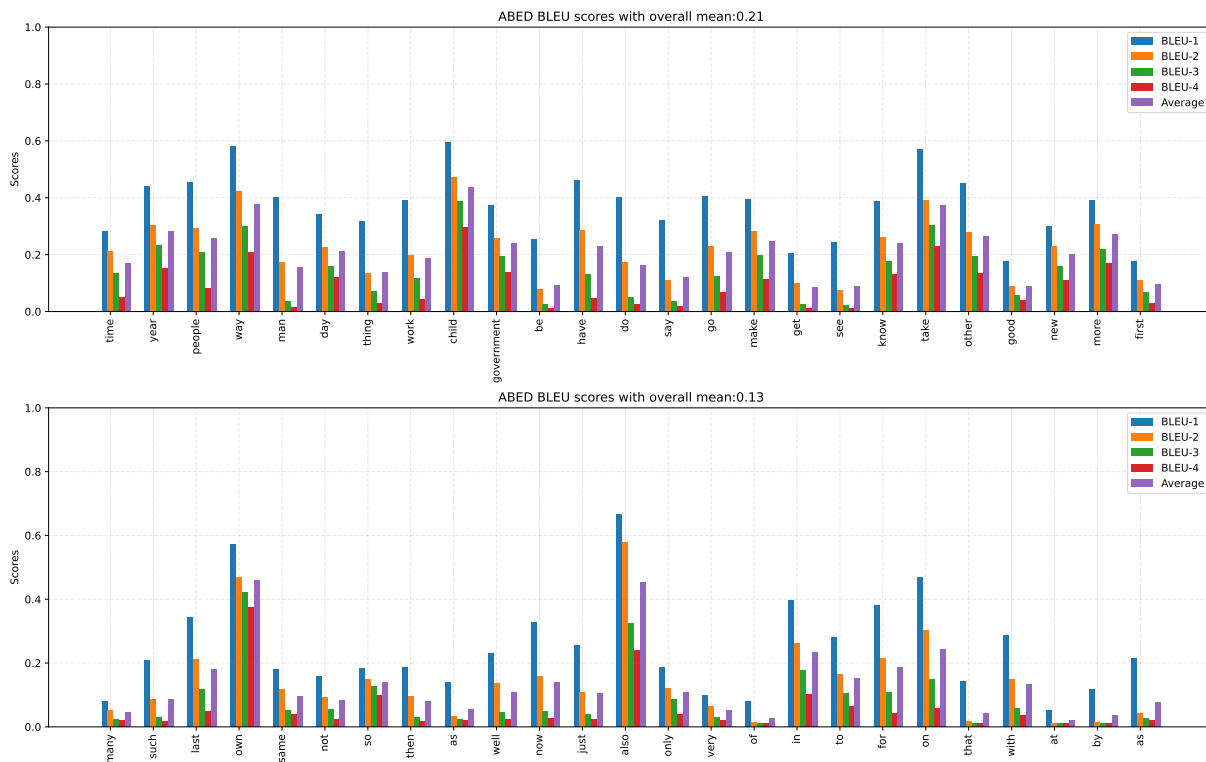


Figure 2: BLEU scores for ABED category. Each bar represents five scores, from left to right: BLEU-1 to BLEU-4, and the average. The top figure shows the first 25 lemmas, while the bottom figure displays the remaining 25. The overall mean for the data is indicated in the title of both figures.

receive lower scores. In fact, ETYA had the longest candidate text and obtained the lowest BLEU score among the five selected lexicographical items.

5.2 ROUGE Scores

This section contains our most important ROUGE results and interpretations of these. We calculate ROUGE-1 (R1), ROUGE-2 (R2), and ROUGE-L (RL) with recall, precision, and F1 scores for each method. On the figures, these values are denoted with their first letter. For example, ROUGE-1 recall is abbreviated to R1-R.

ROUGE scores consistently indicate that bigrams (R2) have the lowest scores compared to unigrams and RL for all of our lemmas, and all five selected lexicographical items. Remarkably, all n -grams scores for the lemma *also* achieve a perfect score of 1.0, as illustrated in Figure 3. This is due to the 100% match between the candidate and reference texts. ChatGPT and OALD provide the same definition, *in addition; too*, with identical punctuation. The trend of R2 scores being the lowest is consistent across all of our ROUGE score charts.

6. Evaluation

In this section, a summary of the results obtained from manual analysis and similarity score tests are presented. The manual analysis included a thorough evaluation of the

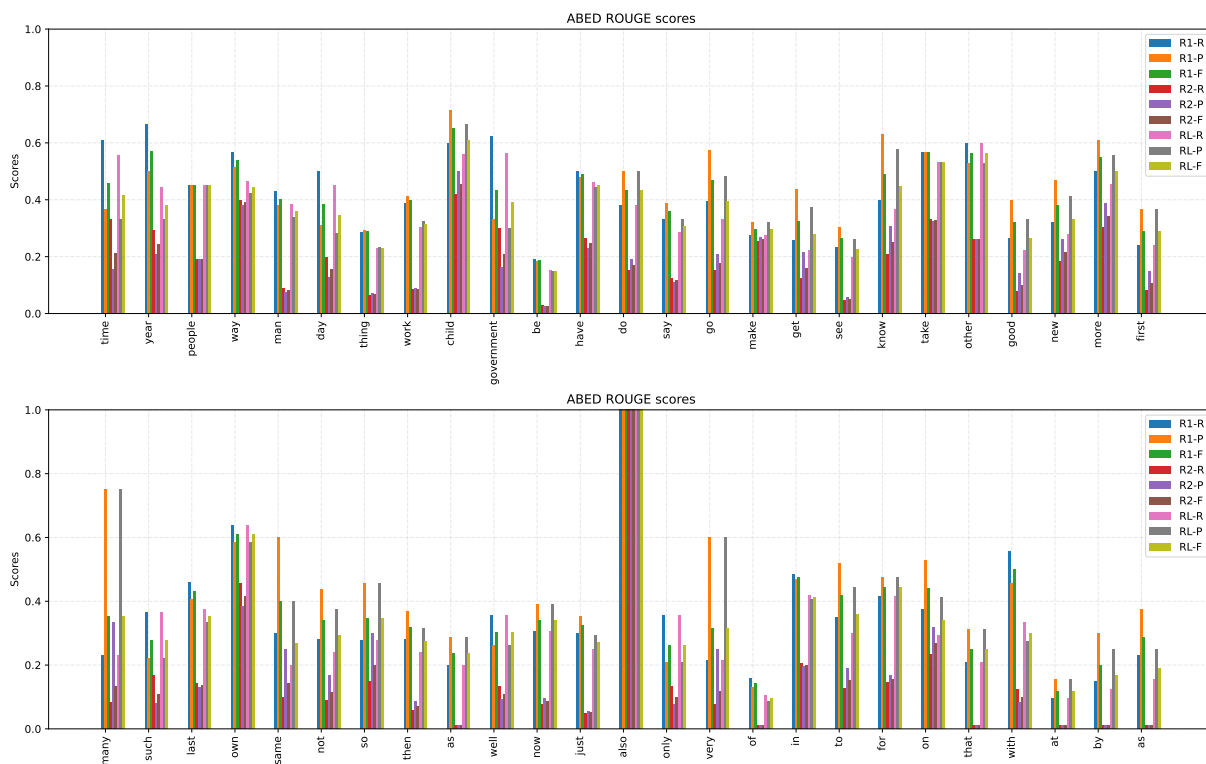


Figure 3: ROUGE scores in the ABED category. We visualize nine values for each lemma. In order from left to right, these are ROUGE-1 (R1), ROUGE-2 (R2), and ROUGE-L (RL) for each we provide recall (R), precision (P), and F1 (F) scores.

capabilities of ChatGPT and OALD. This was done by comparing the percentages of each item class between the two sources. Additionally, the analysis presented the total average score for all item classes.

The average capabilities of ChatGPT and OALD for each item class are presented in Table 10. ChatGPT has an average score of 68% in providing dictionary information for the 50 chosen lemmas. This is 11% higher than the average score of OALD, which is 57%. Both ChatGPT and OALD can provide 100% of related information for the LES and SYI item class. ChatGPT has a higher average score than OALD for all item classes except OTI, where OALD has a perfect average score of 100%.

The similarity scores of both BLEU and ROGUE suggest that higher scores are attained when candidate texts are evaluated at the unigram level, with those containing only one word unit reaching a perfect score of 1.0. Conversely, longer word units tend to receive lower scores, as demonstrated by the lower scores of lexicographical items ABED and ETYA. Of the five chosen items, AUSA holds the highest similarity scores followed by RA and WAA, while ETYA has the lowest scores indicating the least similarity to the reference text.

7. Conclusions and future work

The paper compares the abilities of ChatGPT and OALD for lexicographical purposes, specifically focusing on microstructural elements. The study finds that ChatGPT performs better on average than OALD in providing information related to lexicographical items,

Item Class	Average Score	
	ChatGPT	OALD
LES	100%	100%
PPOI	57%	33%
MOI	71%	21%
SYI	100%	100%
SYSI	71%	49%
SEMI	67%	38%
PRAI	30%	13%
OTI	50%	100%
<i>Total</i>	68%	57%

Table 10: The comparison of ChatGPT and OALD in providing information related to each item class yielded average scores. Both platforms achieved 100% for LES and SYI. ChatGPT had overall higher average scores than OALD in all item classes, except for OTI.

indicating its potential as a learner’s dictionary. However, ChatGPT has limitations such as the absence of contextual information and limited interactivity, which are important aspects of learner’s dictionaries. The paper also measures the similarity between the data generated by ChatGPT and OALD using BLEU and ROUGE metrics. While single words show high similarity between the two tools, responses consisting of multiple words differ significantly, suggesting variations in phrase construction and data presentation. The study acknowledges the need for further research on ChatGPT as a learner’s dictionary, including potential prompts for lexicographical tasks, the development of evaluation criteria, comparisons with other learner’s dictionaries, and assessment of response accuracy for different lexicographical items. Despite the limitations, the paper concludes that ChatGPT shows promise as a language learning tool and an efficient lexicographic aid for EFL learners.

8. References

- Araci, D. (2019). Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.
- Bahdanau, D., Cho, K. & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.
- Brants, T., Popat, A.C., Xu, P., Och, F.J. & Dean, J. (2012). Large language models in machine translation. US Patent 8,332,207.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33, pp. 1877–1901.
- Chen, B. & Cherry, C. (2014). A systematic comparison of smoothing techniques for sentence-level BLEU. In *Proceedings of the ninth workshop on statistical machine translation*. pp. 362–367.
- Dathathri, S., Madotto, A., Lan, J., Hung, J., Frank, E., Molino, P., Yosinski, J. & Liu, R. (2019). Plug and play language models: A simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164*.

- Davies, M. & Gardner, D. (2013). *A frequency dictionary of contemporary American English: Word sketches, collocates and thematic lists*. Routledge.
- Engelberg, S. & Lemnitzer, L. (2009). *Lexikographie und Wörterbuchbenutzung: Stauffenburg-Einführungen*. Stauffenburg (Tübingen).
- Herbst, T. (1996). On the way to the perfect learners' dictionary: a first comparison of OALD5, LDOCE3, COBUILD2 and CIDE. *International Journal of Lexicography*, 9(4), pp. 321–357.
- Hornby, A.S. (2019). Oxford advanced learner's dictionary. URL <https://www.oxfordlearnersdictionaries.com/definition/english/>.
- Ivančič, I. & Fabijanić, I. (2017). Structural Development of Oxford Advanced Learners' Dictionary. *Journal of Literature and Art Studies*, 7(5), pp. 588–607.
- Kant, N., Puri, R., Yakovenko, N. & Catanzaro, B. (2018). Practical text classification with large pre-trained language models. *arXiv preprint arXiv:1812.01207*.
- Lin, C.Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, pp. 74–81. URL <https://aclanthology.org/W04-1013>.
- OpenAI (2023). ChatGPT. <https://openai.com/blog/chatgpt>.
- Oxford Text Archive (2007). British National Corpus. <http://ota.ox.ac.uk/desc/2554>. [dataset].
- Oxford University Press (n.d.). The Oxford 3000™ by CEFR level. Online. URL <https://www.oxfordlearnersdictionaries.com/wordlist/english/oxford3000-cefr/>. Accessed on March 30, 2023.
- Papineni, K., Roukos, S., Ward, T. & Zhu, W.J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. pp. 311–318.
- Pilault, J., Li, R., Subramanian, S. & Pal, C. (2020). On extractive and abstractive neural document summarization with transformer language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. pp. 9308–9319.
- Poesia, G., Polozov, O., Le, V., Tiwari, A., Soares, G., Meek, C. & Gulwani, S. (2022). Synchromesh: Reliable code generation from pre-trained language models. *arXiv preprint arXiv:2201.11227*.
- Rikk, R., Várkonyi, T.A., Borsi, Z.R., Pintér, B. & Gregorics, T. (2022). Generating Algorithmic Patterns from Semi-structured Input Using a Transition-Based Neural Network. In *Intelligent Systems and Applications: Proceedings of the 2022 Intelligent Systems Conference (IntelliSys) Volume 2*. Springer, pp. 819–834.
- White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J. & Schmidt, D.C. (2023). A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.
- Wiegand, H.E. (1989). Der Begriff der Mikrostruktur: Geschichte, Probleme, Perspektiven. In *Handbücher zur Sprach- und Kommunikationswissenschaft*, volume 1. Berlin: Walter de Gruyter, pp. 409–462.
- Wiegand, H.E., Beißwenger, M., Gouws, R.H., Kammerer, M., Storrer, A. & Wolski, W. (2010). Wörterbuch zur Lexikographie und Wörterbuchforschung. *Berlin/New York*.
- Yasunaga, M., Ren, H., Bosselut, A., Liang, P. & Leskovec, J. (2021). QA-GNN: Reasoning with language models and knowledge graphs for question answering. *arXiv preprint arXiv:2104.06378*.

A. Similarity Scores Results

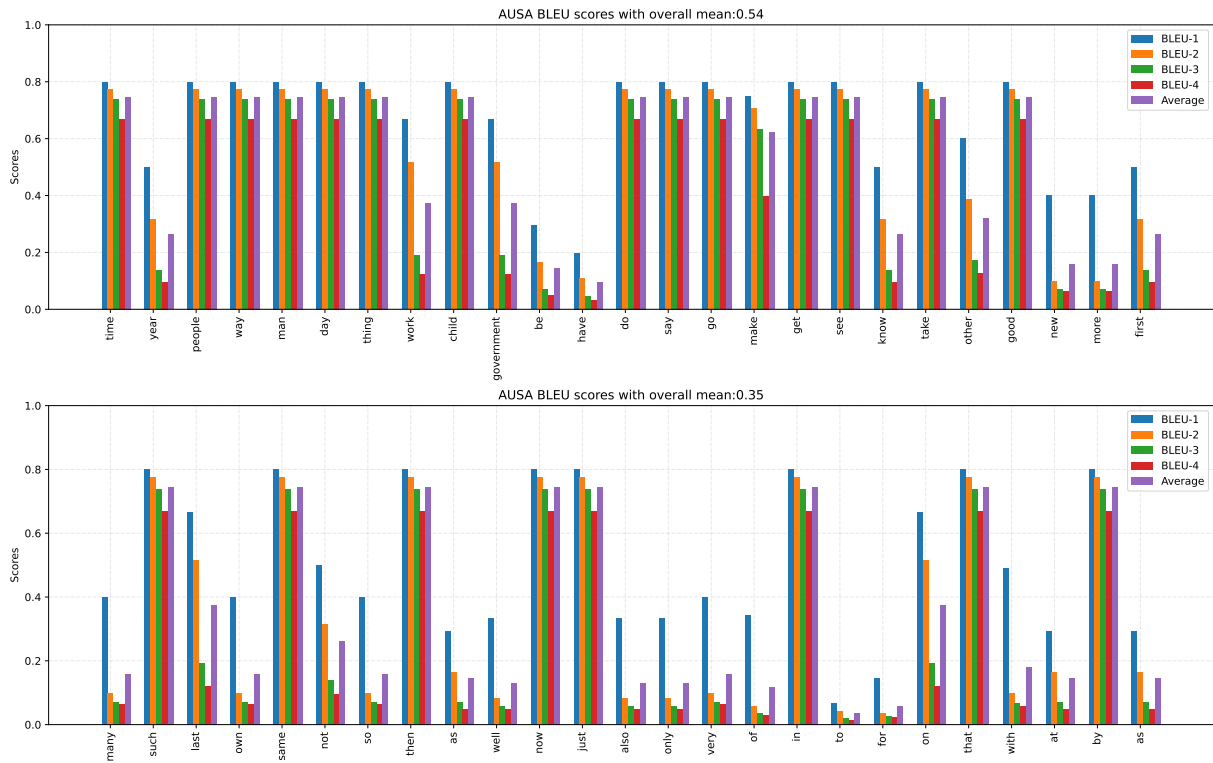


Figure4: BLEU scores for AUSA category. The score results indicate that the more n-grams present in both candidate and reference texts, the lower the score. Furthermore, the data for AUSA contains a comparable amount of word units. As a consequence, The bars from BLEU 1 to 4 for the majority of the chosen lemmas on the graph show quite similar scores.

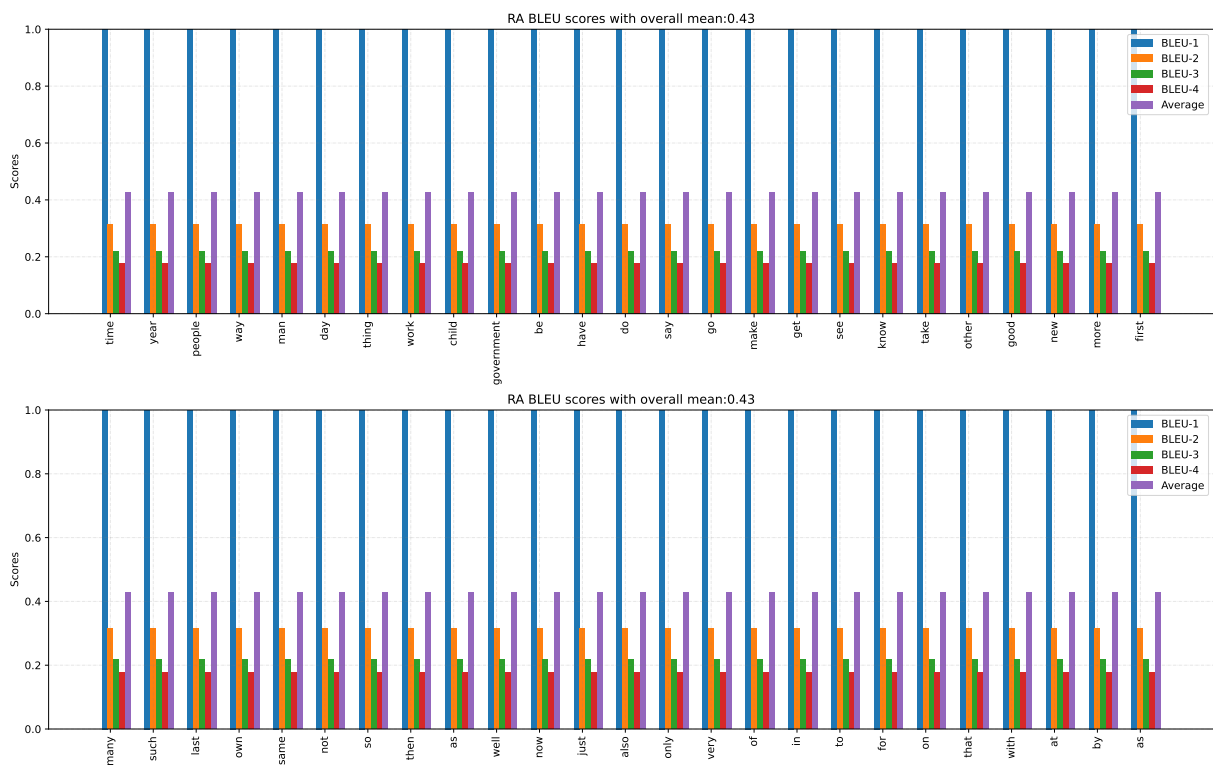


Figure 5: BLEU scores for RA category. Since the data in RA comprises a single word unit that also functions as a lemma sign, the BLEU 1 score is perfect at 1.0, signifying a complete match between the candidate and reference texts. Moreover, all lemmas attain equivalent scores across all BLEU scores from 1 to 4.

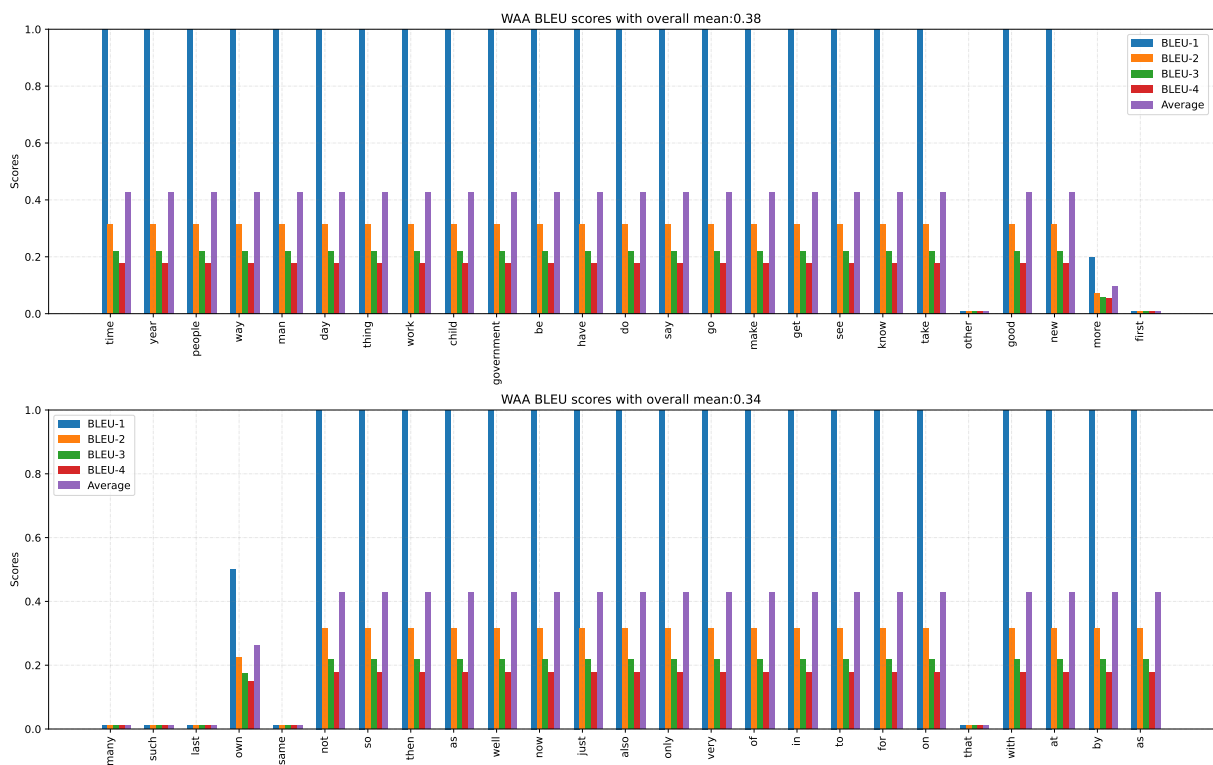


Figure 6: BLEU scores for WAA category. WAA’s data comprises a single word unit, leading most lemmas to achieve a perfect unigram score of 1.0. However, certain lemmas nearly attain a score of 0.0 for the same BLEU 1 score, as ChatGPT and OALD assign them different parts of speech.

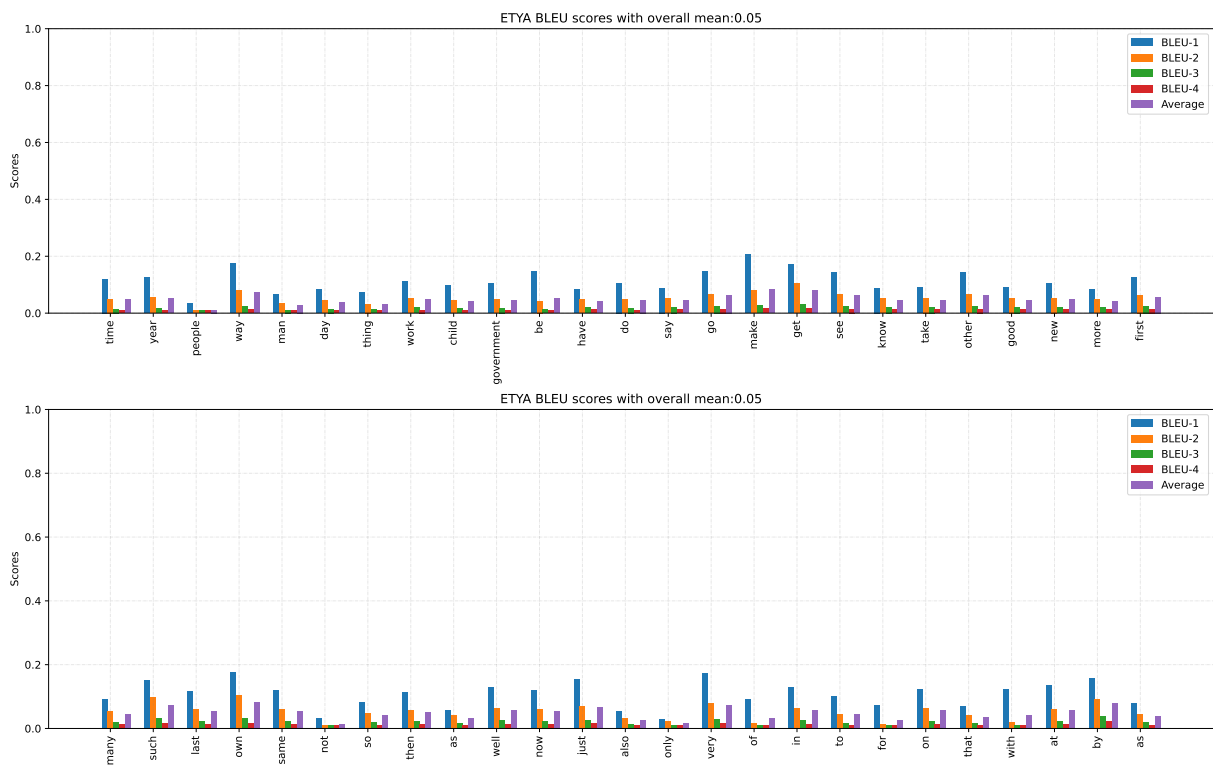


Figure 7: BLEU scores for ETYA category. ETYA contains the most word units among all the selected lexicographical items. The data related to ETYA from both ChatGPT and OALD refer to common origins of the lemmas. However, the formulation of data differs significantly, leading to a considerably lower overall score in this category. When users look up etymological information using ChatGPT, they will still receive the same information pertaining to the lemma.

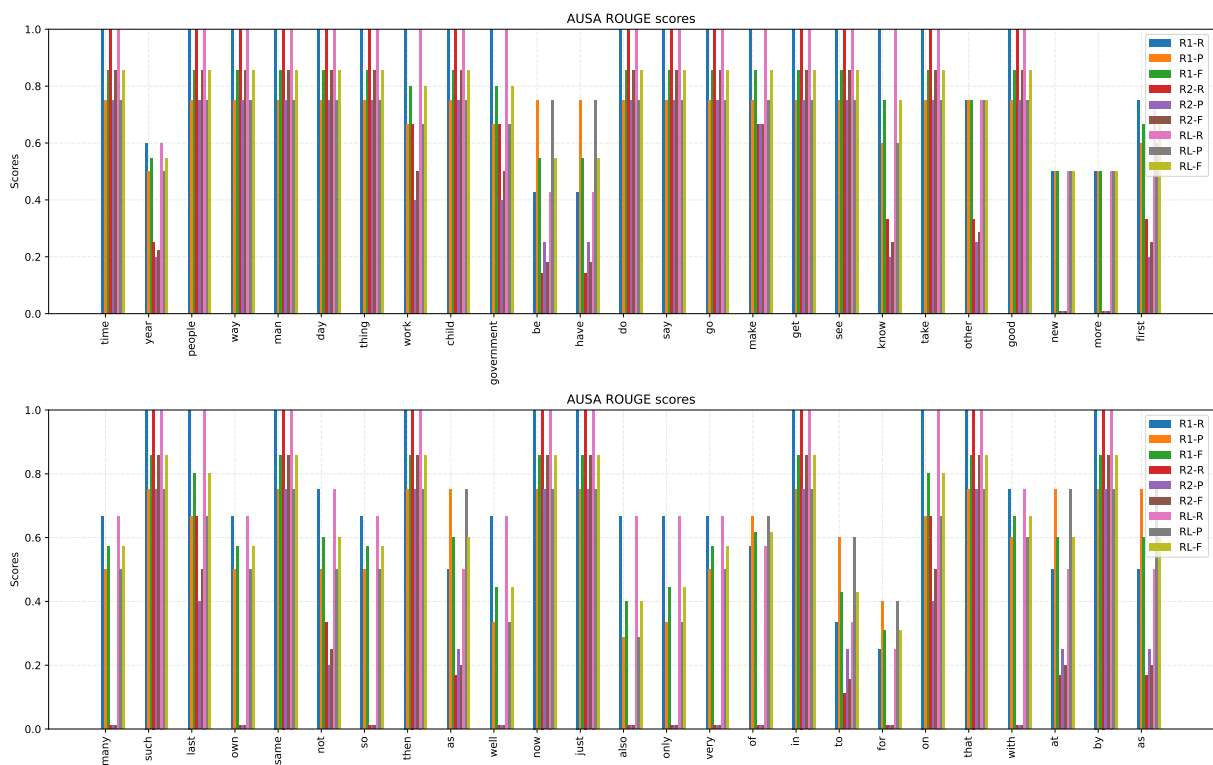


Figure 8: ROUGE scores in the AUSA category. Upon examining the charts, it is apparent that there is a consistent trend in the Recall (R) bars for R1, R2, and RL, with almost all bars reaching a perfect score of 1.0. This trend is particularly notable in the context of our analysis of AUSA data, where we observe high overall similarity scores for all the lemmas.

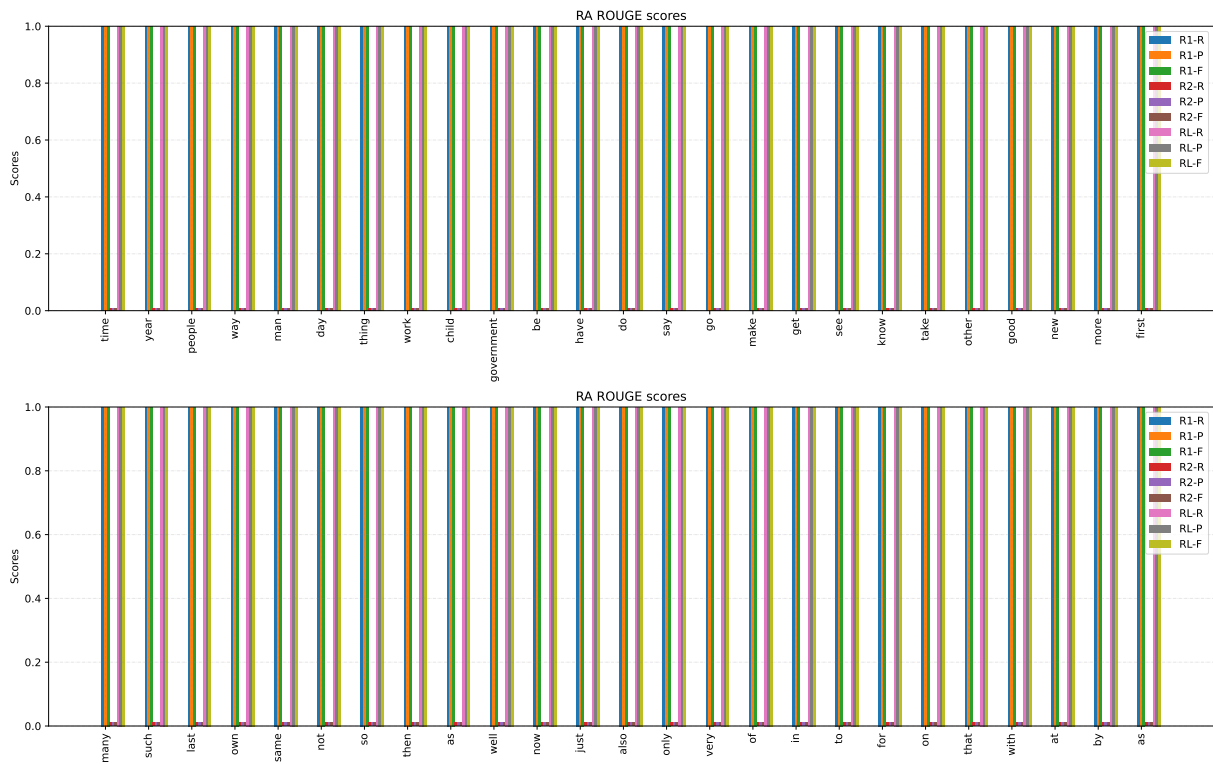


Figure 9: ROUGE scores in the RA category. It is evident that R2 score is not applicable for the data belonging to this category since it consists of only unigrams and not bigrams. Therefore, since the longest word units (RL) are also unigrams, all the lemmas achieve a perfect match score of 1.0 for R1 and RL.

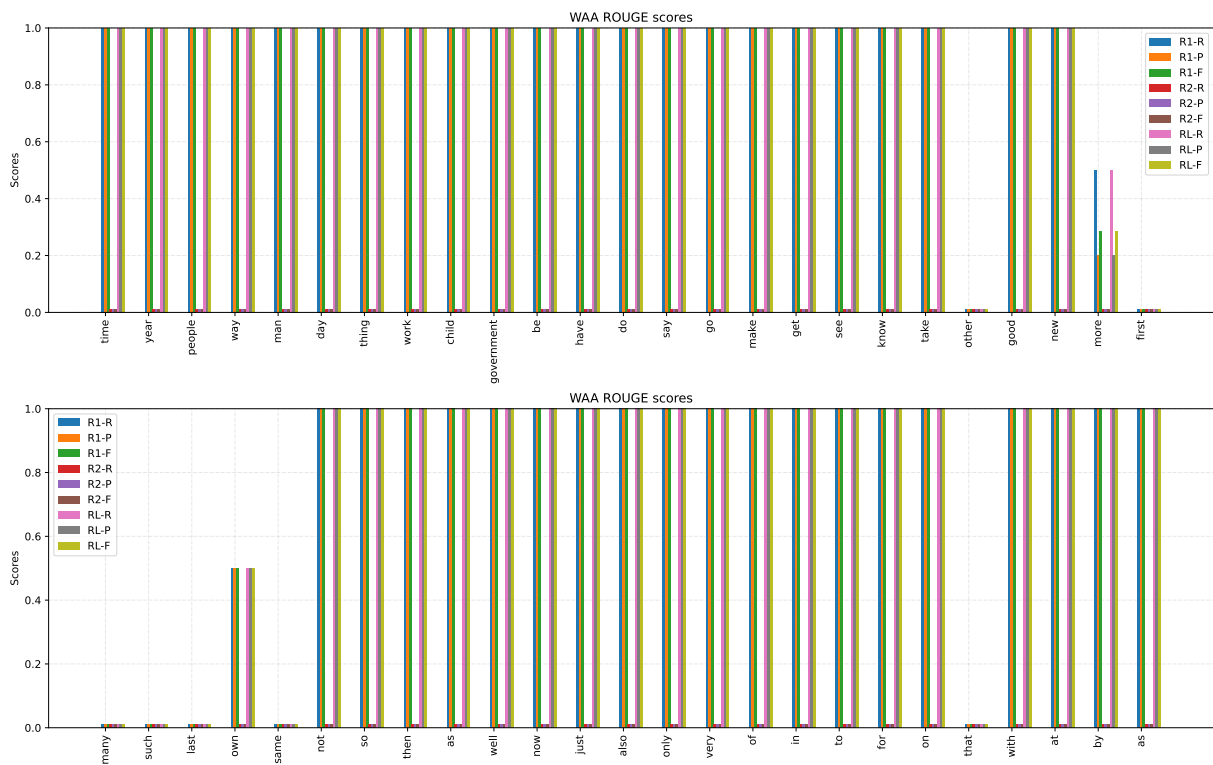


Figure 10: ROUGE scores in the WAA category. The majority of the data in this category comprises of one-word units. As a result, the majority of our lemmas reach a perfect score of 1.0 for R1 and RL. R2 scores are not applicable. However, some of our lemmas receive a score of 0.0 in R1 and RL. This is due to the fact that ChatGPT and OALD provide different part-of-speech information for these lemmas.

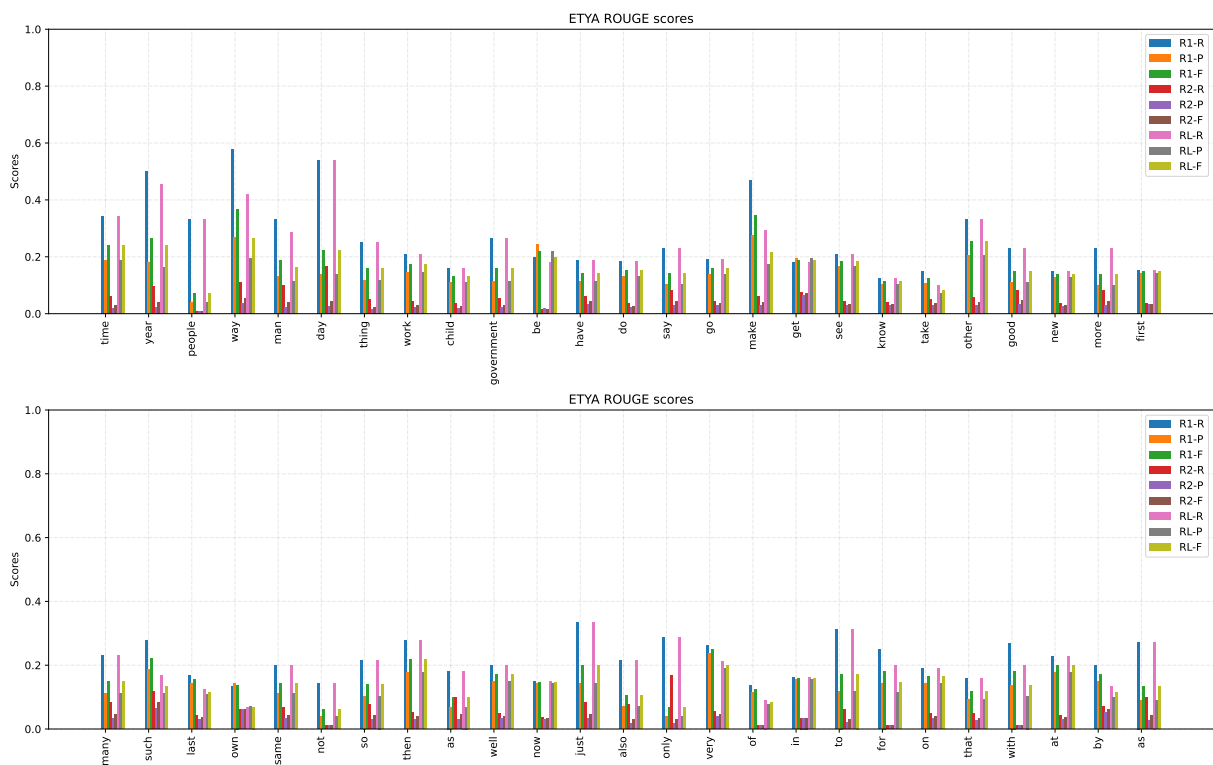


Figure 11: ROUGE scores in the ETYA category. The data in this category contains the highest number of word units, but the bar charts for all lemmas show scores of no more than 0.6, with the majority scoring less than 0.2. This suggests a significant difference between the etymological data in the reference and candidate texts.

Thesaurus of Modern Slovene 2.0

Špela Arhar Holdt^{1,2}, Polona Gantar¹, Iztok Kosem^{1,3}, Eva

Pori¹, Marko Robnik-Šikonja², Simon Krek^{1,3}

¹ Faculty of Arts, University of Ljubljana, Aškerčeva ulica 2, 1000 Ljubljana, Slovenia

² Faculty of Computer and Information Science, University of Ljubljana, Večna pot 113, 1000 Ljubljana, Slovenia

³ Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia

E-mail: Spela.ArharHoldt@ff.uni-lj.si, Apolonija.Gantar@ff.uni-lj.si, Iztok.Kosem@ff.uni-lj.si, Eva.Pori@ff.uni-lj.si, Marko.RobnikSikonja@fri.uni-lj.si, Simon.Krek@ijs.si

Abstract

This paper describes the improvement of the Thesaurus of Modern Slovene from version 1.0 to 2.0. The Thesaurus is a digitally-born, automatically created resource that provides fast access to open data on modern language use and is gradually improved through editing and user participation. The initial version 1.0 lacked metadata, dictionary labels, and semantic information, but was well-received by users. However, a user study identified priorities for improvement, which were addressed in the upgrade funded by the Slovenian Ministry of Culture in 2021-2022. The project aimed to upgrade the dictionary interface design, establish protocols for labeling negative vocabulary, pilot the automatic extraction of antonyms, and supplement the dictionary with semantic indicators for 2,000 entries. This paper presents the upgraded Thesaurus, the methodology for each enhancement, and the challenges and solutions of lexicographic work. The Thesaurus serves as an example of lexical data reuse, interconnectivity, and user involvement, with insights useful for other language communities pursuing similar initiatives.

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

1. Introduction

Thesaurus of Modern Slovene, first published in 2018,¹ introduced the concept of a responsive dictionary: a digitally-born, automatically created language resource that provides fast access to open data on modern language use and is gradually improved through editing, which involves both lexicographic work and user participation (Arhar Holdt et al., 2018: 404). The most defining characteristic of the responsive model is its ability to quickly and flexibly respond to both language change and the feedback provided by the community: in the case of the Thesaurus, users can contribute by

¹ Thesaurus of Modern Slovene 1.0 is available in the interface at <https://viri.cjvt.si/kolokacije/eng/> and as a database at <http://hdl.handle.net/11356/1166>. Thesaurus 2.0 is available in beta version: <https://viri.cjvt.si/sopomenke-beta/slv/>.

adding suggestions of missing synonyms and by up- or downvoting existing synonym candidates.

Thesaurus of Modern Slovene 1.0 consists of 105,473 keywords and 368,117 synonyms. It was automatically generated using pre-existing resources: the Oxford@-DZS Comprehensive English-Slovenian Dictionary and the Gigafida reference corpus of written Slovene (Logar et al., 2012). The extraction of data relied on the co-occurrence of words in translation strings of the Oxford-DZS Dictionary. The next step utilized a method that combined balanced co-occurrence graphs and the Personal PageRank algorithm to divide synonyms into subgroups and rank them based on their degree of semantic relatedness (Krek et al., 2017).

The data published in Thesaurus 1.0 was not lexicographically post-processed. The entries and synonym candidates were presented in a form of lemmata (without part-of-speech or other metadata that would help disambiguate between forms), semantic descriptions were replaced by automatically obtained semantic clusters, and the data also lacked dictionary labels, apart from domain ones. Despite these limitations, the community found the new resource and the concept of a responsive dictionary useful (Arhar Holdt 2020: 470), and statistics show the consistent widespread use of the Thesaurus ever since it was published.

However, continuous development is an integral part of the responsive model, and the aforementioned user study also identified priorities for the first upgrade. The upgrade was funded by the Slovenian Ministry of Culture in 2021–2022 and included upgrading the dictionary interface design; ensuring transparent editorial protocols for evaluating user suggestions; piloting the automatic extraction of antonyms and facilitating crowdsourcing of antonyms through the dictionary interface; adding dictionary labels for extremely offensive (hateful) and vulgar vocabulary and allowing users to also provide dictionary labels when contributing synonyms and antonyms; and finally, supplementing the dictionary database with the description of sense distribution including short definitions of senses known as semantic indicators for 2,000 entries. In the following sections, we describe the database and interface improvements and conclude with plans for future dictionary development.

2. Database Improvements

2.1 Data Cleaning and Import into the Digital Dictionary Database

The first step of the project was to import the data from the Thesaurus of Modern Slovene database into the Slovene Digital Dictionary Database (Kosem et al., 2021a), which would allow for the interlinking, easier editing, and optimized reusability of lexical information. To achieve this goal, we had to undertake a series of technical and editing procedures on the Thesaurus data.

Firstly, we extracted synonym pairs containing domain dictionary labels, which were then reviewed, corrected, or upgraded to correspond to the labeling system used in the Digital Dictionary Database. Secondly, we used the results of previously conducted crowdsourcing campaigns aimed at removing noise from the database (Čibej & Arhar Holdt, 2019). We removed 8,878 problematic entries, such as noisy and redundant multi-word units (e.g. *zeleni pas – zeleni pas med vozišči*, ‘green belt – green belt between the lanes’; *akrobat na vrvi – plesalka na vrvi*, ‘a male tightrope acrobat – a female tightrope dancer’).

Lastly, we addressed the issue of headwords and synonyms not containing part-of-speech information in Thesaurus 1.0. This led to homonymous headwords with synonyms placed together. We disambiguated such cases and semantically separated the synonyms (4,560 units in total) accordingly. For instance, the adverb *blago – zmerno, nežno, rahlo* (‘mildly – moderately, gently, slightly’) vs. the noun *blago – tekstil, material* (‘fabric – textile, material’).

2.2 Semantic Analysis and Sense Division

We selected 2,000 headwords by merging the headword lists from the Thesaurus of Modern Slovene 1.0 (Krek et al., 2018), the Collocation Dictionary of Modern Slovene 1.0 (Kosem et al., 2019), and the Comprehensive Slovenian-Hungarian Dictionary 1.0 (Kosem et al., 2021b), considering relevant parameters such as part-of-speech categories, single or multiple senses, dictionary labels, and potentially offensive vocabulary. We equipped the headwords with semantic indicators,² primarily sourced from the Comprehensive Slovenian-Hungarian Dictionary and supplemented with newly prepared indicators. The synonym candidates for these headwords were then attributed to the corresponding senses. Our original plan was for lexicographers to use the localized and adapted version of Lexonomy,³ but the testing phase revealed slow data classification and challenging workflow management. As a result, we revised the process by exporting data from the dictionary database to a tabular form (Google Sheets), editing and validating the data, and finally importing it back into the database.

In the process of our work, we developed guidelines for the classification of synonym candidates under the appropriate senses. The guidelines provided relevant information

² Semantic indicator is one of the three segments of semantic information included in the CJVT dictionary resources. Along with labels and explanations, the semantic indicator aims to define the meaning of a word concisely and clearly in relation to its other meanings. The primary purpose of the indicators is to create a sense menu, a feature introduced in CJVT dictionary resources such as Collocations 1.0 (<https://viri.cjvt.si/kolokacije/eng/>) and Comprehensive Slovenian-Hungarian Dictionary 1.0 (<https://viri.cjvt.si/slovensko-madzarski/eng/>).

³ <https://lexonomy.cjvt.si/>

for the classification process, including the presentation of the data and the main steps for classifying synonymous material and verifying synonymy.

To ensure the accuracy of the classification, we checked the usage of words in various corpora such as the reference corpus of written Slovene Gigafida 2.0 (Krek et al., 2020), the Monitor corpus of Slovene Trendi (Kosem et al., 2022), the JANES corpus of Slovene user-generated content version 1.0 (Fišer, 2020), and KAS corpus of academic Slovene 2.0 (Žagar et al., 2021). In some difficult cases, we also investigated collocate overlap among the synonyms to help determine which senses they should be attributed to. For this task, we used the Collocation Dictionary of Modern Slovene 1.0, Thesaurus of Modern Slovene 1.0, Gigafida 2.0, and the Sketch Engine tool's Sketch Diff function (Kilgarriff et al., 2014) (see Section 2.5). An example of a headword with distributed synonyms is presented in Table 1.

Headword	Senses ⁴	Synonyms
hiteti (glagol)	[1: pri dejavnosti]	pohiteti, brzeti, drveti, dirjati
	[2: o premikanju]	brzeti, drveti, leteti, dirjati, teči, divjati, hitro hoditi, rezati jo, drobencljati, drobencati, planiti, vrveti, dreti, ubirati jo, poditi se, sukati se, švigniti
	[3: minevati]	brzeti, drveti, leteti, hitro minevati, teči

Table 1: Classification of synonym candidates according to the senses of the headword *hiteti* ('to hurry').

If a word's usage could not be confirmed in our resources, we did not consider it a synonym. False candidates appeared in our data due to the methodology we used for the creation of the Thesaurus, where we exported synonym candidates from the Oxford®-DZS Comprehensive English-Slovenian Dictionary (see Section 1). For example, for the headword *brat* ('a brother'), we included synonyms, such as [1: sorodnik, 'a relative']: *bratec* ('a little brother'); [2: pripadnik skupine, 'a group member']: *prijatelj* ('a friend'); but not *kolega*, *sodelavec* ('colleague, coworker'), as the word *brat* is not used in this sense in Slovene.

⁴ The presented semantic indicators and synonyms for the verb 'to hurry' are roughly equivalent to the English concepts of [1: activity]: to hasten, to accelerate, to race, to rush [2: movement]: to accelerate, to rush, to fly, to dash, to run, to rampage, to walk quickly, to stride, to scurry, to scamper, to leap, to throng, to swarn, to run around, to twist, to dart, and [3: time]: to accelerate, to rush, to fly, to pass quickly, to run.

Anticipatedly, we encountered numerous borderline candidates, and in such cases, we chose to prioritize inclusion over exclusion. Our decision was based on the expectation that future projects would involve further data cleaning and editing. Additionally, our responsive dictionary concept aims to provide users with as much data as possible, facilitating a vast choice of options. Hence, we also included certain candidates with semantic proximity or similarity, for example, *juha – enoločnica, obara* (‘soup – stew, casserole’) or *kavč – divan, zofa, postelja* (‘couch – loveseat, sofa, bed’).

During the classification process, we identified opportunities to propose changes to the existing semantic classification. This included suggesting an additional sense, dividing an existing sense, combining two senses, or changing the semantic indicator. In some cases, the semantic indicators themselves contained one of the synonyms, resulting in repetition within the string, e.g. *besneti* = ‘to be furious’ [1: *jeziti se*, ‘to get angry’]: *peniti se, divjati, jeziti se, kipeti, vreti* (‘to foam, to rage, to get angry, to boil, to seethe’). We marked these cases, which enabled us to review them after the project and enhance the creation of semantic indicators in the future.

2.3 Dictionary Labels

In Thesaurus 1.0, headwords and their synonyms lacked any explicit information on usage, stylistic and pragmatic value, except for a limited set of domain labels. A survey conducted with 671 respondents revealed that more than one-third of users (37%) found the absence of dictionary labels problematic (Arhar Holdt 2020: 472). There were two main issues with the absence of labels. Firstly, automatically generated headwords and synonym candidates appeared without labels or usage warnings even in highly problematic cases, such as the word *buzi* with synonyms *peder, buzerant, toplovodar, homič*, all of which are derogatory expressions for ‘a gay man’. Secondly, users added marked vocabulary as legitimate synonyms, such as for the marked word *južnjak* (‘southerner’), where users suggested similarly marked words like *jugovič, južni brat, jugič, trenirkar*. In some of these cases, users even added a note on usage next to the proposed synonym candidate. Therefore, incorporating a labeling system into the dictionary became a top priority, and it also made sense to upgrade the interface and provide users with the option to label their suggestions in a more systematic way.

We based our labeling system on the dictionary style guide used for developing lexicographic resources in the Digital Dictionary Database at the CJVT, for instance, the Comprehensive Slovenian-Hungarian Dictionary (Kosem et al., 2021b). Negative vocabulary is labeled with three distinctive labels (Table 2) used for hate speech elements (labeled as *hateful*), elements of rudeness and offensiveness (*coarse*), and elements of negative evaluation or connotation (*expresses a negative attitude*). Each label is depicted in the interface with an icon and accompanied by an explanation of the potential impact the use of the labeled word can have (see Section 3).

The labels were assigned manually to the headwords and their synonyms during the lexicographic process, taking into account the distribution of senses and their descriptions, as described in Section 2.2. Arhar Holdt et al. (2023) provide further examples and a detailed explanation of the labeling choices.




Label	Icon	Explanation
hateful		This word can be used to express a hostile or intolerant attitude towards an individual or social group.
coarse		This word can seem coarse or inappropriate to many language users due to social and moral norms. Using the word can make people feel uncomfortable, upset, or offended.
expresses a negative attitude		This word may not be neutral. The word can be used to ridicule, express disapproval, or criticize certain characteristics of individuals, objects, or actions.

Table 2: Labels for negative vocabulary, icons in Thesaurus 2.0, and their explanations.

2.4 Automatic Extraction and Selection of Antonyms

We prepared a prototype sense-antonym extraction methodology based on machine learning using word embeddings and large pre-trained language models. We decided to test using word-sense information in order to prepare antonyms separately for each sense. We started with candidate antonyms without sense information obtained from lexical sources. Our methodology is composed of three approaches. In the first approach, we construct a sense-antonym dataset and cluster a set of contextual embeddings (for words in contexts) to produce sense-clusters; we then assign candidate sense-antonyms to their nearest clusters. In the second approach, meant to obtain antonyms without prespecified antonym-candidate pairs, we first fine-tune a large language model using a dataset of antonyms in context to predict if two words in a context (i.e. a sentence), are antonyms. The third approach is a traditional lexical approach based on a dictionary and WordNet.

For the clustering-based approach, we first constructed a dataset of sense-antonyms used in sentences. For that purpose, we used 2,852 antonym candidates without sense information and extracted examples of their use from the Comprehensive Slovenian-Hungarian Dictionary (Kosem et al., 2021b). We formed the contextual embeddings for each candidate word by concatenating the last four layers of the CroSloEngual BERT model (Ulčar & Robnik-Šikonja, 2020) as recommended by Devlin et al. (2019). To produce the pairs of sense-antonyms on the resulting 3072-dimensional vectors, we used

the k-nearest neighbor method which, for a given word, found the nearest sense-cluster. A portion of candidate pairs was manually checked.

For the classification approach, we first constructed a dataset of antonyms in context, similarly to the process described above. The dataset was split into a training, validation, and test set. We fine-tuned the CroSloEngual BERT language model on the training set. The model achieved 90% precision and 61% recall on the test set. We published the datasets, containing a total of 79,229 records with information about word pairs in specific senses, as an open-access collection on the CLARIN.SI repository (Pegan et al., 2022). Only a fraction of this dataset was manually checked.

To complement the machine learning approaches, we tried a lexical approach producing 4,734 antonym candidates using a bilingual dictionary (English-Slovene) and antonym information available in English Wordnet. Each candidate pair was accompanied by its part-of-speech tags.

All three approaches showed promising results and will undergo further evaluation. In Thesaurus 2.0, we have included only the 2,544 antonym pairs that were manually checked.

2.5 Comparing Collocates of Synonyms

One of the more innovative features of the first version of the Thesaurus of Modern Slovene was the comparison of collocations of the headword and its synonyms. The comparison was made by using the Sketch Diff function in the Sketch Engine tool (Kilgarriff et al., 2014), with specific parameters set for extraction, such as minimum frequency and grammatical relations used for different word classes. The comparison was available only for single-word headwords and single-word synonyms, and only when both compared items were frequent enough to produce the Sketch Diff output. The collocation comparison contained three different sections: joint (collocates typical for both the headword and the synonym), and two individual (collocates used with the headword or the synonym only). Each section included up to 20 collocates; four columns with up to five collocates each. In general, one column per grammatical relation was used, and only the most productive relations were presented in two columns; for example, for the comparison of adjectives, the relation *adjective + noun* was allocated two columns (more examples in Arhar Holdt et al. 2018: 403, 408).

In Thesaurus 2.0, both the data and its presentation received an upgrade. This is a direct result of two factors: a) the change in the methodology used for automatic collocation extraction, and b) the findings of the evaluation study of the informative nature of the collocates provided by the selected grammatical relations. The main change in the methodology of collocation extraction was moving it from POS-tagged to parsed corpus data, and consequently considerably improving the reliability of the results (Krek et al., 2021). This also meant that additional syntactic structures could

be considered for the inclusion into collocational comparison. Relatedly, a study (Arhar Holdt, 2021) was conducted on the informative nature and reliability of the collocations in grammatical relations used in version 1.0. The study did not merely provide an evaluation of existing grammatical relations but also included recommendations on which ones should be removed from the comparison feature, and which new ones should replace them. The two most important changes occurred at the level of (a) nouns, where we replaced the pattern *noun + preposition + noun* (e.g. *program za prihodnost*, ‘program for the future’) with a more productive *noun + noun in genitive* (e. g. *izvedba programa*, ‘the implementation of the program’), and (b) adjectives, where we replaced *adjective + preposition + noun* (e.g. *pozitiven za gospodarstvo*, ‘positive for the economy’) with *adjective + AND + adjective* (e.g. *pozitiven in optimističen*, ‘positive and optimistic’).

We have also made a change in the way we obtain collocations for the three sections of the comparison. As we wanted to focus more on the typicality of the collocations, the sections became “joint”, “more typically used with the headword”, and “more typically used with the synonym”. In this way, we excluded from the comparison the less typical (and often more infrequent) collocates, which were causing some problems in version 1.0. For instance, a collocation that was highly characteristic of one item, but also appeared with the other item, could only be included in the joint list. However, in most cases, it would be excluded from the final list due to numerous other collocates that showed a similar level of typicality of usage with both items. The visualization of collocates in the Thesaurus 2.0 is presented in Section 3.

2.6 Editorial Protocols for Evaluating User-suggested Synonyms

As previously noted, users can suggest synonyms and, as of version 2.0, antonyms to the Thesaurus of Modern Slovene. This feature enables dictionary users to actively participate in the development of openly available language infrastructure in a democratic manner. They can contribute to the expansion and refinement of the Thesaurus, helping to make it more comprehensive.

To encourage participation, user suggestions are displayed in the interface immediately after they are submitted, along with the user’s chosen username. At this stage, the suggestions are not subject to editorial evaluation but are visually distinguished from the rest of the synonyms and not automatically added to the openly accessible database. Instead, the evaluation process occurs during dictionary upgrades, where lexicographers carefully consider the user suggestions according to editorial guidelines. This approach ensures that user input is taken into account while also maintaining the consistency and quality of the database.

To develop the guidelines, 972 user-suggested synonym pairs were evaluated by a team of six lexicographers and classified as suitable, unsuitable, or conditionally suitable for inclusion in the database. The evaluation was complemented by a larger study that

involved selected user groups (e.g. teachers, translators, and proofreaders) performing the same task (Gapsa, 2023). This study provided valuable insight into the preferences of users compared to lexicographers and the differences between the conditions for synonym selection reported by different groups (Gapsa & Arhar Holdt, 2023).

Based on our analysis, we have developed a protocol for evaluating user suggestions. Firstly, we check whether the proposed word or phrase appears in authentic language use by consulting the corpora already mentioned in Section 2.2. Secondly, we assess whether the suggestion is appropriately categorized under the relevant headword, which we determine by studying the use of the word in resources. Thirdly, we consider any proposals for dictionary labels made by users and add the appropriate label if needed. Finally, we consider feedback on the suggestion from other users, based on their upvotes and downvotes. If there is any uncertainty about the suitability of a user-suggested synonym for inclusion in the database, we err on the side of caution and do not include it. It is worth noting that even if a proposal is not included in the database, it remains accessible via the dictionary interface. We would only remove entries from the interface in rare cases when they are deemed malicious (see also Section 2.3).

Currently, there are 60,976 user-suggested synonyms awaiting evaluation for potential inclusion in the Thesaurus. The implementation of the editorial guidelines will be part of the next edition, and the guidelines will be continuously improved in the process. In the meantime, we have upgraded the interface in version 2.0 to enable users to contribute dictionary labels to proposed synonyms and antonyms. Our evaluation process has shown that many user suggestions include regionally specific, slang, or jargon terms that could benefit from a label, both to facilitate the editorial process and to assist other users in understanding the terms.

3. Interface Improvements

In addition to enhancing the content of the Thesaurus, we also placed a significant emphasis on developing the user interface. The designer has created a library of redesigned interface elements, which has enabled all elements within CJVT dictionary resources to share a consistent visual appearance and logical structure, including colors, icons, typography, and element formatting such as search, toggle, share, and user engagement. Additionally, the implementation of a responsive font size provides improved accessibility for users who may need to adjust zoom levels for different reasons. Working closely with the team, the designer has also introduced a new design for the dictionary headers, footer, and “About” section, as well as interface features not present in version 1.0, including antonyms, sense-separated entries, icons for negative vocabulary, and user mechanisms for adding dictionary labels and including suggested synonyms and antonyms under the relevant sense.

Thesaurus of Modern Slovene 2.0 offers two different entry layouts. Figure 1⁵ depicts the layout of the automatically generated entry, while Figure 2 shows an entry with sense division and manually arranged synonyms (Section 2.2). In both layouts, new metadata has been included: the part-of-speech label, an indicator of the headword's frequency in the reference corpus Gigafida 2.0, and an indicator of the entry layout type - either automatically generated or sense-divided. Both layouts also feature a tab for antonyms, although only a limited number of headwords currently have antonyms (Section 2.4).

The layout of the automatically generated entry for the word *jedrnat* ('concise') is depicted in Figure 1. The extracted synonym candidates are presented in two sections: core synonyms that are semantically closer are on a white background, while less related near synonyms are on a grey background. The entry also includes a section for user-suggested synonyms, with the option to add more suggestions. The image shows one such user suggestion: *kratek in jedrnat* ('short and concise'), which was added in version 1.0. The Antonym section currently lists one antonym, *dolgovezen* ('verbose'). Finally, the section for user-added antonyms is currently empty.

⁵ Currently, the interface is only available in Slovene, but a translation to English is in the works and will be available before the official launch.

The screenshot shows the 'jedrnat' web application interface. At the top, there is a red header with the 'cjvt sopomenke 2.0' logo, a search bar containing 'jedrnat', and navigation links for 'O viru', 'Skupnost', and 'Slovenščina'. Below the header, the word 'jedrnat' is displayed with a subtitle 'pridevnik / pogostost: ●●○○○ / strojno pripravljeno geslo / 2022-11-15'. The main content area is divided into two sections: 'Sopomenke' and 'Protipomenke'. The 'Sopomenke' section contains a grid of related terms: 'kratek', 'vsebinsko poln', 'zgoščen', 'strnjen', 'kompakten', 'lakoničen', 'koncizen', 'pregnanten', 'majhen', 'hiter', 'odločen', and 'bežen', 'redkobeseden', 'odrezav'. Below this grid is a section for 'Uporabniško dodane sopomenke' with a '+ Prispevajte svoj predlog' button. A 'v1.0' label is present, and an example entry 'kratek in jedrnat' by Tatjana J. is shown. Statistics for 'jedrnat' are provided: 'Število jedrnih sopomenk: 8. | Število bližnjih sopomenk: 6. | Število uporabniško dodanih sopomenk: 1.'. The 'Protipomenke' section contains the entry 'dolgovezen' and another '+ Prispevajte svoj predlog' button. Statistics for 'protipomenke' are: 'Število protipomenk: 1. | Število uporabniško dodanih protipomenk: 0.'.

Figure 1: The layout of the automatically generated entry *jedrnat* ('concise').

Figure 2 presents the layout of the sense-divided entry for the word *baba* ('a broad'). This word can be used to refer to a woman and express either a negative or a positive attitude, or to refer to a man and express a negative attitude in the sense of 'a coward'. We chose this example as it includes both the icon for a coarse word (*lajdra*, 'a slut') and a hateful word (*peder*, 'a faggot'). Clicking on the icon opens a longer explanation of the potential impact that the use of the labeled word can have (see Table 2).

As mentioned in Sections 2.3 and 2.6, we have upgraded the user-suggestion protocol to allow users to add dictionary labels to their proposed words or phrases. In the sense-divided entries, they can also attribute the suggestion to a corresponding sense (Figure 3). The default option is for the user's suggestion to be *without label*, while other options are available in the drop-down menu. In Thesaurus 2.0, labels for *hateful*, *coarse*, and *expresses a negative attitude* are available upon clicking, along with a box where users can type in any other possible label. The meaning and usage of these labels are explained and illustrated with examples, which will help achieve a certain level of consistency in user labeling (information is available upon clicking the icon (i) in Figure

3). It is expected that users will interpret and use these labels differently than lexicographers would in some cases. User suggestions will thus be valuable not only for supplementing the open-access dictionary database but also for the analyses of the perception of the labeling system.

The screenshot shows the dictionary entry for 'baba' on the website 'cjvt.sopomenke.si'. The entry is divided into three sections based on gender and relationship:

- 1 ženska** | lahko izraža negativen odnos: babura, mačka, tečnoba, coprnica, ta stara, lajdra, večša, babnica.
- 2 ženska** | lahko izraža pozitiven odnos: mrha, ta prava babnica.
- 3 o moških** | izraža negativen odnos: mevža, peder, reva, boječka, šleva.

Below the main entry, there is a section for 'Uporabniško dodane sopomenke' (User-added synonyms) with a '+ Prispeljite svoj predlog' button. It shows a list of user suggestions for the word 'baba' with their respective usernames and the gender label 'ženska'.

Figure 2: The layout of the sense-divided entry *baba* ('a broad').

The screenshot shows the 'Uporabniško dodane sopomenke' (User-added synonyms) form for the word 'brat'. The form includes fields for 'Uporabnik' (test-username), 'Sopomenka' (test-suggestion), and 'Slovarska oznaka' (Brez oznake). A dropdown menu for 'Izberi pomen' is set to '1 sorodnik'. The form also displays a list of user suggestions for the word 'brat' with their respective usernames and the gender label 'sorojenec'.

Figure 3: Adding potentially labeled synonyms for the noun *brat* ('a brother').

From the initial layout, users can click selected synonyms to open the collocate comparison view. Figure 4 presents this view for the pair *program* and *plan* (‘a program - a plan’). As mentioned in Section 2.5, firstly the collocates that appear with both words are presented, followed by collocates that typically occur with only one of the words. For example, *razvojni načrt* and *razvojni program* (‘development plan, development program’) are both typical collocations, while *kulturni*, *nacionalni*, *študijski* tend to collocate with *program* and *prostorski*, *poslovni*, *lokacijski* with *načrt* (‘cultural, national, study program’ and ‘spatial, business, location plan’).

The screenshot shows the 'program' collocate comparison view. The interface includes a search bar with 'program', a sidebar with a list of synonyms, and three main panels of collocates.

Sopomenke

- plan
- plani
- plan
- načrt
- plani
- načrti
- razpored
- shema
- blagovna skupina
- ekonomija
- linija
- ekonomija
- nastop
- glasba
- spored
- urnik
- platforma
- računalništvo
- volilna platforma
- gledališki list
- manifest
- kanal
- programska oprema
- računalništvo
- softver
- računalništvo
- software
- računalništvo
- rdeči karton
- šport
- rumeni karton
- šport

Besede, s katerimi se pojavljata tako program kot načrt

razvojni	okviri	imeti v	pripraviti
učni	izvajanje	biti na	predstaviti
sanacijski	priprava	potekati po	sprejeti
Operativni	izdelava	vključiti v	pripravljati
investicijski	osnutek	uvrstiti v	izdelati

Besede, s katerimi se pojavlja predvsem program

kulturni	del	sodelovati v	izvajati
nacionalni	vodja	nastopiti v	ponujati
študijski	urednik	iti za	oblikovati
Izobraževalni	razvoj	poskrbeti za	spremljati
poseben	financiranje	imeti na	ponuditi

Besede, s katerimi se pojavlja predvsem načrt

prostorski	sprememba	biti v	imeti
poslovni	uresničitev	iti po	prekrižati
lokacijski	sprejetje	teči po	narediti
finančni	podlaga	zgraditi po	spremeniti
akcijski	dopolnitev	graditi po	kovati

Figure 4: Collocate comparison for *program* and *plan* (‘a program - a plan’).

4. Conclusion and Future Work

In this paper, we presented an upgraded version of the Thesaurus of Modern Slovene, which was developed to address the lack of openly available synonym data for modern Slovene. Our work serves as a benchmark for other languages that face similar issues, demonstrating the importance of data reusability and user involvement in language infrastructure development. To address contemporary needs, such as the desire of dictionary users to participate in the development of the language infrastructure, we integrated machine processes and user suggestions into our workflows. This integration

allows for the incorporation of user feedback, ensuring that openly accessible language data is readily available.

The upgraded version of the Thesaurus addressed some of the most significant shortcomings of the previous version. While the project had limitations in scope and not all of the database could be manually edited, Thesaurus 2.0 sets a clear direction for future development. The newly established guidelines for preparing sense divisions, labeling negative vocabulary, and evaluating user suggestions provide a solid foundation for the expansion of the database. In version 3.0, our attention will remain focused on the automatic extraction of antonyms, which has displayed promising results and requires further evaluation and more extensive implementation. Additionally, more detailed work is planned for the selection and visualization of collocations. With the development of the Collocation Dictionary of Modern Slovene (Kosem et al., 2018), the data is improving, and new possibilities for utilization will soon be available.

5. Acknowledgements

The authors acknowledge that the project Empirical foundations for digitally-supported development of writing skills (J7-3159) and the programmes Language Resources and Technologies for Slovene (P6-0411) and Slovene Language – Basic, Contrastive, and Applied Studies (P6-0215) were financially supported by the Slovenian Research Agency. The project Upgrading fundamental dictionary resources and databases of CJVT UL was funded by the Ministry of Culture of the Republic of Slovenia in the period 2021–2022.

6. References

- Arhar Holdt, Š. (2023, in press). Negativno zaznamovano besedišče v Slovarju sopomenk sodobne slovenščine 2.0. *Slovenščina 2.0*.
- Arhar Holdt, Š. (2021). Kolokacije v Slovarju sopomenk sodobne slovenščine: evalvacija podatkov in predlog za izboljšavo. In I. Kosem (ed.) *Kolokacije v slovenščini*. Ljubljana: Znanstvena založba Filozofske fakultete, pp. 269–296. Available at: <https://ebooks.uni-lj.si/zalozbaul/catalog/view/318/465/6967-1>
- Arhar Holdt, Š. (2020). How users responded to a responsive dictionary: the case of the Thesaurus of Modern Slovene. *Rasprave Instituta za hrvatski jezik i jezikoslovlje*, 46(2), pp. 465–482. doi: 10.31724/rihjj.46.2.1
- Arhar Holdt, Š., Čibej, J., Dobrovoljc, K., Gantar, P., Gorjanc, V., Klemenc, B., Kosem, I., Krek, S., Laskowski, C. & Robnik-Šikonja, M. (2018b). 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*. Ljubljana: Znanstvena založba Filozofske fakultete, pp. 401–410. Available at: <https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/118/211/3000-1>
- Devlin, J., Chang, M.W., 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*. Minneapolis, Minnesota. Association for Computational Linguistics, pp. 4171–4186. Available at: <https://aclanthology.org/N19-1.pdf>
- Čibej, J. & Arhar Holdt, Š. (2019). Repel the syntruders! A crowdsourcing cleanup of the thesaurus of modern Slovene. A crowdsourcing cleanup of the thesaurus of modern Slovene. In I. Kosem et al. (eds.) *Proceedings of the eLex 2019 conference, Electronic lexicography in the 21st century: Smart lexicography*. Sintra, Portugal. Brno: Lexical Computing, pp. 338–356. Available at: https://elex.link/elex2019/wp-content/uploads/2019/09/eLex_2019_19.pdf
- Fišer, D., Ljubešić, N. & Erjavec, T. (2020). The Janes project: language resources and tools for Slovene user generated content. *Language resources and evaluation*, 54(1), pp. 223–246. doi: 10.1007/s10579-018-9425-z
- Gapsa, M. (2023, preprint). “But why??” Evaluation of user-suggested synonyms in the Thesaurus of Modern Slovene. Research Square. doi: 10.21203/rs.3.rs-2775161/v1
- Gapsa, M. & Arhar Holdt, Š. (2023). How Lexicographers Evaluate User Contributions in The Thesaurus of Modern Slovene in Comparison to Dictionary Users. In *Electronic lexicography in the 21st century. Invisible Lexicography. Proceedings of the eLex 2023 conference*. Brno, Czech Republic.
- 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(1), pp. 7–36. doi: 10.1007/s40607-014-0009-9
- Kosem, I., Čibej, J., Dobrovoljc, K., Erjavec, T., Ljubešić, N., Ponikvar, P., Šinkec, M. & Krek, S. (2022). *Monitor corpus of Slovene Trendi 2022-10*, Slovenian language resource repository CLARIN.SI, <http://hdl.handle.net/11356/1681>.
- Kosem, I., Krek, S. & Gantar, P. (2021a). Semantic data should no longer exist in isolation: the Digital Dictionary Database of Slovenian. In Z. Gavriilidou, L. Mitits, S. Kiosses (eds.), *EURALEX XIX, Congress of the European Association for Lexicography, Lexicography for inclusion*. Komotini: Democritus University of Thrace, pp. 81–83. Available at: https://euralex2020.gr/wp-content/uploads/2021/09/EURALEX2020_BookOfAbstracts-Preview-1.pdf
- Kosem, I., Bálint Čeh, J., Ponikvar, P., Zaranšek, P., Kamenšek, U., Koša, P., Gróf, A., Böröcz, N., Harmat Császár, J., Szijártó, I., Šantak, B., Gantar, P., Krek, S., Roblek, R., Zgaga, K., Logar, U., Pori, E., Arhar Holdt, Š. & Gorjanc, V. (2021b). *Comprehensive Slovenian-Hungarian Dictionary 1.0*, Slovenian language resource repository, CLARIN.SI, <http://hdl.handle.net/11356/1453>.
- Kosem, I., Gantar, P., Krek, S., Arhar Holdt, Š., Čibej, J., Laskowski, C., Pori, E., Klemenc, B., Dobrovoljc, K., Gorjanc, V. & Ljubešić, N. (2019). *Collocations Dictionary of Modern Slovene KSSS 1.0*, Slovenian language resource repository CLARIN.SI, <http://hdl.handle.net/11356/1250>.
- Kosem, I., Krek, S., Gantar, P., Arhar Holdt, Š., Čibej, J., Laskowski, C. A. et al. (2018). Collocations dictionary of modern Slovene. In J. Čibej et al. (eds.)

- Proceedings of the 18th EURALEX International Congress: lexicography in global contexts*. Ljubljana: Ljubljana University Press, Faculty of Arts, pp. 989–997. Available at: <https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/118/211/3000-1>
- Krek, S., Gantar, P., Kosem, I. & Dobrovoljc, K. (2021). Opis modela za pridobivanje in strukturiranje kolokacijskih podatkov iz korpusa. In Š. Arhar Holdt (ed.) *Nova slovnica sodobne standardne slovenščine*. Ljubljana: Znanstvena založba Filozofske fakultete, pp. 160–194. Available at: <https://ebooks.uni-lj.si/zalozbaul/catalog/download/325/477/7320-1?inline=1>
- Krek, S., Arhar Holdt, Š., Erjavec, T., Čibej, J., Repar, A., Gantar, P., Ljubešič, N., Kosem, I., Dobrovoljc, K. (2020). Gigafida 2.0: the reference corpus of written standard Slovene. In N. Calzolari (ed.) *LREC 2020: Twelfth International Conference on Language Resources and Evaluation: May 11-16, 2020, Palais du Pharo, Marseille, France*. Paris, ELRA - European Language Resources Association, pp. 3340-3345. Available at: <http://www.lrec-conf.org/proceedings/lrec2020/LREC-2020.pdf>
- Krek, S., Laskowski, C., Robnik Šikonja, M., Kosem, I., Arhar Holdt, Š., Gantar, P., Čibej, J., Gorjanc, V., Klemenc, B. & Dobrovoljc, K. (2018). *Thesaurus of Modern Slovene 1.0*, Slovenian language resource repository CLARIN.SI, <http://hdl.handle.net/11356/1166>.
- Krek, S., Laskowski, C. & Robnik Šikonja, M. (2017). From translation equivalents to synonyms: creation of a Slovene thesaurus using word co-occurrence network analysis. In I. Kosem et al. (eds.) *Proceedings of eLex 2017: Lexicography from Scratch*. Leiden, Netherlands, pp. 93–109. Available at: <https://elex.link/elex2017/wp-content/uploads/2017/09/paper05.pdf>.
- Logar Berginc, N., Grčar, M., Brakus, M., Erjavec, T., Arhar Holdt, Š. & Krek, S. (2012). *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba*. Ljubljana: Trojina, zavod za uporabno slovenistiko; Fakulteta za družbene vede. doi: 10.4312/9789610603542
- Pegan, J., Robnik-Šikonja, M., Kosem, I., Gantar, P., Ponikvar, P. & Laskowski, C. (2022). *Slovenian datasets for contextual synonym and antonym detection*, Slovenian language resource repository CLARIN.SI, <http://hdl.handle.net/11356/1694>.
- Ulčar, M. & Robnik-Šikonja, M. (2020). FinEst BERT and CroSloEngual BERT: less is more in multilingual models. In P. Sojka (ed.) *Proceedings of Text, Speech, and Dialogue: 23rd International Conference*. Brno, Czech Republic. Cham: Springer, pp. 104–111. doi: 10.1007/978-3-030-58323-1_11
- Žagar, A., Kavaš, M., Robnik Šikonja, M. (2021). Corpus KAS 2.0: Cleaner and with New Datasets. In *Proceedings of the 24th International Multiconference – IS2021 (Slovenian Conference on Artificial Intelligence)*. Available at: <https://doi.org/10.5281/zenodo.5562228>

Trawling the corpus for the overlooked lemmas

Nathalie Hau Sørensen¹, Nicolai Hartvig Sørensen¹, Kirsten Lundholm Appel¹, Sanni Nimb¹

¹The Society for Danish Language and Literature, Chr. Brygge 1, 1219 Copenhagen K
E-mail: nats@dsl.dk, ka@dsl.dk, nhs@dsl.dk, sn@dsl.dk

Abstract

Lemma selection is a significant part of lexicographic work, also in the case of the online Danish Dictionary (DDO), a corpus-based monolingual dictionary updated twice a year based on the prior identification of good lemma candidates by means of statistical corpus methods as well as introspection. All low frequent word forms have until now been discarded in the statistical process, but in this paper, we present a method to also identify lemma candidates among these. Our hypothesis is that some words are too inconspicuously mundane to be noticed by introspection and at the same time so infrequent that they are overlooked by statistical measures. The method is based on different automatic measures of “lemmaness” by means of language models, character n-grams, statistical calculations and the development of a compound splitter based on information in the DDO. We evaluate the method by comparing the generated list with the lemmas included in the online DDO since 2005. Two trained DDO lexicographers furthermore evaluate words from the top as well as the bottom of the list. Though there is room for improvement, we find that our method identifies a large number of lemma candidates which otherwise would have been overlooked.

Keywords: Neology detection; lemma selection; low frequent words

1. Introduction

Lemma selection is a significant part of lexicographic work, also in the case of the online Danish Dictionary (DDO, Det Danske Sprog- og Litteraturselskab (2023a)), a corpus-based monolingual dictionary updated twice a year based on the prior identification of good lemma candidates primarily by means of statistical corpus methods. The corpus is extended monthly and consists of texts from the past 40 years, in total around 1.1 billion tokens. All low frequency word forms have until now been discarded in the automatic process of finding lemma candidates, but in this paper, we present a method to also identify low frequent ones among the discarded and noisy data. The method is based on different automatic measures of “lemmaness” by means of language models (word2vec, NER), character n-grams as well as statistical calculations and the development of a compound splitter based on lemma compound information registered in the DDO. We evaluate the method by comparing the resulting list of lemma candidates with the lemmas having been included in the online DDO since 2005. Two trained DDO lexicographers furthermore evaluate parts of the top as well as the bottom of the list.

The Danish Dictionary DDO was initially published as a printed dictionary 2003-2005, at that time describing the senses of 60,000 lemmas. Since 2009 the dictionary has been published online and today it describes more than 100,000 lemmas. The online publication gives us the opportunity to continuously update the content and thereby reflect the changes in the Danish vocabulary. Currently, the online DDO is updated twice a year. In Nimb

et al. (2020), we describe a corpus-based method of detecting new senses, new collocations as well as new fixed expressions of already included DDO lemmas; a method that has already resulted in many revised entries in the updates. The dictionary also publishes a substantial number of new lemmas in the updates, including both neologisms and words which were not included in the printed version due to space restrictions. Related to this part of the editorial work, we are interested in developing automatic methods to supplement our existing procedures of detecting lemma candidates. In spite of basing it on statistical corpus methods, we are aware that among the low frequency words in the corpus, some good candidates are still being overlooked, i.e. lemmas which become more and more relevant in the current phase of extending the DDO from 100,000 lemmas to up to 200,000 lemmas ¹. They hide among the large amount of noisy data which is discarded in our statistical methods, and which at a quick glance contains far more undesirable noise than good lemma candidates – meaning that the lexicographers would have to check the data manually word by word.

Lemma selection principles vary across different dictionary projects. In the case of the DDO, the selection is highly based on well-developed statistical corpus methods resulting in monthly lists of good candidates to choose from, supplemented by lemma suggestions from users, new words in the Danish orthographic dictionary *Retskrivningsordbogen* published by the Danish Language Council every year, and of course the editors’ own notifications of relevant words not yet covered by the DDO, e.g. as a benefit of the editorial work with the Danish Thesaurus where lexical gaps in DDO were sometimes discovered (Lorentzen & Nimb, 2011). In any case, the overall criteria of lemma selection in the DDO project is always a certain representation in the DDO corpus. For words with a frequency lower than 50/1,000,000,000, not only corpus frequency but also other aspects should always be considered by the editor. The occurrences should appear in different texts published over a number of years, normally at least three years. In some cases, when a very low frequency lemma (i.e. a lemma occurring less than 10 times in the corpus) is used outside the domain of newswire, e.g. within specific domains, or in daily life, maybe also by specific age groups or in oral rather than written language, occurrences found by searching the internet are included.

The automatic method we suggest is inspired by the introspective judgments of low frequency words which have until now been randomly discovered by the DDO lexicographers. It aims at identifying lemma candidates in the large amount of noisy low frequency corpus word forms by measuring “lemmaness” based on a combination of NLP techniques. We evaluate the method by comparing the identified lemma candidates with the lemmas added to the DDO dictionary over the last 15 years, but also by introspective judgments of the results carried out by two experienced DDO lexicographers with the knowledge of the specific editorial principles of the DDO project.

1.1 Background

We know for a fact that low frequency lemmas are relevant to include in the DDO. The user log reveals that users do query low frequency words that are not yet in the dictionary (Trap-Jensen et al., 2014). When applying the DDO in a sense annotation task on 2000 sentences from Wikipedia in the ELEXIS project (Martelli et al., 2023; Pedersen et al.,

¹ the predecessor, the ODS dictionary published 1918-1955 contains around 220,000 Danish lemmas

2023), we also found a surprisingly high number of lemma candidates among the word forms that were not represented in the dictionary, even though they only occurred once in the text. Among these were also candidates that our normal detection procedures had not discovered. Our hypothesis is that some words are simply too inconspicuously mundane to be noticed by introspection and at the same time so infrequent that they are overlooked by statistical measures. We are also aware that our statistical corpus methods depend on the quality of the corpus. The ideal corpus contains a broad collection of different text genres. However, the DDO-corpus mainly contains newswire because Danish texts from other genres have turned out to be difficult to obtain due to copyright issues. In a corpus with a lion's share of newswire texts, some mundane words used in daily life may be underrepresented.

We have previously studied the relation between corpus frequencies and lexicographic relevance in DDO (Trap-Jensen et al., 2014). Where all words represented among the top 100,000 most frequent forms were indeed well established in the language and for sure relevant DDO-lemmas, it turned out that frequency was less useful as a criterion when it came to the identification of relevant lemmas among the rest of the corpus words. When examining Spanish neologisms with corpus frequency and perception surveys, Freixa & Torner (2020) also found that even though frequency is an important factor to determine the degree to which a word is institutionalised, there are other factors (e.g. loan words, transparency of derivations and compounds) in play when considering whether a word should be included in the dictionary.

1.2 Related work

In Halskov & Jarvad (2010) a previous attempt to automatically identify neologisms in Danish is described. Due to the lack of available NLP tools for Danish at the time, the method did not yield very good results seen from a DDO perspective. Some of the problematic areas were named entities and compounds. There is a very high amount of the latter in Danish, also among low frequency words and new lemmas. An analysis of neologisms in the DDO updates showed that 52% of the neologisms are in fact compounds (Trap-Jensen, 2020). It is a challenge to distinguish ad hoc compounds from the more established compounds of which the senses are relevant to describe in the DDO dictionary, not only automatically but also by introspection. However, in the experiment in Halskov & Jarvad (2010), the aim was primarily to identify simplex neologisms, which are more important to include in the Danish orthographic dictionary than easily spelled compounds.

A popular approach to neologism detection is the use of exclusion lists (i.e. list of words that are already in the dictionary or otherwise beneficial to remove from the investigated data). In this approach, a corpus is preprocessed before the exclusion list is used to remove the already included lemmas. A series of filters and postprocessing steps can then be applied to get a list of potential neologisms or lemma candidates. The *NeoCrawler* (Kerremans et al., 2012) removes noisy tokens by using character trigrams to calculate whether a token is a probable English word, and we apply a similar technique on the Danish data (see section 2.3.5).

The exclusion list approach relies on the quality of the pre- and post-processing steps. In a corpus, many of the unknown words (words not registered already in a dictionary) are not neologisms but instead named entities, spelling errors and derivatives. In Langemets

et al. (2020), which describes an experiment with detecting Estonian neologisms, they conclude that only 10% of the words in an automatically derived list of lemma candidates are in fact good candidates to include in the dictionary. The results can be explained by a high number of derivatives and semantically transparent compounds in the final candidate list as well as the poor quality of NLP tools for Estonian.

Alternatively, the exclusion list approach can be combined with machine learning. For instance, Falk et al. (2014) uses supervised machine learning to identify neologisms in a French newswire corpus. For each unknown word, they extract a range of features related to form, spelling, and theme. The result is a ranked list representing a word’s probability of being a neologism according to a trained model. However, supervised machine learning approaches require that we have manually annotated data to train on which can be time-consuming to obtain. Since we use a simple weighted average, we avoid the need of manually annotated training data.

In this work, we present an automatic method for lemma selection based on both exclusion lists and a scoring mechanism that imitates the editorial principles of the DDO lemma selection of low frequency lemmas. The main focus is not only to detect neologisms in the traditional sense, but also to detect the overlooked lemmas in the entire DDO corpus with texts from 1982-2022, i.e. lemmas that could have been added to the corpus-based dictionary since the project was initiated in 1992. Like the approaches mentioned above, we use a series of preprocessing and filtering steps to remove the worst noise. We take a similar approach to Falk et al. (2014) by also extracting a range of features. Each feature corresponds to a post-processing step, however we do not remove candidates on the basis of only one of these. Instead, we calculate a combined score as a weighted average of each feature. The idea is that in the cases where a feature is not realised, or represents an error, another feature might balance the score.

In the next section we describe the method and the various steps involved in it, initiated by a description of the corpus that we use in our experiments. In section 3, we evaluate the results. Section 4 discusses some pitfalls of our approach, and finally we conclude in section 5.

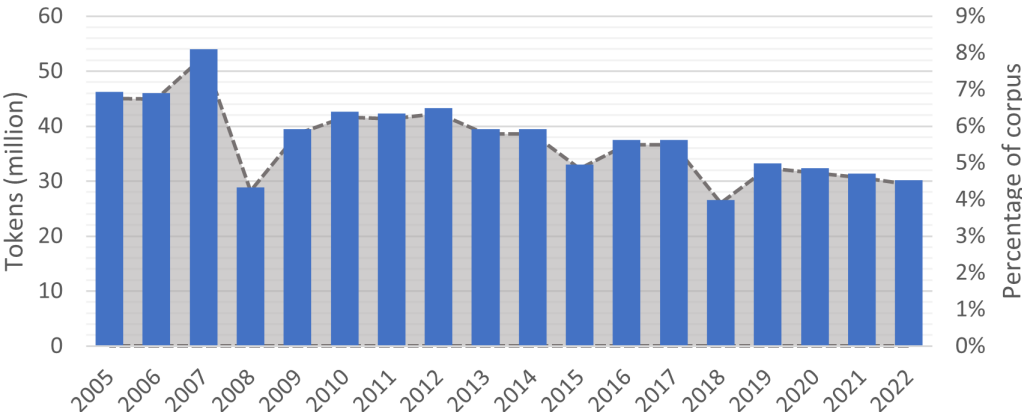


Figure 1: Distribution of tokens across the 18 years present in our corpus. Tokens are measured in millions.

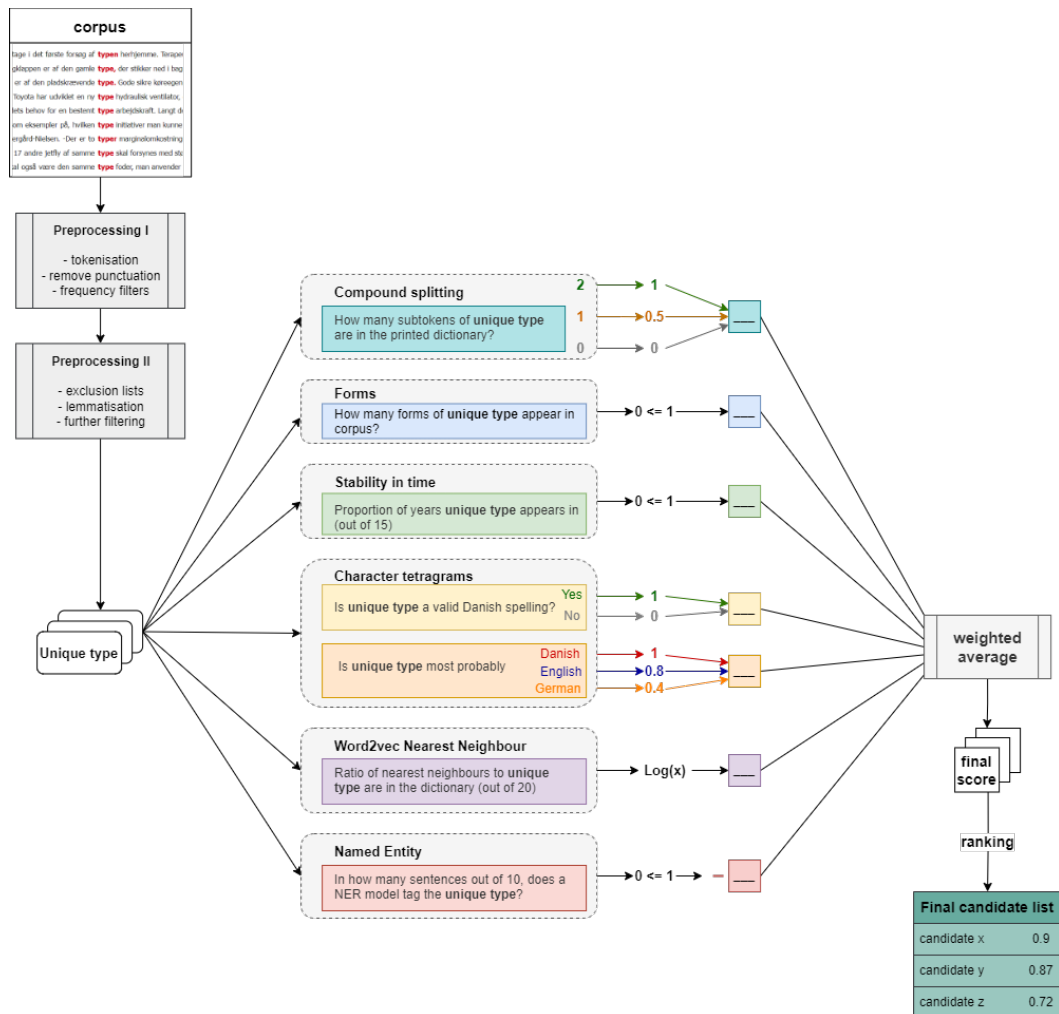


Figure 2: Overview of the automatic method from start to finish. The process begins with the corpus which is then preprocessed by two modules. Next, we extract features for each remaining word. Finally, the features are combined into a lemma score which the data is ranked after.

2. Methodology

Since the very beginning, the DDO dictionary project has been highly based on the access to high quality corpus data. The initial task in the project in 1991-1993 was to create the first part of today’s DDO corpus specifically for the subsequent compilation of the dictionary in the years 1994-2005. At the time, the corpus contained 40 million running words of Danish produced between 1983 and 1992 and contained both written and spoken Danish from a wide range of media and genres (Norling-Christensen & Asmussen, 2012). It has since been extended with two similar batches from around 2000 and 2010 respectively and a batch of texts from Wikipedia in 2017. Since 2005 it has been extended with newswire texts every month. It currently holds roughly 1.1 billion running words.

In our main corpus the genre composition varies over time making word frequency comparisons across different years unreliable. We have therefore chosen a subset containing only one genre to eliminate this issue. For our current experiment we use a part of the corpus composed exclusively of newswire texts from 2005-2022. It contains a total of

683,277,792 running words, and the texts are roughly equally distributed across the time period ranging from 26,640,430 words in 2018 to 53,958,056 words in 2007. The distribution is shown in figure 1. A subdivision into years allows us to track changes over time, and the homogeneous nature of the corpus makes the different years directly comparable, as the text domain remains constant across the entire corpus.

The starting point of the automatic method is a division of the lemmas in the online DDO into two parts, one containing only the lemma entries from the first printed edition of the dictionary from 2003-2005, another one containing the lemma entries in the dictionary added from 2005 to now. In the experiment, we assume that we have no knowledge of the new lemmas. Instead, we use the information as a 'lemma selection gold standard' when we evaluate the results. Hereafter, we refer to the original 2003-2005 printed DDO edition with 60,000 lemma entries as *the printed dictionary* and refer to the current online edition containing 100,000 lemma entries as *the online dictionary*.

The method includes three main steps which are visualised in figure 2. In the first step, we apply a cleaning and filtering process to reduce the number of investigated tokens. In step two we assign scores to each string using different scoring mechanisms. Each score assignment aims at representing the information used by the lexicographer in the process of selecting lemma candidates. In the last step, the final score is calculated using a weighted average of the scores, and the list is sorted accordingly. The following subsections explain the individual steps in detail.

2.1 Preprocessing I

The first step is to extract all the word forms from the corpus which will constitute the initial list of candidates. Thus, we convert the corpus into a raw text format, and the text is tokenised simply using space as a separating character. We then preprocess the data by lowercasing all tokens and removing all punctuations except hyphens. Next, we implement a script which counts the frequency of all types for all years and groups the data by type.

To further reduce the list of candidates, we apply two filters. First, we only keep the types with a frequency of 10 or above, as this is the lower limit used by the editors for low-frequency words. Secondly, we choose only the types that do not appear in the years 2005 to 2007. We primarily add this filter to reduce the number of candidates, as some of the tools we have developed require a lot of computing power. The raw list contains more than 600,000 word types, but the reduced list contains only 79,944 word types. This ensures that we focus on word types that tend to be new in the corpus - although the relatively low frequency of many of the word types naturally makes it less certain that they are actually neologisms. In principle, we are equally interested in candidates that were previously overlooked by the editors, and in the final dataset we will include all word types.

At this step it still contains a high number of word forms that are unlikely to be good lemma candidates, as shown in Table 1 which contains several proper nouns, German words and unidentifiable strings, and only three possible lemma candidates out of 12 tokens. If we sort the list by frequency as seen in Table 2, we easily conclude that frequency alone is not a useful criterion when identifying lemma candidates. Even though some are very good candidates (e.g., *coronakrise*, and *covid-19*), we also find named entities among the

Word	Total	2005-2007	2008	2009	..	2018	2019	2020	2021	2022
stickstoffdioxid (possibly German)	11	0	1	0	..	7	2	0	0	0
hoé (unknown)	29	0	1	0	..	1	0	2	7	2
personalekrævende ('staff-intensive')	10	0	1	0	..	0	0	0	2	0
cleantech-virksomheder ('cleantech firms')	31	0	1	5	..	0	0	0	1	0
-kursus ('course (suffix)')	11	0	1	0	..	1	0	2	1	0
kosin (unknown)	18	0	2	0	..	4	0	0	0	11
grega (unknown)	12	0	1	0	..	0	0	0	0	0
zurückfallen (German word)	13	0	2	0	..	0	2	2	2	0
husfliddk (web url)	11	0	1	1	..	1	1	4	0	2
amap (proper noun)	11	0	2	2	..	0	0	0	4	0
renneberg (proper noun)	22	0	1	3	..	0	0	0	1	1
hotspot-indsatsen ('the hotspot effort')	10	0	1	1	..	0	0	0	0	0

Table 1: Random types in the corpus that do not occur in 2005-2007.

Word	Total	2005-2007	2008	2009	..	2018	2019	2020	2021	2022
covid-19	11141	0	0	0	..	0	0	5761	4049	1331
coronakrisen ('the corona crisis')	14385	0	0	0	..	0	0	9742	3579	1064
brexit	10098	0	0	0	..	1254	3151	933	596	354
coronavirus ('corona virus')	8898	0	0	2	..	0	0	6260	2013	594
instagram (proper noun)	7831	0	0	0	..	644	997	1033	1136	1130
macron (proper noun)	7029	0	0	0	..	1026	1001	674	620	1305
-fhv (unknown)	6142	0	1	286	..	516	765	492	82	125
ipad	4565	0	0	0	..	166	238	145	133	155
coronapandemien ('the corona pandemic')	3980	0	0	0	..	0	0	1396	1583	1001
e-mailfinansritzaudk (an url)	3780	0	0	0	..	0	0	0	0	0
bnb (proper noun, 'b'n'b')	3737	0	0	0	..	1	3	0	0	0
radio24syv (proper noun)	3301	0	0	0	..	317	991	190	158	81

Table 2: Most frequent types in the corpus that do not occur in 2005-2007.

high frequency tokens. Additionally, these candidates are already identified by the normal corpus methods in the DDO project. Interestingly, there are unidentifiable strings even among the high frequency types (e.g. *-fhv*).

2.2 Preprocessing II

To further improve the list of candidates, we apply a second, more extensive preprocessing step which also includes the comparison of the data with lists of already registered word forms in other resources. The goal is to remove as many of the tokens that are highly unlikely to be lemma candidates as possible. First, we remove all numbers and URLs. Next, we use the previously kept hyphens to find a very common type of tokenisation error in Danish texts which are caused by the spelling rule of not writing the full form of a compound when listing it next to another one containing the same word component (e.g. we remove *morgen-* which occurs in the phrase *morgen- og aftenritual* ('morning and evening ritual')), corresponding to *morgenritual og aftenritual*. We also remove tokens starting with a hyphen (e.g. *-kursus* and *-fhv*).

To remove proper names, e.g. *Renneberg* and *Quintus*, we use lists of registered personal and place names published by *Danmarks Statistik*². We are aware that some personal names are also common appellatives in Danish³ but assume that these are already present in the dictionary. Names of organisations are identified and processed in a later step (see 2.3.3).

In a similar way, the list of inflected forms of lemmas appearing in the printed DDO is used to identify and remove all inflected forms of lemmas in the printed DDO. The list of inflected forms of lemmas added to the DDO since 2005 (the online dictionary) is, however, not used to lemmatise the remaining data, it is instead kept for evaluation purposes (see section 3). We use the CSTLEMMA lemmatiser (Jongejan & Dalianis, 2009) as an alternative⁴, and sum up the corresponding frequencies at lemma level (keeping also a list of the original tokens). This step greatly reduces the size of the dataset, and at the same time allows us to treat inflected forms of the same lemma in a similar manner.

Since the focus of the experiment is low frequency lemma candidates, we finally remove tokens and lemmas with a total frequency above 100 across all years; these are likely to have already been identified by our standard corpus methods. We also remove all words that only occur in texts from one or two years out of the 15 years of corpus data that we investigate, taking into consideration the editorial criteria of representation in texts from at least 3 years.

After these preprocessing steps, the list of lemma candidates is reduced from 79,976 to 36,172 candidates, which is still a very high amount of data which - even though it includes a high percentage of lemma candidates *middelklassedrenge* 'middle class boys', *lektiehjælper* 'private tutor', *fejltankning* ('misfuelling'), *envejsbil* ('one-way-car'), and *bodyage* (English loan) - still also contains a lot of named entities (e.g. *okmans*, *grunerwidding*, *jammerbugts*) as well as other noise, such as tokenisation errors (*erlyd*, and *dkandidat*).

2.3 Measuring "lemmaness"

In order to improve the list, we develop a measure of "lemmaness" which we call the *lemma score*. The lemma score is a weighted average of several subscores related to stability in time, adaptation to Danish orthography and morphology, and last but not least semantic similarity to known lemmas in the dictionary. By measuring these features, we reflect a large number of the criteria used by the lexicographer when selecting a lemma for the dictionary.

2.3.1 Stability in time

In the current list, we know that all words are represented in texts from at least three years, but we assume that an even more widespread representation correlates with the suitability to be included as a lemma in a dictionary. We do not take the frequency in each year into account, but simply assign a score if the word occurs just once. The 'stability in

² 'Statistics Denmark', a government authority: <https://www.dst.dk>

³ *Sten* 'stone' is a Danish male first name

⁴ We are interested in developing a method that can also be used for future detection of overlooked lemmas. Therefore, we do not want to base the lemmatisation only on the online dictionary

time’ score is the number of years the candidate occurs (in any form) divided by the total 15 investigated years (2005-2022).

2.3.2 Form

A word’s adaptation to Danish morphology highly indicates that it has been lexicalised in the language and that it is highly unlikely to be a named entity. Therefore, the previous lemmatisation is used to count the number of forms in the (lemmatised) groups of words, and assign a score according to the number. We do not take PoS into account, which in turn may discredit candidates with inherently fewer word forms in Danish (e.g. adverbs have fewer forms than nouns and adjectives). However, only a few candidates appear with more than three word forms. We assume that the disparity has minimal influence on the final score.

2.3.3 Named Entity

Even though many named entities were already removed from the list during the filtering process by use of data from *Danmarks Statistik*, many names of organisations, products, foreign personal nouns as well as creative artist nouns are still present in the candidate list. To automatically detect these, we use the ScandiNER model⁵. The model expects a sentence as input and it outputs NER-tags with their respective position in the sentence. Therefore, the model cannot identify whether a word is a named entity when seeing it in isolation. To circumvent the problem, we randomly select up to ten sentences from the corpus. Each sentence is then tagged with the model and we compare the tag position with the candidate’s position in the sentence. If the two positions overlap, we increase the named entity score. The final named entity score is the total percentage of sentences where the named entity tag overlaps with the candidate position out of all selected sentences. The named entity score is subtracted from the total score.

2.3.4 Compound splitting

As mentioned above, the automatic analysis of the high number of compounds in Danish is a challenge. We chose to develop the compound splitter *DSLSplit*⁶ in order to be able to split the many compounds in the most likely word components, so that we are able to analyse these based on existing lemma information in our resources. The compound splitter is characterised by having two modes, a ”careful” and a ”brute” one. We base the ”careful” mode on CharSplit - a German n-gram based compound splitter (Tuggener, 2016) that we have adapted to Danish. The ”brute” mode also uses probabilities to estimate the most likely split. It was trained using the manually added information on a part of the compounds in the DDO (30,211 compounds). This data turned out to be far from sufficient and it was therefore supplemented with automatically generated compound information based on the retro-digitised historic Danish dictionary *Ordbog over det danske Sprog* (ODS, Det Danske Sprog- og Litteraturselskab (2023c)). 168,321 compounds were identified automatically, modified to modern orthography and used as training data, even though

⁵ Available at <https://huggingface.co/saatrupdan/nbailab-base-ner-scandi>

⁶ DSLSplit and a more detailed description can be found <https://github.com/dsldk/dslsplit>

they contain some compounding errors. The brute method does not always correctly identify the first of the two compound elements when they are joined by an "e", or an "s" (these letters are also very often final or initial letters of simplex lemmas in Danish). In the present work, we run our compound splitter in *mixed* mode, meaning that the splitter first tries the "careful" approach, and if this method doesn't find a probable split, the "brute" method is applied.

Scores are assigned to the components of the compounds depending on whether they are lemmas in the printed DDO or not. If both are, we assign the highest score (1) to the candidate. If only one component is included in the dictionary, the candidate is assigned a low score (0.5) If none of the components are found in the dictionary or if the candidate cannot be split by the algorithm, it receives no score (0).

2.3.5 Language features

With the naked eye, it is evident that some candidates on the list do not follow the typical spelling of Danish words, either due to tokenisation or spelling errors, or because the word is of foreign origin. To identify this part of the data from the list of lemma candidates, we calculate the likelihood of a character sequence being in accordance with the standard Danish spelling. We apply the tetra-gram model in LexiScore⁷. The model is trained on the list of inflected word forms from the online DDO dictionary containing 641,971 words. Additionally, Laplace smoothing (k=100) is applied to offset the low-frequency tetra-grams. To determine the probability of a character sequence belonging to the target language, we multiply the probabilities for each tetra-gram, assigning a very low probability for any tetra-gram not found in the list of words (specifically, 1e-20). To avoid penalising longer words, the base probability is normalised after the length of the sequence. We set the probability threshold to 0.0001. A candidate above this threshold gets a Danish validity score of 1, while a candidate below the threshold is deemed invalid and gets a score of 0.

Some of the words of foreign origin are highly relevant lemma candidates. We checked 1363 new lemmas in the DDO (lemmas included in 2019-21, both neologisms and older words), and 8% contain non-Danish orthography (e.g. *gefühl, betting, aftersun, ajvar*), and we know for a fact that many neologisms in Danish are loanwords from especially English, but also German. LexiScore also contains tetra-gram models for English (trained on the Moby Crosswords word list⁸) and for German (trained on the German Aspell dictionary⁹). We also include an extra Danish model trained exclusively on head words in DDO. This allows us to compare the probability of a character sequence being the most typical for either of the three languages. The language origin feature score is shown in Table 3.

Highest probability	Danish	DDO head	English	German
Language origin	1	0.9	0.8	0.4

Table 3: Language origin feature score

⁷ The source code of LexiScore is available at <https://github.com/dsldk/LexiScore>. New languages are easily added from simple word lists.

⁸ Available at <https://www.gutenberg.org/files/3201/files/>

⁹ Available at <https://ftp.gnu.org/gnu/aspell/dict/0index.html>

2.3.6 Semantic model

Synonyms as well as near synonyms of already included lemmas are very likely to be good lemma candidates. For instance, the new lemma *daikon* ‘daikon, type of Japanese radish’ was added to the online dictionary in 2021, having a sense very close to three lemmas already included in the printed DDO, namely *kinaradise* (‘Chinese radish’), *radise* (‘radish’), as well as *ræddike* (‘black radish’).

One way of checking whether a word is a good lemma candidate or not is to investigate whether it is semantically related to any of the already included lemmas. Word embedding models like word2vec (Mikolov et al., 2013) build on the assumption that a word’s meaning can be estimated from its distribution in text, in line with the distributional hypothesis summed up in the famous line ‘you shall know a word by the company it keeps’ (Firth, 1957). Or, put differently: Semantically similar words typically appear in the same context. A word embedding model creates a vector representation (i.e. a word embedding) of each token in a corpus. By computing the distance between two embeddings, we are able to estimate their semantic similarity. In order to find similar lemmas to the already known ones from the dictionary, we simply need to search through the semantic space created by the word embedding model.

In the experiment we use a model that we have previously trained on texts in the DDO-corpus published before October 2019 when the corpus contained over 1 billion raw tokens, 7.17 million word types, and 2.79 million sentences. We use a model that is trained similarly to the one in Sørensen & Nimb (2018) with opensource Gensim package for Python (Řehůřek & Sojka, 2010), and use the model to find the 20 nearest neighbours of a lemma candidate in our list. The candidate is assigned a value if the neighbour is included as lemma in the DDO. We calculate an overall semantic feature score based on the number of neighbours being DDO lemmas, however with a logarithmic function to decrease the influence of having more than 5. We find that it shouldn’t count drastically more having a higher number than this.

2.4 The final lemma score

For each candidate, the features are combined into the final lemma score through a weighted average. The weights are set manually depending on how reliable we find each feature. We adjusted the weights after inspecting the lemma score in a preliminary experiment on only the first ten years of the corpus data. For instance, we found it beneficial to lower the initial weight of the compound feature to lower the advantage of compound candidates. We further discuss this problem in section 4. In descending order, the final weights are: semantic feature (0.6), origin language (0.5), compound (0.4), form (0.2), stability in time (0.1), valid Danish spelling (0.1), and named entity score (-0.8). Finally, the candidate list is ranked after the final lemma score. Examples of candidates and their respective lemma scores and rank is visible in Table 4.

3. Evaluation

We evaluate the final candidate list in two ways. First, we compare the list with the updates from the online dictionary carried out since 2005 in order to see to which degree the automatic method is able to identify the lemmas which have in fact been selected by

Candidate	Semantic	Origin	Compound	Form	Time	Spelling	Named	Ent	Lemma score	Rank
havesaks	0.9	1	1	0.1	0.5	1	0		1.47	3
campingstol	0.65	0.8	1	0.1	0.44	1	0		1.34	176
filologi	0.77	1	0.5	0.1	0.56	1	0		1.24	771
olieforum	0.1	1	1	0.1	0.63	1	1		0.38	28937
fwd	0	0	0	0.1	0.19	0	1		-0.69	36166

Table 4: Selection of candidates with their respective lemma score and rank.

the lexicographers in the last 17 years based on the existing methods in the DDO project. But we also want to evaluate the quality of the remaining candidates on the list. We expect to find a high number of lemma candidates not yet added to the dictionary and would like to measure the recall accordingly. Therefore, two experienced DDO lexicographers manually check a random selection of words from the top and bottom of the list.

3.1 Comparison with dictionary updates from 2005-2022

The online dictionary has been updated with approximately 38,000 new lemma entries since 2005, 30% with morphological information, 70% also with sense descriptions. We assume that all entries added to the dictionary are deemed to be good candidates by the editors and hence can be used as a gold standard in the evaluation. We will refer to the lemmas included in the dictionary updates as *true lemmas*.

In the final list of 36,172 candidates, we find 1550 true lemmas which account to less than 5% of the candidates. Although it seems like a low coverage, we have to consider that we have removed the candidates with the highest frequency and only include the candidates that did not appear in 2005-2007 (see section 2.2). In figure 3, the cumulative sum of true lemmas in the candidate list is presented. From the figure, we can see that the ratio of true lemmas is highest in the top section of the list. In fact, 10% of the true lemmas are covered by the top 1% of the highest scoring candidates, and 50% of the true lemmas by the top 17%. Additionally, we see almost no true lemmas in the bottom section of the list as the curve on figure 3 levels out after around rank 25,000.

3.2 Qualitative analysis

The majority of the words in our generated list of lemma candidates have not (yet) been included in the DDO, and therefore cannot be evaluated by a comparison with the ‘gold standard’ of entries established in the DDO since 2005. To evaluate the quality of these, we instead extract a subset to be annotated by two DDO lexicographers. We consider a lemma to be correctly identified as a lemma candidate, i.e. a true positive, when at least one of the two lexicographers find it relevant to present on a lemma candidate list. The subset is composed of a random selection of 394 candidates; 199 from the top 1500 (the best candidates), and 195 among the last 1500 (the worse candidates) (after having removed all the true lemmas)¹⁰. The subset is shuffled to obscure the rank of the candidates. On this subset, we calculate the recall to be 92.6%

¹⁰ Originally, we extracted 200 from each selection. However, some of the candidates appeared in the dictionary in another form, e.g., *alterego* appeared as *alter ego*.

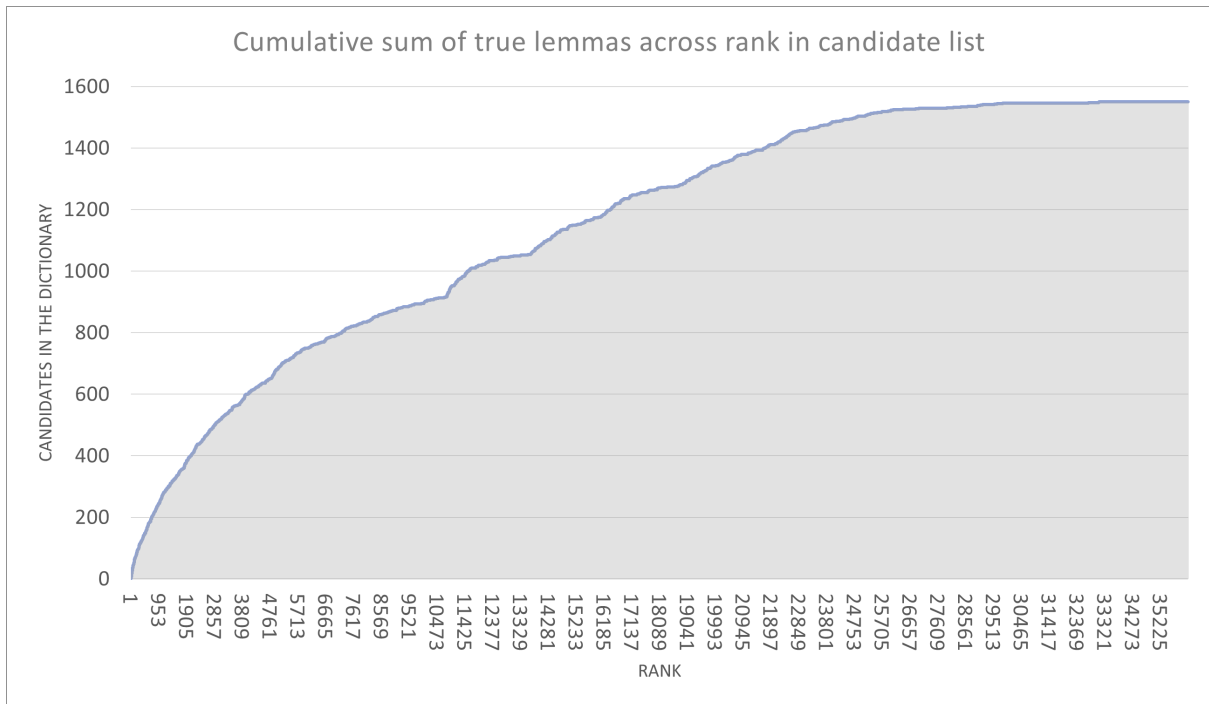


Figure 3: Cumulative sum of the true lemmas (i.e. lemmas added to the dictionary since 2005) across the rank in the final candidate list

Among the 199 top candidates, 94% (188) are estimated to be a relevant candidate by at least one of the two annotators. The 6% (11) non-relevant fall into four categories: ad hoc compounds (e.g. *accelerationssamfund* ‘acceleration society’ and *haveoplevelse* ‘garden experience’), named entities (e.g. the bakery company *Lagkagehuset*), tokenisation errors (*haftafgørende* → *haft afgørende*), and domain specific terms (*hjertertaber* ‘lit. heart loser’, a bridge term). The ad hoc compounds constitute more than half.

Among the candidates with the lowest score, 92% (180) are annotated to be irrelevant lemmas by at least one of the two annotators. The remaining 15 words were judged to be relevant. When we take a closer look at these, four received a low score in the automatic process because of lemmatisation errors (e.g. *facebookven* falsely lemmatised as *facebookkv*), and the remaining part received a low score because they had either features resembling a named entity, or (in 7 cases) are (entirely or partly) words of foreign origin (*twitterfeed*, *techcrunch*), (*tuaregoprører* ‘tuareg rebels’), or maybe even a combination of both.

4. Discussion

In spite of the promising results, we are aware of some pitfalls in the approach that we will address in the following subsections.

4.1 The impact of the corpus composition

A limitation of our study is the composition of the corpus. The corpus is a newswire corpus and does not reflect the language of the average speaker. First, the text is both written and edited by professional writers and large groups of language users are not represented, e.g. children, teenagers, as well as adults from other professions. Secondly, the texts in it

only cover a specific range of topics that are considered newsworthy, and they represent only a rather formal style. This means that a great deal of words and expressions used in everyday language are most likely missing from the corpus and thus will still not be uncovered by our methods.

On the other hand, if we had access to a more balanced corpus, our method may be trivial as the corpus frequencies would be more representative. Still, it could be interesting to investigate whether our method can be used in other domains to uncover even more lemma candidates. We are particularly interested in a corpus that contains more everyday Danish. In future work, we are therefore considering including Danish internet forums. Here, we need to consider whether the weights can be directly transferred to other text types and whether we need to examine a new type of feature connected to the domain.

4.2 Balancing of the weights

One of the challenges in the experiment was to balance the weights. This was done manually, and we aimed at finding the optimal balance between the features. Although the current weights successfully distribute a large number of relevant candidates among the top ranks, we see space for improvements in the middle range of the list. This becomes evident when studying the true lemma curve on figure 3 since the curve suddenly increases around rank 10,000. From the qualitative analysis in section 3.2, we also know that the named entity feature has too high a negative impact on some of the relevant lemmas. Likewise, the language features may penalise foreign words too much. The question is how many missed lemmas we can accept. Adjusting the feature scores may cause more noise to appear higher up in the list. We need to further investigate the impact of the individual features to refine our approach in the future.

One idea is to split our data into a training and test set. The weights can then be set on the training set before we evaluate on the test set. Even though we have a good standard in the form of the dictionary updates, these only give us the positive cases. We still need to collect a sample of non-lemma cases, e.g. cases of words that were actively discarded by the lexicographers. Simply using the words that are not in the dictionary would not be representative as they may be good, overlooked lemmas. Now that we have gathered some information about the characteristics of irrelevant words, we are able to conduct more experiments.

4.3 Are compounds too prominent?

The compound feature rewards compounds of which the components are already included in the printed dictionary. Thus, we might overemphasise compounds at the expense of simplex words and derivations. The question is how problematic this is for the lemma selection process. Compounds are prominent in Danish neologisms. A study by Trap-Jensen (2020) estimates that 52% of neologisms in recent DDO updates are compounds. However, the percentage of compounds is much higher if we look at both actual neologisms and words that were previously overlooked or fell outside the scope of the lemma selection when editing the printed dictionary. In the last three updates of DDO (November 2021, June 2022, and November 2022 (Det Danske Sprog- og Litteraturselskab (2023b))) a total of 938 lemmas were added to the dictionary. Of these, 738 are compounds, equivalent to

78.7%. Thus, it is not unreasonable to expect our method to also identify a high number of compounds. In addition, DDO describes a number of common derivational affixes and suffixes, and the current compound splitter is therefore able to split certain derivations. In the future, it is worth exploring whether we can update the compound splitter to differentiate between actual compounds and derivations and thereby give them different scores.

With the high number of compounds, we also face the problem of ad hoc compounds. The automatic compound feature cannot distinguish compounds that have been established in the language from ad hoc ones. Since ad hoc compounds constitute more than half of the high ranking irrelevant words in the qualitative analysis, they seem to cause a more general problem to our method. We simply lack more information about the characteristics of compounds which are produced on the fly. For instance, are certain words more productive than others as components in ad hoc compounds? This is something we plan to investigate further.

4.4 The "newness" criterion

A large number of the highest ranking candidates are in fact not new in the language, although we have disregarded word types that occur in 2005-2007. The main purpose of disregarding these word types was to reduce the size of the dataset rather than filtering it for actual neologisms. When the editors of DDO include more lemmas in the dictionary, they are not only searching for neologisms, but also previously overlooked words like *julekugle* (eng. 'Christmas ornament') and *havesaks* (eng. 'garden shears'). Therefore, the scope of the study is wider than just neologisms in a strict sense. Nevertheless, we plan to expand the method on all word types in the entire newswire corpus (i.e. words that also occur in 2005-2007) to identify even more overlooked lemma candidates.

4.5 Are foreign words unfairly penalised?

We introduced LexiScore in order to automatically identify and filter out noise coming from web addresses, failed tokenisation, and also non-Danish texts from the corpus. However, many neologisms in Danish are direct borrowings, especially from English. Some domains are also naturally described by means of loan words, e.g. culinary terms like *nduja* 'italian pork sausage'. In the corpus, we find many examples of full sentences from other languages, for instance through a direct quote. For a foreign word to be considered a loan in Danish, it has to occur in a context with a majority of Danish lemmas. In our experiment we are limited to only study the character composition without being able to include the context. A better way to calculate the language origin feature might be to look at the combined probability of the candidate and its closest context (+/- two or more words) to see whether they belong to a specific language. Alternatively, LexiScore can be used to identify and subsequently remove longer sequences of English and German texts during the preprocessing steps to reduce the number of foreign words in the data.

5. Conclusion

The lemma score method that we have presented is a useful contribution to the task of identifying the new lemmas to be added to the DDO dictionary. The approach has enabled

us to effectively sort out a large amount of irrelevant words in the extracted corpus data so that only a minimum of noise is left. Where a manual inspection in the initial candidate list (before the scoring) showed that only roughly every 20th word was relevant, we now find relevant words in up to 94% of the cases in the top of the list. The fact that we find half of the lemmas added since 2005 in the top 17% of the generated list also proves the high quality of the method. In the daily lexicographic work it means that it is now a manageable task for the lexicographers to manually inspect the list. In this way our method speeds up the process of lemma selection in the DDO project significantly.

The greatest influence on the scores was provided by the word embeddings, and by the automatic identification of Named Entities. We find that the use of word embeddings and Named Entity recognition allows for a more efficient and accurate selection process. We believe that especially the idea of combining the scores ensures the good quality of the results. Instead of using each feature as a filter to remove noise, we consider all features at once to get a complete picture of each word's potential for inclusion. Thus, it is not detrimental if a word gets a low score in one feature if the scores in the other features are high enough to counterbalance the score.

Another advantage is that the method does not require an annotated dataset to train a supervised machine learning model. Since the list is going to be manually processed by the lexicographers, we don't need a very high accuracy of the ranking. We have shown that a weighted average yields good results when an annotated dataset is not available.

In conclusion, we believe that the data we have obtained is highly useful for the DDO lexicographers, as it allows them to select lemmas in a more efficient and objective manner which in turn also leads to a higher quality dictionary.

6. Acknowledgements

We would like to express our sincere gratitude to several individuals and organisations who have contributed to this work. First and foremost, we would like to thank senior editor Jonas Jensen for his invaluable help with the annotation of the data as well as with the proof-reading and commenting on the paper, and senior editor Thomas Troelsgård for supplying us with a comprehensive list of all inflected forms of lemmas in the written version of *Den Danske Ordbog*.

Finally, we would like to thank *The Carlsberg Foundation* and the Danish Ministry of Culture for their generous grants, which have enabled us to continue our work on the dictionary. Without their support, this project would not have been possible.

7. References

- Det Danske Sprog- og Litteraturselskab (2023a). Den Danske Ordbog. <https://ordnet.dk/ddo>. Accessed on April 21, 2023.
- Det Danske Sprog- og Litteraturselskab (2023b). Nyeste ord i DDO. <https://ordnet.dk/ddo/nyeste-ord-i-ddo>. Accessed on April 21, 2023.
- Det Danske Sprog- og Litteraturselskab (2023c). Ordbog over det danske Sprog. <https://ordnet.dk/ods>. Accessed on April 21, 2023.

- Falk, I., Bernhard, D. & Gérard, C. (2014). From non word to new word: Automatically identifying neologisms in French newspapers. In *LREC-The 9th edition of the Language Resources and Evaluation Conference*.
- Firth, J. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*, pp. 10–32.
- Freixa, J. & Torner, S. (2020). Beyond frequency: On the dictionaryization of new words in Spanish. *Dictionaries: Journal of the Dictionary Society of North America*, 41(1), pp. 131–153.
- Halskov, J. & Jarvad, P. (2010). Manuel og maskinel excerpering af neologismer. *NyS, Nydanske Sprogstudier*, (38), pp. 39–68.
- Jongejan, B. & Dalianis, H. (2009). Automatic training of lemmatization rules that handle morphological changes in pre-, in-and suffixes alike. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. pp. 145–153.
- Kerremans, D., Stegmayr, S. & Schmid, H.J. (2012). The NeoCrawler: Identifying and retrieving neologisms from the internet and monitoring ongoing change. *Current methods in historical semantics*, 73, p. 59.
- Langemets, M., Kallas, J., Norak, K. & Hein, I. (2020). New Estonian Words and Senses: Detection and Description. *Dictionaries: Journal of the Dictionary Society of North America*, 41(1), pp. 69–82.
- Lorentzen, H. & Nimb, S. (2011). Fra krydderkage til running sushi – hvordan nye ord kommer ind i Den Danske Ordbog. *Nye ord, Sprognævnets Konferencereserie*, 1, pp. 69–85.
- Martelli, F., Navigli, R., Krek, S., Kallas, J., Gantar, P., Koeva, S., Nimb, S., Sandford Pedersen, B., Olsen, S., Langemets, M., Koppel, K., Üksik, T., Dobrovoljc, K., Ureña-Ruiz, R., Sancho-Sánchez, J.L., Lipp, V., Váradi, T., Györffy, A., Simon, L., Quochi, V., Monachini, M., Frontini, F., Tiberius, C., Tempelaars, R., Costa, R., Salgado, A., Čibej, J., Munda, T., Kosem, I., Roblek, R., Kamenšek, U., Zaranšek, P., Zgaga, K., Ponikvar, P., Terčon, L., Jensen, J., Flörke, I., Lorentzen, H., Troelsgård, T., Blagoeva, D., Hristov, D. & Kolkovska, S. (2023). Parallel sense-annotated corpus ELEXIS-WSD 1.1. URL <http://hdl.handle.net/11356/1842>. Slovenian language resource repository CLARIN.SI.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.
- 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.
- Norling-Christensen, O. & Asmussen, J. (2012). The Corpus of the Danish Dictionary. *Lexikos*, 8(1). URL <https://lexikos.journals.ac.za/pub/article/view/955>.
- Pedersen, B.S., Nimb, S., Olsen, S., Troelsgård, T., Flörke, I., Jensen, J. & Lorentzen, H. (2023). The DA-ELEXIS Corpus - a Sense-Annotated Corpus for Danish with Parallel Annotations for Nine European Languages. In *RESOURCEFUL 2023 - Proceedings of the second workshop on Resources and Representations for Under-Resourced Languages and Domains, NoDaLiDa 2023*. Forthcoming.
- Řehůřek, R. & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, pp. 45–50. <http://is.muni.cz/publication/884893/en>.

- Sørensen, N.H. & Nimb, S. (2018). Word2Dict - Lemma Selection and Dictionary Editing Assisted by Word Embeddings. In *Proceedings of the 18th EURALEX International Congress: Lexicography in Global Contexts*. pp. 819–827.
- Trap-Jensen, L. (2020). Language-Internal Neologisms and Anglicisms: Dealing with New Words and Expressions in The Danish Dictionary. *Dictionaries: Journal of the Dictionary Society of North America*, 41(1), pp. 11–25.
- Trap-Jensen, L., Lorentzen, H. & Sørensen, N.H. (2014). An odd couple – Corpus frequency and look-up frequency: what relationship? *Slovenscina 2.0*, 2(2), pp. 94–113.
- Tuggener, D. (2016). *Incremental coreference resolution for German*. Ph.D. thesis, University of Zurich.

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

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

Institute of Mathematics and Computer Science, University of Latvia, 29 Raina Blvd.,
Riga, LV-1459, Latvia
E-mail: peteris@ailab.lv

Abstract

In this paper, we describe our findings from developing the lexicographic platform Tēzaurs.lv, extending it from a traditional explanatory dictionary into a multifunctional resource for structured lexical data. Tēzaurs.lv is the largest Latvian dictionary with more than 390,000 entries, which emerged as a compilation from nearly 300 prior dictionaries and other sources. Recently, it has been extended with Latvian WordNet data, effectively making it also a synonym dictionary and a translation dictionary. Each entry can contain multiple lexemes with their grammatical information and inflection tables, enabling search on inflection forms and spelling variants.

For the new requirements, we have developed a lexical database system and a collaborative online editor toolkit, which are also used for two other major Latvian dictionaries. While previously the data model and tools were based on what the end user would see in a dictionary entry, the current infrastructure is designed with a highly structured lexical data model. This avoids duplication and helps to ensure consistency if entries or word senses are edited or merged, and it supports the usage of this data in computational linguistics.

Keywords: lexicography; platform; data model; Latvian

Introduction

Tēzaurs.lv is the largest Latvian electronic dictionary with more than 390,000 entries, which was initiated as a consolidated compilation from approximately 300 dictionaries (Spektors et al., 2016) and other sources, and has recently been extended and developed with the addition of Latvian WordNet (Paikens et al., 2022) data (6,610 synonym sets) and 75,400 manually curated corpus examples for specific senses. All entries contain at least one lexeme and one sense defined. 118,000 lexemes contain appropriate inflectional paradigms to provide inflectional tables and the ability to search by inflectional forms.

Tēzaurs.lv emerged around 2009 as a side result of a larger research project in computational linguistics. The dictionary was encoded in an ad-hoc text format which annotated the beginning of each element with a two-letter code. As the dictionary grew in size, a set of consistency verification scripts was developed (Danovskis, 2014), but there was no multi-user editing support, and everything had to be done by a single editor in the single authoritative copy of the dictionary. New releases were published four times per year by pre-processing the dictionary data from the in-house format to static HTML and loading the pre-rendered entries into a simple database with a thin front-end application on top. It provided basic search and display functionality, however, it lacked in-depth search

functionality, e.g., it was possible to search only by the headword of the entry, and not by derivatives, multi-word expressions (MWEs), glosses, etc.

Until 2020, Tēzaurs.lv had a flat structure, resembling a printed dictionary with lots of duplicated information (for example, each MWE was usually described multiple times – in each constituent word entry – but with potentially different definitions), which also turned out to be a major obstacle for further enhancement of the dictionary.

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

The dictionary editing tools and the whole infrastructure have also undergone major changes over the course of Tēzaurs.lv development.

The goal for this undertaking was a web-based multi-user multi-dictionary application with a centralized database as a single source of truth, which supports dictionary creation and editing, as well as dictionary publishing. For these needs, we have developed a lexical database system and an editor toolkit which, besides the Tēzaurs.lv dictionary itself, is used also for two other major Latvian dictionaries: Dictionary of Standard Latvian (LLVV, retro-digitised)¹ and Dictionary of Contemporary Latvian (MLVV, continuously updated)². We also considered using TLex³ or Lexonomy⁴ (Rambousek et al., 2021), but we were worried about that the large amount and rather complex structure of already existing Tēzaurs.lv data might make these solutions slow and hard to maintain. From newer development it is worth mentioning Lexmart⁵ (Simões et al., 2019), however it was not available yet when work on Tēzaurs.lv platform started, and it works on top of an XML database, which does not fit our plans for using a fine-granular data model.

In Section 1, we describe the Tēzaurs.lv online platform and the features relevant for its end-users. Section 2 describes the data model used for the multifaceted lexicographic data, and Section 3 describes the tools supporting the lexicographic workflow.

1. Tēzaurs.lv online platform

The Tēzaurs.lv lexicographic platform is developed as a web application which supports collaborative dictionary editing as well as dictionary publishing.

In the editor mode, the application works directly on the atomic data stored in the database. Data consistency is ensured via backend validations, database constraints and

¹ <https://llvv.tezaurs.lv>

² <https://mlvv.tezaurs.lv>

³ <https://tshwanedje.com/tshwanelex/>

⁴ <https://www.lexonomy.eu/>

⁵ <http://lexmart.eu/>

transactions. In the publishing mode, it works in read-only mode on pre-generated data (complete entries for fast response time and reverse indices for search support), which are created in the quarterly release preparation process.

From the data point of view, the published version is an enriched read-only copy (snapshot) of the dictionary database state in the moment of publishing. From the application point of view, the published version utilizes a subset of the same data procedures and view templates as the editor's view, thus ensuring consistency between both views.

1.1 User Interface

1.1.1 Entry View

The central element of the interface is the view of an Entry (see Figure 1). It consists of the Heading, one or more Lexeme blocks and one or more Sense blocks, and ends with a list of the lexical sources for this entry. Lexeme blocks may have inflection information, Sense blocks may have several sub-blocks: usage examples, related senses, translations, MWEs. To make the presentation mode compact, the inflection tables in lexeme blocks and all sub-blocks in the sense blocks are expandable but initially collapsed. All blocks may have verbalization of grammatical and usage information.

Pavasara versija 2023
390 186 šķirklji

Līdzīgi šķirklji:
forma šķirklī [labt](#)

Pameklēt plašāk

Apkaime
labrīts
Labrjogana
labrocība
labrocis
labrocīts
labs
labsajūta
labsirdība
labsirdīgs
labsirdīgums
labsirdis

MLVV
labs

LLVV
labs

labs
labs ipašības vārds Locīšana

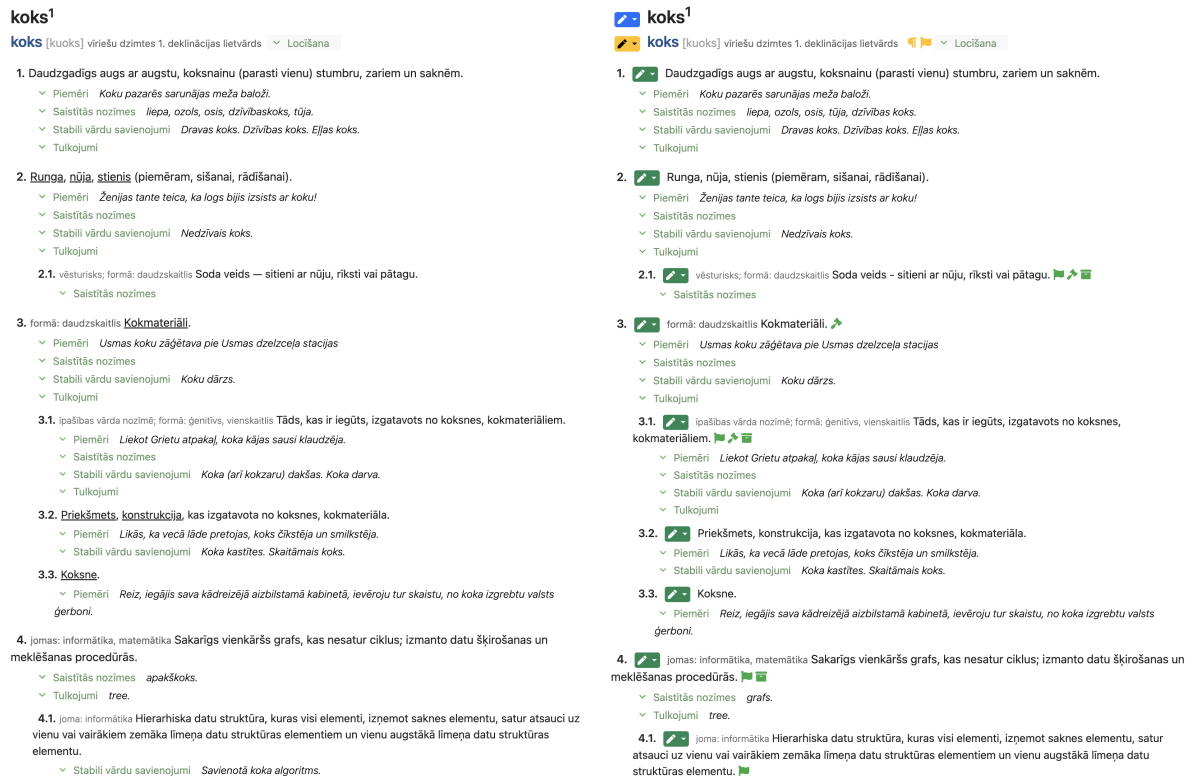
- Tāds, kuram piemīt morāles normām atbilstošas īpašības, kura attieksme (pret citiem) ir iejūtīga, labsirdīga, labvēlīga.
 - ▼ Saistītās nozīmes
 - ▼ Tulkojumi *valuable*.
 - Tāds, kas atbilst morāles normām, tāds, kurā izpaužas iejūtība, [labsirdība](#), [labvēlība](#).
 - ▼ Stabili vārdu savienojumi *Labais tonis*.
 - lietvārda nozīmē; formā: vīriešu dzimte, pamata pakāpe, noteiktā galotne; retāk formā: vīriešu dzimte, pamata pakāpe, nenoteiktā galotne **Tas, kas atbilst morāles normām, ideālam; šī ideāla īstenojums.**
- Tāds, starp kuriem ir saskaņa, saprašanās, draudzīgas attiecības (par cilvēku grupu); tāds, ar kuru ir šādas attiecības.
 - Tāds, kurā izpaužas šādas īpašības.
 - lietvārda nozīmē; formā: pamata pakāpe, noteiktā galotne **Cilvēks, ar kuru saista draudzība, mīlestība.**
 - Tāds, kas izturas miermīlīgi, nav nikns (par dzīvnieku).
- Tāds, kas prasmīgi, apzinīgi veic savu darbu, pienākumu.
- Tāds, kas atbilst, ir piemērots noteiktām izmantošanas prasībām (piemēram, par priekšmetu, telpu, vietu, parādību).

labs šķērssaites:
atvasinājums [nelabs](#)
skatīt arī [labā](#)

Figure 1: Public view of a Tēzauris.lv entry. The upper panel contains a search bar, the left side panel contains a box of related entries, a box of neighbours, and the boxes of results of the search in two other dictionaries. The right side panel contains a box of other links to related entries, an entry text in the middle.

At the top of the entry area is the header with the search box. On both sides of the entry area there are side bars with navigational items. The left side bar is devoted to the neighbourhood navigation, and the right side bar to the larger distance navigation.

The application user interface is constructed from a set of templates, which are rendered on the appropriate data fetched from the database.



(a) Public view of an entry *koks:1*.

(b) Editor’s view of the same entry *koks:1*.

Figure 2: Public and editor’s view of the same entry, with all blocks collapsed.

The editor’s view uses the same templates as the reader’s view to ensure WYSIWYG⁶, augmenting it with some additional icons for extra information and for initiating editing actions, thus ensuring that editors see as close as possible the look of the final entry (see Figure 2).

Where possible, editor tools allow to choose attribute values from pre-filled drop-down lists to ensure data consistency, as shown in Figure 3. The platform also provides means to link to multiple external data sources. Currently it is possible to add usage evidence from the Latvian National Corpora Collection (Saulite et al., 2022) and links to the Princeton WordNet (Paikens et al., 2023).

To ease the adding of corpora links, the editor’s view can be switched to a view where in addition to the meanings also any examples found in corpora are also visible, as shown in Figure 4. A separate editing window has been created for marking Wordnet links, where you can create links to both Latvian language wordnet synsets and Princeton wordnet synsets (Figure 5).

The platform hosts multiple (currently three) dictionaries with slightly different entry standards and requirements. The differences are mostly covered via conditional fragments

⁶ What You See Is What You Get

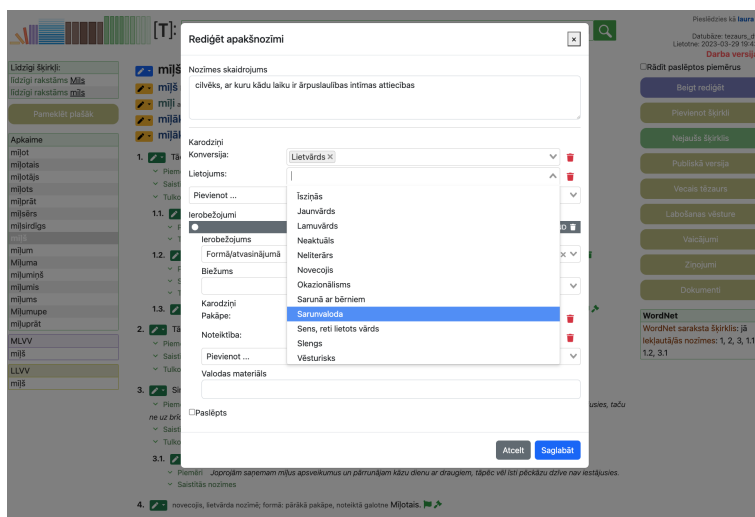


Figure 3: Editing dialog for a Sense, with a flag value selection dropdown opened.

in the templates or conditionally selected sub-templates. More dictionaries could be easily added.

Figure 4: Editor's view of a Tēzaurus.lv entry in mode which allows adding examples. The upper left side shows the list of word senses. The newly added example browser is on the right.

1.1.2 Search

The goal is to provide a simple and unified interface for different types of users through a single input field that can handle all their needs, much like a search bar on a modern web browser.

Show all Synset ×

bērnelis₁, bērņuks₁, ķipars₂, ute₃, bērns₁, knīpa₁, kverpis₁

bērnelis₁ Bērns.
bērņuks₁ Bērns.
ķipars₂ humoristiska ekspresīvā nokrāsa Bērns.

ute₃ sarunvaloda Bērns.

bērns₁ zēns vai meitene (aptuveni līdz 14 gadu vecumam).
knīpa₁ Maza meitene, mazs zēns.
kverpis₁ Bērns.

SYNONYM DICTIONARY

bērns - mazulis, mazais, bērņuks, bērnelis, ķipars

TRANSLATIONS

bērns - preadolescent, wean, bairn, youngling, babe, child, children, tad, kid, trick, baby, infant, fruit of the womb

LINKS

EXTERNAL LINKS:

(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"

HYPONYMS:

kniveris₁, mazpuika₁, puika₁, puisāns₁, puiskins₁, zēns₁, zeņķis₁, zeperis₁, puisis_{1,2}

kniveris₁ Zēns, puisēns.
mazpuika₁ Zēns.
puika₁ Zēns.
puisāns₁ Zēns.
puiškins₁ Zēns.
zēns₁ Vīriešu dzimuma bērns (aptuveni līdz 11 gadiem); arī pusaudzis.
zeņķis₁ Zēns, aptuveni skolas vecumā; arī pusaudzis.

meitene
 Only senses
 With subsenses
 English
 English*

SENSES

jaunmeitene₁
Jauniete.

Ganu meitene₁
ganumeita.

zemniekmeitene₁
Meitene, kas aug zemnieku ģimenē; arī zemniekmeita.

zvejniekmeitene₁

SYNSETS

jaunekle₁, jauniete₁, meitene₂, skuķis_{1,1}, meitēns₂, jaunmeita₁, mamzele₁

jaunekle₁ Jauniete.
jauniete₁ Sieviete vecumā starp pusaudzes un brīduma gadiem.
meitene₂ Jauniete.
skuķis_{1,1} Nepieredzējusi, arī nenopietna jauniete.
meitēns₂ Meitene (2).
jaunmeita₁ Jauna meitene.
mamzele₁ novecojis Jaunkundze.

LINKS

EXTERNAL LINKS:

(n) girl, miss, missy, young_lady, young_woman, fille

a young woman; "a young lady of 18"

HYPERONYMS:

sieva₂, sieviete₁
sieva₂ Sieviete.
sieviete₁ Cilvēku dzimuma būtne, kuras organisma morfoloģiskās un fizioloģiskās īpašības ir piemērotas bērnu dzemdēšanai; pieaugusi šāda cilvēku

Figure 5: Synset edit view for a sense of the word *bērns* ‘child’. The upper section contains synset information: included senses, synonyms and word translations on the left, a list of linked synsets by type on the right. The lower section allows to add new links by searching within Tēzaurs.lv or Princeton WordNet. The image shows Tēzaurs.lv results for a search query *meitene* ‘girl’. The first result column contains a list of senses that are not yet in Latvian WordNet. The second column contains a list of Latvian WordNet synsets. By clicking on any of these synsets, a list of all its links is displayed in the third column.

If a match to the search prompt is found, the corresponding entry is opened in the main area.

On the left side, several boxes may be shown each containing links to neighbour entries of various kinds of neighbourhood (see Figure 6a): homonym entries, homoforms in other entries, entries having inflectional forms similar to the search prompt, similarly spelt words, alphabetical neighborhood with adjacent entries of the same type (words, MWEs, word parts), and search word in other dictionaries.

The right side is reserved for graph relations between entries.

If in the first pass no satisfying entry (i.e., matching the search prompt) has been found, a deeper search is performed, looking also into the glosses etc., and the search results are presented grouped by match place and type, as shown in Figure 6b.

Homonīmi
koks ¹
koks ²
koks ³

Līdzīgi šķirkļi:
līdzīgi izrunājams Gogs
līdzīgi izrunājams kogs
pamatleksēma šķirkli koki
rakstības variants šķirkli kokss

Pameklēt plašāk

Apkaime
kokpiepe
kokpiesis
kokpile
kokpits
kokpūkaune
kokroze
kokrūpniecība
koks
Koksa
koksagīzs
koksagra
koksaldāja
koksankilometrs
koksartrīts
koksartrolistēze

MLVV
koks ¹
koks ²

LLVV
koks ¹
koks ²

Paplašinātā meklēšana

Meklējam **inteleks**.

▼ Atrasts vārdos (1):

- [inteleks:1](#)

▼ Atrasts vārdu savienojumos (1):

- mākslīgais [inteleks](#)

▼ Atrasts skaidrojumos (12):

- **racionālists** Cilvēks, kam raksturīga saprātīga attieksme pret īstenību; cilvēks, kura rīcību nosaka tikai vai galvenokārt [inteleks](#), prāts.
- **AI Mākslīgais [inteleks](#)** (angļu "Artificial Intelligence").
- **MI mākslīgais [inteleks](#)**.
- **polidispondilisms** Nepareiza vairāku skriemeļu attīstība, kuras dēļ attīstās punduraugums, vājš [inteleks](#) un turku seglu malformācija.
- **afāzija** Nespēja runāt vai runāto saprast, kaut gan nav traucēts [inteleks](#) un valodas aparāts; smadzeņu garozas saslimšana, kurā traucēta valodas impulsu koordinācija.
- **lisps** Programmēšanas valoda, gk. neskaitlisko uzdevumu risināšanai (loģiskais izvedums, dabīgās valodas, mākslīgais [inteleks](#)).
- **prāts** Psihisko norišu un personības īpašību kopums, kas rada iespēju apzināties, domāt, saprast un veidot īstenības atspoguļojumu, izzināt priekšmetu un parādību vispārīgās un būtiskās īpašības, gūt un izmantot pieredzi; arī [inteleks](#).
- **hondrodipplāzija** Skrimšļa veidošanās traucējumi ar disproporcionālu punduraugumu; mazs augums; liela galva ar dziļi ievilkto deguna sakni; mazi pirksti, rokas un kājas; liels vēders, dzija naba; dzimumorgānu attīstība normāla; [inteleks](#) labi attīstīts.
- **pasekls** Tāds, kam ir samērā, arī mazliet ierobežots [inteleks](#), intereses, jūtas.
- **racionāls** Tāds, kura attieksmi pret dzīvi nosaka tikai vai galvenokārt prāts, [inteleks](#) (par cilvēkiem).
- **sejas teleangiektāzija** teleangiektāzija bērniem ar hipofīzes priekšējās daivas attīstības traucējumiem (autosomāli recesīva pārmantošana); punduraugums, hipogonādisms, garīgā un fiziskā atpalcība; pirmais muža gadus uz sejas rodas teleangiektātiskā erītema, kas atgādina sarkano vilkēdi (Sauls starojuma ietekmē pastiprinās); uz lūpām bieži bulozi izsitumi; ilgstoši saglabājas infantila balss; [inteleks](#) normāls.
- **Binah** Viena no trijām Zefirotām, kas veido Kosmiskā Koka intelektuālo daļu, Dieva [inteleks](#).

(a) Results of basic search for the prompt *koks*. (b) Results of extended search for the prompt *inteleks*.

Figure 6: Basic and extended search.

Additionally, the word is looked up in other dictionaries hosted on the Platform, and in case of success the links are provided for opening the corresponding entry in sibling dictionaries.

1.2 Inflections and morphology

We generate and display inflection tables for lexemes that have their morphological paradigm specified, as illustrated in Figure 7. The inflections are generated by an external morphology engine (Pretkalniņa & Paikens, 2018) and fetched via an API call. The engine returns wordforms and certain lexical flags which are then used for both generation of inflection tables and searching for inflected word forms.

2. Data model of the dictionary

While previously the data model and tools were based on what the end user would see in a dictionary entry (document based model where the documents were dictionary entries), the current infrastructure is designed with a focus on a maintainable structured model – a graph which consists of lexical entities and links between them, thus avoiding duplication and enabling persistent links that stay consistent even if word senses are edited or moved. For example, multi-word entities used to be listed separately in the entry of words referring to it, duplicating the data with some accidental variation, but now both entries include the same entity.

darīt

darīt 3. konjugācijas darbības vārds; transitīvs ↕ Locišana

Īstenības izteiksme:

	Tagadne		Pagātne		Nākotne	
	Vsk.	Dsk.	Vsk.	Dsk.	Vsk.	Dsk.
1. pers.	<i>daru</i>	<i>darām</i>	<i>darīju</i>	<i>darījām</i>	<i>darišu</i>	<i>darīsim</i>
2. pers.	<i>dari</i>	<i>darāt</i>	<i>darīji</i>	<i>darījāt</i>	<i>dariši</i>	<i>darīsiet, darīsīt</i>
3. pers.		<i>dara</i>		<i>darīja</i>		<i>darīs</i>

Pavēles izteiksme: *dari* (vsk. 2. pers.), *dariet* (dsk. 2. pers.)

Atstāstījuma izteiksme: *darot* (tag.), *darišot* (nāk.)

Vēlējuma izteiksme: *darītu*

Vajadzības izteiksme: *jādara*

Figure 7: Expanded block with inflection information for the verb *darīt*

This design permitted us to start with a more classical, entry oriented data structure, coming from the paper age, and incrementally move towards a graph oriented data model.

This highly structured approach simplifies exporting data for various purposes. Currently, we have TEI⁷ for most dictionary data and LMF⁸ for WordNet related data. Additionally, an export to the PINI tool (Barzdins et al., 2020) has been developed for marking word senses in literary texts.

2.1 Core structure

The new data model of the dictionary (see Figure 8) consists of entities (Entry, Lexeme, Sense, Example, Synset) and links between them. The main root elements in the data model are Entry and Synset.

The nature of the new data model is a hybrid between a document-based and a graph-based model: some of the relations are represented as graph edges between entities (many-to-many, either symmetric or asymmetric), some others are based on one-to-many relations (Entry ← Sense, Entry ← Lexeme, Synset ← Sense, Sense ← Subsense, etc.). Symmetric links are used in WordNet between synsets to represent relations “anthonymy” and “similar”. Asymmetric links depict also the direction of a relation, e.g., “A derivativeOf B” tells that A is derived from B.

2.1.1 Entry

An Entry roughly corresponds to the entry in a traditional dictionary. However, most of the lexicographic information is delegated to other entities, and the entry itself serves just as the point joining these entities together. Lexemes, senses, examples, and sources of lexicographic information all are attached to the Entry. An entry can have a link to another entry if it is a derivative of a word with its own senses (**derivativeOf**) or an entry of a MWE that contains this word (**hasMWE**). In rare cases, a **seeAlso** link is used between an entry and another entry to indicate some kind of relationship between the words that

⁷ Text Encoding Initiative, P5 Guidelines, Chapter 9: Dictionaries, available: <https://tei-c.org/release/doc/tei-p5-doc/en/html/DI.html>

⁸ Global WordNet Association, guidelines for formats: <https://globalwordnet.github.io/schemas/#xml>

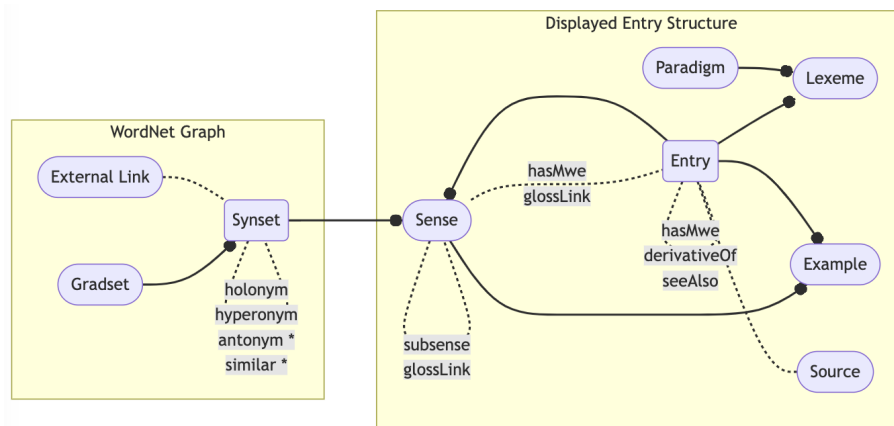


Figure 8: Conceptual data model. Continuous line: one-to-many relation (bullet denotes the singular end of the relation); dotted line: a many-to-many relation between entities, which can be symmetric (with *) or asymmetric.

are not in the derivation relations, nor in the WordNet-defined relations between word meanings.

2.1.2 Lexeme

Each Lexeme is built around one lemma, together with all information related to it. An entry has one or more lexemes attached to it. A lexeme may have a paradigm assigned to it. Paradigm has its own table in the database. Each lexeme has a type: main lexeme, spelling variant or derivation lexeme. Even though there is no direct database relation linking senses with lexemes, the lexeme type provides us information about how the senses in the entry are related to the lexemes in the same entry. By default, a sense in the entry provide definition for all main lexemes and spelling variants, but not derivation lexemes. However, with the use of Restriction (see 2.2.3) any given Sense can be targeted to a specific derivation lexeme.

2.1.3 Sense

An entry has one or more Senses and subsenses attached to it. The main content of a sense is its textual description (gloss). In order to fine-tune the meaning described, a sense may have one or more examples attached to it and one or more MWE entries linked to it, in which the word is used in this specific sense. Senses can be organized in a two level hierarchy – senses and subsenses. A gloss of a sense can have “anchor” links that create a link from a word used in a gloss to another entry or to a particular sense of another entry. “Anchor link” is asymmetrical. Examples of word usage where the word is used in a specific sense may be attached to the meaning. Sense is also an element in creating the Latvian WordNet, so a link to the WordNet synset can be made from it.

2.1.4 Synset

A WordNet core element is a Synset, which can be composed of either one or more Senses, usually coming from different dictionary entries. WordNet links are drawn between synsets, not directly between senses. Two or more synsets can be involved in a larger set “Gradset”. A synset can have one or more External Link attached, showing the relation between Latvian synset and a related entity in other, external lexicographical resources. Currently, synsets are linked to the corresponding Princeton WordNet synset, but links to other resources may be added later.

2.1.5 Example

An Example consists of a text fragment together with information about its origin. Normally, examples are attached to a sense, but the structure supports also examples on the entry level.

2.2 Supplementary non-tabular data

In addition to the core tabular data structure, each data item (entity or link) can be enriched with supplementary non-tabular information in the form of structured JSON data. This approach enables the inclusion of more detailed, complexly structured information, while also allowing for uncomplicated data model extension without altering the database structure. Certain elements of the JSON data (**Flags**, **StructuralRestrictions**) are predefined, while the data can be easily expanded with new components (e.g., **Pronunciations**, **Etymology**, **Normative**, **ImportNotices**, **sketchEngineTokenNum**, etc.). This approach also keeps open the option to move some parts of the JSON data over to relational database tables if such optimization needs should arise.

2.2.1 Paradigms

In this model, Paradigm is a named category that defines a set of inflection rules and flags which can be assigned one for each lexeme. Lexeme inherits certain properties, such as part of speech or grammatical gender, from the assigned paradigm, but each of these inherited properties can be overridden by flags defined at the lexeme level. Currently around one third of all single-word lexemes has a paradigm assigned, however, the desirable future state would be to have the paradigms for most if not all single word lexemes.

Providing paradigm for lexeme ensures that the morphological analyzer (Paikens et al., 2013) can be used to generate all inflectional forms for given word. These forms are further used to improve dictionary search and to generate inflection tables shown to the user.

2.2.2 Flags

A part of the payload information in data items is structured as Flags. A flag is a key-value pair, where the key is a descriptive name, and the value can be a string or a list of strings. The definition of a flag type usually contains a set of permitted values, however, free entry values are also supported; additionally, a flag type definition prescribes cardinality (single

value, or list of values) and scope (for which entity types the flag can be used). Flag type definitions are stored in the database, thus enabling to provide convenient UI components for entering/editing the flag assignments.

Currently, there are 64 predefined flag types, with 470 predefined values in total. Both flag types and flag values can be marked as deprecated, thus supporting evolution of the flag-set. Some examples of flag types from different aspects are: *POS*, *Conjugation*, *Transitivity*, *Tense*, *Pronoun type*, *Domain*, *Style*, *Language*, *Dialect features*.

2.2.3 Restrictions

In order to represent additional contextual grammatical or usage restrictions on some entities, the Restriction data structure has been created. These restrictions can be attached to any entity, besides the set of flags. These restrictions describe things like the fact that certain sense in the entry is used only for certain wordforms of the lexeme, or when lexeme is used in certain grammatical structure (see example in Figure 9) or that a certain lexeme might be used only in some of the forms its inflectional paradigm formally prescribes.

Each such restriction consists of 3 parts: restriction type, restriction frequency and restriction's "value" – a set of attribute-value flags. Restriction type broadly classifies all restrictions in several groups by their functioning (see below). Restriction frequency loosely describes how often this restriction is applicable, and currently their values are, e.g., **always**, **often**, **rare** and **unspecified** as inherited from retro-digitised paper dictionaries. However, we envision possible improvement here by switching to data-backed frequencies. The third part of the restriction structure is set of flags. Flags here are the same as described above (see Section 2.2.2) and they describe actual properties we want to restrict by, e.g., **Case=Nominative**. This part can also contain a free-text string that describes some kind of language material – either some phrase or certain word form.

Currently, the platform supports following 6 Restriction types:

- **togetherWith** – denotes usage together with certain parts-of-speech, lexemes or forms
- **inStruct** – denotes usage in certain structures, e.g., exclamation sentences
- **inForm** – denotes selection of certain inflectional forms or derivative of the lexeme in question
- **wordbuildingPart** – one of the restriction types meant specifically for entries describing parts of the words: this restriction is used for describing the other parts of the compound to be made
- **wordbuildingResult** – another one of the restriction types meant specifically for entries describing parts of the words: this restriction is used for describing resulting compound
- **overallFrequency** – for cases when certain sense is described as *rare* or *often* we decided to use restriction with this type and appropriate restriction frequency, but without any flags, not to duplicate restriction frequencies as flag values

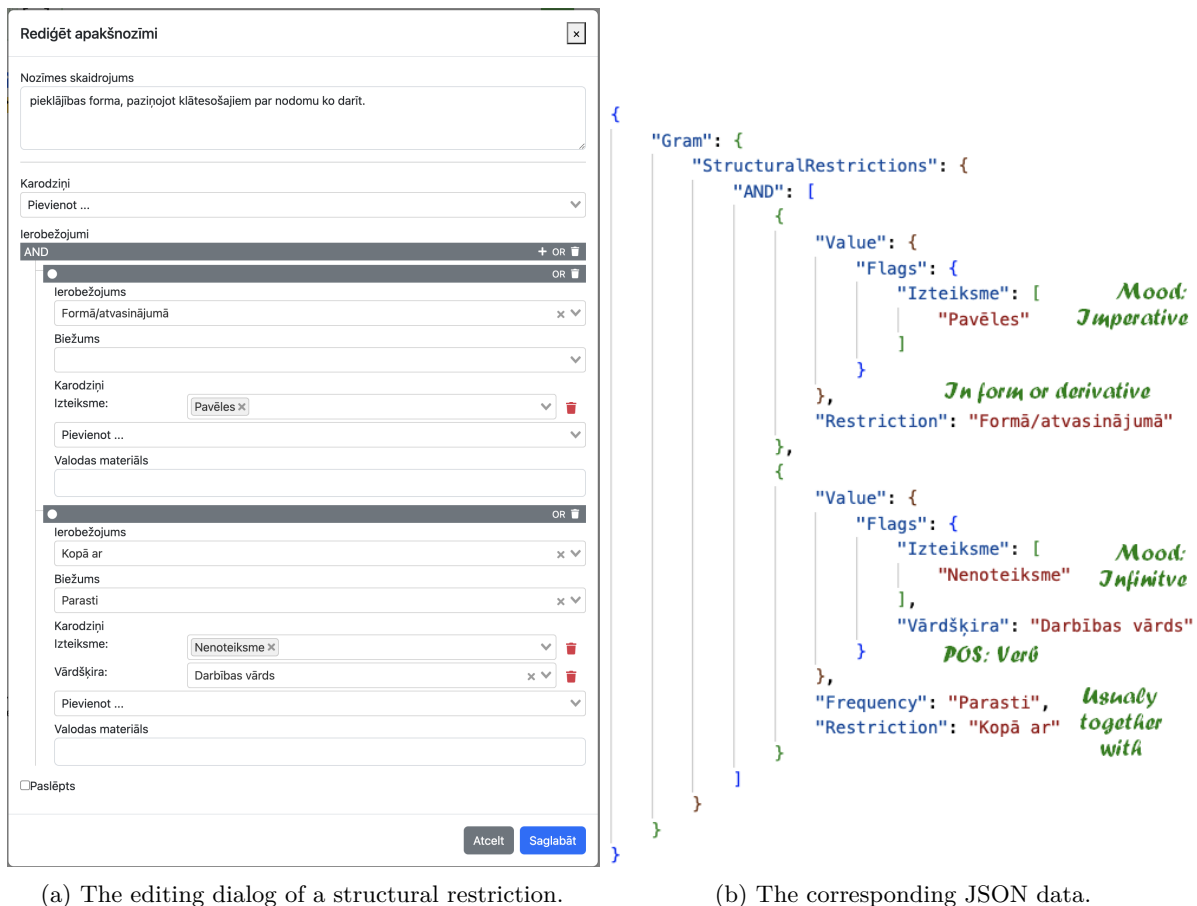


Figure 9: The restriction for sense 1.2 of the word *atļaut*: this sense is expressed with imperative form and usually in a phrase with an infinitive verb.

2.3 The Technical Stack

The application uses *node.js*⁹ as the application host, with *express.js*¹⁰ as the HTTP server, and *pug.js*¹¹ as the template engine for server side rendering. Mixins are extensively utilized for rendering repetitive components in the interface.

*vue.js*¹² forms are used as self-contained independent components providing property editing dialogues for each entity type. These forms communicate directly with the backend, which is responsible for data validation and persistence.

*PostgreSQL*¹³ is used as the database engine. Each entity type and link type has its own table in the database. Currently, the data model consists of approximately 40 database tables, including administrative and supporting tables. Operations of larger scale, such as the merging of two entries, are implemented as database procedures. Change logging is also implemented as a database trigger function.

⁹ <https://nodejs.org/en>

¹⁰ <https://expressjs.com/>

¹¹ <https://pugjs.org>

¹² <https://vuejs.org/>

¹³ <https://www.postgresql.org/>

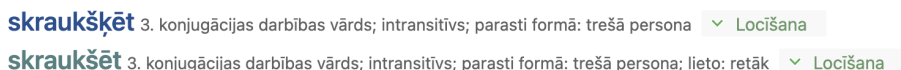
The dictionary application is being deployed on *Ubuntu Linux*¹⁴ as a *Docker*¹⁵ container, with *nginx*¹⁶ as a reverse proxy engine in front of it, and *PostgreSQL* installed directly on the server.

In publishing mode the application is currently serving a moderate workload of up to 200K requests per day. If future increases in workload would cause performance issues, there are several easy yet unexplored optimization possibilities. Additionally, if necessary, load balancing can be implemented across multiple servers. In editing mode, the app supports multiple named editors who can work concurrently.

3. Tools for lexicographic workflow

3.1 Verbalization of structured data

The grammatical and usage information for the dictionary entities (entries, lexemes, senses, and examples) is stored as structured data in the form of paradigms, flags and structural restrictions. The verbalization module generates a human-readable textual representation of this information. Verbalization builds upon atomic rules for simple flags, enhanced with aggregation rules for logical expressions, overriding rules for specific over general, prioritization rules for placing the most important facts at the beginning, etc. Sample results of verbalization are shown in Figure 10.



skraukšēt 3. konjugācijas darbības vārds; intransitīvs; parasti formā: trešā persona ▾ Locīšana
skraukšēt 3. konjugācijas darbības vārds; intransitīvs; parasti formā: trešā persona; lieto: retāk ▾ Locīšana

Figure 10: Verbalization results (in gray) for two lexemes.

3.2 Queries / Reports

To support the lexicographers, a query subsystem has been created, which allows to define reusable queries for finding entities satisfying a specific lexicographic criteria as well as data validation queries. This module enables users to define and reuse queries and presents the results as navigable tables or lists of dictionary items (an example see in Figure 11). The system supports both pure SQL queries and SQL+code queries. Currently, the system comprises around 100 queries of varying complexity.

3.3 Interface for bulk-editing

To support bulk-editing of some aspects in the lexicographer's work which cannot be fully automated, a special module has been created which presents a list of micro-tasks to the editor, who can select one of the quick choices, or open the entry for regular editing in the unclear cases (see Figure 12).

¹⁴ <https://ubuntu.com/>

¹⁵ <https://www.docker.com/>

¹⁶ <https://www.nginx.com/>

Sinonīmkopu saites: hiponīmija

Skaits: 3564

1 2 ... 4

Sākums	Saite	Beigas
(tēvu tēvu (arī tēvutēvu, tēvtēvu) laiki, arī seni laiki ₁)	hiponīms:hiperonīms	(bijušais _{1,3} , pagātne ₁)
(degt zilās ugunīs (arī liesmās, arī ar zilām liesmām) ₁)	hiponīms:hiperonīms	(degt ₁)
(sacelt kājās ₁)	hiponīms:hiperonīms	(pamodināt ₁)
(Rīgas pods ₁)	hiponīms:hiperonīms	(mērs ₁ , mērvienība ₁)
(dzīvsudraba (arī ūdens) staba (arī stabiņa) milimetrs ₁)	hiponīms:hiperonīms	(mērs ₁ , mērvienība ₁)
(dzīves (arī mūža) draudzene, laulātā draudzene ₁ , laulene ₁ , sieva ₁ , sievone ₁ , vecā ₃ , vecene ₂)	hiponīms:hiperonīms	(otrā puse ₂)
(manējā ₁)	hiponīms:hiperonīms	(dzīves (arī mūža) draudzene ₁ , laulātā draudzene ₁ , laulene ₁ , sieva ₁ , sievone ₁ , vecā ₃ , vecene ₂)
(ganu rags (arī taure) ₁)	hiponīms:hiperonīms	(pūšaminstrument ₁)
(zviedziens ₂)	hiponīms:hiperonīms	(smejas ₁ , smiekl ₁)

Figure 11: Result fragment of a query for hyponymy links – first and last columns contain link endpoints, middle column displays link type (direction).

Talka

Talkas vadlīnijas [šeit](#)

Skaits: 103



Šķirkļvārds	Leksēma tagad	ar mazo	Pogas
Atsevišķā studentu rota	Atsevišķā studentu rota	atsevišķā studentu rota	abc Abc Abc+Tx Abc+Vv
Bovera begonija	Bovera begonija	bovera begonija	abc Abc Abc+Tx Abc+Vv
Cedoārijas kurkuma	Cedoārijas kurkuma	cedoārijas kurkuma	abc Abc Abc+Tx Abc+Vv
Celza oreocerejs	Celza oreocerejs	celza oreocerejs	abc Abc Abc+Tx Abc+Vv
Conoska ligustrs	Conoska ligustrs	conoska ligustrs	abc Abc Abc+Tx Abc+Vv
Dalienu kņazistes	Dalienu kņazistes	dalienu kņazistes	abc Abc Abc+Tx Abc+Vv
Debesu manna	Debesu manna	debesu manna	abc Abc Abc+Tx Abc+Vv
Degēra morēnas	Degēra morēnas	degēra morēnas	abc Abc Abc+Tx Abc+Vv

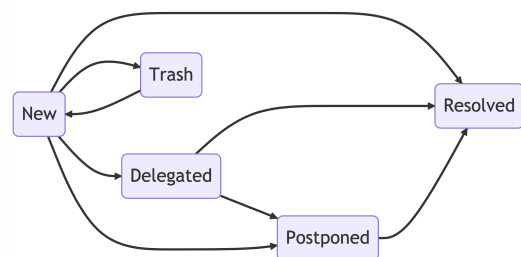
Figure 12: Bulk-editing interface. Task: capitalization of multi-word expressions; table columns contain entryword, current lexeme, task specific automatically provided potential correction and buttons for answering.

3.4 Collecting statistics

In the publishing mode, the application reports statistics on successful entry requests, as well as on failed (entry not found) entry requests. The log of not found requests is utilized to further enhance the dictionary content.

3.5 Collecting user feedback and suggestions

The platform provides a system that enables end-users to provide feedback and suggestions for any dictionary entry. All suggestions entered are stored in the database, and a simple workflow is provided to facilitate the feedback processing (see Figure 13a).



(a) Workflow for user feedback processing.

(b) A new feedback message embedded in the editor's view of an entry.

Figure 13: User feedback processing.

All feedback can be viewed either in a list by workflow state, or in the related entry as an embedded block with message and action buttons (see Figure 13b).

3.6 Keeping change history

In the editing mode, the application records all changes made to the dictionary at the entity level. This includes information such as who made the change, when it was made, the type of operation performed, and the data before and after the change. These logs (see Figure 14) enable editors to trace the change history and resolve any errors or misunderstandings that may arise during the editing process.

Šķirkļa izmaiņu vēsture

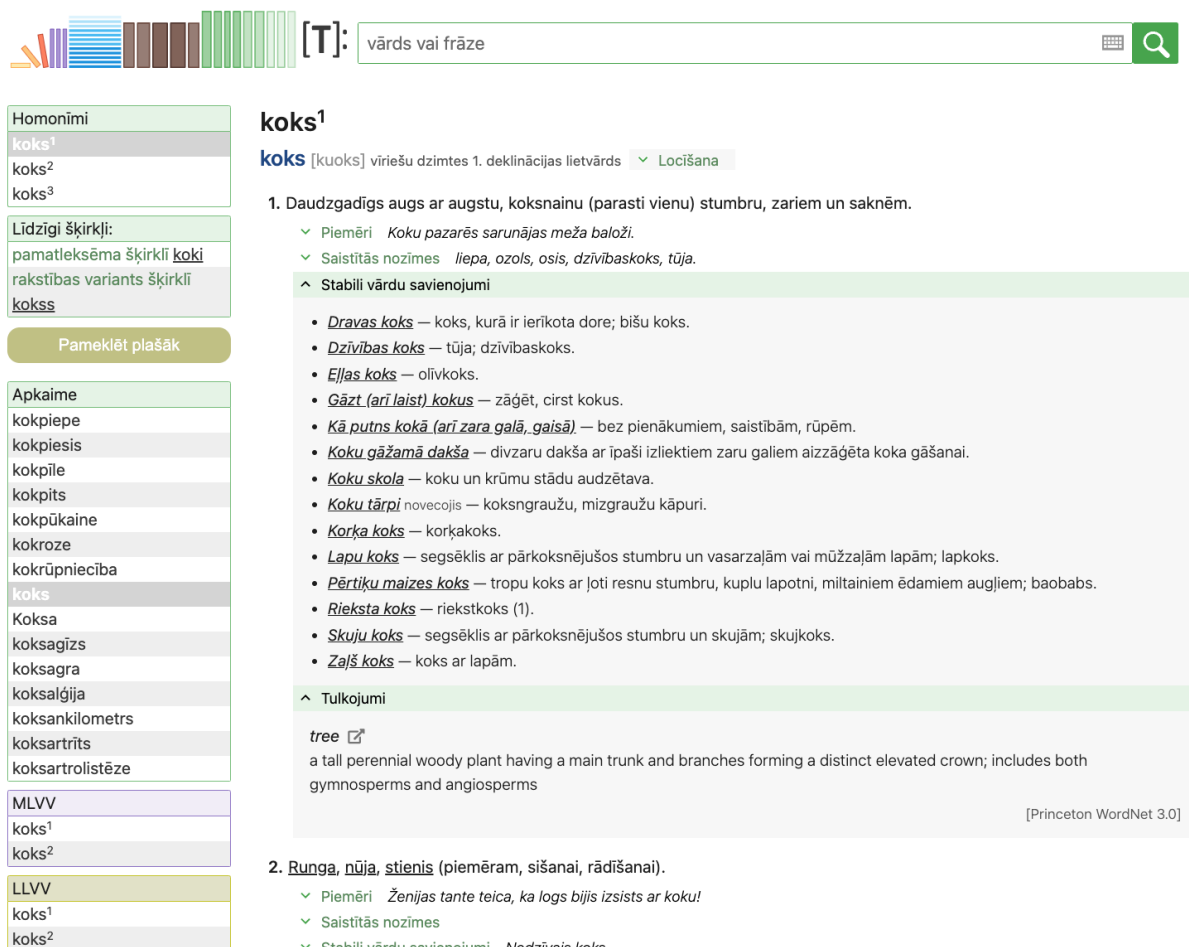
Šķirkli **nākt**:1 kopš datu ielādes ir veiktas 17 izmaiņas.

Kad	Kurš	Operācija	Kur	Ko
2023-03-17 17:19	laura	UPDATE	senses	Kļūt (kādam), nonākt (kādā stāvoklī,...
2022-12-09 15:38	agute	UPDATE	senses	Virzīties šurp, arī ierasties, lieko...
2022-06-09 15:46	script	UPDATE	senses	Braukt šurp (ar transportlīdzekli, p...
2022-06-09 15:16	script	UPDATE	lexemes	nākt
2021-08-24 11:08	agute	UPDATE	senses	Dzimt.
2021-06-28 12:35	agute	UPDATE	senses	Maksāt (2).
2020-12-04 02:16	system	UPDATE	senses	Maksāt (2).
2020-12-04 02:16	system	UPDATE	senses	Tikt ievietotam, novietotam (kur).
2020-12-04 02:16	system	UPDATE	senses	Veidoties, iestāties, iesākties (pie...
2020-12-04 02:16	system	UPDATE	senses	Tikt iegūtam (piemēram, par ražu, pa...
2020-12-04 02:16	system	UPDATE	senses	Tikt sūtītam šurp, atrasties ceļā šu...
2020-12-04 02:16	system	UPDATE	senses	Pakāpeniski rasties, kļūt intensīvāk...
2020-12-04 02:16	system	UPDATE	senses	Pakāpeniski iestāties (par parādībām...
2020-12-04 02:16	system	UPDATE	senses	Tuvoties, pakāpeniski iestāties (par...
2020-12-04 02:16	system	UPDATE	senses	Virzīties, braukt šurp (par transpor...
2020-12-04 02:16	system	UPDATE	senses	Peldēt šurp (parasti uz nārstošanas ...
2020-12-04 02:16	system	UPDATE	senses	Lidot šurp (parasti par putniem, kuk...

Figure 14: List of change history for the verb *nākt*

4. Use cases

The primary role of Tēzaur.lv is as a multi-functional online dictionary designed to meet the needs of a diverse range of users. We provide search results based on inflectional forms and spelling variants, as well as links to phonetically similar, alphabetically adjacent, and semantically linked words. For the entries of Latvian WordNet we also provide translations allocated to specific senses, which helps language learners and translators. To manage the extensive information, we use openable/closable blocks to display the data, rather than hiding it. We have also redesigned the interface with the understanding that the amount of data may continue to expand in the future. This allows each user to explore the data as deeply as they desire (see Figure 15). This approach makes Tēzaur.lv a valuable resource for a wide range of end-users, including language learners, students, translators, and the general population.



[T]: vārds vai frāze

Homonimi

- koks¹
- koks²
- koks³

Līdzīgi šķirkļi:

pamatleksēma šķirkli **koki**
rakstības variants šķirkli **kokss**

Pameklēt plašāk

Apkaime

- kokpiepe
- kokpiesis
- kokpile
- kokpits
- kokpūkaune
- kokroze
- kokrūpniecība
- koks
- Koksa
- koksagizs
- koksagra
- koksalgija
- koksankilometrs
- koksartrīts
- koksartrolietēze

MLVV

- koks¹
- koks²

LLVV

- koks¹
- koks²

koks¹

koks [kuoks] vīriešu dzimtes 1. deklinācijas lietvārds Locīšana


1. Daudzgadīgs augs ar augstu, koksnainu (parasti vienu) stumbru, zariem un saknēm.

- ✓ Piemēri *Koku pazarēs sarunājas meža baloži.*
- ✓ Saistītās nozīmes *liepa, ozols, osis, dzīvibaskoks, tūja.*

^ **Stabili vārdu savienojumi**

- *Dravas koks* — koks, kurā ir ierīkota dore; bišu koks.
- *Dzīvības koks* — tūja; dzīvibaskoks.
- *Eļļas koks* — olīvkoks.
- *Gāzt (arī laist) kokus* — zāgēt, cirst kokus.
- *Kā putas kokā (arī zara galā, gaisā)* — bez pienākumiem, saistībām, rūpēm.
- *Koku gāžamā dakša* — divzaru dakša ar īpaši izliektiem zaru galiem aizzāgēta koka gāšanai.
- *Koku skola* — koku un krūmu stādu audzētava.
- *Koku tārpi* novecojis — koksngraužu, mizgraužu kāpurī.
- *Korķa koks* — korķakoks.
- *Lapu koks* — segsēklis ar pārkoksnējušos stumbru un vasarzaļām vai mūžzaļām lapām; lapkoks.
- *Pērtiņu maižes koks* — tropu koks ar ļoti resnu stumbru, kuplu lapotni, miltainiem ēdamiem augļiem; baobabs.
- *Rieksta koks* — riekstkoks (1).
- *Skuju koks* — segsēklis ar pārkoksnējušos stumbru un skujām; skujkoks.
- *Zaļš koks* — koks ar lapām.

^ **Tulkojumi**

tree 

a tall perennial woody plant having a main trunk and branches forming a distinct elevated crown; includes both gymnosperms and angiosperms

[Princeton WordNet 3.0]

2. **Runga, nūja, stienis** (piemēram, sišanai, rādīšanai).

- ✓ Piemēri *Ženijas tante teica, ka logs bijis izsists ar koku!*
- ✓ Saistītās nozīmes
- ✓ Stabili vārdu savienojumi *Nedzīvais koks.*
- ✓ Tulkojumi

2.1. vēsturisks; formā: daudzskaitlis Soda veids — sitieni ar nūju, riksti vai pātagu.

- ✓ Saistītās nozīmes

Figure 15: The public view of the entry *koks* ('tree') with opened interface panels for MWEs and translations.

Secondly, the created platform provides wide functionality for dictionary editors. Despite the fact that the information to be included in the dictionary can be quite extensive and

structurally complex (as in cases with restrictions), according to Tēzaurs.lv editorial team, the dictionary editor is quite convenient, intuitive and user-friendly. Furthermore, it can be adapted for the creation and development of other dictionaries with minimal modifications, as demonstrated by its successful use with two other dictionaries. Key advantages and benefits:

- The system efficiently handles even relatively large dictionaries
- Editor authentication/authorization, change history tracking, and parallel work capabilities (with a limitation that multiple editors cannot edit the same entity (lexeme, sense, etc.) at the same time)
- Queries and reports enable data review from diverse angles, with future plans to support more advanced searches (e.g., using regular expressions)
- Where feasible, data entry utilizes predefined lists to minimize errors by editors, and incorporates routinely updated automatic error checks to accommodate emerging requirements
- User feedback storage and processing, including content error reporting as well as suggestions for new entries or clarifications, is fully integrated into the system, eliminating the need for e-mail communication
- Offers a visually appealing web-based interface, reducing compatibility issues across different operating systems
- Streamlines the creation of interfaces for bulk data processing and facilitates regular cleanups to address more uniform issues (e.g., sorting MWEs by types such as toponyms, taxons, other proper names, etc.)

4.1 Support for multiple dictionaries

Currently, the platform is used for 3 different dictionaries already published online: for Tēzaurs.lv, for the actively developed MLVV (Dictionary of Modern Latvian) (Zuicena, 2012), and for a retro-digitised version of the earlier authoritative dictionary of Latvian – LLVV (Dictionary of Standard Latvian)¹⁷. All 3 dictionaries are served from the same application code, pointing to different databases. The only differences between them are the used style sheets and some conditionally excluded features that are not required for a specific dictionary.

In general, the platform is designed to be extensible and adaptable for a multitude of future uses, too. We are starting to develop a Latgalian dictionary on this platform, and hope in future to add several retro-digitised dictionaries. While we focus more on Latvian, the core platform itself is language independent as long as someone translates the user interface, updates the flag sets, the morphological analyser integration, the verbalization, and if target language uses similar dictionary structure.

5. Conclusions and future work

The choice of a hybrid data model has permitted to evolutionary move from an entry-oriented view towards more graph-oriented data structures, as well as to support dictionaries of various level of formalization and improve the formalization of Tēzaurs.lv itself.

¹⁷ <https://llvv.tezaurs.lv>

When building such a diverse and multi-functional lexical resource, precise and task specific tool support turned out quite crucial even if our team is quite small. The searching and error-checking abilities of the developed system majorly improved both the speed of content creation and the quality of the result.

Another advantage of this approach is its flexibility to enable supplementary micro-tools which build upon the shared core data model. Currently the synthesizer of morphological forms is integrated in this manner, but in future we would like to add other micro tools as well, such as generation of pronunciation samples. We have also recently started a project on extending Tēzaurus.lv with additional lexicographic data (namely, etymology and derivation links) and this platform enables us to include them as extra information in a shared lexical resource, instead of creating a separate resource like Derinet (Vidra et al., 2019) which afterwards could diverge from the continuously maintained dictionary.

Future platform improvements include extending the graph related features of the data model both on the data model level and on the visualization level. We plan to extend the reach of this work by publishing the source code of the platform under the GPL licence. We also plan to use Tēzaurus.lv platform to host even more dictionaries, both including other retro-digitised dictionaries and providing a platform for Latvian researchers to create new digital dictionaries.

We hope that this experience will be useful for other researchers building lexical resources and tools for maintaining them.

6. Acknowledgements

This work was supported by the Latvian Council of Science, project “Advancing Latvian computational lexical resources for natural language understanding and generation” (LZP-2022/1-0443) and State Research Programme project “Research on Modern Latvian Language and Development of Language Technology” (VPP-LETONIKA-2021/1-0006). We also thank the anonymous reviewers for their input in improving this paper.

7. References

- Barzdins, G., Gosko, D., Cerans, K., Barzdins, O.F., Znotins, A., Barzdins, P.F., Gruzitis, N., Grasmanis, M., Barzdins, J., Lavrinovics, U., Mayer, S.K., Students, I., Celms, E., Sprogis, A., Nespore-Berzkalne, G. & Paikens, P. (2020). Pini Language and PiniTree Ontology Editor: Annotation and Verbalisation for Atomised Journalism. In *Proceedings of the 17th Extended Semantic Web Conference (ESWC): Posters and Demos*. URL https://preprints.2020.eswc-conferences.org/posters_demos/paper_281.pdf.
- Danovskis, G. (2014). *Leksikogrāfa rīks kļūdu meklēšanā*. Kvalifikācijas darbs, Latvijas Universitāte.
- Fellbaum, C. (ed.) (1998). *WordNet: An electronic lexical database*. MIT Press.
- Paikens, P., Grasmanis, M., Klints, A., Lokmane, I., Pretkalnina, L., Rituma, L., Stade, M. & Strankale, L. (2022). Towards Latvian WordNet. In *Proceedings of the 13th Language Resources and Evaluation Conference (LREC)*. pp. 2808–2815. URL <http://www.lrec-conf.org/proceedings/lrec2022/pdf/2022.lrec-1.300.pdf>.
- Paikens, P., Gruzitis, N., Rituma, L., Nespore, G., Lipskis, V., Pretkalnina, L. & Spektors, A. (2019). Enriching an Explanatory Dictionary with FrameNet and PropBank Corpus

- Examples. In *Proceedings of the 6th Biennial Conference on Electronic Lexicography (eLex)*. pp. 922–933. URL https://elex.link/elex2019/wp-content/uploads/2019/09/eLex_2019_52.pdf.
- Paikens, P., Klints, A., Lokmane, I., Pretkalnina, L., Rituma, L., Stade, M. & Strankale, L. (2023). Latvian WordNet. In *Proceedings of the 12th Global Wordnet Conference (GWC2023)*. To be published.
- Paikens, P., Rituma, L. & Pretkalnina, L. (2013). Morphological analysis with limited resources: Latvian example. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA)*. URL <http://stp.lingfil.uu.se/nodalida/2013/pdf/NODALIDA24.pdf>.
- Pretkalniņa, L. & Paikens, P. (2018). Extending Tēzaurus.lv Online Dictionary into a Morphological Lexicon. In *Human Language Technologies—The Baltic Perspective*. IOS Press, pp. 120–125.
- Rambousek, A., Jakubíček, M. & Kosem, I. (2021). New developments in Lexonomy. In *Electronic lexicography in the 21st century. Proceedings of the eLex 2021 conference*. pp. 455–462. URL https://elex.link/elex2021/wp-content/uploads/2021/08/eLex_2021_28_pp455-462.pdf.
- Saulite, B., Dargis, R., Gruzitis, N., Auzina, I., Levane-Petrova, K., Pretkalnina, L., Rituma, L., Paikens, P., Znotins, A., Strankale, L., Pokratniece, K., Poikans, I., Barzdins, G., Skadina, I., Baklane, A., Saulespuren, V. & Ziedins, J. (2022). Latvian National Corpora Collection – Korpuss.lv. In *Proceedings of the 13th Language Resources and Evaluation Conference (LREC)*. pp. 5123–5129. URL <http://www.lrec-conf.org/proceedings/lrec2022/pdf/2022.lrec-1.548.pdf>.
- Simões, A., Salgado, A., Costa, R. & Almeida, J. (2019). LeXmart: A Smart Tool for Lexicographers. In *Proceedings of eLex 2019 conference*. pp. 453–466. URL https://elex.link/elex2019/wp-content/uploads/2019/09/eLex_2019_25.pdf.
- Spektors, A., Auzina, I., Dargis, R., Gruzitis, N., Paikens, P., Pretkalnina, L., Rituma, L. & Saulite, B. (2016). Tezaurus.lv: the largest open lexical database for Latvian. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*. Portoroz, Slovenia, pp. 2568–2571.
- Vidra, J., Žabokrtský, Z., Ševčíková, M. & Kyjánek, L. (2019). DeriNet 2.0: Towards an All-in-One Word-Formation Resource. In *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology*. Prague, Czechia: Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, pp. 81–89. URL <https://aclanthology.org/W19-8510>.
- Zuicena, I. (2012). Ilustratīvā materiāla atveide “Mūsdienu latviešu valodas vārdnīcā”. *Baltistica*, (8), pp. 181–188. URL <http://www.baltistica.lt/index.php/baltistica/article/view/2120/2026>.

Novel Slovenian COVID-19 vocabulary from the perspective of naming possibilities and word formation

Senja Pollak¹, Ines Voršič², Boris Kern³, Matej Ulčar⁴

¹ Jožef Stefan Institute, Ljubljana, Slovenia

² Faculty of Education and Faculty of Arts, University of Maribor, Slovenia

³ ZRC SAZU, Fran Ramovš Institute of the Slovenian Language, Ljubljana Slovenia and University of Nova Gorica, School of Humanities, Nova Gorica, Slovenia

⁴ University of Ljubljana, Faculty of Computer and Information Science, Ljubljana, Slovenia
E-mail: senja.pollak@ijs.si, ines.vorsic@um.si, boris.kern@zrc-sazu.si, matej.ulcar@fri.uni-lj.si

Abstract

We analyze a sample of novel Slovenian vocabulary related to COVID-19, focusing on naming possibilities and word-formation processes. We grouped previous descriptions of COVID-19 vocabulary and extended the list with a semi-automated selection based on embedding-based keyword expansion. In terms of naming possibilities, the analysis shows that a large majority of COVID-19 lexemes were created through derivation, showing high productivity and language vitality in Slovenian, and that a smaller number of examples are set phrases and neosemantisms, as well as explicit borrowings, whereas calques were not a productive strategy. From the point of view of the word-formation system, it is mainly possible to distinguish infix compounds, ordinary derivatives, and compositions. The most productive substructure is the root morpheme *korona*, which produces most of the infix subordinate compounds, but also higher-order adjectival derivatives (e.g., *koronski* ‘corona’) and compositions (e.g., *protikoronski* ‘anti-corona’). Otherwise, infix subordinate compounds turn out to be the most productive word-formational type. The most productive derivatives are adjectival and nominal derivatives with the suffix *-ost*, and these are also the ones that show the most frequently confirmed combinatorics of the suffix *-en-* + *-ost*.

Keywords: COVID-19; embeddings; naming possibilities; word formation; formant combinatorics

1. Introduction

The COVID-19 pandemic has fundamentally altered our reality—and with it our linguistic reality. In this article, we extract and analyze a sample of novel Slovenian vocabulary related to COVID-19, focusing on naming possibilities and word-formation processes. Expansion of the lexicon due to pandemics is not specific to COVID-19. As summarized by Gustilo et al. (2021), the words *epidemic* and *pandemic* are related to the seventeenth-century plague in Europe, and *quarantine* was first used in the fourteenth century to describe the forty-day period during which ships were in isolation before landing during the Black Death.

This changed linguistic reality has been characterized in particular by the penetration of a large number of terms into common usage that have thus been subjected to determinologization, the emergence of neologisms, occasionalisms, and neosemantisms that have been subjected to a process of accelerated Slovenization, and the imposition of particular lexical variants.

Identifying and analyzing new vocabulary is of high importance from several aspects. First, from the lexicographic perspective, adding novel lexis is of crucial importance, both in terms of the synchronic description of a language, as well as from the point of view of the historical character of a special pandemic era. The goal of this study is not only to identify novel vocabulary, but especially to understand different naming possibilities and word-formation processes, which are signs of language vitality and are highly interesting from a linguistic perspective.

Our work groups existing descriptions of COVID-19 vocabulary and applies natural language processing methods to extend the dataset by extracting candidates for novel Slovenian vocabulary related to COVID-19. Specifically, we trained a fastText embedding model on a dataset of COVID-19 news articles from the initial period of the COVID-19 pandemic, and then, using seed words related to COVID-19 and keyword expansion via embedding of nearest neighbors, we extended this initial set. The resulting material is used for manual analysis of COVID-19 keywords in terms of naming possibilities and word-formation processes.

This article is structured as follows. Section 2 presents the background, including naming possibilities and word-formation processes in Slovenian, followed by related work in Section 3. Next, we introduce the methodology, including the natural language processing approach used for lexicon extraction, in Section 4. In Section 5, we present the findings and analysis from the perspective of naming possibilities and word-formation processes. Finally, the conclusions and plans for further work are presented in Section 6.

2. Background: naming possibilities and word formation in Slovenian

2.1 Naming possibilities

A naming typology for Slovenian lexemes was proposed by Ada Vidovič Muha (2013: 23–25) in her work *Slovensko leksikalno pomenoslovje* (Slovenian Lexical Semantics). Below, we present a slightly adapted naming typology. When a new denotatum or a need for a new denotation arises in a language, we can 1) search for naming possibilities in the language itself or 2) borrow from a foreign language. When looking for possibilities in the language itself, the possibilities are 1.1) a simplex or a set

phrase (e.g., *roka* ‘hand’, *osebna izkaznica* ‘ID’), 1.2) a derivative (e.g., *nalivnik* ‘fountain pen’), or 1.3) a neosemantic term (e.g., *hrošč* ‘bug’). In the case of borrowings from a foreign language, this borrowing may be 2.1) disguised as a calque (e.g., *kolidž* ‘college’, *strežnik* ‘server’), or 2.2) expressed (*powerpoint* ‘Powerpoint (presentation)’, *halloween* ‘Halloween’).

Further, Vidovič Muha’s (2013) typology of naming possibilities in the language itself is further divided into two groups: simplex words (e.g., *roka* ‘hand’) and non-simplex words, which are further divided into set phrases, derivatives, and neosemantisms. In the case of acquisition from foreign languages, she distinguishes between disguised borrowings with groups of a) denotational and b) semantic calques (from classical languages, from a lingua franca, and from other languages), and explicit borrowing, with a) non-adapted, b) semi-adapted, and c) systemic acquisition. In citation, the word retains its form in the source language; in semi-quotation, there is a partial adaptation to the recipient language (especially in inflexion); and, in systemic acquisition, the word is completely integrated into the formal system of the recipient language.

2.2 Word formation in Slovenian

As a branch of linguistics, word formation is used to analyze the vitality of a language’s lexicon and to chart the course of linguistic development. Moreover, word formation facilitates the formation of new words at two levels: linguistically described, predictably formative, and transformative processes as well as systemically unpredictable word-formation patterns. Modern Slovenian word-formation theory includes derivation by suffixation, derivation by prefixation, and compounding, among the traditional word-formation processes. Current systemic word formation is briefly presented by Plemenitaš, Stramljič Breznik & Voršič (2020), who conclude that in Slovenian suffixation is the most productive word-formation pattern, with more than 300 suffixes used for creating nouns, adjectives, and verbs. The majority of suffixes are used for the formation of nouns, which can be masculine, feminine, or neuter. Adjectives can be formed with approximately 70 affixes. Verbal word-formation, on the other hand, uses only 15 suffixes. Prefixation, including foreign prefixes, uses 14 nominal, 4 adjectival, 20 nominal and adjectival, and around 40 verbal prefixes (Toporišič, 2000: 142–234).

The most productive word-formational process in Slovenian is nominal suffixation. Research shows that the majority of Slovenian words are still formed through nominal suffixation (Stramljič Breznik, 2005). Nouns can be derived from verbs, adjectives, and other nouns through suffixation or prefixation. Denominal derivation of nouns also includes derivation via prepositional phrases. Verbs can be derived from nouns, adjectives, interjections, other verbs, and prepositional phrases through suffixation or prefixation. Prefixation in verbs typically involves deverbal derivation

from bases containing prepositional phrases or from prepositional verbs. Adjectives can be derived from nouns, verbs, adverbs, and other adjectives mainly through suffixation. Prefixation in adjectives typically involves denominal derivation via prepositional phrases. Adverbs can be derived from nouns, adjectives, verbs, other adverbs, and prepositional phrases. Adverbial derivation mainly uses suffixation, yielding the semantic categories of place, time, manner, and quantity.

There are also word-formation patterns that are unpredictable from a formative and transformative point of view; that is, non-systemic formations cannot be assigned to a syntactic stem or be unequivocally morphemized into a word-formation stem and an affix, given the unpredictability of the number and the fact that affixation words can be different parts of speech. Non-systemic formations are also distinguished from systemic formations by their function. Their central purpose is not a naming necessity or a lexical gap, but a striving for originality and attractiveness. As such, they deliberately break the laws of word formation. Among the non-systemic word-formation processes, the following are recognized in current Slovenian: blends, back formations, abbreviations, bicapitalizations, and graphoderivatives; that is, formations enriched with graphic elements (Voršič, 2010).

3. Related work

3.1 Studies of COVID-19 vocabulary in Slovenian

The COVID-19 pandemic period has had a profound impact on our lives, and it has certainly had a linguistic impact—and Slovenian is of course no exception. There are a multitude of linguistics articles on this impact for individual languages, but in this article we limit ourselves to a brief summary of works analyzing current COVID-19 vocabulary in Slovenian.

The lexicographic portal Fran.si, published by the Fran Ramovš Institute of the Slovenian Language, launched the subpage Fran.si, COVID-19 Version (7.1) in early April 2020. It “brings together the most important new and previously published dictionary entries and language advice, provided by the Language Advising Service, in the context of the COVID-19 epidemic and the novel coronavirus. In addition to current vocabulary and orthographic and terminological notes, the material also contains a thematic overview of the history and etymology of pandemic-related vocabulary, as well as contagion-related terms from Slovenian dialects.” The most relevant part of the material, which was also included in this research, is new vocabulary from *Sprotni slovar slovenskega jezika* (Growing Dictionary of the Slovenian Language, GDSL).

This is a dictionary that contains living, newer words not yet registered in dictionaries, and at the same time contains the latest, emerging meanings of words

already registered. “The core of the dictionary consists of words not yet registered in dictionaries and whose use has been confirmed by corpus material in recent years, supplemented by suggestions from language users. Because these suggestions are usually relatively up to date, the glossary also contains words that are not (yet) present in current (time-limited) corpora of Slovenian, but whose use has been registered in other (especially digital) sources” (Krvina, 2023). The vocabulary described can be categorized into two groups: a) the lexicon already established whose use increased significantly or which acquired new meanings during the pandemic, and b) the lexicon that was newly created upon the emergence of the novel coronavirus and the associated pandemic. Among the words in the first group, the word *korona* should be mentioned first. This word was already introduced in *Slovar slovenskega knjižnega jezika* (Dictionary of the Slovenian Standard Language 2, SSKJ 2) with the meanings “1. *music* ‘a semicircle with a dot, indicating an indefinite prolongation of a note or a pause [i.e., *fermata*]’, and 2. *astronomy* ‘a layer of the Sun’s atmosphere, which passes into interplanetary space’.” While the word *koronavirus* ‘coronavirus’, which came into Slavic languages as an integral English borrowing (Będkowska-Kopczyk & Łaziński, 2020), is already recorded in SSKJ 2 with the meaning of ‘virus of the family *Coronaviridae*’, the word *korona* has acquired a new meaning from a dictionary point of view and is explained in GDSL (*Sprotni slovar slovenskega jezika*) as ‘a coronavirus, in particular the highly contagious SARS-CoV-2, or a disease characterized by inflammation of the upper respiratory tract, in the severe form a pneumonia, caused by this virus’. It is also used adjectivally to mean ‘that which is related to this virus or to the economic, social, or health consequences of an epidemic of the disease caused by this virus’ (as a synonym of the derivative *koronski* ‘corona’), and it can be considered a converse derivative in a word-forming sense. Otherwise, *korona* is one of the most productive bases for neologisms (see also Stramljič Breznik, 2021). The lexeme *koronavirus* also spread in phrasal usage during the pandemic as *novi koronavirus* ‘a coronavirus, in particular the highly contagious SARS-CoV-2, or a disease characterized by inflammation of the upper respiratory tract, in the severe form a pneumonia, caused by this virus [i.e., novel coronavirus]’. Otherwise, there are two more synonymous terms for the disease; namely, *SARS-CoV-2* (severe acute respiratory syndrome coronavirus 2) and *COVID-19*. Both are also listed in GDSL.

In addition to the formations mentioned above, GDSL also records other words that increased in frequency or acquired a new or at least broader meaning during the coronavirus outbreak. These include *bolezen* ‘illness’, in the phrases *pridružena bolezen* and *spremljajoča bolezen* ‘comorbidity, usually a chronic disease, which is already present in the patient at the onset, the start of treatment for another disease’); *distanca* ‘distance’, in the phrase *socialna distanca* ‘1. weak or less intense interaction between the usually dominant group and other groups of people due to personal, social, economic differences [i.e., social distance]; 2. avoiding non-necessary physical contact with others in order to prevent spreading of the virus’; *govorec* ‘one that represents an authority, organization, or individual by presenting its views and

decisions in public [i.e., spokesperson]'; *helikopterski* 'helicopter' in the phrase *helikopterski denar* 'money granted by a state or a community of states in particular for individuals or companies in times of adversity or emergency to promote economic growth'; *izolacija* 'isolation', in the phrase *kohortna izolacija* 'isolation in which several patients with the same pathogen are placed in the same room'; *obhajilo* 'communion', in the phrase *duhovno obhajilo* 'a union with Christ without ingesting the consecrated host by focusing on him and longing for him'; *omejevanje* 'limitation', in the phrase *omejevanje socialnih stikov* 'avoidance of non-essential physical contact with others to prevent the spread of infection'; *pacient* 'patient', in the phrases *številka ena* and *ničti pacient* 'whoever in a particular area or country is the first to contract a communicable, usually highly contagious disease [i.e., patient zero]'; and *testirati* 'to carry out a procedure to determine the presence of a disease-causing agent or of a particular substance, active ingredient [i.e., to test]'. Among the formations already recorded in SSKJ 2 is *samoizolacija* 'isolation, closing oneself off from others, usually of one's own accord [i.e., self-isolation]'. The same formation with the meaning 'quarantine in the face of very probable infection, which individuals spend at home by order or choose to do so themselves because of their responsibility towards others' (GDSL) is cited in word-formation analysis as a higher-stage derivative of the verb *samoizolirati se* 'to self-isolate'. This, interestingly enough, is a verbal infix compound not previously recorded in the dictionary.

Another study on the coronavirus pandemic vocabulary was performed by Stramljič Breznik (2021). In her study, she analyzed the development of newly created words and "tried to evaluate the importance of word-formative processes involved in their coinage" (Stramljič Breznik, 2021: 321). The analysis includes new vocabulary found on the subpage Fran.si, COVID-19 Version (7.1), as well as random searches for lexis from various websites and "less formal social media" (Stramljič Breznik, 2021: 321). In the article, the author focuses on the compounds with the constituent *korona* 'corona', which are particularly problematic in Slovenian from an orthographic point of view—there is a dilemma as to whether to write them as one word or as two (e.g., *koronavirus* or *korona virus* 'corona virus'). The material the analysis is based on shows that in 74% of cases writing them together is chosen, which is quite surprising considering that the general tendency in the language for this type of compound is to write them separately. This was undoubtedly influenced by the advice of the Language Advising Service, which advocated writing these as compounds (Weiss & Dobrovoljc, 2020), published as early as March 2020. In this article, special attention is paid to word-formation productivity of new COVID-19 vocabulary and analysis of COVID-19 occasionalisms. COVID-19 vocabulary contains a large number of expressive lexemes, which the author explains by the fact that the epidemiological situation has brought with it a series of mental, social, and economic hardships that have also had a linguistic impact.

Last but not least, Voršič (2022) focused on ad-hoc formations with an expressive association with the various social consequences of the pandemic. Ad-hoc formations

are words that are formed through the most current word-formation processes and reflect the creative flexibility of a language. The study moves away from neologisms formed due to lexical gaps, focusing instead on ad-hoc formations with an expressive association with the various social consequences of the coronavirus pandemic (Voršič, 2022: 265).

3.2 COVID-19 and word formation

The closest research to this article is a corpus study by Gustilo et al. (2021), who focused on meanings and word-formation processes of pandemic-related lexemes across English varieties, also leveraging a news corpus. They identified COVID-19 terms in the News on the Web (NOW) English corpus and classified them as compounds or blends. Specifically, they differentiated between revitalized compounds, blends, and new formations.

Word formation related to COVID-19 neologisms was also addressed in Asif et al. (2020), who analyzed neologisms related to COVID-19 from various text sources, including social media, where in terms of word-formation processes the neologisms most frequently corresponded to compounds, abbreviations, acronyms, and blends. Compounding, blending, and affixation were the most frequent word-formation processes identified in a study focusing on COVID-19 by Akut (2020), and Al Salman and Haider (2021) identified compounding, clipping, blending, acronyms, and other dual word-formation processes in their study of lexemes from various online sources.

3.3 Natural language processing and COVID-19 vocabulary

Natural language processing methods have been used previously for analysis of COVID-19 texts and related vocabulary. For example, Lei et al. (2021) studied the emergence of COVID-19 neologisms in Chinese, based on data from the Baidu Index. They followed the dynamics of the usage of various words for COVID-19, grouped into five categories (official, stigmatizing, English abbreviations, etc.), during the first six months of the pandemic. Wang and Huang (2021) compared the usage of terms relating to contact prevention and social distancing in Chinese and English in two cities, Hong Kong and Guangzhou. They analyzed how cultural differences affected the evolution of social distancing terms during the COVID-19 pandemic in the two cities.

Another line of research focuses on named entity recognition. Truong et al. (2021) present a COVID-19 domain-specific dataset for Vietnamese, with annotated named entities, including epidemic-specific entity types, and they implement several baselines for this task. COVID-19 was also one of the corpora included in the Third Slavic Cross-Lingual Challenge on Recognition, Normalization, Classification, and Linking of Named Entities across Slavic Languages (Piskorski et al. 2021).

There are also works exploring semantic shifts related to COVID-19. Montariol et al. (2021) propose a novel semantic change detection method using contextual embedding cluster distribution comparison and apply it to a corpus of COVID-19 news. In another work, Kellert and Zaman (2022) introduced a novel dataset focusing on lexical change triggered by the COVID outbreak and compare various types of analyses capable of capturing linguistic change; namely, relative frequency analysis, n -gram analysis, lexical change analysis based on word embeddings, and topic modeling. They show that changes of word distributions in topics can provide insights into changes in words' pragmatic meanings.

4. Methodology

Our methodology consists of the following steps: training the embedding model on a news corpus, seed word selection related to COVID-19, expanded COVID-19 candidate vocabulary extraction via embedding nearest-neighbor extraction, vocabulary filtering, and manual selection of the final COVID-19 dataset used for fine-grained analysis in form of naming possibilities and word formation.

First, we selected words related to COVID-19 to be used either directly for analysis, or for embedding-based expansion. We used the list of COVID-19 vocabulary from GDSL (Krvina, 2014–). Next, we used the COVID-19 vocabulary from the CJVT Language Monitor (Kosem et al., 2021) and, third, the list of COVID-19 vocabulary of occasional words collected by Voršič (2022). The resulting joint list contains 186 unique keywords (or key phrases) for embedding-based expansion.

Next, we trained a fastText word embedding model (Bojanowski et al., 2017) on a large Slovenian corpus of 144,352 articles about COVID-19, described in detail in Pollak et al. (2021). The corpus, collected by the EventRegistry service (Leban et al., 2014), contains articles from news portals that contain at least one of the following keywords: *covid*, *koronavirus*, *sars-cov-2*, *covid19*, *covid-19*, *korona virus*, *koronavirusna*, or *koronavirusen* and cover the early pandemic period in Slovenia, between January 1st and December 31st, 2020. We selected a domain-specific corpus for output as closely related to COVID-19 as possible. The fastText model was chosen because of the size of our corpus and as it uses subword information, which makes it the most suitable for morphologically rich languages, and especially given the neologism detection and word-formation perspectives under investigation. Also, unlike the models from the BERT family (Devlin et al., 2019), the output tokens are not tokenized, which is important for our analysis.

For each word (or multiword expression) in the seed vocabulary, we first extracted its 200 nearest neighbors from the fastText embeddings model. Next, we lemmatized the tokens and used Levenshtein distance-based filtering to avoid extracting words that were too similar. We decided to perform lemmatization after the embedding-expansion step because we did not want to fully rely on lemmatization,

which can be unreliable when neologisms are in question. We filtered the extracted candidates by removing those that do not contain any letter of the Slovenian or English alphabet, as well as all words that were already included in the Slovenian lexicon Sloleks (Dobrovoljc et al., 2019) because we were only interested in novel vocabulary. In the end, we kept the most related neighbors for each seed word and grouped them in a joint list by removing duplicates.

Altogether, 4,947 lemmas were extracted. In our study, we analyzed 843 lemmas that occurred at least 5 times in our corpus. We selected the lemmas according to the criteria of direct relevance to COVID-19. As a result, 66 relevant lemmas were identified.

The word lists and categorisations are available at: https://kt.ijs.si/data/elex_covid.zip

5. Analysis and results

5.1 Analysis of naming possibilities

The analysis of naming possibilities included 149 lexemes. In addition to the 66 lexemes resulting from our embedding-based expansion process, 29 lexemes were added from GDSL by Krvina (2014–) and 54 from the COVID-19 vocabulary of occasional words collected by Voršič (2022).

We followed our typology presented in Section 2, distinguishing between: 1) a search for naming possibilities in the language itself, including 1.1) a simplex or a set phrase, 1.2) a derivative, 1.3) a neosemantic formation, when a lexeme acquires new meaning, and 2) borrowing from a foreign language, including 2.1) disguised borrowing or calques and 2.2) expressed borrowing; for a detailed description, see also Section 2. A schematic overview is presented in Figure 1.

The results show that 85.9% of the lexemes were created by word-formation processes (i.e., derivatives in Figure 1), showing high language productivity and language vitality. Among the naming possibilities derived from Slovenian, set phrases follow at 6% (e.g., *socialna distanca* ‘social distance’, *omejevanje socialnih stikov* ‘limiting social contacts’, and *pridružena bolezen* ‘associated illness’), and neosemantisms account for 1.3% (e.g., *govorec* ‘spokesperson’ and *testirati* ‘to test’). Explicit borrowings, on the other hand, account for 6.7% (e.g., *covid-19*, *korona* ‘corona’, *lockdown*, *webinar*), and we did not identify any calques in our examples.

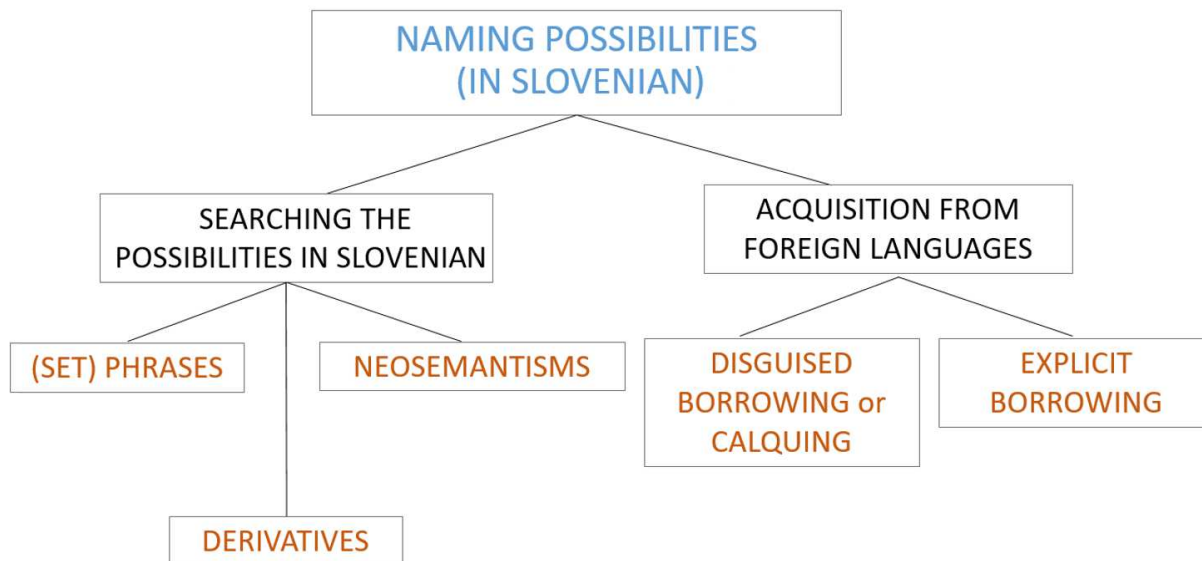


Figure 1: Naming possibilities in Slovenian (Vidovič Muha 2000, 2023)

5.2 Analysis of word-formation processes

5.2.1 Analysis of neologisms

A more detailed word-formation analysis included a total of 77 lexemes. 62 examples from our embedding-based extension method (out of the total list of 66 lexemes, four instances were not kept for analysis due to the fact that they were explicit borrowings from English and were not formed using word-formation processes in Slovenian) and 15 neologisms were included in the already confirmed GDSL, but were originally not kept in the embedding expansion results (because they did not appear above the selected frequency threshold).

The analysis of this dataset (see Table 1) shows that the most frequent are systemic formations (94.8%), out of which the most productive are interfixal compounds, followed by ordinary derivatives by suffixation and ordinary derivatives by prefixation, modificational derivatives by suffixation, derivatives from a prepositional phrase, coordinate interfixal-suffixal compounds, and subordinate interfixal-suffixal compounds. Compared to the systemic formations, the percentage of systemically unpredictable formations is much lower (5.2%), with abbreviations, blend words, and bicapitalizations.

Word-formation type	%	Example and gloss
<u>Systemically predictable formations</u>		
Interfixal compounds	41.56	<i>koronavirus</i> ‘coronavirus (n.)’
Ordinary derivatives by suffixation	27.27	<i>koronavirusni</i> ‘coronavirus (adj.)’
Ordinary derivatives by prefixation	15.58	<i>asimptomatski</i> ‘asymptomatic’
Modificational derivatives by suffixation	3.90	<i>gripca</i> ‘little flu (diminutive)’
Derivatives from a prepositional phrase	2.60	<i>brezkontakten</i> ‘contactless’
Coordinate interfixal-suffixal compounds	2.60	<i>nosno-žrelni</i> ‘nasopharyngeal’
Subordinate interfixal-suffixal compounds	1.30	<i>visokorizičen</i> ‘high-risk’
<u>Systemically unpredictable formations</u>		
Abbreviations	2.60	<i>DSO < dom starejših občanov</i> ‘retirement home’
Blends	1.30	<i>infodemija</i> ‘infodemics’
Bicapitalizations	1.30	<i>OstaniDoma</i> ‘StayHome’
Total	100	

Table 1: Categorization of examples by word-formation type.

Systemic derivatives

We first focus on systemic formations; that is, those that are formed in accordance with the word-formation rules of Slovenian. The systemic formations are categorized as compounds, derivatives by suffixation, and derivatives by prefixation.

Among the systemically predictable formations related to the coronavirus pandemic, the most frequent are nominal interfixal compounds; namely, those containing the prefix *korona-* in the first part. For example, in addition to the aforementioned borrowed word *koronavirus* ‘coronavirus’, these are neologisms of the type *koronačas* ‘coronetime’, *koronahumor* ‘coronahumor’, and *koronazakon* ‘coronalaw’. Alternatively, the non-adapted term *covid-* is also productive for interfixal compounds, but, whereas compounds with the first component substituted are consistently written together in the material, compounds with the first borrowed component *covid-* in the first part can be written either together (e.g., *covidbolnišnica* ‘covid hospital’ and *covidoddelek* ‘covid ward’) or with a hyphen (e.g., *covid-redar* ‘covid checker’, *covid-pozitiven* ‘covid-positive’). Among the noun interfix compounds in the collected material, there are also compounds with borrowed prefixoids (e.g., *alfakoronavirus* ‘alphacoronavirus’, *kiberkriminallec* ‘cyber criminal’) and unborrowed prefixoids (e.g., *samokarantena* ‘self-quarantine’). Adjectival compounds are also confirmed (e.g., *novopotrjen* ‘newly confirmed’,

novookužen ‘newly infected’) and to a lesser extent verbal interfixal compounds; for example, *samoizolirati (se)*, *samoosamiti (se)* ‘to self-isolate’. It is noticeable that among the more productive formations, especially those with the constituents *novo-* ‘new(ly)’ and *samo-* ‘self’ are the most productive. At the same time, among the compounds, there are those with both an acronymic (e.g., *PCR-metoda* ‘PCR method’, *PCR-test*) and a numeric (e.g., *10-dneven* ‘ten-day’, *14-dneven* ‘fourteen-day’) constituent in the first part. Among the subordinate interfixal-suffixal compounds, only the adjectival formation *visokorizičen* ‘high-risk’ is present. Coordinate interfixal-suffixal compounds are also rare; only the adjectival formations are attested, namely *nosno-žrelni* ‘nasopharyngeal’, *ustno-nosni* ‘oral-nasal’. Certain formations, such as *videopovezava* ‘video connection’, *14-dneven* ‘fourteen-day’, and *nosno-žrelni* ‘nasopharyngeal’, are not really neologisms because they were already in use before, but their frequency increased sharply during the pandemic.

Interfixal compounds are followed in frequency by ordinary derivatives. Here we can again mention the lexeme *korona* ‘corona’, which is productive of the adjectival derivative *koronski* ‘corona(l)’ and the nominal derivative *koronik* ‘corona-positive person’. The lexeme *koronavirus* ‘coronavirus’ is the basis of the derivative *koronavirusni* ‘coronavirus (adj.)’. Otherwise, adjectival formations are the most frequent among the derivatives (e.g., *prebolevniški* ‘convalescent’), whereas derivatives from the noun stem *pandemija* ‘pandemic’—for example, *pandemičen* ‘pandemic (adj.)’ and *pandemski* ‘pandemic (adj.)’—are used synonymously. The derivative *samoizoliran* ‘self-isolated’ is derived from the verbal stem *samoizolirati* ‘to self-isolate’. Among the noun derivatives, there are verb derivatives (e.g., *oksigenacija* ‘oxygenation’, *predihavanje* ‘ventilation’) and adjectival derivatives with the suffix *-ost* (e.g., *asimptomatičnost* ‘asymptomaticity’, *brezkontaktnost* ‘the state of being contactless’). Another interesting derivative formation is the suffix *-izem* ‘-ism’ (i.e., *starizem* ‘ageism’). We also trace adverbial derivatives—in synonymous use, the derivatives of the adjectival stem are *asimptomatično* ‘asymptomatically’ and *asimptomatsko* ‘asymptomatically’. Another set is derivatives from a prepositional phrase, for which only two formations with the native prefix *brez-* ‘non-’ or ‘-less’ are found in the material; namely, the adjectival formations *brezstičen* ‘contactless’ and *brezkontakten* ‘contactless’, which are in a synonymous relationship. Among the modificational derivatives, the verbal formations *prekuževati* ‘to develop herd immunity’ and *predihavati* ‘to ventilate’ and the nominal diminutive form *gripca* ‘little flu’ appear. The prepositional phrase formations and the modifying derivatives thus prove less productive.

Finally, we also observe nominal and adjectival ordinary derivation by prefixation with borrowed prefixes (e.g., *superprenašalec* ‘superspreader’, *antitelesa* ‘antibodies’, *asimptomatičen* ‘asymptomatic’) and non-borrowed prefixes (e.g., *neinvaziven* ‘non-invasive’, *nekužen* ‘non-contagious’, *protikorona* ‘anti-corona’).

Non-systemic derivation

In the context of less frequent systemically unpredictable formations, the abbreviations *DSO* < *dom starejših občanov* ‘retirement home’ and *PKP* < *protikoronski paket* ‘anti-corona package’ have been confirmed. Next, we find bicapitalizations (e.g., *OstaniDoma* ‘StayHome’), and blends (*infodemija* < *informacija* ‘information’ + *epidemija* ‘epidemic’ (‘infodemic’, an epidemic of false, misleading information)).

5.2.2 Analysis of occasionalisms

Occasionalisms are words composed for a specific purpose, with low frequency, and they are only at the periphery of the language system. The ad hoc formations are words that are the manifestation of the most current word-formation processes and reflect the creative flexibility of language, which is always a reflection of society, what is happening in it, and the social changes that are taking place.

To complement the word-formation description of the COVID-19 vocabulary in Section 5.2.1, which is based on material either from the confirmed vocabulary from GDSL or corpus-grounded (extracted words appearing at least five times in our corpus), in this section we discuss occasionalisms, which are an important counterpart to the analysis of neologisms. These were used in the seed list part of our vocabulary consisting of occasionalisms by Voršič (2022), collected from various social media sources, or were found in the embedding-based results, but did not match the set frequency threshold. The categorization is based on the work by Voršič (2022).

Among the systemic derivatives, the most productive are verb derivatives from proper nouns: *beović-iti* ‘to speak like Bojana Beović’ (‘to speak in such a way that you leave people in suspense’ referring to Bojana Beović, who was the head of the Medical Chamber of Slovenia); similarly, *kaciniti* ‘to speak like Kacin’ (‘to explain instructions in a mischievous and mildly threatening manner’, referring to Jelko Kacin, who was the main governmental director of public relations). Such derivatives also give rise to higher-stage nominal derivatives; that is, *beovičenje* ‘speaking like Beović’, *kacinjenje* ‘speaking like Kacin’.

In contrast to the results in Section 5.2.1, non-systemic derivatives are much more common in the context of ad hoc vocabulary. The essential characteristic of non-systemic formations is the indeterminacy of the syntactic stem and the impossibility of morphemization, but also the unpredictability of the number and different parts of speech of stem words that are merged into a neologism. The fact that most of the ad hoc formations are non-systemic formations is not surprising because these are words formed for stylistic effect and as a more attractive parallel to the existing lexemes. In the seed words by Voršič (2022), there are examples such as *kapitalizolacija** ‘capitalisolation’ < *kapital* ‘capital’ + *izolacija* ‘isolation’ (‘you can go to work but you can’t hang out with your friends’), or hashtags such as

#*OstaniZdrav* ‘StayHealthy’, where each component in the keyword is often capitalized, and so these formations could therefore also be defined as sets of bicapitalizations.

The high productivity among ad hoc formations is shown by blends. These are a more recent type of formation, formed by the compounding and back formation of two, or more rarely several, independent words that are expressively overlapping at a certain point. Sicherl and Žele (2018: 76) point out two basic conditions that have to be met to justify this type of formation; namely, 1) the overlap must be semantically recognizable, meaningful, and stylistically effective, and 2) the degree of back formation of individual subordinate words must be adapted to the pronunciation possibilities in the given language and determined by the creator. Stylistic effectiveness, wit, and jocularity are features also highlighted by Bugarski (2002: 217), who treats blend words as a distinctly sociolinguistically motivated word-formation process. Thus, blend words are formed on purpose to achieve a certain stylistic effect or with the intention of influencing (Sicherl & Žele, 2018: 82).

This is also reflected in the blend words analyzed, grouped into the following types: a) blend words in which the first part of the first sub-word and the whole of the second sub-word are joined; for example, *covinek** ‘covidbend’ < *covid* + *ovinek* ‘bend’ (‘loosely avoiding the oncoming person while walking’); b) blend words, in which the whole of the first sub-word and the last part of the second sub-word are joined together; for example, *koronačitnice* ‘coronavacation’ < *korona* ‘corona’ + *počitnice* ‘vacation’ (‘vacation in the time of corona’); c) blend words, in which the central part of the overlap is shared by the two base words and they overlap in this part; for example, *covidiot* < *covid* + *idiot* (‘a person that ignores the measures’); d) blend words in which part of the second subword is inside the first: *opravljičilo** ‘escusetale’ < *opravičilo* ‘excuse’ + *pravljica* ‘fairy tale’ (‘an obviously a made-up reason when the police stop you in the next municipality’; *natednovanje** ‘strainweaking’ < *nategovanje* ‘straining’ + *teden* ‘week’ (‘which is every day for another 14 days and then we’ll see’); and e) a special type of blend word structured by a mental association with current social conditions, already mentioned by Sicherl and Žele (2018): near homonyms (paronyms). The wide selection confirms that these were particularly common during the coronavirus pandemic: *dombola** ‘homeraffle’ < *dom* ‘home’ + *tombola* ‘raffle’ (‘a raffle to see which parent can go for a walk alone and who stays at home with the children’).

5.2.3 Word-formation combinatorics

Here we discuss the material from the perspective of formant combinatorics. Slovenian, like other Slavic languages, is characterized by a rich morphemic structure of words, which is the result of multi-stage formation; for example, from the adjective *star* ‘old’ the noun *starost* ‘age’ is formed in the first stage, from it the adjective

starosten ‘age’ in the second stage, from it the noun *starostnik* ‘old man’ in the third stage, and from it the possessive adjective *starostnikov* ‘old man’s’, which is the fourth stage. This example demonstrates the associativity of the four suffixal forms: *-ost* + *-en* + *-ik* + *-ov*, and the associativity of the forms is to be understood as the ability of different word-forming elements to coexist in the context of a multi-stage formation, taking into account the meaning-formation aspect.

As shown by recent vocabulary referring to the outbreak of the coronavirus pandemic, the most productive word-forming type is the interfixal compound, with the constituent *korona* ‘corona’ in the first part proving to be particularly productive (e.g., *koronačas* ‘corona time’, *koronabedak* ‘corona idiot’, *koronazakon* ‘corona law’). Otherwise, *korona* is productive for adjectival derivation (e.g., *koronski* ‘corona(l)’), and compounding (e.g., *protikorona* ‘anti-corona (n.)’, *protikoronski* ‘anti-corona (adj.)’). The second most frequent are common derivatives, which is not surprising because the word-formative derivation is the most common derivation process for Slovenian (Plemenitaš et al., 2020). Derivatives are also the word-formation type that most clearly demonstrates the sociability of affixes. This indicates, for example, the nominal compound *rizičn-ost* ‘riskiness’, derived from the adjective *rizič-en* ‘risky’. In this case, it is the combinatorics of the suffix forms *-en* + *-ost*. The same combinatorics is also confirmed in the case of higher-stage derivatives from the noun *simptom* ‘symptom’: the adjective *simptomatičen* ‘symptomatic’ from the first-stage noun *simptom-atika* ‘symptomatics’ is productive in the material for the compound *asimptomatičen* ‘asymptomatic’, from which the two adjectives are derived, irrespective of the prefix *a-*, according to the pattern *simptomatičen* (adj.) → *simptomatičn-o* (adv.), or *simptomatičen* (adj.) → *simptomatičn-ost* (n.) forming the adverb *asimptomatičn-o* ‘asymptomatically’ and the noun *asimptomatičn-ost* ‘asymptomaticness’ at the same stage. As a synonym of the adjectival formations *simptomatičen* ‘symptomatic’, the adjective *simptomatski* ‘symptomatic’ appears, from which the adverb *simptomatsk-o* ‘symptomatically’ is formed, and in the material the adverb *asimptomatsk-o* ‘asymptomatically’ is derived as a higher-stage derivative from the compound *a-simptomatski* ‘asymptomatic’. *Simptomatičn-o* ‘symptomatic’ is the adjective from which the adverb *simptomatsk-o* ‘symptomatically’ is formed. The companionability of the suffixal forms with the combinatorics *-en* + *-ost* is also expressed in the derivative *brezkontaktn-ost* ‘contactlessness’, which is a derivation of the derivative of a prepositional phrase: *kontakt* ‘contact’ → *brez-kontakt-en* ‘contactless’ → *brezkontaktn-ost* ‘the state of being contactless’, and in the verbal derivative from the prepositional phrase, productive for the formation of nouns with the meaning of properties: *pre-boleti* ‘to recover (perfective verb)’ → *prebol-el* ‘recovered’ → *prebolel-ost* ‘recovery from disease; *pre-boleti* ‘to recover (v. pf.)’ → *prebol-eva-ti* ‘to recover (v. impf.)’ → *prebolev-en* ‘recovering’ → *prebolevn-ost* ‘recovers from disease’. The adjectival derivative *prebolev-en* ‘recovering’ is also used to form the noun *prebolevn-ik* ‘person that has recovered’, which is a stem of the adjective *prebolevni-ški* ‘relating to persons that have recovered’.

When observing word-formation combinatorics, we are mainly dealing with the combinatorics of two suffixal forms; namely, the adjectival suffix *-en* and the nominal suffix *-ost*. Only some of the formations show a combinatorics of three suffixes, namely *-atika + -en + -ost* and *-en + -ik + -ški*.

Among the ad hoc formations, only the verbal noun derivatives are confirmed within the systemic formation, which give rise to the verbal noun derivatives with the word-formative meaning of action, and they show the combinatorics of the suffixal forms *-iti* (e.g., *beović-iti* ‘to speak like Beović’) + *-enje* (e.g., *beovičenje* ‘speaking like Beović’) and *-ovati* (e.g., *krekovati* ‘to speak like Milan Krek’) + *-anje* (e.g., *krekovanje* ‘speaking like Krek’).

6. Conclusions and further work

This article analyzed COVID-19 Slovenian vocabulary from the perspective of naming possibilities and word formation, including formant combinatorics. We grouped various sources with COVID-19 vocabulary and used natural language processing techniques to expand this and acquire additional vocabulary. The results of our study have an impact on understanding various naming possibilities and word-formation processes in Slovenian, and, on the applied side, 41 newly identified words will be proposed for expansion of the current description of COVID-19 vocabulary in the Growing Dictionary of the Slovenian Language (ed. Krvin 2014–). The analysis shows that a large majority of lexemes were created through word-formation processes, whereas set phrases, neosemantisms, and explicit borrowings were much less frequent, and no calques were identified in the examples analyzed.

From the point of view of the word-formation system, systemic derivatives were the most frequent formation process, among them interfixal compounds, ordinary derivatives by suffixation, and derivatives by prefixation. The analysis also confirms the finding of Stramljič Breznik (2021) that the most productive substructure is the root morpheme *korona*, which produces most of the subordinate interfixal compounds, but also higher-stage adjectival derivatives by suffixation and by prefixation. In addition, we also analyzed occasionalisms, which are mainly non-systemic and employ the blending strategy. In terms of formant combinatorics, we found that the most productive derivatives are adjectival and nominal derivatives with the suffix *-ost*; these are also the ones that show the most frequently confirmed combinatorics of the suffix *-en- + -ost*.

Because this is only a preliminary study performed on a small sample and on the vocabulary extracted from existing resources and from the corpus from 2020, one must note that the sample might not be representative. The main goal was to detect various types of naming possibilities and word-formation processes. In the future, we plan to update the study using either the updated domain COVID-19 corpus, or by

adding material from the Slovenian Monitoring corpus (Kosem, 2022; Kosem et al., 2022). Whereas this study focused on single-word terms, further work also covering multiword expressions would be of interest.

7. Acknowledgements

This article was written as part of the project Formant Combinatorics in Slovenian (J6-3134) funded by the Slovenian Research and Innovation Agency (ARIS). We also acknowledge ARRS funding through the core programs Knowledge Technologies (P2-0103) and The Slovenian Language in Synchronic and Diachronic Development (P6-0038).

8. References

- Akut, K. B. (2020). Morphological Analysis of the Neologisms during the COVID-19 Pandemic. *International Journal of English Language Studies*, 2(3), pp. 1–7. Available at: <https://doi.org/10.32996/ijels.2020.2.3.11>.
- Al-Salman S., & Haider A.S. (2021). COVID-19 Trending Neologisms and Word Formation Processes in English. *Russian Journal of Linguistics*, 25(1), pp. 24–42. doi: 10.22363/2687-0088-2021-25-1-24-42
- Asif, M., Zhiyong, D., Iram, A., & Nisar, M. (2021). Linguistic Analysis of Neologism Related to Coronavirus (COVID-19). *Social Sciences & Humanities Open* 4. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3608585.
- 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.
- Będkowska-Kopczyk, A. & Łaziński, M. (2020): COVID-19 Vocabulary in Slavic. In Marc L. Greenberg (ed). *Encyclopedia of Slavic Languages and Linguistics Online*. Available at: <https://referenceworks.brillonline.com/browse/encyclopedia-of-slavic-language-s-and-linguistics-online>.
- Bugarski, R. (2002). *Nova lica jezika*. Belgrade: Čigoja štampa.
- Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 (Long and Short Papers), pp. 4171–4186. Minneapolis, MN: Association for Computational Linguistics.
- Dobrovoljc, K., Krek, S., Holozan, P., Erjavec, T., Romih, M., Arhar Holdt, Š., Čibej, J., Krsnik, L., & Robnik Šikonja, M. (2019). *Morphological lexicon Sloleks 2.0*, Slovenian Language Resource Repository CLARIN.SI, ISSN 2820-4042. Available at: <http://hdl.handle.net/11356/1230>.

- Fran, različica covid-19 (7.1)*, (2020). Available at:
https://fran.si/o-portalu?page=Covid_19_2020.
- Gustilo, L., Pura, C. M., & Biermeier T. (2021). Coronalexicon: Meanings and Word-Formation Processes of Pandemic-Related Lexemes across English Varieties. *International Journal of English Language Studies* 27(4), pp. 1–15. Available at:
https://www.researchgate.net/publication/356999244_Coronalexicon_Meanings_and_Word-formation_Processes_of_Pandemic-related_Lexemes_across_English_Varieties.
- Kosem, I. (2022). Trendi—A Monitor Corpus of Slovene. In *Proceedings of the XX EURALEX International Congress*. Mannheim, Germany, pp. 230–239. Mannheim: IDS-Verlag.
- Kosem, I., Čibej, J., Dobrovoljc, K., & Ljubešić, N. (2022). Spremljevalni korpus Trendi: metode, vsebina in katalogizacija besedil. In *Proceedings of the Conference on Language Technologies and Digital Humanities*, Ljubljana, Slovenia, pp. 86–92. Ljubljana: Inštitut za novejšo zgodovino.
- Kosem, I., Čibej, J., Gantar, P., Arhar Holdt, Š., Krek, S., Laskowski, C., Robnik Šikonja, M., Klemenc, B., Dobrovoljc, K., Gorjanc, V., Repar, A., & Ljubešić, N. (2021). *Sledilnik 1.0: Language Monitor*. Available at: viri.cjvt.si/sledilnik.
- Krvina, D. (ed.). (2014–). *Sprotni slovar slovenskega jezika*. Available at:
www.https://fran.si/iskanje?FilteredDictionaryIds=132&View=1&Query=%2A.
- Krvina, D. (2023). *Sprotni slovar slovenskega jezika.: Uvod*. Ljubljana: Založba ZRC. Available at:
https://fran.si/132/sprotni-sprotni-slovar-slovenskega-jezika/datoteke/Sprotni_Uvod.pdf
- Leban, G., Fortuna, B., Brank, J., & Grobelnik, M. (2014): Event Registry: Learning about World Events from News. In *Proceedings of the 23rd International Conference on World Wide Web (WWW '14 Companion)*, pp. 107–110. New York: Association for Computing Machinery.
- Lei, S., Yang, R., & Huang, C.-R. (2021) Emergent Neologism: A Study of an Emerging Meaning with Competing Forms Based on the First Six Months of COVID-19. *Lingua*, 258: 103095. Available at:
<https://www.sciencedirect.com/science/article/pii/S002438412100067X>.
- Piskorski, J., Babych, B., Kancheva, Z., Kanishcheva, O., Lebedeva, M., Marcińczuk, M., Nakov, P., Osenova, P., Pivovarov, L., Pollak, S., Přibáň, P., Radev, I., Robnik-Sikonja, M., Starko, V., Steinberger, J., & Yangarber, R. (2021). Slav-NER: The 3rd Cross-Lingual Challenge on Recognition, Normalization, Classification, and Linking of Named Entities across Slavic Languages. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pp. 122–133, Kyiv, Ukraine: Association for Computational Linguistics.

- Plemenitaš, K., Voršič, I., & Stramljič Breznik, I. 2020. Derivational Networks in Slovene. In L. Körtvélyessy, A. Bagasheva, & P. Štekauer (eds.) *Derivational networks across languages*, pp. 105–113. Berlin: De Gruyter Mouton.
- Pollak, S., Martinc, M., Pelicon, A., Ulčar, M., & Vezovnik, A. (2021). Covid-19 v slovenskih spletnih medijih: analiza s pomočjo računalniške obdelave jezika. In *Pandemična družba: Slovensko sociološko srečanje*, Ljubljana, pp. 260–268. Ljubljana: Slovensko sociološko društvo.
- Sicherl, E., & Žele, A. (2018). Prekrivanke v slovenščini z vidika vpliva angleškega jezika. *Jezik in slovstvo*, 63(1), pp. 75–88. Available at: <https://www.jezikinslovstvo.com/pdf.php?part=2018%7C1%7C75%E2%80%93388>.
- Stramljič Breznik, I. (2005). Kvantitativne lastnosti slovenskega tvorjenega besedja v poskusnem besednodružinskem slovarju za črko B 'Quantitative properties of Slovene complex words in the test dictionary of word families for entries starting in B'. *Slavistična revija* 53 (4): 505–520.
- Stramljič Breznik, I. (2021). Pandemija koronavirusa – zunajjezikovni dejavnik jezikovne ustvarjalnosti. In P. Kowalski (ed.) *Słowotwórstwo w przestrzeni komunikacyjnej (= Prace Slawistyczne, Slavica 151)*, pp. 307–322. Warsaw: Instytut Slawistyki Polskiej Akademii Nauk. Available at: https://ispan.waw.pl/ireteslaw/bitstream/handle/20.500.12528/1927/Pawe%20c5%82_Kowalski_%28red%29_S%c5%82owotw%c3%b3rstwo_w_przestrzeni_komunikacyjnej.pdf?sequence=1&isAllowed=y.
- Škerget, S. (2020). Novotvorjenke, nastale v času pandemije koronavirusa. *Liter jezika: literarno-jezikoslovna revija* 11: 14. Available at: http://literjezika.ff.um.si/?page_id=371.
- Toporišič, J. (2000). *Slovenska slovnica*. Maribor: Obzorja.
- Vičar, B. (2021). Koronabesedje. *Naši izviri: glasilo Občine Miklavž na Dravskem polju* 21(105), pp. 33–34.
- Vidovič Muha, A. (2013). *Slovensko leksikalno pomenoslovje*. Ljubljana: Znanstvena založba Filozofske fakultete.
- Voršič, I. (2010). Sistemske in nesistemske tvorjenke v tiskanih oglasih. *Jezikoslovni zapiski*, 16(1), pp. 107–120. Available at: <http://www.dlib.si/details/URN:NBN:SI:DOC-5AGEWGAH>.
- Voršič, I. (2022). Priložnostne tvorjenke kot odraz dobe koronavirusa. In G. Nikolovski (ed.) *Slavistična prepletanja*, 3, pp. 253–266. Maribor: Univerza v Mariboru, Univerzitetna založba. Available at: <https://press.um.si/index.php/ump/catalog/book/668>.
- Wang, X., & Huang, C. (2021). From Contact Prevention to Social Distancing: The Co-Evolution of Bilingual Neologisms and Public Health Campaigns in Two Cities in the Time of COVID-19. *SAGE Open*. Available at: <https://journals.sagepub.com/doi/10.1177/21582440211031556>.

Weiss, P., & Dobrovoljc, H. (2020). Kako pisati in sklanjati izraze “koronavirus” in bolezen “covid-19” ali “koronavirusna bolezen 2019”. *Jezikovna svetovalnica*. Available at: <https://svetovalnica.zrc-sazu.si>.

Automating derivational morphology for Slovenian

Tomaž Erjavec¹, Marko Pranjič^{1,4}, Andraž Pelicon¹, Boris Kern²,
Irena Stramljič Breznik³, Senja Pollak¹

¹Jožef Stefan Institute, Jamova cesta 39, Ljubljana, Slovenia

²ZRC SAZU, Fran Ramovš Institute of the Slovenian Language and University of Nova Gorica, School of Humanities, Novi Trg 4, Ljubljana, Slovenia and University of Nova Gorica

³University of Maribor, Faculty of Arts, Koroška cesta 160, Maribor, Slovenia

⁴Jožef Stefan International Postgraduate School, Jamova cesta 39, Ljubljana, Slovenia

E-mail: {tomaz.erjavec,marko.pranjic,andraz.pelicon,senja.pollak}@ijs.si,
boris.kern@zrc-sazu.si

Abstract

In this paper, we focus on computational approaches for supporting derivational word formation analysis in Slovenian. The main contributions are two-fold: first, we derive word formation rules and chains from given examples of the trail volume of a derivational dictionary and apply them to larger lexicons from two Slovenian resources; and second, we propose the first morphological segmenter for Slovenian. More specifically, based on the digitised trail volume (words starting with *b*) of the derivational dictionary of Slovenian, we extracted suffixal word-formation rules, and applied them to two lexicons of Slovenian, Sloleks and the one extracted from the metaFida corpus, to acquire new word formation instances for each chaining rule. The study of word-formation chains is relevant because it gives us an insight into word-formation mechanisms and productivity. The results show that when the derived chaining rules were applied to Sloleks, 21.95% to 31.58% of derivational chains are correct. In contrast, when the chaining rules were applied to the metaFida lexicon, the results are very noisy, with an extremely low percentage of correct chains. Next, motivated by the fact that morphological segmentation is a prerequisite for determining the structure of word formation chains and the need for more general analysis on the level of morphemes, we implemented the first automated morphological segmentation models for Slovenian. The supervised model is based on BiLSTM-CRF and achieves F1-Score of 83.98%, which is significantly higher than the two implemented unsupervised baselines, Morfessor and MorphoChain, to which we the model is compared.

Keywords: derivational morphology; word formation; automated morphological segmentation; derivational dictionary; morphological chains

1. Introduction

Word formation is a branch of linguistics which helps to analyse the lexical vitality of a given language and also shows trends of language development. Slovenian is characterised by an extremely rich morphemic structure of words, a result of multistage formation: e.g. in the first stage, the adjective *mlad/young* yields the noun *mladost/youth*, which in turn yields the adjective *mladosten/youthful* in the second stage, which yields the noun *mladostnik/adolescent*, yielding the possessive adjective *mladostnikov/adolescent's* in the fourth stage. The example shows the compatibility of four suffixal formants: *ost + -en + -ik + -ov*. The compatibility of formants is considered as the ability of different word-formational formants to co-exist within the multistage formation, taking into account the semantic extension aspect. Our paper contributes to the goal of better understanding the

characteristics of word-formation and semantic extension mechanisms in the contemporary Slovenian language, by determining the systemic predictability of formation in terms of compatibility of formants, with a focus on suffixal formants.

While there were some linguistic descriptions of Slovenian word formation (Vidovič Muha, 1988; Toporišič, 2000), including the description of formation of words in several stages that enables the linguistic investigation of multistage word-formation in Slovenian (see Breznik (2004); Kern (2010, 2020)), there is a lack of corpus-based grounding of theoretical findings. In the field of natural language processing, several researchers (Ruokolainen et al. (2013); Cotterell et al. (2015, 2019); Zundi & Avaajargal (2022); Peters & Martins (2022)) addressed the problem of morphological analysis, but there is no such study for Slovenian.

The main contributions are two-fold: first, we derive word formation rules and chains from given examples of the trail volume of the derivational dictionary BBSJB, and apply them to larger lexicons from two Slovenian resources; and second, we propose the first morphological segmenter for Slovenian. While the tasks are of different nature, they both contribute to the final goal of analysing word formation processes and their combinatorics in Slovenian. In the first case, we applying the rules derived from the existing database, and in the second one, we do not get specific rules, but get more general segmentation rules, which are less sensitive to the noisy corpora and are an underlying component of various systems for analysing word formation processes. In the derivation of word formation rules, we currently concentrated on suffix-adding rules only, as they are by far the most common in Slovenian, while in the segmentation task, the approach is more general and also other affixes are considered.

The basis for our study was the already existing Trail volume (headwords starting with the letter *b*) of the derivational dictionary of Slovenian (BBSJB) (Breznik, 2004). The dictionary gathers words in word families centred around a root, and inside those presents sequences of derivations, also split into constituent morphemes and giving the part-of-speech of the source and derived words. We leveraged BBSJB constructing morphological rules and chains (e.g., for *boj/a* 'fight' → *bojevati* → *bojevanje*). The derived rules can then be applied to infer examples from novel corpora, with the goal of comprehensive and corpus-grounded linguistic description of derivational processes, beyond the currently available dataset consisting of letter *b* headwords only. Moreover, BBSJB was leveraged for constructing a dataset for training and evaluation of morphological word segmentation, which is a prerequisite for determining the structure of word formation chains, also beyond the ones described in the rules derived from the BBSJB data. The work was performed in the scope of the project Formant Combinatorics in Slovenian.

The paper is structured as follows. After presenting related work in Section 2, Section 3 describes the resources used in our study (BBSJB, Sloleks and metaFida). Next, we present the methodology of rule-based chain extraction (Section 4.1) and morphological segmentation (Section 4.2), including two unsupervised and one supervised model. Section 5 contains the results of the rule-based chain evaluation and compares different morphological segmentation approaches and is followed by the conclusion and plans for future work (Section 6).

2. Related work

Work on automatic induction of rules for Slovenian lemmatisation has already been researched Slovenian a while ago Erjavec & Džeroski (2004), where Inductive Logic Programming was used to derive rules that compute the lemma of a word given its word-form and part-of-speech tag. This work was then followed up with approaching the same task but using so called Ripple Down Rules (Juršič et al., 2010). But while at first glance the two approaches could be also used to predict derivational rules, there is a considerable difference between inflectional and derivational morphology, as a word-form will always have a lemma, while a word will not necessarily yield a derivation, nor will a potentially derived word necessarily be such, i.e. both the source and target words in a derivational process must be attested in a lexicon. It should also be noticed that there also exists an automatically derived but manually checked set of morphological rules (Arhar Holdt et al., 2020) that relate entries in the Sloleks morphological lexicon (Dobrovoltj et al., 2022). While we also use this lexicon in our experiments, the rules themselves, again, cover only inflection, and are therefore not useful for work on derivational morphology. Rules for morphologically related words have been designed and applied to Sloleks in Čibej et al. (2020). The resource contains only word pairs, not entire chains, and automatic segmentation was performed without evaluation of the method.

Beyond the Slovenian natural language processing landscape, there are several directions. For Croatian, a closely related language, CroatianCroDeriV (Filko et al., 2019; Šojat et al., 2014) was developed, a language resource that contains data about morphological structure and derivational relatedness of verbs. Focusing on derivational processes from computational methods' perspective (see e.g. (Vylomova et al., 2017)). Evaluation of word embeddings by Gladkova et al. (2016) evaluates the processes in the scope of analogy tasks, and shows that derivational morphology is significantly more difficult to model than inflectional. Works by Lazaridou et al. (2013); Cotterell & Schütze (2018); Hofmann et al. (2020a) for example, attempt to predict a derived form given a corresponding base form. In recent research, Hofmann et al. (2020b) leverage pre-trained Neural Network Language Models and propose DagoBERT (Derivationally and generatively optimized BERT) for generation of derivationally complex words.

Morphological segmentation is a task closely related to the analysis of derivational morphology. Although the resulting segmentation does not provide explicit rules for word formation, the output of automatic morphological segmentation is a chain of morphemes. The task of morphological segmentation has generated considerable scientific attention, with several shared tasks (e.g. SIGMORPHON Batsuren et al. (2022), MorphoChallenge Kurimo et al. (2010)) being organized. For baselines in our work, we selected Morfessor and MorphoChain methods. Morfessor is a family of probabilistic machine learning methods for morphological segmentation from text data, and Morfessor 2.0 (Smit et al. (2014)), while MorphoChain Narasimhan et al. (2015) is an unsupervised model used for morphological segmentation that integrates orthographic and semantic views of words. In one of the earlier studies, Cotterell et al. (2015) designed a machine learning system for joint morphological segmentation and morpheme tagging which directly models morphotactics. Ruokolainen et al. (2013), which is also the foundation of our supervised model, addressed the task of morphological segmentation as a character-based sequence labelling task. The authors modelled the sequence labelling task with a Conditional Random Field (CRF) model. The joint BiLSTM-CRF models, introduced by Huang et al. (2015), were later

successfully used for a number of sequence tagging tasks such as part-of-speech tagging and named entity recognition. Several recent studies have achieved state-of-the-art results by using Transformer-based encoder-decoder models (Zundi & Avaajargal (2022); Peters & Martins (2022)). However, these models usually require relatively large amount of labeled data to properly converge. Further, due to large number of training parameters, such models are prohibitively expensive for training and inference in terms of computational power.

3. Resources

In this section we overview the data used in our experiments, starting with trial volume of the Derivational Dictionary of Slovenian, which was the essential resource for the study, and following with the two subsidiary resources, namely the morphological dictionary of Slovenian called Sloleks, and the metaFida corpus of Slovenian.

3.1 The Derivational Dictionary of Slovenian

The basis for our study was the already digitised Word-family dictionary of Slovenian, Trial volume for headwords beginning with letter *b*, or Besednodružinski slovar slovenskega jezika, Poskusni zvezek za iztočnice na B Stramljič Breznik (2005), BSSJB in short. The dictionary gathers words in word families centred around a root, and inside those presents sequences of derivations, also split into constituent morphemes, together with their type (e.g. suffix, prefix, compound, etc.) and giving the part-of-speech of the source and derived words. The trial volume contains 666 word families and 11,194 derivations. The trial volume was first converted from its source encoding into TEI Lex0 Romary & Tasovac (2018)¹, which is a TEI-based XML schema and guidelines for encoding dictionaries and other lexical resources, developed in the scope of the DARIAH research infrastructure.

The trial dictionary in Lex0 contains all the information from the source, and, additionally, a conversion of the part-of-speech and lexical properties of the entry from the source (Slovenian) labels to Universal Dependencies morphological features (de Marneffe et al., 2021) and to its MULTEXT-East morphosyntactic description (MSD) (Erjavec, 2012), which makes the resource better compatible with other Slovenian lexical and corpus resources. It also introduces a taxonomy of morpheme types, and links the morphemes to it.

Figure 1 gives an example of the encoding and content of a typical word family (here for the word *baba*, in this case giving only its first derived word (*babaj*, which can also be further nested, to give higher order derived words. It is thus possible to make sequences (chains) of the derivations, also giving the order number (level) of each derived word. The root word of a word-formation family (*baba* in the example) will always be level 0, while *babaj* will be 1.

To simplify processing for our experiments, we converted the TEI Lex0 format into a TSV file, which contains all the information relevant to our experiments, in particular the ID of the word family, the ID of the entry, the lemma, its level, the chain of words, of lexical properties, and morpheme types, all starting from the family root. The dictionary in this format was then used as the starting point for all further experiments.

¹ <https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html>

```

<entry xml:lang="sl" type="mainEntry" xml:id="bssj36337">
  <form type="lemma">
    <orth type="headword">baba</orth>
    <pron>bába</pron>
  </form>
  <gramGrp type="orig">
    <gram type="pos">sam.</gram> <gram type="gender">ž</gram></gramGrp>
  <gramGrp type="UD" xml:lang="en">
    <gram type="pos">NOUN</gram> <gram type="other">Gender=Fem</gram></gramGrp>
  <gramGrp type="MTE" xml:lang="en"><gram type="other">Nc</gram></gramGrp>
  <form type="morphemes">
    <orth><seg n="1" ana="#root1Morph">baba</seg></orth>
    <pron><seg n="1" ana="#root1Morph">bába</seg></pron>
  </form>
<entry xml:lang="sl" type="mainEntry" xml:id="bssj45627">
  <form type="lemma">
    <orth type="headword">babaj</orth>
    <pron>babáj</pron>
  </form>
  <gramGrp type="orig">
    <gram type="pos">sam.</gram> <gram type="gender">m</gram></gramGrp>
  <gramGrp type="UD" xml:lang="en">
    <gram type="pos">NOUN</gram> <gram type="other">Gender=Masc</gram></gramGrp>
  <gramGrp type="MTE" xml:lang="en">
    <gram type="other">Ncm</gram>
  </gramGrp>
  <form type="morphemes">
    <orth><seg n="1" ana="#root1Morph">bab</seg>
    <seg n="2" ana="#suffix1Morph">aj</seg></orth>
    <pron><seg n="1" ana="#root1Morph">bab</seg>
    <seg n="2" ana="#suffix1Morph">áj</seg></pron>
  </form>
</entry>
...
</entry>

```

Figure 1: Start of an example entry form BSSJB.

3.2 The Sloleks lexicon

Sloleks 2.0 Dobrovoljc et al. (2022) is a Slovenian morphological lexicon, which gives for each entry, inter alia, the lemma of the word, its complete inflectional paradigm (the word-forms paired with their MULTEXT-East morphosyntactic description and Universal Dependencies morphological features), and the frequency of occurrence of each triplet in the Gigafida reference corpus of Slovenian Krek et al. (2019). As the lexicon contains over 100,000 hand-verified lemmas, it is a large and clean resource for finding pairs of lemmas, where one is a derivation of the other.

As a preprocessing step we extracted all the lemmas from Sloleks, together with their lexical features (e.g. *Ncf* for *Noun*, *type="common"*, *gender="feminine"*) and the frequency of occurrence in Gigafida. This lexicon contains 85,398 lemmas, somewhat less than Sloleks, as derivationally non-productive entries, such as numerals, abbreviations, pronouns etc. were not included.

3.3 The metaFida corpus

The metaFida 0.1 corpus (Erjavec, 2022) is the prototype edition of the collection of 34 Slovenian corpora, which are available on the concordancers of the CLARIN.SI research infrastructure, and was made so that linguists can analyse the Slovenian language using a single resource. The corpus is, by definition, the largest corpus of Slovenian, with over 3,6 billion words and so seemed as a good candidate to collect ever more lemmas than are available in Sloleks.

As each word in metaFida is marked up with its MULTEXT-East MSD, we extracted from it a lexicon identical in format to the Sloleks lemma lexicon, i.e. a list of lemmas, accompanied by their lexical features and the frequency of occurrence in metaFida; the lexicon was also filtered in the same manner as the Sloleks one. This gave us a lexicon with 1,229,345 lemmas.

While the metaFida lexicon most likely covers the lexis of Slovenian very well, certainly much better than Sloleks, it also, as the huge number of lemmas makes obvious, contains a large portion of noise, from encoding errors and typos, to errors in automatic lemmatisation and PoS tagging. This noise is not very noticeable at the token level, but in a lexicon each mistake can produce a new lexical entry.

4. Methodology

4.1 Rule-based Chain Extraction

In our first experiment the goal was to explore the possibility of using the existing information in the trail volume of Derivational Dictionary of Slovenian (BSSJB) to induce, on the basis of the Sloleks and metaFida lexicons, entries for word families not present in BSSJB.

The method relies on the headwords, morpheme segmentation and Universal Dependencies part-of-speech labels present for each entry in BSSJB. We take pairs of entries (source and derived word), and construct pairs or rules ("deep" and "surface" rules) that map the source word to the derived word, the former formulated as a sequence of morphemes, and the latter as regular expressions. For example, if we take the entry *boj-evati/to fight* (*VERB*) from which the entry *bojev-anje/figthing* (*NOUN*) is derived, we construct the deep rule *VERB:X-evati* \rightarrow *NOUN:X-anje* and pair it with the surface rule describing the derivation as a minimal transformation on surface forms, in this case "*VERB:X+ti* \rightarrow *NOUN:X+nje*". It should be noted that we concentrate on rules that operate on suffixes only.

Such rules are then also gathered into chains, as presented in the dictionary (e.g. for *boj/a fight*" (*NOUN*) \rightarrow *boj-evati* \rightarrow *bojev-anje*). We currently concentrated on suffix rules only, with BSSJB yielding 1,641 distinct rules and 1,649 chains.

We next applied the constructed surface rules to part-of-speech / lemmas pairs from the Sloleks and metaFida lexicons. For each entry in the lexicons we try to apply the left-hand part of the regular expression of all surface rules to it, also taking into account the part-of-speech, and, if successful, construct the target word. If the target word with the correct part-of-speech also exists in the lexicon, we have found a potential derivational

pair, and can assign to it the the deep derivational rule. Once the complete lexicon has been processed, we also connect, as far as possible, the found pairs into chains.

With this, the initial set of morphological rules and chains from BSSJB and consisting only of roots starting with the letter *b* is extended to words starting with all the others letter of the alphabet (e.g. *izklic* → *izklicevati* → *izklicevanje*). Of course, not all the found pairs and chains - esp. for the metaFida lexicon - are valid derivations or derivation chains, but the derived resource could offer a good starting point for manual verification. Using the described method, we gathered 117,769 potential pairs and 32,823 potential chains from From Sloleks, while from the metaFida lexicon we get 1,549,644 potential pairs and 496,486 potential chains.

4.2 Morphological Segmentation

In this section we describe the unsupervised and supervised models used for morphological segmentation. We also present the dataset we constructed for the task of morphological segmentation based on BSSJB. We used this dataset to train and evaluate our supervised model as well as for the evaluation of unsupervised models. While in general supervised models tend to perform better, this is sensitive on the size of the training data, especially for deep learning models. Therefore, we are interested in whether unsupervised models trained on large amount of data outperform supervised models with relatively small labeled training dataset (around 10,000 examples).

4.2.1 Datasets

We have generated a gold standard dataset for morphological segmentation based on the Derivational Dictionary of Slovenian (BSSJB), more specifically on the morphological sequence chains with which we enriched the original version of the dictionary (see Section 3.1). As described in Section 4.1, the morphological chains contain only the information on the latest derivational suffix at each level. For example, the word *babeževanje* has the corresponding morphological chain *baba* → *bab-ež* → *babež-evati* → *babežev-anje*. In order to train a supervised automatic morphological segmenter, we had to preprocess the morphological chains to obtain a gold label segmentation of the word for all the levels (e.g. *bab-ež-ev-anje*). The preprocessing was done programatically using simple rule-based approach. Since Slovene is morphologically complex and words frequently omit certain phonemes as they are derived, the rule-based approach produced a small amount of faulty segmentations. In order to limit the amount of noise in the training set, such examples were removed from the data, resulting in 210 words being omitted from the dataset. Some of the words present in the dictionary were reflexive verbs and were therefore recorded with a reflexive pronoun (e.g. *babiti se*). Since the reflexive pronouns themselves are not derivational morphemes we decided to remove them from the dataset during preprocessing. The resulting dataset contains 9,883 words and their gold standard morphological segmentations. This dataset was used for training the supervised approaches, and for evaluation of supervised and unsupervised segmentation models.

In addition, for unsupervised methods, we also used Sloleks and metaFida, which are described in Sections 3.2 and 3.3. The metaFida corpus was additionally cleaned up by removing all words containing characters not found in Slovene alphabet (namely, *x*, *y*, *w*,

q), all words that contain a sequence of a single character repeated successively 3 or more times, all words shorter than 3 letters, and all words occurring less than 4 times in the corpus. Due to entries in the Sloleks lexicon being manually verified, we did not do any data cleanup or preprocessing.

4.2.2 Morfessor

Morfessor is a family of probabilistic machine learning methods for morphological segmentation from text data. The underlying model is trained such that it optimizes for maximum a posteriori (MAP) estimate of models parameters Θ given the training data D :

$$\Theta_{MAP} = \arg \max_{\Theta} p(\Theta)p(D|\Theta) \quad (1)$$

During training, for each word all possible two segment combinations are evaluated. The segmentation that produces the lowest cost is selected and the same procedure is recursively applied to the resulting segments. During inference, a variation of Viterbi algorithm is used to produce the segmentation with the lowest cost. In this work, a *Morfessor 2.0* Smit et al. (2014) variant of the model was used.

We induce two Morfessor models, one using the words from Sloleks lexicon and one using words from metaFida corpus. Due to entries in the Sloleks lexicon being manually verified, we train the Morfessor model on this data in a type-based training regime that assigns equal frequency for each word in the corpus. In metaFida, we have actual count of occurrences for each word in corpus. In contrast to the approach taken with Sloleks, we train the model in log-token based training regime where number of occurrences of words are modified to use logarithm of the raw count instead. While frequency-based weighting in metaFida serves as a regulariser for the noise inherent to the dataset, we did not opt for this strategy for Sloleks, as the resource is clean and manually verified.

4.2.3 MorphoChain

Introduced in Narasimhan et al. (2015), MorphoChain is an unsupervised model used for morphological segmentation that integrates orthographic and semantic views of words. On the orthographic level, several features are used to estimate how the affixes are reused, how the words are changed when new morphemes are added to the chain and whether a sequence of morphemes exists in the corpus. Semantic comparison between words uses an additional list of word vector representations, like those produced by deep-learning models.

The model was configured by specifying letters from Slovene alphabet, by lowering the minimum morpheme size to 2, and specifying word vectors to be used for the semantic features. We use existing, publicly available, Slovene fastText word vectors described in Ljubešić & Erjavec (2018).

As in previous section, we induce one model using words from Sloleks lexicon and one using words from metaFida corpus.

4.2.4 BiLSTM-CRF - Tagging of morphological segments

Following Ruokolainen et al. (2013), we model the problem of morphological segmentation as a sequence labelling problem on the character level.² Each character c in a target word w is labeled with a label from the label set $y \in \{START, B, M, E, S, STOP\}$. From this set, a character labeled with B represents a character at the beginning of the morpheme, M represents character in the middle, and E represents the character at the end of the morpheme. A label S is used for characters that are morphemes by themselves. For example, a word *bankomat* with a ground-truth segmentation *bank-o-mat* will be transformed to labels as $[B, M, M, E, S, B, M, E]$. The special labels *START* and *STOP* are added to the beginning of each word in order to constrain the model further.

Similarly to the original work, we model the sequence tagging problem with a Conditional Random Field (CRF) model and use it to train a morphological segmenter for Slovenian language. The main advantage of the CRF model is that it models the output sequence by considering dependencies between output variables. The CRF models the conditional probability of a sequence of labels \hat{y} with respect to the input sequence \hat{x} as follows:

$$P(\hat{y}|\hat{x}, w) = \frac{\exp(\sum_t^T \hat{w} * F(y_{t-1}, y_t, \hat{x}, t))}{\sum_{y' \in Y} \exp(\sum_t^T \hat{w} * F(y'_{t-1}, y'_t, \hat{x}, t))} \quad (2)$$

where t represents the position of the character in the sequence, T denotes the total length of the sequence, w denotes the parameter vector and F represents the feature function. During inference, we find the sequence of labels that maximizes the conditional probability from Equation 2 using Viterbi algorithm. We modify the original work however by employing a BiLSTM network as a feature function. The advantage of this approach is that the feature functions are not a preset mapping from words to features but are trained jointly with the CRF model. Furthermore, the use of the BiLSTM network allows us to effectively use feature information from the past n sequence steps when assigning the tag to the $n+1$ -th character.

The input to our model is a word to be segmented, split into separate characters with two special *START* and *STOP* characters added to the beginning and the end of the character sequence. Each character in the sequence is then embedded into a shared embedded space \mathbb{R}^e where e denotes the dimensionality of the embeddings. The embedded sequence of characters is then modelled by a BiLSTM network serving as a feature extractor which transforms the input at each step to \mathbb{R}^h where h represents the hidden size of the BiLSTM. The output from the last step of the BiLSTM network is then linearly transformed to \mathbb{R}^l where l is the dimensionality of the label set. Dropout is applied on the input to the linear transformation as a form of regularization. This output is then used as the emission scores for the Conditional Random Field which outputs the tag at the next step.

In our experiments, we set the embedding size e as 50 and the hidden size h of the BiLSTM as 25 while the dropout probability is set to $p = 0.2$. Training of the model is performed in batches with the batch size set to 32. For efficient computation, the lengths of the

² While we do not have permission for sharing the segmented data, the code for the segmentation method is public: https://gitlab.com/Andrazp/automating_derivational_morphology_for_slovenian

sequences were padded with padding tokens to the same length. The maximum size of the sequence was set to 30 characters which corresponds to the longest word in the dataset. For training, the training fold of the dataset is split into training and evaluation sets in 90%–10% proportion. During initial experiments, we have observed slow convergence of the model especially in the early stages of training. For this reason, we let the model train for 100 epochs. After each epoch, the performance of the model was evaluated on the evaluation set to prevent overfitting.

For final evaluation, we use 5-fold cross-validation, where we repeat the training procedure five times, each time evaluating on different fold of the data. We construct 5-fold cross-validation data by arranging words into folds such that all entries sharing the same root of the word are in the same fold. This is achieved by first collecting words into groups according to their root. We form a multiway number partitioning optimization problem Graham (1969) such that word groups are assigned to 5 bins in a way that minimizes differences between number of words between each bin. This optimization problem is solved using a greedy Longest-processing-time-first (LPT) algorithm³. In this way we ensure two important properties of our training data. One, each fold contains approximately equal number of words⁴. Two, closely related words that are derived from the same root are always assigned to the same fold. Using such constructed folds, the model is always evaluated on the words containing roots unseen during training and this enables us to test the performance of the model when applied on new words.

5. Evaluation

In this section we present the evaluation and results of rule-based chain extraction and of the machine learning-based models for morphological segmentation. Section 5.1 presents the results of chain extraction Sloleks and metaFida corpora and analyses the most common manually extracted chains. Section 5.2 presents results of the three approaches for automatic morphological segmentation of words.

5.1 Rule-based Chain Evaluation

In Table 1, we present frequencies of chain lengths on each dataset. The columns for the BSSJB dataset contains lengths from gold standard data, while Sloleks and metaFida columns contain statistics for inferred chains on each corpus. Words in the BSSJB have chains with length from 0 (root words) to 6, with the most common length being 1, ie. words composed of just a root and a single additional morpheme. Regarding Sloleks and metaFida, if the chain extraction method returned a chain with less than two rules, the resulting chain was discarded to reduce the amount of noise. For this reason, some values in the table are missing. Even if this is taken into account, there is a large discrepancy between statistics of the inferred chains and chains found in the gold standard BSSJB dictionary.

Some chains occur more often than others. In Table 2 we see ten most frequent rule chains inferred on Sloleks and metaFida, with a relative frequency of words in the corpus

³ The implementation is available in the PRTPY library: <https://github.com/erelsgl/prtpy>

⁴ A perfect division that assigns each fold the same number of words is not possible due to a total amount of entries in the dataset not being divisible by 5.

Length	BSSJB freq.	Sloleks freq.	metaFida freq.
0	5.92%	-	-
1	38.23%	-	-
2	34.15%	87.14%	85.75%
3	15.31%	12.49%	13.86%
4	5.23%	0.36%	0.39%
5	1.07%	0.01%	<0.01%
6	0.10%	0.00%	0.00%

Table 1: Comparison between distributions of chain lengths. Column for the BSSJB dataset shows distribution of chain lengths on manually annotated data, while Sloleks and metaFida show distributions of inferred chains. Due to noise reduction, inferred chains with the length less than 2 were discarded, therefore the statistics are not directly comparable with BSSJB.

explained by these rules. All chains show a combination of two morphemes. The most frequent chain in both corpora is *NOUN* → *ADJECTIVE* (-en) → *ADVERB* (-o), see the following examples:

- *abeceda* 'alphabet' → *abeceden* 'alphabetical' → *abecedno* 'alphabetically'
- *časť* 'honour' → *časten* 'honourable' → *častnosť* 'honourability'
- *didaktika* 'didactics' → *didaktičen* 'didactic' → *didaktično* 'didactically'

Among the most productive rules are the first-stage adjectives on -en, which form the base for the second-stage nouns on -ost, -a and -ik and the verbs on -eti. For example:

- *absurd* 'absurd (noun)' → *absurden* 'absurd (adjective)' → *absurdnosť* 'absurdity'
- *žito* 'cereal (noun)' → *žitén* 'cereal (adjective)' → *žitnica* 'a grain silo'
- *dež* 'rain' → *dežen* 'rainy' → *dežník* 'umbrella'
- *ľad* 'ice' → *leden* 'icy' → *ledeneti* 'to freeze'

In the Sloleks corpus, three chains with a non-noun simplex (non-derivative from) stand out:

- VERB:X → ADJ:X-en → NOUN:X-ost (e.g., *ganiti* 'to move (emotionally)' → *ganjen* 'moved' → *ganjenost* 'emotions from being moved'),
- ADJ:X → VERB:X-ati → NOUN:X-anje,
- ADJ:X → NOUN:X-ik → NOUN:X-ica

On the other side, in the metaFida corpus there is only one chain with a non-noun simplex:

- VERB:X → NOUN:X-0 → ADJ:X-en

In order to evaluate the accuracy of the rule chain derivations we tasked a single expert in linguistics and word formation to manually verify the correctness of the entire inferred rule

Most Common Rules	Frequency
SloLeks	
NOUN:X → ADJ:X-en → ADV:X-o	3.49%
NOUN:X → ADJ:X-en → NOUN:X-ost	2.82%
VERB:X → ADJ:X-en → NOUN:X-ost	2.70%
NOUN:X → ADJ:X-ski → ADV:X-o	2.25%
NOUN:X → VERB:X-ati → NOUN:X-anje	2.06%
ADJ:X → VERB:X-ati → NOUN:X-anje	1.86%
NOUN:X → NOUN:X-a → NOUN:X-ica	1.79%
NOUN:X → ADJ:X-en → NOUN:X-ica	1.76%
ADJ:X → NOUN:X-ik → NOUN:X-ica	1.75%
NOUN:X → ADJ:X-en → NOUN:X-ik	1.60%
metaFida	
NOUN:X → ADJ:X-en → ADV:X-o	3.37%
NOUN:X → ADJ:X-en → NOUN:X-ost	2.00%
NOUN:X → NOUN:X-0 → ADJ:X-en	1.82%
NOUN:X → ADJ:X-en → VERB:X-eti	1.72%
NOUN:X → ADJ:X-ski → ADV:X-o	1.69%
VERB:X → NOUN:X-0 → ADJ:X-en	1.61%
NOUN:X → ADJ:X-en → NOUN:X-a	1.56%
NOUN:X → NOUN:X-a → NOUN:X-ica	1.52%
NOUN:X → ADJ:X-en → NOUN:X-ik	1.38%
NOUN:X → VERB:X-ati → NOUN:X-anje	1.38%

Table 2: Ten most common rule chains inferred on SloLeks and metaFida with a relative frequency of words in the corpus explained with this rules.

chain (in future work, we plan to extend this part by conducting inter-annotator agreement experiments). For each corpus, we randomly select words and rule chain that explains the formation of selected word. The words are selected such that all words from a particular corpus in the verification dataset have distinct rule chains. The rule chains were randomly chosen with probability of being selected proportional to the logarithm of frequency of this chain occurring in the vocabulary. Due to very small number of words exhibiting longer rules, as per Table 1, we selected 100 examples for rules of size 2, 100 examples for rules of size 3, and all available examples for each rule of sizes 4 and 5. In total, this procedure selected 233 words from SloLeks and 264 words from metaFida. Next, we exclude words starting with *b*, as the examples could be identical to the ones in the BBSJB gold standard. The results of manual evaluation are presented in Table 3. The results are relatively low, especially on metaFida which is a noisy data. There are several sources of mistakes, including: semantically unrelated words (e.g., *diva* 'diva' → *divji* 'wild'), incorrect order in word-formation chain (e.g., *krsten* 'baptismal' → *krst* 'baptism'), incorrect simplex, i.e. non-derivative form (e.g., *zobati* 'to nibble' (instead of *zob* 'tooth') → *zoben* 'dental').

Corpus	Chain length	Sample size	Correct	Accuracy
Sloleks	2	94	25	26.60%
Sloleks	3	82	18	21.95%
Sloleks	4	19	6	31.58%
metaFida	2	98	5	5.10%
metaFida	3	92	3	3.26%
metaFida	4	42	0	0.00%
metaFida	5	1	0	0.00%

Table 3: Results of the manual verification of rule chains inferred on Sloleks and metaFida.

5.2 Morphological Segmentation Evaluation

5.3 Evaluation metrics

In this section we present the results achieved by the inferred models on the task of morphological segmentation. For all models, we report precision, recall, F_1 score, and accuracy. We define F_1 score analogous to Ruokolainen et al. (2013). Each correctly predicted split between two morphemes in a word adds to the true positives (TP), under-splitting and over-splitting count towards false negatives (FN) and false positives (FP), respectively. As an example, for a ground truth segmentation *bank-o-mat* and a prediction *ban-ko-mat*, we have one false negative prediction (*bank•omat*, this split is not detected), one false positive (*ban•komat*, the split is added by the model but not present in the gold standard), and one true positive prediction (*banko•mat*).

The F_1 score is then defined as follows:

$$F_1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \quad (3)$$

We define accuracy of the model as the fraction of words with completely correct segmentation, or alternatively, as a probability estimate of the model returning a correct segmentation. Although this metric is usually not used in semantic segmentation task (cf. Batsuren et al. 2022, Narasimhan et al. 2015), we consider it highly relevant and intuitive for model comparison.

The results for unsupervised methods are provided on the entire dataset (see Section 4.2.1), while for the supervised method, where 5-fold cross-validation was used, the results are presented as an average score across all training runs, together with a standard deviation between them.

5.4 Evaluation results

We present the results in the Table 4.

Among the unsupervised baseline models, there is a consistent difference in F_1 score when model is inferred on metaFida vs. on the Sloleks corpus. Although the metaFida corpus

Model	Precision	Recall	F_1 score	Accuracy
Morfessor 2.0 (mFida)	63.99%	22.64%	33.45%	15.97%
Morfessor 2.0 (Sloleks)	40.53%	34.19%	37.09%	13.90%
MorphoChain (mFida)	62.42%	23.88%	34.54%	15.33%
MorphoChain (Sloleks)	63.33%	34.86%	44.97%	20.90%
BiLSTM-CRF	83.45% (± 0.9)	84.58% (± 2.7)	83.98% (± 1.2)	47.73% (± 1.7)

Table 4: Results on the inferred models on the task of morphological segmentation.

used in training is significantly larger (x6.5), the F_1 score is consistently improved with Sloleks corpus, especially the recall component of the metric. The MorphoChain model consistently outperforms the Morfessor 2.0 model on F_1 score. This is to be expected as MorphoChain model also includes semantic information when resolving the morphemes of the word.

The supervised BiLSTM-CRF approach shows the strongest performance on our dataset. All metrics show consistent performance over the folds as indicated by very low standard deviations which shows the model is not sensitive to the variability in the training data. An advantage of this model is that even though it is supervised it can be effectively trained on smaller amounts of labeled data.

6. Conclusion and future work

With this work we tackle the problem of automating the derivational morphology for Slovenian language with two complementary approaches. With one approach, we induce a model on annotated data of derivational dictionary and produce rules that explain transformation from a base word to a derived one. With the other approach, we induce a model for morphological segmentation and evaluate it on the derivational dictionary.

Although the extraction of rule chains provides a richer information about word formation, the accuracy of our approach is not satisfactory when evaluated on a random selection of words from Sloleks lexicon and metaFida corpus. Results on the metaFida corpus are significantly worse than those inferred on the Sloleks lexicon. One explanation for this is the amount of noise present in each dataset. Entries in the Sloleks lexicon were manually verified, which is not the case for metaFida corpus. This opens up a topic to be explored in future work, how to improve the rule-based chain extraction by incorporating the probabilistic estimates derived from the word frequencies, or even the semantic similarity of words as used in MorphoChain (Narasimhan et al. (2015)).

Morphological segmentation was explored by evaluating both unsupervised and supervised models, and evaluated on a dataset constructed from the derivational dictionary. Unsupervised models were induced on both Sloleks lexicon and metaFida corpus, while the supervised model was induced and evaluated on the constructed dataset using the 5-fold cross-validation. All unsupervised approaches have very low values of F_1 score and accuracy, but those results are comparable with results reported in related work (cf. Batsuren et al. (2022)).

The supervised approach based on the BiLSTM-CRF model achieves higher scores compared to the unsupervised approaches which is to be expected as it is trained on the BSSJB dataset with supervision. While care has been taken to prevent the model from overfitting on the root of the word and capitalizing on this during evaluation, the model is able to learn better patterns as the training and test set come from the same data distribution.

For future work, we will evaluate the BiLSTM-CRF model on other out-of-distribution datasets to fully gauge its performance in a practical setting. Furthermore, the current training and test data contain only words starting on letter *b*. While we assume the rules for morphological derivation are general across the vocabulary of a language, we would like to test the model on a more varied vocabulary to gauge the impact of this particular bias of our dataset. We also plan to leverage automated morphological segmentation for deriving novel rules from the actual corpora, which will enable to analyse word formation processes and formant combinatorics beyond the rules described in the BSSJB trial data. The developed methods have high potential for faster and corpus driven approaches to creation of contemporary derivational dictionaries.

7. Acknowledgements

This article was written in the framework of the project Formant Combinatorics in Slovenian (J6-3134) funded by the Slovenian Research Agency (ARRS). We also acknowledge the ARRS funding through the core programmes Knowledge Technologies (P2-0103) and The Slovenian Language in Synchronic and Diachronic Development (P6-0038).

8. References

- Arhar Holdt, Š., Čibej, J., Laskowski, C. & Krek, S. (2020). *Morphological patterns from the Sloleks 2.0 lexicon 1.0*. Slovenian language resource repository CLARIN.SI. <http://hdl.handle.net/11356/1411>.
- Batsuren, K., Bella, G., Arora, A., Martinovic, V., Gorman, K., Žabokrtský, Z., Ganbold, A., Dohnalová, Š., Ševčíková, M., Pelegrinová, K., Giunchiglia, F., Cotterell, R. & Vyloмова, E. (2022). The SIGMORPHON 2022 Shared Task on Morpheme Segmentation. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Seattle, Washington: Association for Computational Linguistics, pp. 103–116. URL <https://aclanthology.org/2022.sigmorphon-1.11>.
- Breznik, I.S. (2004). *Besnodružinski slovar slovenskega jezika, Poskusni zvezek za iztočnice na B (Word-family dictionary of Slovenian, Trial volume for headwords beginning with letter B)*. Maribor: Slavistično društvo.
- Čibej, J., Arhar Holdt, Š. & Krek, S. (2020). List of word relations from the Sloleks 2.0 lexicon 1.0. URL <http://hdl.handle.net/11356/1386>. Slovenian language resource repository CLARIN.SI.
- Cotterell, R., Kirov, C., Hulden, M. & Eisner, J. (2019). On the Complexity and Typology of Inflectional Morphological Systems. *Transactions of the Association for Computational Linguistics*, 7, pp. 327–342. URL https://doi.org/10.1162/tacl_a_00271. https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00271/1923163/tacl_a_00271.pdf.
- Cotterell, R., Müller, T., Fraser, A. & Schütze, H. (2015). Labeled Morphological Segmentation with Semi-Markov Models. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*. Beijing, China: Association for Computational Linguistics, pp. 164–174. URL <https://aclanthology.org/K15-1017>.

- Cotterell, R. & Schütze, H. (2018). Joint semantic synthesis and morphological analysis of the derived word. *Transactions of the Association for Computational Linguistics*, 6, pp. 33–48.
- de Marneffe, M.C., Manning, C.D., Nivre, J. & Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, 47(2), pp. 255–308. URL https://doi.org/10.1162/coli_a_00402. https://direct.mit.edu/coli/article-pdf/47/2/255/1938138/coli_a_00402.pdf.
- Dobrovoljc, K., Krek, S., Holozan, P., Erjavec, T., Romih, M., Arhar Holdt, Š., Čibej, J., Krsnik, L. & Robnik-Šikonja, M. (2022). *Morphological lexicon Sloleks 2.0*. Slovenian language resource repository CLARIN.SI. <http://hdl.handle.net/11356/1230>.
- Erjavec, T. (2012). MULTEXT-East: Morphosyntactic Resources for Central and Eastern European Languages. *Language Resources and Evaluation*, 46(1), pp. 131–142. URL <https://doi.org/10.1007/s10579-011-9174-8>.
- Erjavec, T. (2022). *Corpus of combined Slovenian corpora metaFida 0.1*. Slovenian language resource repository CLARIN.SI. <http://hdl.handle.net/11356/1746>.
- Erjavec, T. & Džeroski, S. (2004). Machine learning of morphosyntactic structure: lemmatizing unknown Slovene words. *Applied artificial intelligence*, 18, p. 17–41.
- Filko, M., Šojat, K. & Štefanec, V. (2019). Redesign of the Croatian derivational lexicon. In *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology*. Prague, Czechia: Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, pp. 71–80. URL <https://aclanthology.org/W19-8509>.
- Gladkova, A., Drozd, A. & Matsuoka, S. (2016). Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In *Proceedings of the NAACL Student Research Workshop*. pp. 8–15.
- Graham, R.L. (1969). Bounds on Multiprocessing Timing Anomalies. *SIAM Journal on Applied Mathematics*, 17(2), pp. 416–429. URL <https://doi.org/10.1137/0117039>. <https://doi.org/10.1137/0117039>.
- Hofmann, V., Pierrehumbert, J. & Schütze, H. (2020a). Predicting the growth of morphological families from social and linguistic factors. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020. ACL.
- Hofmann, V., Pierrehumbert, J.B. & Schütze, H. (2020b). DagoBERT: Generating derivational morphology with a pretrained language model. *arXiv preprint arXiv:2005.00672*.
- Huang, Z., Xu, W. & Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*.
- Juršič, M., Mozetič, I., Erjavec, T. & Lavrač, N. (2010). LemmaGen: multilingual lemmatisation with induced Ripple-Down rules. *Journal of universal computer science*, 16(9), p. 1190–1214.
- Kern, B. (2010). Stopenjsko besedotvorje. *Slavistična revija*, 58, p. 35–348.
- Kern, B. (2020). Obrazilna kombinatorika v besedotvornih sestavih glagolov čutnega zaznavanja, p. 67–79.
- Krek, S., Erjavec, T., Repar, A., Čibej, J., Arhar Holdt, Š., Gantar, P., Kosem, I., Robnik-Šikonja, M., Ljubešić, N., Dobrovoljc, K., Laskowski, C., Grčar, M., Holozan, P., Šuster, S., Gorjanc, V., Stabej, M. & Logar, N. (2019). *Corpus of Written Standard Slovene Gigafida 2.0*. Slovenian language resource repository CLARIN.SI. <http://hdl.handle.net/11356/1320>.
- Kurimo, M., Virpioja, S., Turunen, V. & Lagus, K. (2010). Morpho Challenge 2005-2010: Evaluations and Results. In *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*. Uppsala, Sweden: Association for Computational Linguistics, pp. 87–95. URL <https://aclanthology.org/W10-2211>.

- Lazaridou, A., Marelli, M., Zamparelli, R. & Baroni, M. (2013). Compositionally derived representations of morphologically complex words in distributional semantics. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. pp. 1517–1526.
- Ljubešić, N. & Erjavec, T. (2018). Word embeddings CLARIN.SI-embed.sl 1.0. URL <http://hdl.handle.net/11356/1204>. Slovenian language resource repository CLARIN.SI.
- Narasimhan, K., Barzilay, R. & Jaakkola, T. (2015). An Unsupervised Method for Uncovering Morphological Chains. *Transactions of the Association for Computational Linguistics*, 3, pp. 157–167. URL <https://aclanthology.org/Q15-1012>.
- Peters, B. & Martins, A.F.T. (2022). Beyond Characters: Subword-level Morpheme Segmentation. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Seattle, Washington: Association for Computational Linguistics, pp. 131–138. URL <https://aclanthology.org/2022.sigmorphon-1.14>.
- Romary, L. & Tasovac, T. (2018). TEI Lex-0: A Target Format for TEI-Encoded Dictionaries and Lexical Resources. In *Conference of the Text Encoding Initiative: TEI as a Global Language*. Tokyo. <https://doi.org/10.5281/zenodo.2613594>.
- Ruokolainen, T., Kohonen, O., Virpioja, S. & Kurimo, M. (2013). Supervised Morphological Segmentation in a Low-Resource Learning Setting using Conditional Random Fields. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*. Sofia, Bulgaria: Association for Computational Linguistics, pp. 29–37. URL <https://aclanthology.org/W13-3504>.
- Smit, P., Virpioja, S., Grönroos, S.A. & Kurimo, M. (2014). Morfessor 2.0: Toolkit for statistical morphological segmentation. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Gothenburg, Sweden: Association for Computational Linguistics, pp. 21–24. URL <https://aclanthology.org/E14-2006>.
- Šojat, K., Srebačić, M., Tadić, M. & Pavelić, T. (2014). CroDeriV: a new resource for processing Croatian morphology. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), pp. 3366–3370. URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/1074_Paper.pdf.
- Stramljič Breznik, I. (2005). Kvantitativne lastnosti slovenskega tvorjenega besedja v poskusnem besednodružinskem slovarju za črko B. *Slavistična revija*, 53(4), p. 505–520. URL <http://www.dlib.si/details/URN:NBN:SI:DOC-EWKNYRGH>. Bibliografija: str. 518-520 Summary.
- Toporišič, J. (2000). *Slovenska slovnica*. Maribor: Obzorja.
- Vidovič Muha, A. (1988). *Slovensko skladenjsko besedotvorje ob primerih zloženik*. Ljubljana: Znanstvena založba Filozofske fakultete, Partizanska knjiga.
- Vylomova, E., Cotterell, R., Baldwin, T. & Cohn, T. (2017). Context-aware prediction of derivational word-forms. *arXiv preprint arXiv:1702.06675*.
- Zundi, T. & Aavaajargal, C. (2022). Word-level Morpheme segmentation using Transformer neural network. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Seattle, Washington: Association for Computational Linguistics, pp. 139–143. URL <https://aclanthology.org/2022.sigmorphon-1.15>.

Using lexicography for learning mathematics

Theresa Kruse¹, Ulrich Heid¹, Boris Girnat¹

¹University of Hildesheim, Universitätsplatz 1, 31141 Hildesheim, Germany

E-mail: {kruset, heidul, girnat}@uni-hildesheim.de

Abstract

Students struggle with the transition from school to university in mathematics. One reason is that at school, mathematics tends to be presented as an ensemble of calculations rather than as a network of concepts. We plan to investigate how lexicography and e-dictionary construction can help students in this transition. In the paper, we introduce the concept of a seminar that uses lexicographic methods in first-year mathematics courses. In the seminar, students will be provided with basic lexicographic knowledge and thus enabled to discuss the newly learned concepts and the relations that hold between them. We also present the lexicographic concept of the resource to be developed in the course: We focus on its article structure and its access structure and describe both in terms of the function theory of lexicography. We suggest innovative access structures which can support the acquisition of mathematical concepts as well as of mathematical terminology. The article structures are based on an ontology structure of the subject matter domain with different kinds of concepts and relations between them.

Keywords: terminology; learning; mathematics

1. Introduction

In mathematics, students struggle especially with the transition from school to university (Geisler & Rolka, 2021). One of the reasons might be that, at school, mathematics tends to be presented as an ensemble of calculations rather than concepts. Thus, students have to learn that mathematics is basically a building constructed of definitions, theorems, and relations between them. We plan to investigate to which extent a lexicographic approach to e-dictionary construction can help in this transition.

In this paper, we present the concept of a seminar accompanying a regular lecture for first-year students in mathematics. In the seminar, the students collaboratively create a lexical resource on the concepts and terminology that they learn in the lecture. In the following, we discuss lexicographic methods as well as the design of the lexical resource to be created in the seminar.

Our contribution shows the concept of the planned dictionary as well as the structure of the seminar which is intended to accompany an introductory lecture in mathematics. In Sections 2 and 3, we present related work and the subject matter area. In Section 4, we describe the prerequisites based on the function theory. In Section 5, we present the lexicographic concept, and in Section 6, the concept of our planned seminar. We conclude in Section 7.

2. Related Work

About twenty years ago, Cubillo (2002) already used lexicography with chemistry students. To support them in understanding and learning concepts of chemistry as well as the pertaining terminology, they were invited to create their own (printed) dictionaries of the field. However, this exercise was not backed up by any lexicographic introduction or training. Since then, electronic dictionaries took root and almost replaced printed dictionaries in several fields (cf. Fuertes-Olivera, 2016).

Kruse & Heid (2020) present a concept of how to structure the mathematical terminology of graph theory for a lexicographic purpose. They establish the following conceptual categories: types of graphs (e.g. *Petersen graph*), parts of graphs (e.g. *edge*, *node*), properties (e.g. *bipartite*), activities (e.g. *(to) map*), theorems (e.g. *four color theorem*), mappings (e.g. *isomorphism*), algorithms (e.g. *Dijkstra's algorithm*). Between concepts of these classes, one or more of the following relations may hold: equivalence, synonymy, hypernymy/hyponymy, holonymy/meronymy, pertainymy, antonymy, mediality, analogy, alternative, attributivity, mapping, eponymy. A similar inventory of concept types and relations may be used in our project.

The lexicographic function theory was developed over several years and is presented by Tarp (2008) and Fuertes-Olivera & Tarp (2014) in its current form. The theory provides a framework to describe the usage situations of a dictionary based on the users' needs. The users can be characterized by their lexicographic knowledge, their terminological knowledge, their expertise level in a special field, and their language level. The users can be in communicative, cognitive, operative, or interpretative situations. The combination of user-profiles and situations leads to different needs which can be fulfilled by a dictionary and which motivate the dictionary design. Below, we analyze the lexicographic needs of first-year mathematics students and thus motivate our dictionary design (cf. Section 4).

Tall & Vinner (1981) and Vinner (1991) introduce the theory of concept image and concept definition in the didactics of mathematics. The concept image denotes non-verbal associations a learner has with a certain term. These associations are always influenced by personal experience and thus continuously re-shaped. It is difficult to exactly determine the concept image of a learner for a particular concept. It can only be expressed by the concept definition, i.e. by how a learner verbalizes a certain concept. Concept image and concept definition always interact. Following this theory, learning is the process of the development and the evolution of concept images and concept definitions. An electronic dictionary might support students in this process as it contains concept definitions that contribute to shaping the concept images.

3. Subject Matter Area

We focus on a lecture that is a general introduction to mathematics, intended for teacher students. In this course, students learn the concepts of algebraic structures like groups, rings, and fields as well as vector spaces and matrices. Aspects of these concepts are also used in engineering, economics, and natural sciences.

In the following, we introduce some mathematical concepts which are the basis for the examples used in Section 5. As the introduction of all the axioms necessary to properly

introduce the concepts from a mathematical perspective goes beyond the scope of this paper, we rather give a general description of the concepts.

A *relation* is – mathematically speaking – a set of ordered pairs. We use some examples to illustrate what that means. Common examples are the less-than-relation $<$, the divisibility-relation $|$ or the equality-relation $=$. We show some properties such relations can have. The first property we look at is *reflexivity* which means that the relation exists between two same elements, which only applies for divisibility and equality as $a|a$ and $a = a$ but not for less-than as $a < a$ is not true. Symmetry means that, if the relation holds for a and b , it also holds for b and a . This is only true for the equality as from $a = b$ it follows that $b = a$, but it is not true for less-than as $a < b$ does not imply $b < a$, and not for divisibility because $a|b$ only implies $b|a$ if $a = b$ but not in general. Another common property is transitivity which holds for all three examples: From $a < b$ and $b < c$, one can conclude that $a < c$ and similarly $a|b$ and $b|c$ implies that $a|c$; finally $a = b$ and $b = c$ implies $a = c$. If a relation is reflexive, symmetric, and transitive it is called an *equivalence relation*, which is only the case for equality in our examples.

Further possible properties of relations are among others *left-total*, *right-total*, *left-unique*, and *right-unique*. We do not discuss them here in detail as they require a broader mathematical basis but we introduce some terminology which is derived from these concepts. A *function* as taught in high school is a left-total and right-unique relation. If the function is also right-total it is called *surjective* and if it is left-unique it is called *injective*. If the function is surjective and injective it is called *bijective*. These terms are used in the examples in Figures 1 and 2 in Section 5.

4. Intended Usage Situations

In introductory university courses in mathematics, students have to learn concepts, the relations between them, and typical phrases of the specialized language of mathematics. At school, however, mathematics tends to be presented as an ensemble of calculations rather than concepts. Thus, students have to learn that mathematics is basically a building constructed of definitions, theorems, and relations between them. The course at hand consists of a lecture and related tutorial lessons. At the end of the course, students have to pass a written examination. Each year about 150 students have to attend the lecture in the first year of their teacher program. In the following, we describe the students' needs by relying on the function theory by Tarp (2008) and Fuertes-Olivera & Tarp (2014).

The intended users speak German at a first language level as they are studying in a German Bachelor's program. We also assume that they have an advanced level of English due to their school education. We regard them as laypeople in both, their mathematical concept knowledge and their mathematical language knowledge, as they are in their first year of study. Even if they have reached a certain degree in school mathematics which gives them useful background knowledge, we can safely assume this categorization, as academic mathematics highly differs from school mathematics in most cases.

Furthermore, we assume them to be acquainted with using online resources as general sources of information but are only beginning to rely on lexicographic tools for mathematics, as such resources are not commonly used in mathematical school education. While Wikipedia as a kind of lexicographic tool is often used by students in mathematics

(Henderson et al., 2017; Anastasakis & Lerman, 2022), we assume that they only begin using it in the course of their studies as Wikipedia presents mathematics in a way it is taught at universities but not in schools. As we work with first-year students, we have the possibility of changing or even shaping their habits in our seminar: Investigations show that Wikipedia articles do not always provide the highest quality information (Jayakody & Zazkis, 2015; Selwyn & Gorard, 2016; Dunn et al., 2019) and that they may not always be easy to understand for non-experts (Kruse & Heid, 2022).

The learners are in our case mainly in a systematic cognitive situation following the terminology by Tarp (2008) and Fuertes-Olivera & Tarp (2014). There might be smaller sporadic cognitive situations as well as short communicative situations but we neglect the latter two for our conceptualization as the main goal of the course is to provide mathematical knowledge, i.e. shaping the concept image as well as learning the corresponding concept definitions from a formal perspective. Cognitive situations with the need to consult an electronic dictionary might thus occur in the following ways: attending a lecture, watching a learning video, discussing with fellow students, working on tasks, or reading a script or a textbook.

The concept image not only consists of discrete concepts but certain relations occur between them on a conceptual level and are expressed in the concept definitions as semantic relations on a linguistic level. In a formal domain like mathematics, these two levels of relations are almost completely identical. Nevertheless, linguistic relations between terms also appear, e.g. synonymy: Several expressions denote the same abstract concept and should be presented by the same concept image. For example, students have to learn that the symbols $\{\}$, \emptyset and the term *empty set* all refer to the same concept, namely a set without any elements in it.

From these user prerequisites as well as their usage situations the following user needs evolve which should be fulfilled by lexicographic assistance: The most common need is to look up the definition of a given term. In this context, not only the formal definition but also further information on the usage of the term is useful, i.e. in the form of concrete examples. In some cases, users might need algorithms for carrying out certain calculations, e.g. the Euclidean algorithm to find the greatest common divisor of two natural numbers.

A similar need affects not only one but two terms as users might be interested in their relation; for example, if they denote the same concept (e.g. *node* and *vertex*) or if they exclude each other (e.g. *positive integers* and *negative integers*). Conversely, it might be the case that a user has the right concept in mind but does not know the term which is used for it. Another example need is that users contextualize definitions in the concepts they have already learned, e.g. a *tree* is defined as a graph that does not contain any cycles. The learning of the new concept *tree* requires knowledge of the concepts *graph* and *cycle*. From a user perspective, it might be interesting to find out for two given terms if their combination yields a new term. In all these cases, the dictionary should be able to provide assistance.

5. Lexicographic concept

Based on the users, their situations, and their needs, we present a dictionary concept, in particular regarding the article structure and the access structure. Further, based on the

idea that the dictionary content is developed by the students, it is a semi-collaborative dictionary at the first stage which might be used as a resource with only indirect user involvement later on by other students (cf. Abel & Meyer, 2016).

5.1 Article structure

The content given in an article (i.e. in a dictionary entry) depends on the type of the particular lemma. Building on the work by Kruse & Heid (2020) we use four concept categories: OBJECT, PROPERTY, THEOREM, METHOD. OBJECT comprises all kinds of mathematical entities with mappings, parts, and types as sub-categories. Examples are *set* or *group*. In the category PROPERTY, we comprise all properties these entities could have, e.g. *complete* or *bijection*. THEOREM are all kinds of mathematical statements, like propositions, lemmas, or theorems themselves. For our conceptualization, we do not differentiate if the theorem has been proven yet. The theorems make statements about the elements of the categories PROPERTY and OBJECT. The last category is called METHOD and comprises algorithms as well as mathematical strategies for proving.

Between the elements of the categories, different semantic relations exist. Some of them have been already pointed out in the description of the categories. The relations can exist between members of the same and of different categories. We work with the following relations:

- OBJECT₁ is hypernym of OBJECT₂
- OBJECT can have PROPERTY
- OBJECT has always PROPERTY
- THEOREM is about OBJECT
- THEOREM is about PROPERTY
- THEOREM₁ implies THEOREM₂
- METHOD is based on THEOREM
- METHOD can find OBJECT (with PROPERTY)
- METHOD₁ and METHOD₂ have same goal
- PROPERTY₁ implies PROPERTY₂
- PROPERTY₁ excludes PROPERTY₂
- PROPERTY₁ and PROPERTY₂ can co-exist

The list above is also visualized in Table 1. It should be read by starting with one of the items in the leftmost column; the item above the relation field is the second object of the relation; e.g. *THEOREM is about OBJECT*. The table only covers relations on the conceptual level. Further relations on the linguistic level can appear, like synonymy. Additionally, between two entities from the category OBJECT more relations than indicated here are possible like holonymy/meronymy or antonymy. The selection criteria that define which of them should be included in the dictionary will be developed in the seminar (cf. Section 6)

In an article, the names of concepts that are in a certain relation to the lemma are given in addition to the definition. For example for an OBJECT, the dictionary article gives the following information: hypernyms, hyponyms, facultative properties, mandatory properties, theorems about it, and methods how to calculate it. To avoid overloading the mathematics students with lexicographical terminology, we suggest using the general

	METHOD	OBJECT	PROPERTY	THEOREM
METHOD	has same goal as	can find	can find	is based on
OBJECT	can be found by	is hypernym/hyponym of/...	can be / is always	is mentioned in
PROPERTY	can be found by	is always attached to / can be attached to	implies / excludes / can co-exist with	is mentioned in
THEOREM	is basis for	is about	is about	implies

Table 1: Possible relations between concepts to be indicated in the microstructure

language paraphrases of the relations given above and using them directly as structural indicators.

It needs to be decided and evaluated how the article should be presented: For example, if it should be shown in one of the rather classical electronic views like panel view, tab view, explorer view or print view (Koplenig & Müller-Spitzer, 2014) or if more innovative forms should be used which give a better visualization of the network-like structure of mathematical conceptualizations (e.g. by means of knowledge graphs), as proposed in EcoLexicon (cf. e.g. León-Araúz et al., 2019). In Figure 1, we show such a presentation, focused on a single lemma, namely the term *equivalence relation*. It might also be possible to let users switch between different view formats as it might depend on the user and their particular need in a given situation which view fits best.

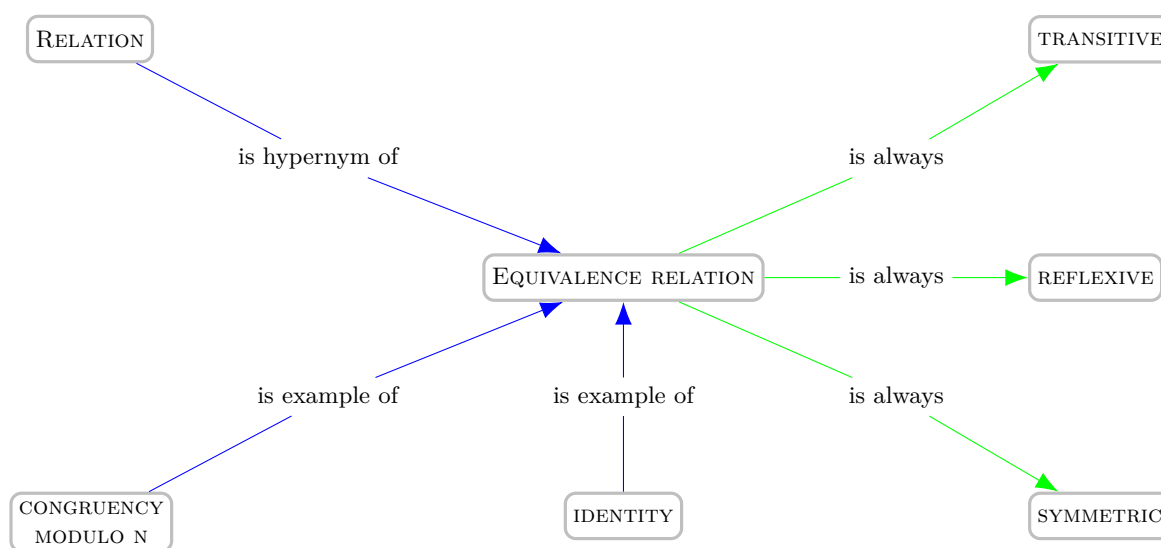


Figure 1: Example article for *Equivalence relation*

In addition to these conceptual categories, there is a category of domain-specific phraseology (e.g. *if and only if*, *q.e.d.*, *corollary*). But as this cannot be really integrated into the concept net it should be provided as part of the outer features.

5.2 Access structure and access paths

To satisfy the user needs we described in Section 4, likely more than one access structure will be needed, and students may use several types of access paths in combination. The example of the graph-based article structure can be the starting point of a graph-based access structure. It should allow users to zoom in and out of the graph. In addition to the graph-based access structure, an input-based search should be implemented as well a navigation.

The input-based search can be used to find definitions and examples for a given term. In other cases, the navigation or the graph-based search are probably useful aids. For example, if someone has a concept in mind but does not know (or does not remember) the appropriate term, they can navigate through the graph until they arrive at the right term. In some cases also a full-text search might help as well as an access structure using general language. To that end, the names of the concepts can be associated internally with quasi-synonyms from general language which allow users to find them. A search for *is equal to* or *is the same as* could then point the user to lemmas such as *isomorphic*, *identical*, or *equivalent*. Either by the graph-based structure or by the navigation it should be also possible to name two concepts and get the relation between them as a result. An example of such an excerpt from the concept net is shown in Figure 2.

6. Seminar concept

The dictionary as it is conceptualized here is not isolated but integrated into the lecture, as it is a task for the students to write articles of this semi-collaborative dictionary. Additionally, they can have their own private dictionary each, comparable to an individual flashcard set. Thus, the writing of the articles is a fixed part of the seminar accompanying the lecture. This individual student work is accompanied by sessions of the seminar in which the students can discuss their results.

We plan to give the students basic lexicographic training and access to a dictionary writing system that is optimized for the construction of specialized dictionaries, in particular for mathematics. Therein, they can note the concepts they have learned and indicate the semantic relations between these concepts. In the seminar, the students also learn basic lexicographic knowledge to be able to appropriately use the provided tool.

When building their personal e-dictionaries during the course, we introduce the students to a routine for including new terms:

1. Collect the new terminology and phraseology from your lecture notes and from the literature you worked with last week.
2. Choose a category for each term. If there are theorems that only have a number but no name, choose an appropriate name for them.
3. Find relations between the new concepts from the established relations.
4. Connect the new terms to the ones already learned.

If there are terms the students have difficulties allocating a category to, this will be discussed in the seminar. This empirical validation helps to improve the category system.

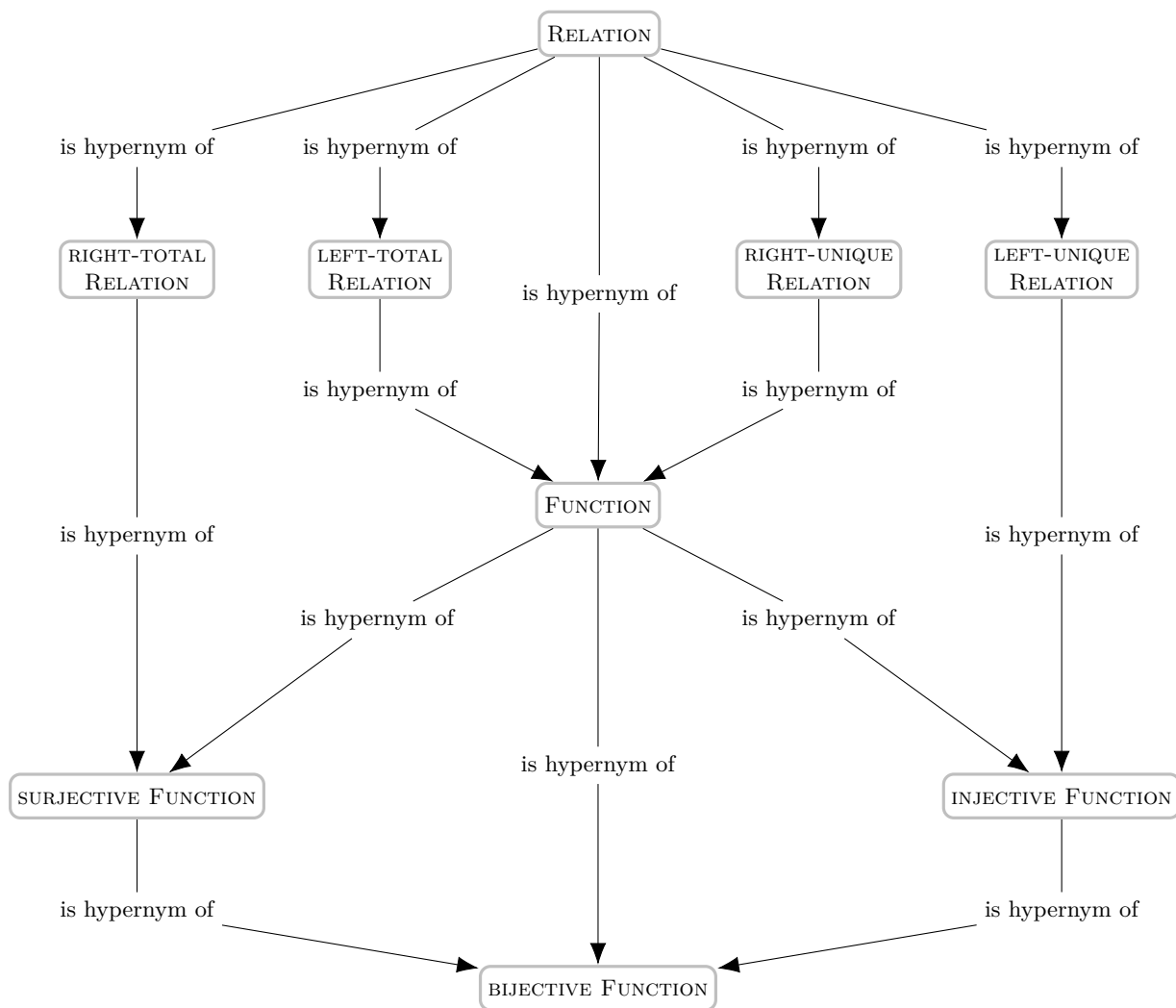


Figure 2: Extract from the network of concepts

New categories may be added to the conceptualization. The same applies to difficulties in assigning the relations between the terms.

Concurrently, the lexicographic structuring of the data helps the students to gain a deeper understanding of mathematics which in turn supports the acquisition of the content as it addresses the constructivist dimension of learning (Girnat & Hascher, 2021).

The dictionary writing system to be used has to fulfill certain requirements for the project. It needs to be easy to use as the students should be able to focus on learning the mathematical concepts rather than being distracted by the software. This also implies the inclusion of mathematical formulae by clicking or drag-and-drop as not all of the students – especially in the first year – have enough knowledge in scientific word processing, e.g. with \LaTeX . Additionally, it should be possible to search through the entries but also to navigate through them by use of the categories and relations, in order to use them in other articles. Further possible extensions are the export of flashcards and tagging for individual learning progress. We aim at an open source framework to be independent of economic interests and to allow students to continue using the system in the further course of their studies.

7. Conclusion and future work

In this paper, we presented a concept for a lexicographic resource that can be used in the process of learning mathematics. As a next step, we will implement a prototype of such a resource and use it in a lecture and a seminar with students to evaluate it. The implementation of the dictionary tool will likely be done by using existing frameworks that can be combined with the learning platform used in the courses. Choosing and establishing an appropriate system is the next step in the project.

Concerning the evaluation, we plan to compare the students in our proposed seminar with a group of students who attended a regular seminar with the same content. Both groups will be tested on their mathematical knowledge as well as on their mathematical beliefs (Pehkonen & Törner, 1996).

8. References

- Abel, A. & Meyer, C.M. (2016). Nutzerbeteiligung. In *Internetlexikografie. Ein Kompendium*. Berlin/Boston: De Gruyter, pp. 249–290. URL <https://doi.org/10.1515/9783050095615-009>.
- Anastasakis, M. & Lerman, S. (2022). Tool-Use profiles in Undergraduate Mathematics. *International Journal of Science and Mathematics Education*, 20(4), pp. 861–879. URL <https://doi.org/10.1007/s10763-021-10196-9>.
- Cubillo, M.C.C. (2002). Dictionary Use and Dictionary Needs of ESP Students: An Experimental Approach. *International Journal of Lexicography*, 15(3), pp. 206–228. URL <https://doi.org/10.1093/ijl/15.3.206>.
- Dunn, P.K., Marshman, M. & McDougall, R. (2019). Evaluating Wikipedia as a Self-Learning Resource for Statistics: You Know They’ll Use It. *The American Statistician*, 73(3), pp. 224–231. URL <https://doi.org/10.1080/00031305.2017.1392360>.
- Fuertes-Olivera, P.A. (2016). A Cambrian Explosion in Lexicography: Some Reflections for Designing and Constructing Specialised Online Dictionaries. *International Journal of Lexicography*, 29(2), pp. 226–247. URL <https://doi.org/10.1093/ijl/ecv037>.
- Fuertes-Olivera, P.A. & Tarp, S. (2014). *Theory and Practice of Specialised Online Dictionaries. Lexicography versus Terminology*. Berlin/Boston: De Gruyter. URL <https://doi.org/10.1515/9783110349023>.
- Geisler, S. & Rolka, K. (2021). “That Wasn’t the Math I Wanted to do!”—Students’ Beliefs During the Transition from School to University Mathematics. *International Journal of Science and Mathematics Education*, 19(3), pp. 599–618. URL <https://doi.org/10.1007/s10763-020-10072-y>.
- Girnat, B. & Hascher, T. (2021). Beliefs von Schweizer Schülerinnen und Schülern zum konstruktivistischen und instruktivistischen Lernen im Mathematikunterricht der Sekundarstufe I – Ergebnisse eines Large-Scale-Assessments zur Überprüfung mathematischer Grundkompetenzen (ÜGK) 2016. *Unterrichtswissenschaft*, 49(4), pp. 525–546. URL <https://doi.org/10.1007/s42010-021-00136-5>.
- Henderson, M., Selwyn, N. & Aston, R. (2017). What works and why? Student perceptions of ‘useful’ digital technology in university teaching and learning. *Studies in Higher Education*, 42(8), pp. 1567–1579. URL <https://doi.org/10.1080/03075079.2015.1007946>.
- Jayakody, G. & Zazkis, R. (2015). Continuous problem of function continuity. *For the learning of mathematics*, 35(1), pp. 8–14. URL <https://flm-journal.org/Articles/1FFA CAA6ADCA155A39C19F33EBFEB.pdf>.

- Koplenig, A. & Müller-Spitzer, C. (2014). Questions of design. In C. Müller-Spitzer (ed.) *Using Online Dictionaries*, volume 145 of *Lexicographica. Series Major*. Berlin/Boston: De Gruyter, pp. 189–204. URL <https://doi.org/10.1515/9783110341287.189>.
- 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.
- Kruse, T. & Heid, U. (2022). Learning from Students. on the Design and Usability of an E-Dictionary of Mathematical Graph Theory. In A. Klosa-Kückelhaus, S. Engelberg, C. Möhrs & P. Storjohann (eds.) *Dictionaries and Society. Proceedings of the XX EURALEX International Congress*. Mannheim: IDS-Verlag, pp. 480–493.
- León-Araúz, P., Reimerink, A. & Faber, P. (2019). EcoLexicon and by-products: Integrating and reusing terminological resources. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 25(2), pp. 222–258. URL <https://doi.org/10.1075/term.00037.leo>.
- Pehkonen, E. & Törner, G. (1996). Mathematical beliefs and their meaning for the teaching and learning of mathematics. *ZDM*, 28(4), pp. 101–108.
- Selwyn, N. & Gorard, S. (2016). Students' use of Wikipedia as an academic resource — Patterns of use and perceptions of usefulness. *The Internet and Higher Education*, 28, pp. 28–34. URL <https://doi.org/10.1016/j.iheduc.2015.08.004>.
- Tall, D. & Vinner, S. (1981). Concept image and concept definition in mathematics with particular reference to limits and continuity. *Educational Studies in Mathematics*, 12(2), pp. 151–169. URL <https://doi.org/10.1007/bf00305619>.
- Tarp, S. (2008). *Lexicography in the Borderland between Knowledge and Non-Knowledge: General Lexicographical Theory with Particular Focus on Learner's Lexicography*. Max Niemeyer. URL <https://doi.org/10.1515/9783484970434>.
- Vinner, S. (1991). The Role of Definitions in the Teaching and Learning of Mathematics. In D. Tall (ed.) *Advanced Mathematical Thinking*, volume 11 of *Mathematics Education Library*. Dordrecht: Springer, pp. 65–81. URL https://doi.org/10.1007/0-306-47203-1_5.

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

Geda Paulsen^{1, 2}, Ahti Lohk³, Maria Tuulik¹, Ene Vainik¹

¹ Institute of the Estonian Language, Roosikrantsi 6, Tallinn 10119, Estonia

² Uppsala University, Thunbergsvägen 3 L, Uppsala 75126, Sweden

³ Tallinn University of Technology, Ehitajate tee 5, 19086 Tallinn, Estonia

E-mail: geda.paulsen@eki.ee, geda.paulsen@moderna.uu.se, ahti.lohk@taltech.ee, maria.tuulik@eki.ee, ene.vainik@eki.ee

Abstract

In this study, we discuss the process of developing a multi-parameter application – the adjective similarity calculator (ASC) – that determines the relative adjectivity of a word or a word form. The tool relates the statistical summary of a word (form)’s corpus behaviour to the most typical and central aspects of the Estonian adjective: the adjectival corpus profile. To establish this profile, we use close-context patterns characterising adjectives and detectable in the corpus (see the experiments in Tuulik et al. 2022, Paulsen et al. 2022, and Vainik et al., 2023). The first prototype of the ASC will be evaluated based on clear cases of adjectives and PoS representatives overlapping with adjectival properties, but also based on words representing more distant classes. The main purpose of the application is to improve lexicographic work in categorisation procedures of the partly overlapping lexical categories to the adjective, particularly in such ambiguous cases as adjectivised participles, nouns and adverbs.

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

1. Introduction

The identification of the boundaries between lexical categories is a common task in part-of-speech tagging and lexicographic procedures. In many languages, these boundaries can be rather blurred. One of the most problematic word classes for lexicographers working with Estonian is the adjective (Paulsen et al., 2019, 188–189), a category overlapping with the noun, verb, adverb, pronoun (see Vainik, Paulsen, Lohk, 2021: 122–123) and ordinal (e.g., Erelt, 2017: 63). Lexicographers need to make decisions about lexicalising participles, a phenomenon common for other languages as well (e.g. English, where participles tend to develop into full-blown adjectives, such as *blessed* and *hammered*). Another phenomenon yielding ambiguity between lexical categories is systematic polysemy (see Langemets 2010, 159–161), emerging as conversional transposition (see Vare, 2006:199), in which a word can be used in another category without changing its form, e.g. *vigur* ‘trick’ (noun); ‘tricksy, prankish’ (adjective).

The prototypical behaviour of a word class can be captured by using corpus data, in the form of a corpus profile gathering the central morphosyntactic patterns characteristic to the category (see Tuulik et al., 2022; Paulsen et al., 2022; Vainik et

al., 2023). Is this profile operational as a template for comparing particular words or word forms? Motivated by this question, we introduce the first working prototype of the Adjective Similarity Calculator (ASC). This multi-parameter application is designed as a tool for lexicographers working with contemporary Estonian. The ASC is based on a statistical summary of a word (form)'s corpus behaviour¹ in comparison to the most typical aspects of the Estonian adjective. To establish the adjectival corpus profile, we use a selection of the most central close-context patterns characterising adjectives and detectable in the corpus (see the experiments in Tuulik et al, 2022, Paulsen et al 2022, and Vainik et al., 2023). To measure the distance of a word from the adjectival profile, we have selected an approach we call conformity assessment, derived from the methods we have tested in our previous studies (see Tuulik et al., 2022; Paulsen et al., 2022; Vainik et al., 2023).

The ASC elaboration process comprises two main optimisation issues: 1) the scope of the overlapping parts of speech targeted by the calculator, and 2) the optimisation of the thresholds of adjectivity on the basis of the results of a statistical analysis. The constituency of the set of automatically searchable test patterns should be applicable to all of the word classes overlapping with adjectives. The second issue involves adjustments to the method we use for calculating the distance of a word from the adjectival profile (see Tuulik et al., 2022; Paulsen et al., 2022; Vainik et al., 2023).

We will begin with a short overview of the Estonian adjective and the theoretical foundations behind the development of the adjectival corpus profile in Section 2. Here we describe the idea behind the statistic and its calculation and explain the similarity estimation method we call conformity assessment. The details of its realisation as a script interacting with the corpus via Sketch Engine API are given in Section 3. Section 4 is devoted to the demonstration of the results illustrated by the examples from seven lexical classes. The results are compared with the decisions made by lexicographers in the EKI Combined Dictionary (CombiDic) and checked against the corpus data using the Sketch Engine tool Word Sketch and concordances. The problems and future directions of development are discussed in Conclusions.

¹ The mechanism of the tool developed in this study can roughly be compared to the Find X function of Sketch Engine, providing additional information about the usage of a word; the solution is described in Kilgarrieff and Rychlý (2008). The Find X function uses frequencies of word forms to determine whether a word is predominantly used in plural or singular, whether a verb appears more in the present participle than in the passive form etc. The difference is, however, that our assessment battery is based on frequency data of a set of corpus patterns, not on frequencies of certain forms of a word.

2. Background

2.1 The Adjectival behaviour and its measurable patterns

In Estonian, there are five main word classes overlapping with adjectives: nouns, verbs, adverbs, pronouns and ordinals. Since the last two represent closed classes, we can say that the classes posing problems for lexicographers are mainly nouns, participles and adverbs. The noun-adjective type is the largest group showing ambiguity in word class², typically via transpositional derivation forming systematic polysemy networks (see Vare, 2006; Langemets, 2010). The second largest type is the adverb-adjective, consisting of words occurring in contexts typical of both classes, such as verbal or nominal modifiers. The transition zone between verbs and adjectives comprises the non-finite forms of verbs: participles³, gerunds and supines. (For a typology of overlapping lexical categories in Estonian, see Vainik et al., 2021.) The determination of the lexicalisation degree of these forms is a challenge for lexicographers and also poses huge problems for automatic morphological analysis.

Hence, there are several lexical categories approaching the morphological, syntactic and semantic properties⁴ of the adjective. Characteristically, the adjective occurs in a sentence together with a noun that it describes or modifies. The morphological characteristics of Estonian adjectives include inflection in case and number, forms of gradation and derivation. Syntactically, the adjective constitutes an adjective phrase by itself or together with its modifier(s). The constructions in which an adjective is most recognisable are those where it occurs as an attribute (1a) or as a predicative (1b). (About the Estonian adjective, see Viitso, 2001: 32–35, 42; Erelt, 2017: 405–406.) The adjective can be modified by an adverb in all of these configurations, exemplified below by the sentence (1b), where the intensifying adverb *täitsa* ‘quite’ precedes the predicative adjective *põnev* ‘exciting’.

- (1a) *Matka-me* *lumis-te-s* *mäge-de-s*.
hike-1PL snowy-PL-INE mountain-PL-INE
‘We hike in the snowy mountains.’

² Based on an analysis of the database on words and forms that are ambiguous in terms of their PoS categorisation, compiled mainly from lexicographic sources (Vainik et al., 2021: 122).

³ Participle endings in Estonian function partly as grammatical and partly as lexical suffixes (see Viht & Habicht 2019: 37); usually, participles are not regarded as independent PoS, except for corpus-tagging systems.

⁴ The semantic properties an adjective typically describes centre around dimension, age, value and colour (Dixon 2006: 3–4); the adjective has no internal temporal structure (expressing states rather than activities and permanent rather than temporary characteristics, see e.g. Fábregas, Marín 2017); the adjective can have semantic valency (Helbig 1992; Haugen 2013).

- (1b) *Film on täitsa põnev.*
 film is quite exciting-NOM
 ‘The film is quite exciting.’

A corpus-based application aimed at the identification of adjectival morphosyntactic patterns must focus on the structures that emerge as the most distinctive, as well as being detectable by the corpus tagging system. In our previous studies, we tested seven adjectival patterns (Tuulik et al., 2022; Paulsen et al., 2022; Vainik et al., 2023). We screened out four patterns⁵ that instantiate a central set of parameters of adjectival corpus behaviour. The selection is based on attributive and predicative constructions, but also the modifiability of an adjective candidate by an intensifying adverb (the abbreviation TW stands for the target word assessed for adjectival behaviour):

- 1) **the attribute pattern** (ATTR), targeting the sequence of the TW immediately preceding a noun. This pattern is based on the tendency of an adjective to modify the noun as an attribute (TW_NOUN), cf. *kollane pall* ‘yellow ball’.
- 2) **the sentence starter pattern** (ATTR/ST) adds a syntactic restriction to the attribute phrase by restricting its location at the beginning of a sentence. This differentiates inter alia verbal participles from adjectivised ones (e.g. past participles in compound tenses require the preceding auxiliary verb *olema* ‘be’).
- 3) **the adverb pattern** (ADV) targets the sequence of ADVERB_TW, a characteristic pattern of the adjectives in the corpus, particularly with scalar adjectives.
- 4) **the predicative pattern** (PRED) combines two sequences: a) the copula verb *olema* ‘be’ followed by the TW, and b) a copula verb followed by an adverb and the TW.

To improve the distinction of adjectival behaviour, we added an inclusive list⁶ of over 66 selected adverbs in queries of the adverb and predicative patterns (see Appendix 1). Hence, the patterns involving adverbs include only the adverbs typically modifying

⁵ We have excluded e.g. the pattern ascertaining the agreement condition from the set of the attribute patterns because it excludes indeclinable adjectives and (also indeclinable) lexicalised past participles. Another pattern characteristic to adjectives left out of the final set is the gradation pattern, because the study of prototypical adjectives showed considerable variation in the occurrence of comparative forms (see Paulsen et al 2022: 89–92). Also, a precondition for the use of the gradation pattern is an automatic generator of comparative forms of any given word, which would considerably increase the “footprint“ of corpus data analysis.

⁶ The list was compiled using the Sketch Engine word list tool, through which the 100 most frequent adjectives were extracted and the 30 most frequent adverbs for each of these adjectives were selected. The adverbs with frequencies of 10 or more were included in the list; some of the less frequent adverbs were included if they clearly expressed properties typical of adjective modifiers (e.g. intensifiers).

adjectives, leaving out, for instance, manner adverbs that predominantly modify verbs.

2.2 Conformity assessment and the estimated ranges of normal variation

The selection of the statistical method to calculate the similarity of a word with the prototypical adjective was based on previous experiments of three methods: conformity assessment⁷, Euclidean distance and cosine similarity (Tuulik et al., 2022, Paulsen et al. 2022a, Paulsen et al., 2022b, Vainik et al., 2023). Since the conformity assessment proved to be the most flexible (making possible the qualitative adjustment of the adjectival ranges of different lexical groups during the testing process) and, unlike the other tested methods, this enabled us to analyse the performance of a target word in different patterns separately⁸, we chose this method as the similarity assessment measure for the ASC.

Conformity assessment allows for the systematic comparison of the relative frequency values of a target phenomenon with the respective measurements of a standard. There is no predefined formula in conformity assessment, and the relevant parameters are estimated and compared one by one. On the basis of the measurements, it is possible to identify the ranges of adjectival behaviour typical for each pattern. This approach relies on the prototype theory and the idea that a lexical class is not a clear-cut phenomenon but shows variance to a certain degree⁹ (about the application of the prototype theory in lexical semantics, see e.g. Berlin & Kay 1969; Geeraerts 1989).

Using this approach, we operated with relative frequencies¹⁰ of a target word's occurrences in the four selected corpus patterns (cf. Section 2.1). We defined a range of adjectival behaviour for every pattern based on the marginal rates of the 100 most central and prototypical adjectives in Estonian (for a detailed description of the setting of ranges and the selection of the sample adjectives, see Paulsen et al., 2022a¹¹). These

⁷ We have used the term *deviation analysis* for this method in our previous studies (Tuulik et al., 2022; Paulsen et al., 2022; Vainik et al., 2023); the shift of perspective from deviation to conformity is for practical reasons: the application assessing a word's adjectivity counts matches of the behaviour of the prototypical adjective within the predetermined ranges of variation; thus, the process concerns compliance with the standard rather than deviation from it.

⁸ This is important regarding the main user group – the lexicographers – who may need to acquire explicit information about the patterns that the target word performs, such as an adjective (or not).

⁹ Our previous study (Vainik et al., 2023) indicated that even words marked as adjectives in dictionaries may differ in how high they score in different patterns. For instance, the actual usage of adjectives tends to incline towards either attributive (ATTR and ATTR/ST) or non-attributive (ADV, PRED) patterns. Hence, the patterns have a co-effect within a predetermined variation space.

¹⁰ For frequency results to be comparable, the absolute frequencies of the corpus pattern occurrences are divided by the word's general lemma frequencies.

¹¹ The sample of prototypical adjectives was randomly selected from lexicographically verified adjectives in the Basic Estonian Dictionary (about the dictionary, see Kallas et al., 2014). Note also that the analysis in Paulsen et al. (2022a) was based on the state-of-the-art ENC

ranges represent the estimation of normal variation for adjectives and define the adjectival corpus profile. The ranges of adjectival behaviour of the four morphosyntactic patterns selected as the basis for the ASD are presented in Table 1:

Patterns	adjectival ranges (relative frequencies)
ATTR	0.246–1
ATTR/ST	0.015–0.193
ADV	0.01–1
PRED	0.036–0.344

Table 1: The ranges of adjectival behaviour, defining the adjectival corpus profile

Although the adjectival ranges primarily drew on the corpus behaviour of the sample of 100 prototypical adjectives¹², we adjusted the ranges qualitatively to improve their ability to differentiate other word classes, particularly participles from adjectives. For example, when setting the range for the adverb pattern (ADV), we excluded the results of highly deviating adjectives (the non-scalar adjectives, e.g. *ühetoaline* ‘one-room (flat)’, *vasak* ‘left’ and *homne* ‘tomorrow’s’) by raising the lower limit. Also, to avoid excluding perfectly clear adjectives (e.g. *haruldane* ‘rare’), we raised the upper limits of the attribute (ATTR) and adverb (ADV) patterns to the maximum (1). Table 2 provides examples where the relative frequency results of the example words are analysed as either a conforming result (1) or non-conforming result (0) to the ranges of adjectival behaviour of the four corpus patterns:

Word	ATTR	ATTR/ST	ADV	PRED	Conforming patterns
<i>uhke</i> ‘proud’	0.473 (1)	0.03 (1)	0.112 (1)	0.19 (1)	4
<i>haihtuv</i> ‘vanishing’	0.72 (1)	0.037 (1)	0.028 (1)	0.028 (0)	3
<i>õnnitletud</i> ‘congratulated’	0.116 (0)	0 (0)	0.041 (1)	0.136 (1)	2

corpus available at that time, the ENC 2019. All calculations done in the present study are based on the ENC 2021 corpus; also, the adjectival ranges have been checked on ENC 2021.

¹² In the testing process of this study, we used the representative sample (N = 100) of prototypical adjectives and two control groups of participles tested in our previous studies (Paulsen et al., 2022; Vainik et al., 2023); as control groups also functioned six samples of word groups representing lexical categories overlapping with the adjective, used in Tuulik et al. (2022).

Word	ATTR	ATTR/ST	ADV	PRED	Conforming patterns
<i>hiir</i> ‘mouse’	0.237 (0)	0.015 (1)	0.007 (0)	0.03 (0)	1
<i>oskama</i> ‘can, know’	0.11 (0)	0.003 (0)	0.005 (0)	0.017 (0)	0

Table 2: Examples of conformity assessment analysis

3. Creating the calculator

3.1 The prerequisites of the ASC

There are basically four main requirements for creating an ASC application:

- 1) knowledge of the normal variation within the patterns of adjectival behaviour;
- 2) an established scale of adjectivity;
- 3) the availability of a morphologically annotated corpus for retrieving the frequency data of patterns and lemmas;
- 4) a script communicating with the corpus and retrieving statistics on the occurrences of the input word in the selected patterns, as well as calculating conformity assessment results.

The first requirement, the ranges of normal adjectival variation for each selected corpus pattern, were presented in Section 2.2 (Table 1). Conformity assessment results in each corpus pattern are the basis for evaluating a word’s closeness to adjectival behaviour. The counts corresponding to the criteria allow us to establish a scale of similarity to the adjectival corpus profile, which brings us to the second requirement of our calculator. The values matching the ranges of adjectivity vary over five degrees, presented in Table 3 (the function of the colours is to facilitate the perception of the values; these colours are also used on the display of the ASC):

Values	Scale
4	very likely
3	likely
2	ambiguous
1	unlikely
0	very unlikely

Table 3: The scale of adjectivity

The third requirement of the ASC is its data source: the ENC 2021 corpus, currently the newest and largest corpus of the Estonian language, with 2.4 billion words (Koppel & Kallas, 2022b). The ENC corpora (Koppel & Kallas, 2022a) are stored in the corpus query system Sketch Engine (Kilgarriff et al., 2004; Kilgarriff et al., 2014). ENC 2021 is pre-tagged, lemmatised, and disambiguated with the estNLTK 1.6.9 program (Laur et al, 2020). This corpus contains eleven sub-corpora¹³.

The fourth requirement of the ASC, a script retrieving the frequency data from ENC 2021 and linking the Sketch Engine system to the application, is described in the next subsection.

3.2 The algorithm

The algorithm¹⁴ we used for evaluating the adjectivity of a given word utilises statistics queried via the corpus query system Sketch Engine’s API¹⁵. First, we will provide an overview of the statistics queried and their query patterns.

To retrieve the necessary frequencies, we queried the Sketch Engine API using a specific set of query patterns. These patterns correspond to various occurrences of the input word in a given text corpus (in our case, ENC 2021). Table 4 displays the query patterns and the corresponding frequencies obtained through the Sketch Engine API.

Identification	Definition	Query
lemma_freq	overall frequency of the input word (lemma)	[lemma = "lemma"]
lemma_S_freq	the frequency of an input word followed by a noun	[lemma = "lemma"] [tag = "S.* "]
s_lemma_S_freq	the frequency of an input word when it is at the beginning of a sentence and followed by a noun	<s>[lemma = "lemma"] [tag = "S.* "]
Dlist_lemma_freq	the frequency of the input word if it is preceded by one of the predefined adverbs	([lemma = "adv1"] [lemma="adv2"] ...) [lemma="lemma"]

¹³ Web 2013, Web 2017, Web 2019, Web 2021, Feeds 2014–2021, Wikipedia 2021, Wikipedia Talk 2017, the Open Access Journals (DOAJ), Literature, the Balanced Corpus and the Reference Corpus.

¹⁴ The code is available at <https://github.com/PRG1978/A-multi-purpose-lexicographic-resource>.

¹⁵ About the communication with the Sketch Engine via automated HTTP requests, see more at <https://www.sketchengine.eu/documentation/api-documentation/>.

Identification	Definition	Query
be_DlistQ_lemma_freq	the frequency of the input word which may be preceded by one of the predefined adverbs preceded by "be" given as the base form	[lemma = "be"] ([lemma="adv1"] [lemma="adv2"] ...)?[lemma="lemma"]

Table 4: Frequency identification and query patterns

The queried statistics include the total frequency of the input word (lemma) and the frequency of the input word as part of a sequence (column “query” in Table 4). The first column shows five identifiers corresponding to the five frequencies obtained from the query pattern in the third column. The “lemma” in quotation marks represents the input word, while the query fragment “([lemma="adv1"]|[lemma="adv2"] | ...)” represents the inclusive list of over 66 adverbs (see Appendix 1). Exceptions in the ASC queries are non-inflected past participles with the endings *-dud*, *-tud* and *-nud*; for those forms, only text words are considered, not lemmas. It is important to note that all data processing is based on the frequencies of the actual occurrences of different PoS-interpretations in the corpus; the PoS of the target words is not predefined.

To estimate the adjectivity of a given testing word, we normalise the frequencies obtained from Table 5 using formulas (1) to (4):

$$lemma_S_norm_freq = lemma_S_freq / lemma_freq \quad (1)$$

$$s_lemma_S_norm_freq = s_lemma_S_freq / lemma_freq \quad (2)$$

$$Dlist_lemma_norm_freq = Dlist_lemma_freq / lemma_freq \quad (3)$$

$$be_DlistQ_lemma_norm_freq = be_DlistQ_lemma_freq / lemma_freq \quad (4)$$

These formulas involve dividing the second to fourth frequencies by the overall frequency of the test word (first row of Table 4). The resulting normalised frequencies are then checked against a set of predefined ranges of adjectival behaviour (see Table 1 in Section 2.2), following the steps of conformity analysis. After that, the corresponding adjectivity rate from the scale of values (as established in Table 3, Section 3.1) is found and displayed on the screen together with frequency data and numeral values in each pattern.

4. The calculator at work

The ASC works on the web address <https://adjcalculator.pythonanywhere.com/>. It can be opened in a separate window of a web browser while working in a dictionary writing system or checking corpus data via Sketch Engine platform. The application is supported by the most common browsers (Microsoft Edge, Mozilla Firefox, Chrome, Safari and Brave).

Figure 1 presents the user interface of the ASC. There is a search box below the title and further below are tabular fields for the results of a query. The user needs to press the “enter” button on the keyboard to start the query.

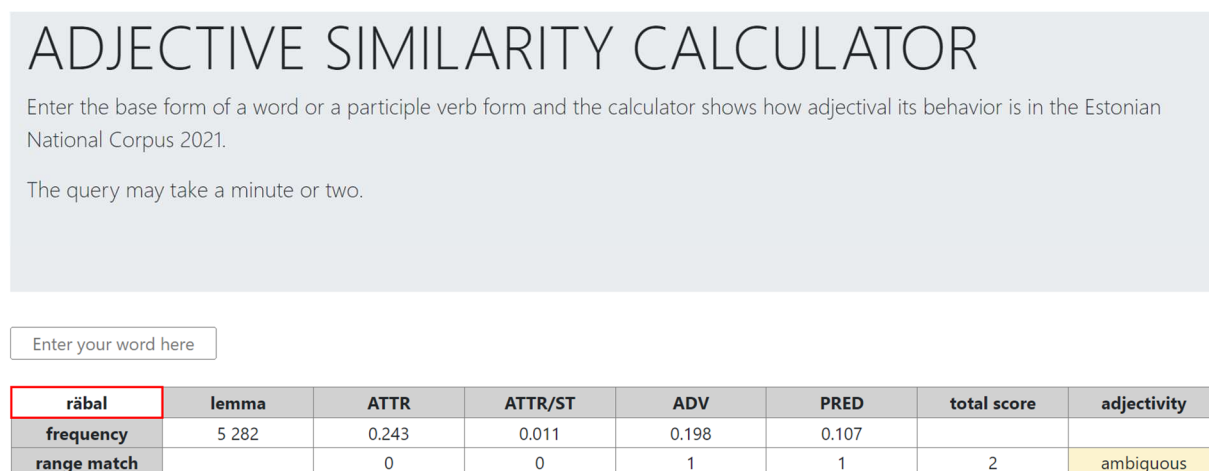


Figure 1: The user interface of the ASC

The word entered in the application – *räbal* ‘rag; miserable, shabby’ – is an existing entry in CombiDic, marked with two PoS tags: as a noun, mostly used in the plural (*räbalad* ‘rags’), and as an adjective (*räbal meeoleolu* ‘shabby mood’). The result of the calculator, the value “ambiguous”, reflects its twofold PoS affiliation. The outcome also shows that its use as an attribute is below the level of prototypical adjectives while the score in the adverb pattern and the role of the predicative match the criteria of typical adjectives. For closer examination of the actual corpus behaviour of this word, one can look at the concordances and/or Word Sketch tool in Sketch Engine.

4.1 Quantitative parameters

A single query by ASC took 3.6–88.4 seconds during the test period of the prototype. Because the ASC retrieves the frequency data via the Sketch Engine API (see Section 3.3), the speed of the ASC is dependent on the smoothness of queries by Sketch Engine. The query time may be shorter if a request has previously been processed.

4.2 Evaluation of the ASC and its results

In this section, we test words from seven different lexical groups with different lexicographic status to demonstrate how the ASC works and to evaluate the results. The examples selected for analysis represent different subtypes of the main word classes and exemplify how the ASC works with both non-ambiguous and ambiguous cases regarding PoS categorisation. The categories examined are adjectives, nouns, verbs,

adverbs, pronouns and numerals (both cardinals and ordinals). The participles, one of the most problematic areas in lexical categorisation, are not analysed in connection with verbs, but receive their own analysis in Section (4.2.4).

The words are checked for their status as a lexical entry in the CombiDic dictionary; the collocational analysis of the results is based on the Sketch Engine tool Word Sketch searching ENC 2021, the corpus the ASC also relies on. The usage examples come from ENC 2021, sometimes shortened to show the most relevant information.

4.2.1 Adjectives

First, we test three adjectives that are headwords in the CombiDic, to see if they match the adjectival profile measured by the ASC. These are the root adjective *ilus* 'beautiful, pretty', the derivative *pöörane* 'frantic, wild', and the indeclinable adjective *eri* 'separate; different'. As the ASC results depicted in (2a–2c) show, all three adjectives achieve the highest results, scoring in all four patterns.

(2a) *ilus* 'beautiful, pretty'

ilus	lemma	ATTR	ATTR/ST	ADV	PRED	total score	adjectivity
frequency	762 326	0.478	0.051	0.123	0.155		
range match		1	1	1	1	4	very likely

(2b) *pöörane* 'frantic, wild'

pöörane	lemma	ATTR	ATTR/ST	ADV	PRED	total score	adjectivity
frequency	24 982	0.625	0.043	0.116	0.088		
range match		1	1	1	1	4	very likely

(2c) *eri* 'separate, different'

eri	lemma	ATTR	ATTR/ST	ADV	PRED	total score	adjectivity
frequency	322 216	0.951	0.039	0.023	0.061		
range match		1	1	1	1	4	very likely

Let us now take a look at two adjectives – perfectly common and validated as adjectives in the CombiDic – categorised as ambiguous by the ASC. These adjectives are *ükskõikne* 'indifferent' and *sõjaline* 'military'. The screenshots of the ASC analyses show the scores concentrating either to the left (2d) or the right side (2e) of the table. These results reflect a division of labour in behavioural profiles among adjectives: there are adjectives that are predominantly used as attributes and those prevalent in the predicative role (see Vainik et al., 2023). Such a differentiation is identified in other languages (for English, see Bolinger 1967, Lassiter 2015: 145) but, to our knowledge, has not yet been investigated in Estonian.

(2d) *sõjaline* ‘military’

<i>sõjaline</i>	lemma	ATTR	ATTR/ST	ADV	PRED	total score	adjectivity
frequency	90 882	0.942	0.029	0.007	0.028		
range match		1	1	0	0	2	ambiguous

(2e) *ükskõikne* ‘indifferent’

<i>ükskõikne</i>	lemma	ATTR	ATTR/ST	ADV	PRED	total score	adjectivity
frequency	18 947	0.242	0.009	0.139	0.115		
range match		0	0	1	1	2	ambiguous

4.2.2 Nouns

The examples of nouns tested for adjectival behaviour are the concrete noun *kala* ‘fish’ (3a), the noun *kool* ‘school’ (3b), with twofold semantic content denoting both a building and an institution, and the abstract noun *armastus* ‘love’ (3c):

(3a) *kala* ‘fish’

<i>kala</i>	lemma	ATTR	ATTR/ST	ADV	PRED	total score	adjectivity
frequency	315 785	0.201	0.014	0.011	0.03		
range match		0	0	1	0	1	unlikely

(3b) *kool* ‘school’

<i>kool</i>	lemma	ATTR	ATTR/ST	ADV	PRED	total score	adjectivity
frequency	1 455 069	0.306	0.017	0.005	0.035		
range match		1	1	0	0	2	ambiguous

(3c) *armastus* ‘love’

<i>armastus</i>	lemma	ATTR	ATTR/ST	ADV	PRED	total score	adjectivity
frequency	380 904	0.172	0.012	0.012	0.055		
range match		0	0	1	1	2	ambiguous

As expected, all three nouns show low results in the ASC; they also vary in the actual realisation of tested corpus patterns. The first of them, *kala*, receives the label “unlikely adjective”, with one matching pattern, the ADV. The most frequent adverbs preceding *kala* are the degree adverbs (*palju* ‘a lot of’, *rohkem* ‘more’, *peamiselt* ‘mostly’), but also *lihtsalt* ‘simply’, and *hoopis* ‘instead; completely’. It is important to note that *rohkem* and *lihtsalt* are not included in the adverb list (cf. Appendix 1) because they are predominantly used as verb modifiers. The predicative pattern is possible but rather infrequent for *kala* (e.g. *hai on kala* ‘a shark is a fish’).

Why does the noun *kool* ‘school’ match the adjectives in the attributive patterns (ATTR and ATTR/ST)? The reason is the fact that, in Estonian, nouns can be used as genitive attributes, which is a frequent pattern for this word, as in the following collocations:

- (3d) *kooli* *söökla* / *õpetaja* / *õpilane*
 school.GEN canteen / teacher / pupil
 ‘the canteen / teacher / pupil of the school’

The abstract word *armastus* ‘love’ shows relatively high results in adverb and predicative patterns. The ranges of the adverb pattern are relatively large, for instance, for manner adverbs (*lihtsalt* ‘simply’); this noun is also modified by the adverbs included on our list of adjectival modifiers. Abstract nouns can be used predicatively as in *Jumal on armastus* ‘God is love’, and *elu on armastus* ‘life is love’.

To test the ASC for more ambiguous cases of PoS manifestation, we examine the words *haige* ‘sick; sick person’ and *lemmik* ‘favourite thing; favourite, dearest’, both tagged as noun and adjective in CombiDic. These words represent productive patterns of nominalisation and adjectivisation: as a result of ellipsis, basically every adjective can employ the syntactic functions typical to nouns (i.e. occur as a subject, object or predicative), and some nouns can be used as modifiers (Vainik et al., 2021: 123). The example of nominalisation, *haige* (3e), is labelled “very likely adjective”, corresponding to the adjective profile in every respect. The adjectivised noun *lemmik* (3f) matches only half of the patterns: apparently, this word still does not behave fully as an adjective.

- (3e) *haige* ‘sick, ill; sick person’

haige	lemma	ATTR	ATTR/ST	ADV	PRED	total score	adjectivity
frequency	286 438	0.286	0.025	0.063	0.101		
range match		1	1	1	1	4	very likely

- (3f) *lemmik* ‘favourite thing; favourite, dearest’

lemmik	lemma	ATTR	ATTR/ST	ADV	PRED	total score	adjectivity
frequency	230 895	0.252	0.014	0.01	0.021		
range match		1	0	1	0	2	ambiguous

4.2.3 Verbs

The verbs selected for illustration represent semantically different areas: the concrete motion verb *kõndima* ‘walk’ (4a) and two cognitive verbs, *nuputama* ‘figure, contrive’ (4b) and *mõtleva* ‘think’ (4c). The results show variation in corpus behaviour, even for the two cognitive verbs; the overall adjectivity assessments are very low (“unlikely adjective”).

- (4a) *kõndima* ‘walk’

kõndima	lemma	ATTR	ATTR/ST	ADV	PRED	total score	adjectivity
frequency	161 891	0.165	0.017	0.007	0.011		
range match		0	1	0	0	1	unlikely

(4b) *nuputama* ‘figure, contrive’

<i>nuputama</i>	lemma	ATTR	ATTR/ST	ADV	PRED	total score	adjectivity
frequency	17 646	0.079	0.005	0.021	0.012		
range match		0	0	1	0	1	unlikely

(4c) *mõtlemä* ‘think’

<i>mõtlemä</i>	lemma	ATTR	ATTR/ST	ADV	PRED	total score	adjectivity
frequency	2 162 728	0.17	0.007	0.009	0.099		
range match		0	0	0	1	1	unlikely

The motion verb *kõndima* ‘walk’ shows one match with the adjectival profile (4a), in the pattern measuring precedence of a noun at the beginning of a sentence (ATTR/ST). Estonian is a pro-drop language; hence, the pronoun before a verb can be omitted and the sentence can start with the verb followed by an adverbial consisting of a noun in a semantic case form:

(4d) *Kõnnin auto-ni / tänavale / sõbra-ga*
 walk-3SG car-TERM / street-ADE / friend-COM
 ‘I walk to the car / to the street / with a friend’

The two cognition verbs receive matches with the adjective profile, too, but in different patterns. The verb *nuputama* ‘figure, contrive’ (4b) often occurs after an adverb, which may coincide with adverbs typically modifying adjectives (e.g. the degree adverbs *natuke* ‘a little’, *palju* ‘a lot’ and *veidi* ‘a bit’). The verb *mõtlemä* ‘think’ scores in the predicative pattern (4c) for the reason typical of verbs: the main aspect contravening the quality of the PRED-pattern is that the copula verb *olema* ‘be’ is also used as the auxiliary verb in present or past tense forms in connection with compound tempus. An example of *mõtlemä* in a perfect tense is given in (4e).

(4e) *Ta on mõelnud töökoha vahetuse-le.*
 He/she be-3SG think-PAST-PART job.GEN shift-ALL
 ‘He/she has been thinking about a job change.’

4.2.4 Participles

One of the target categories for the ASC analysis is participles, constituting a fuzzy area between verbs and adjectives. Here we analyse the present and past personal and impersonal forms of the verb *lootma* ‘hope, expect’ (see 5a–5d). None of these forms are headwords in CombiDic; however, two of them (*loodetav* (5b) and *loodetud* (5d)) receive quite high adjectivity assessments (“likely adjectives”). These results are to be expected, as the forms with higher scores in fact demonstrate both verbal and adjectival usage patterns in corpus data and the forms with lower results are exclusively used in

verbal functions. Compared to the verbs analysed in the previous section, the ASC results show considerable variation.

(5a) *lootev* ‘hoping’

<i>lootev</i>	lemma	ATTR	ATTR/ST	ADV	PRED	total score	adjectivity
frequency	932	0.552	0.001	0.016	0.002		
range match		1	0	1	0	2	ambiguous

(5b) *loodetav* ‘(being) hoped, expected’

<i>loodetav</i>	lemma	ATTR	ATTR/ST	ADV	PRED	total score	adjectivity
frequency	3 450	0.679	0.044	0.004	0.042		
range match		1	1	0	1	3	likely

(5c) *lootnud* ‘(has) hoped, expected’

<i>lootnud</i>	lemma	ATTR	ATTR/ST	ADV	PRED	total score	adjectivity
frequency	9 304	0.087	0.0	0.012	0.368		
range match		0	0	1	0	1	unlikely

(5d) *loodetud* ‘(has been) hoped, expected’

<i>loodetud</i>	lemma	ATTR	ATTR/ST	ADV	PRED	total score	adjectivity
frequency	10 345	0.706	0.001	0.02	0.074		
range match		1	0	1	1	3	likely

Let us now analyse two examples of participles showing results from both extremes of the scale established in Table 3. A participle that might be considered a strong candidate for the status of the headword in the CombiDic is the present participle form *innustav* ‘encouraging, inspiring’. This form does not yet have the status of a dictionary entry, but receives the highest value of adjectivity with the score “very likely adjective” (see 5e). The adjectival usage is also confirmed by the examples in the ENC 2021 corpus.

(5e) *innustav* ‘encouraging, inspiring’

<i>innustav</i>	lemma	ATTR	ATTR/ST	ADV	PRED	total score	adjectivity
frequency	6 345	0.563	0.027	0.09	0.126		
range match		1	1	1	1	4	very likely

There are words or word forms with highly restricted usage, such as the participle *kohustatud* ‘be obliged to’, receiving 0 points in the ASC analysis. This past participle of an impersonal voice form is mainly used in the construction [X *on kohustatud* V_{inf}] ‘X is obliged to V’. Therefore, we can see under-representation in all patterns except the predicative pattern *olema_TW* (‘be’_TW), where this participle demonstrates clear overuse: the result of this pattern exceeds the adjectival ranges of 0.036–0.344, with a result of 0.575. This indicates that the upper limit of this range also functions well. The ASC analysis of *kohustatud* is presented in (5f):

(5f) *kohustatud* ‘obliged to’

kohustatud	lemma	ATTR	ATTR/ST	ADV	PRED	total score	adjectivity
frequency	72 033	0.228	0.0	0.002	0.575		
range match		0	0	0	0	0	very unlikely

4.2.5 Adverbs

We have selected three words representing different types of adverbs: the degree adverb *natuke* ‘a little, slightly’ (6a), the state adverb *sassis* ‘messy; confused’, indicating the physical or mental condition of the participant in an event (6b), and the sentence adverb *kindlasti* ‘certainly’ (6c).

(6a) *natuke* ‘a little, slightly’

natuke	lemma	ATTR	ATTR/ST	ADV	PRED	total score	adjectivity
frequency	793 043	0.248	0.022	0.006	0.096		
range match		1	1	0	1	3	likely

(6b) *sassis* ‘tangled, messy; confused’

sassis	lemma	ATTR	ATTR/ST	ADV	PRED	total score	adjectivity
frequency	13 906	0.246	0.015	0.155	0.151		
range match		1	1	1	1	4	very likely

(6c) *kindlasti* ‘for sure, certainly’

kindlasti	lemma	ATTR	ATTR/ST	ADV	PRED	total score	adjectivity
frequency	1 565 711	0.133	0.005	0.011	0.135		
range match		0	0	1	1	2	ambiguous

The first two adverbs score quite high in the ASC, labelled as “likely adjective”, whereas *kindlasti* ‘certainly’ is rated as “ambiguous”. The degree adverb *natuke* ‘a little’, as expected, conforms to the adjectival behaviour in both attribute patterns (in such collocations as *natuke aega/nalja* ‘a little bit of time/fun’) but is not modified by an adverb itself. As an intensifier, it precedes predicatives and thus occurs after the verb *olema* ‘be’ (*Uudis on natuke enneaegne* ‘The news is slightly premature’).

The fact that the state adverb *sassis* ‘messy; confused’ receives the highest rating, “very likely adjective” is quite predictable, as it belongs to a type of adverbs functionally overlapping with adjectives¹⁶. It is also frequently modified by the intensifying adverbs

¹⁶ The adverbs belonging to this type can also be analysed as (locative) case forms of nouns, e.g. *lokki-s* ‘curly’ [curl-INE] and, as in this example, a base noun (*lokk* ‘curl’) may be detectable. The static locative semantics (inessive and adessive cases) lead to the adjective interpretation; the directional (illative/elative; allative/ablative) forms of the same words (e.g. *lokk-i* ‘into a curly state’ [curl-ILL]) are read as either an adverb or as the respective case forms of nouns but not an adjective (See more in Vainik et al., 2021: 124).

included on the list of adverbs modifying adjectives (*täiesti / lootusetult / veidi sassis* ‘completely / hopelessly / a bit messy’).

The fact that the sentence adverb *kindlasti* ‘certainly’ only receives two points is not a surprise, as this word, particularly at the beginning of a sentence, affects word order by subject-predicate inversion and is typically followed by the predicate of the sentence (see Lindström 14–15).

1.1.1 Pronouns

Estonian pronouns function similarly to nouns, adjectives or numerals (Erelt 2017: 59). Let us test the indefinite pronoun *keegi* ‘someone’ (7a), the compound demonstrative pro-adjective *samasugune* ‘(the) same’ (7b), and the pro-numeral *tosin* ‘dozen’ (7c).

(7a) *keegi* ‘someone’

keegi	lemma	ATTR	ATTR/ST	ADV	PRED	total score	adjectivity
frequency	2 462 250	0.142	0.007	0.006	0.074		
range match		0	0	0	1	1	unlikely

(7b) *samasugune* ‘same’

samasugune	lemma	ATTR	ATTR/ST	ADV	PRED	total score	adjectivity
frequency	232 692	0.544	0.06	0.027	0.265		
range match		1	1	1	1	4	very likely

(7c) *tosin* ‘dozen’

tosin	lemma	ATTR	ATTR/ST	ADV	PRED	total score	adjectivity
frequency	25 710	0.73	0.061	0.021	0.037		
range match		1	1	1	1	4	very likely

The results correspond quite well with the word class the respective pronoun replaces. The pronoun *keegi* receives only one point in ASC and the label “unlikely adjective”, matching only the predicative pattern (see 7a). The proadjective *samasugune*¹⁷ (7b) behaves as a true adjective and scores on the highest level (“very likely adjective”). Surprisingly, at least at first sight, the pronominal *tosin* also receives the maximum score in ASC (see 7c). The usage patterns typical of an Estonian quantifier phrase explain the phenomenon: in the nominative case the quantifier governs its nominal complements by assigning to them the partitive case (*kaks õun-a* [two apple-PART]; see e.g., Erelt 2009: 19). This pattern explains the high score in the attribute pattern of *tosin*; this quantifier is often followed by a noun in partitive case (*tosin kilo/päeva/õuna* ‘dozen kilo/days/apples’). It is also modifiable by degree adverbs (*vähemalt tosin* ‘at least a dozen’, *peaaegu tosin* ‘almost a dozen’) and is used predicatively. All of these patterns contribute to the high outcome and explain inter

¹⁷ The proadjectives are marked as adjectives in CombiDic.

alia why the cardinal numerals generally meet all the requirements of adjectivity set by ASC (cf. example 8c in next section).

4.2.6 Ordinals and cardinals

The Estonian ordinals are basically considered to function as adjectives (Erelt 2017: 63). In the ASC, the ordinal *seitsmes* ‘seventh’ receives the assessment “likely adjective” with three points (see 8a). The result reveals the one condition in which Estonian ordinals do not behave as adjectives: the adverb pattern.

(8a) *seitsmes* ‘seventh’

seitsmes	lemma	ATTR	ATTR/ST	ADV	PRED	total score	adjectivity
frequency	78 166	0.754	0.084	0.007	0.04		
range match		1	1	0	1	3	likely

Interestingly, the ordinals do not score as high as cardinals, a category assumed to belong to the quantifier class. An example of a cardinal is given in (8b); the explanation given for the pronomeral *tosin* ‘dozen’ in the previous section also applies here.

(8b) *seitse* ‘seven’

seitse	lemma	ATTR	ATTR/ST	ADV	PRED	total score	adjectivity
frequency	280 745	0.714	0.049	0.015	0.052		
range match		1	1	1	1	4	very likely

4.3 Discussion of the results

The problem to be solved by the assistance of the ASC is whether to label a particular word or word form in a dictionary as an adjective or not. The quality of the ASC can be estimated by assessing its output of both non-ambiguous and ambiguous representatives of the word classes overlapping with adjectives. Another guiding line is formed by the decisions made by lexicographers so far, as well as a closer examination of the corpus behaviour of the tested words.

Overall, the ASC results indicate that the application works as intended: the rates of the words that are clearly non-adjectival fall into the lower interval in the similarity assessments (from 0 to 2; “very unlikely”, “likely”, or “ambiguous” regarding adjectival behaviour) and the ratings of cases that can be expected to behave to some extent adjectively fall into the upper interval (3–4, with the corresponding rates “likely” and “very likely”). When it comes to the analysis of validated adjectives themselves, we can conclude that almost all tested words received the rating “very likely”, with the highest score of 4.

Exceptions prove the rule, and this is also the case with the ASC. As our previous studies of adjectival behaviour have indicated, at least some of the (perfectly common) Estonian adjectives seem to prefer either attributive or predicative constructions. This

may be the reason why some quite “normal“ adjectives receive only average or even lower scores in the ASC (see Section 4.2.1). The existence and extent of this phenomenon needs closer examination, which is something the ASC can be used as a tool for.

Another factor interfering with the results are constructions typical to other classes than adjectives but (partly) overlapping with the patterns constituting the adjective profile. One question is: how can we rule out genitive attributes, the typical pattern of nouns modifying other nouns? A solution would be to work out some restrictive conditions. However, as the ASC analysis of the example noun *kool* ‘school’ (3b) showed, a noun frequently used in the attributive function still does not receive a summary value high enough to conform to the adjective profile. This outcome can even be seen as a positive aspect – the ASC allows one to study a noun's tendency to function as a genitive attribute.

An additional issue is the interference of other than predicative constructions around the copula verb *olema* ‘be’. There are different construction families clustering around *olema* in Estonian: compound tenses, existential clauses and possessive clauses. Manual checking of the corpus data regarding the words tested in this study has shown that the occurrences still mostly involve predicative clauses.

The inclusion of pronouns and numerals was mostly motivated by idle interest, as this closed class practically does not pose problems of categorisation. Still, the results of the ASC analysis were interesting, for instance, regarding the different behaviours of cardinals and ordinals: strikingly, the ordinals, regarded as adjectives, did not score as highly as the cardinals. Hence, it is surprising that the cardinals outscore ordinals in conforming adjectives: one would have expected that the meaning of a cardinal is not modifiable by scaling adverbs. This tells us, possibly, something about the practical fuzziness of the meanings. There is evidently a need for further studies in this area.

We are aware that the frequency results of the ASC directly depend on the quality of the tagging system, and we recognise that tagging and disambiguation errors affect the analysis. For instance, the morphoanalyser struggles with the form homonymy cases (e.g. *armutud* can be analysed as the nominative plural form of the adjective *armutu* ‘merciless’ or as the past participle impersonal form of the verb *armuma* ‘fall in love’). At any rate, the experienced lexicographer will discover the abnormalities and can check the results in the corpus to avoid problems.

The analysis in this study is solely based on morphosyntactic patterns, but adjectivity also undoubtedly has a distinctive semantic dimension. A direction for future studies could be the inclusion of semantic aspects in the adjectivity assessment battery. In addition, the semantic effect on the attributive-predicative prevalence noted in Section (4.2.1) is an interesting topic to explore further.

5. Conclusions

The ASC is a web-based application accessible to everyone. It takes a word whose similarity to adjectival behaviour is to be measured as input from the user and retrieves corpus data (the frequencies of the word form in requested positions – corpus patterns – and the total frequency of lemma). The tool calculates the relative salience of the instances of patterns and compares the values to the ranges of adjectival behaviour (cf. section 2.2). The ASC provides the outcome both in terms of numerical measures and verbal labels (as described in section 3.1). The calculator can be used to explore the syntactic behaviour of any word.

The constituency of the set of automatically searchable corpus patterns was tested to find the optimal solution, and the thresholds of adjectival behaviour determined on the basis of the results were adjusted. Decisions about previously tried methods for calculating the distance of a word from the adjectival profile (see Tuulik et al., 2022, Paulsen et al., 2022, and Vainik et al., 2023) were made. The ASC described in this study is the prototype of the application; the development process is still ongoing. Consultations with lexicographers who will test the ASC in actual use will be an important part of the further application design.

This study proved that corpus data can be used to establish the prototypical behaviour of a word class by creating a corpus profile of the central close-context patterns characteristic to the category. At least the adjective profile was confirmed to be operational as a template for comparing particular words or word forms. The study also showed that the patterns constituting the profile work in combination: no pattern alone can be used as proof of adjectivity.

6. Abbreviations

Glossing: ADE – adessive case; ALL – allative case; COM – comitative case; GEN – genitive case; INE – inessive case; NOM – nominative case; PART – partitive case; PAST – past tense; PL – plural; SG – singular; TER – terminative case; TRA – translative case.

7. Acknowledgements

This research was supported by the Estonian Research Council grant PRG1978. We thank the anonymous reviewers for their helpful suggestions.

8. References

- Berlin, B. & Kay, P. (1969). *Basic Color Terms: Their Universality and Evolution*. Berkeley: University of California Press.
- Bolinger, D. (1967). Adjectives in English: Attribution and predication. *Lingua*, 18, pp. 1–34. [https://doi.org/10.1016/0024-3841\(67\)90018-6](https://doi.org/10.1016/0024-3841(67)90018-6).

- CombiDic: *The EKI Combined Dictionary*. (2020). Hein, I., Kallas, J., Kiisla, O., Koppel, K., Langemets, M., Leemets T., ..., & Voll, P. Institute of the Estonian Language. Available at <https://sonaveeb.ee>.
- Erelt, M. (2017). Sissejuhatus süntaksisse [Introduction to syntax]. In M. Erelt & H. Metslang (eds.), *Eesti keele süntaks* [The Syntax of Estonian] (pp. 53–89). Eesti keele varamu III. Tartu: Tartu Ülikooli Kirjastus.
- Erelt, M. (2009). Typological overview of Estonian syntax. *STUF – Language Typology and Universals*, 62.
- Fábregas, A. & Marín, R. (2017). Problems and questions in derived adjectives. *Word Structure*, 10 (1), pp. 1–26.
- Geeraerts, Dirk (1989). Prospects and problems of prototype theory. *Linguistics* 27, pp. 587–612.
- Haugen, T. A. (2013). Adjectival valency as valency constructions: Evidence from Norwegian. *Constructions and Frames*, 5 (1), pp. 35–68. DOI: <https://doi.org/10.1075/cf.5.1.02hau>
- Helbig, G. (1992). *Probleme der Valenz- und Kasus-theorie* [Problems in Valency and Case Theory]. Tübingen: Niemeyer.
- Kallas, J., Tuulik, M. & Langemets, M. (2014). The Basic Estonian Dictionary: the first Monolingual L2 learner’s Dictionary of Estonian. In A. Abel, C. Vettori & N. Ralli (Eds.), *Proceedings of the XVI EURALEX International Congress: The User in Focus*, 15–19 July 2014, Bolzano, Bozen (pp. 1109–1119). Bolzano/Bozen: Institute for Specialised Communication and Multilingualism.
- Kilgarriff, A. & Rychlý, P. (2008). Finding the words which are most X. In: *Proceedings of the XIII EURALEX International Congress* (Barcelona, 15-19 July 2008). 2008. p. 433-436.
- 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*, 6–10 July 2004, Lorient, France (pp. 105–116). Lorient: Université de Bretagne Sud.
- 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, pp. 7–36.
- Koppel, K. & Kallas, J. (2022a). Eesti keele ühendkorpuste sari 2013–2021: mahukaim eestikeelsete digitekstide kogu [Estonian National Corpus 2013–2021: the largest collection of Estonian language data.]. *Estonian Papers in Applied Linguistics*, 18, pp. 207–228. <http://dx.doi.org/10.5128/ERYa18.12>.
- Koppel, K. & Kallas, J. (2022b). *Eesti keele ühendkorpus 2021* [Estonian National Corpus 2021]. <https://doi.org/10.15155/3-00-0000-0000-0000-08E60L>.
- Langemets, M. (2010). *Nimisõna süstemaatilise polüseemia eesti keeles ja selle esitus keelevaras* [Systematic polysemy of nouns in Estonian and its lexicographic treatment in Estonian language resources]. PhD thesis. Tallinn: Eesti Keele Sihtasutus.
- Lassiter, D. (2015). Adjectival modification and gradation. – Shalom Lappin, Chris Fox

- (eds.), *Handbook of Contemporary Semantic Theory*. Oxford: Wiley-Blackwell, 143–167. <https://doi.org/10.1002/9781118882139.ch5>
- 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*, May 2020, Marseille, France (pp. 7152–7160). European Language Resources Association (ELRA). Available at: <https://aclanthology.org/2020.lrec-1.0.pdf>.
- Lindström, L. (2005). *Finiitverbi asend lauses. Sõnajärg ja seda mõjutavad tegurid suulises eesti keeles* [The position of the finite verb in a clause: word order and the factors affecting it in spoken Estonian]. PhD thesis. Tartu Ülikooli kirjastus, Tartu.
- Paulsen, G., Vainik, E., Tuulik, M. & Lohk, A. (2019). The Lexicographer’s Voice: Word Classes in the Digital Era. In I. Kosem, T. Zingano Kuhn., M. Correia, J. P. Ferreria, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek & C. Tiberius (eds.), *Proceedings of the eLex 2019 conference: Smart lexicography*, 1–3 October 2019, Sintra, Portugal (pp. 319–337). Brno: Lexical Computing CZ, s.r.o. Available at: https://elex.link/elex2019/wp-content/uploads/2019/09/eLex_2019_18.pdf.
- Paulsen, G., Tuulik, M., Lohk, A. & Vainik, E. (2022a). From verbal to adjectival. Evaluating the lexicalisation of participles in an Estonian corpus. *Slovenščina 2.0*, 10(1), pp. 65–97.
- Paulsen, G., Vainik, E., Tuulik, M. & Lohk, A. (2022b). The morphosyntactic profile of prototypical adjectives in Estonian. Presentation held at the XX EURALEX International Congress 12–16 July 2022 in Mannheim, Germany.
- Tuulik, M., Vainik, E., Paulsen, G., & Lohk, A. (2022). Kuidas ära tunda adjektiivivi? Korpuskäitumise mustrite analüüs [How to recognize adjectives? An analysis of corpus patterns]. *Estonian Papers in Applied Linguistics*, 18, pp. 279–302. <http://dx.doi.org/10.5128/ERYa18.16>.
- Vainik, E., Paulsen, G. & Lohk, A. (2021). A typology of lexical ambiforms in Estonian. In Z. Gavriilidou, M. Mitsiaki & A. Fliatouras (eds.), *Proceedings of XIX EURALEX Congress: Lexicography for Inclusion*, 7–11 September 2021, Alexandroupolis, Greece, Vol. 1, pp. 119–130. Alexandroupolis, Greece: Democritus University of Thrace. Available at: <https://euralex.org/publications/a-typology-of-lexical-ambiforms-in-estonian/>
- Vainik, E., Paulsen, G., Tuulik, M. & Lohk, A. (2023). Towards the Morphosyntactic Corpus Profile of Prototypical Adjectives in Estonian. *Estonian Papers in Applied Linguistics*, 19, pp. 225–244. DOI: <http://dx.doi.org/10.5128/ERYa19.13>
- Vare, S. (2006). Adjektiivide substantivatsioonist ühe tähendusrühma näitel. [On substantivisation of adjectives: Analysing a semantic group] E. Niit. *Keele ehe*. Tartu: Tartu Ülikool. Tartu Ülikooli eesti keele õppetooli toimetised; 30, pp. 205–222.
- Viht, A. & Habicht, K. (2019). Eesti keele sõnamuutmine [The Estonian inflection]. Eesti keele varamu IV. Tartu: Tartu University Press.

Viitso, T.-R. (2003). Structure of the Estonian language: Phonology, morphology, and word formation. In M. Erelt (ed.), *Estonian language* (pp. 9–92). Tallinn: Estonian Academy Publishers.

Appendix 1. Inclusive list of adverbs used as a filter while searching for the ADV pattern (with English translations)

<i>väga</i>	‘very, highly’
<i>üsna</i>	‘quite, fairly’
<i>päris</i>	‘quite, right’
<i>piisavalt</i>	‘enough, sufficiently’
<i>niiivõrd</i>	‘so, insofar as’
<i>suhteliselt</i>	‘relatively, comparatively’
<i>üpris</i>	‘very, much, greatly’
<i>võimalikult</i>	‘possibly, as possible’
<i>suht</i>	‘relatively’ (colloquial)
<i>liiga</i>	‘too, excessively’
<i>äärmiselt</i>	‘extremely, utterly’
<i>küllaltki</i>	‘rather, fairly’
<i>täiesti</i>	‘entirely, wholly’
<i>erakordselt</i>	‘outstandingly, exceedingly’
<i>võrdlemisi</i>	‘comparatively, relatively’
<i>täitsa</i>	‘completely, quite’
<i>tõeliselt</i>	‘positively, truly’
<i>küllalt</i>	‘sufficiently, enough’
<i>ülimalt</i>	‘infinitely, immeasurably’
<i>sedavõrd</i>	‘inasmuch, so’
<i>liialt</i>	‘excessively’
<i>endiselt</i>	‘as before, still’
<i>üllatavalt</i>	‘surprisingly, amazingly’
<i>üksnes</i>	‘merely, only’
<i>igati</i>	‘to the outmost, in every way’
<i>palju</i>	‘much, a lot of, many’
<i>vähem</i>	‘less, fewer’
<i>ääretult</i>	‘boundlessly, infinitely’
<i>väga-väga</i>	‘very, greatly, highly’
<i>vähemalt</i>	‘at least, at any rate’
<i>kuivõrd</i>	‘insofar as’
<i>peamiselt</i>	‘chiefly, principally’
<i>enam-vähem</i>	‘more or less’
<i>tohutult</i>	‘infinitely, vastly’
<i>uskumatult</i>	‘incredibly, unbelievably’
<i>niigi</i>	‘already, as it is’
<i>hästi</i>	‘very, greatly’

<i>peaaegu</i>	‘almost, nearly’
<i>hoopis</i>	‘instead, entirely’
<i>hirmus</i>	‘very, greatly’
<i>mõnusalt</i>	‘pleasurably’
<i>enamvähem</i>	‘more or less’
<i>suuresti</i>	‘greatly, largely, highly’
<i>erinevalt</i>	‘variously, unlike, differently’
<i>kaugeltki</i>	‘by far’
<i>natuke</i>	‘a little’
<i>kindlasti</i>	‘for sure, certainly’
<i>nüisama</i>	‘just so; for nothing’
<i>iseenesest</i>	‘unintentionally, by itself’
<i>jätkuvalt</i>	‘continually’
<i>valdavalt</i>	‘predominantly’
<i>kahtlemata</i>	‘undoubtedly, definitely’
<i>eeskätt</i>	‘primarily, mainly’
<i>absoluutselt</i>	‘absolutely’
<i>tõenäoliselt</i>	‘probably, likely’
<i>meeletult</i>	‘deliriously, wildly’
<i>tõepoolest</i>	‘indeed, actually’
<i>kaunis</i>	‘pretty’
<i>täielikult</i>	‘completely’
<i>eriliselt</i>	‘specially, particularly’
<i>iseäranis</i>	‘particularly, exclusively’
<i>pisut</i>	‘a little, slightly’
<i>ülemäära</i>	‘excessively’
<i>parajalt</i>	‘moderately’
<i>veidi</i>	‘a bit’
<i>mõnevõrra</i>	‘somewhat’

Collocations Dictionary of Modern Slovene 2.0

Iztok Kosem^{1,2}, Špela Arhar Holdt¹, Polona Gantar¹, Simon

Krek^{1,2}

¹ Faculty of Arts, University of Ljubljana, Slovenia

² Jožef Stefan Institute, Slovenia

E-mail: Iztok.Kosem@ff.uni-lj.si, Spela.ArharHoldt@ff.uni-lj.si,

Apolonija.Gantar@ff.uni-lj.si, simon.krek@ijs.si

Abstract

In this paper, we present the Collocations Dictionary of Modern Slovene 2.0, which is a substantial upgrade of the first version, both in terms of content and the interface. The Collocations Dictionary contains 81,445 headwords, nearly 4.5 million collocations, and more than 17 million examples. Relevant findings of user studies and other related research, as well as the development of new methodology for automatic extraction of collocations from corpora, which is based on the syntactically parsed corpus data, have been used to improve the contents of the dictionary. The interface has undergone some important changes such as the immediate view of all the collocations in the entry, and the easy-to-understand three levels of entry completion. In terms of the data storage, a crucial development has been the introduction of the combination of the Digital Dictionary Database, which allows sharing the data among various resources produced at the Centre for Language Resources and Technologies at the University of Ljubljana, and a data warehouse, where all the automatically extracted collocations and additional metadata are stored.

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

1. Introduction

In 2018, the first version of the Collocations Dictionary of Modern Slovene was published (Kosem et al., 2018).¹ The dictionary contained automatically extracted collocations, and their examples, using (at that point) state-of-the-art tools such as Sketch Grammar and GDEX, customised for Slovene (Gantar et al. 2016). A selection of entries was provided in the finalised form, using post-editing methodology.

Over the past four years, a great deal of research related to the Collocations dictionary and the phenomenon of collocations in Slovene has been conducted, from the analysis and improvement of automatic extraction methods, lexicographic workflow, and data modelling, to user experience and participation. A project named Upgrading

¹ Collocations Dictionary of Modern Slovene 1.0 is available as a database at <http://hdl.handle.net/11356/1250>.

fundamental dictionary resources and databases of CJVT UL funded by the Slovene Ministry of Culture in 2021-22 provided the opportunity to implement the improved methods and new solutions into the next version of the Collocations Dictionary.

In this paper, we first present the developments since the launch of version 1.0 of the Collocations Dictionary of Modern Slovene. These developments include the results of various studies with the users of the dictionary and the improvement of collocation extraction methods, as well as the relevance of the latest trends in data storage and resource linking. Then, we look in detail at the new features of version 2.0 of the Collocations Dictionary, including the data extraction (and selection) method, and the inclusion of collocational data into the Digital Dictionary Database for Slovene. Furthermore, we also take a closer look at the changes in the interface, especially in terms of data visualisation and user participation, i.e., the crowdsourcing module. We conclude the paper with a short outline of future plans, both short-term and long-term.

2. Collocations dictionaries

The importance of collocation has been known since Firth's (1957: 11) famous statement "You shall know the word by the company it keeps", and the phenomenon has been analysed in detail since the arrival of large corpora. However, the compilation of collocation dictionaries for languages other than English, and especially the systematic inclusion of collocational information in general language dictionaries is a more recent trend. There are numerous collocations dictionary projects, either completed or ongoing, and we focus on those that have influenced the further development of the Collocations Dictionary of Modern Slovene. The first one to mention is the Estonian Collocations Dictionary (Kallas et al., 2015) which was compiled using the same methodology as we have been using in the compilation of the Collocations Dictionary for Modern Slovene, namely post-editing lexicography (curation of automatically extracted data). The Estonian Collocations Dictionary does differ in certain characteristics, for example, it was aimed at non-native speakers of Estonian, offers definitions only for polysemous words etc. The Estonian Collocations Dictionary is no longer available as a standalone source, as it has been integrated into the EKI Combined Dictionary.²

Similar to the Estonian Collocations Dictionary in terms of target audience is *Woordcombinaties* (Colman and Tiberius, 2018), a Dutch Collocations Dictionary. This is an ongoing project, which is in the process of switching to post-editing methodology, i.e., the selection of collocations is still done manually from the Sketch Engine corpus tool. Currently, the main focus of the dictionary are verbs. The users can choose from three different views: collocations (divided by syntactic structures), examples of use, and patterns (based on the Corpus Pattern Analysis by Hanks, 2004).

² <https://sonaveeb.ee/>

Targeted at native speakers such as the Collocations Dictionary of Modern Slovene is Croatian Web Dictionary – Mrežnik (Hudeček & Mihaljević, 2020a),³ currently available in a demo version (letters A-F). Mrežnik is a general language dictionary with a significant section in each entry dedicated to collocations (Hudeček and Mihaljević, 2020b). Collocations are divided into blocks introduced by collocational questions and phrases, modelled after the *elexiko* project (Haß, 2005; Storjohann, 2005; Klosa, 2015). Methodologically, Mrežnik is more similar to the *Woordcombinaties*, using a combination of manual insertion of collocations into the dictionary-writing system TLex from the Sketch Engine tool (Hudeček & Mihaljević, 2020b).

The reports by the authors of the abovementioned projects, as well as of other similar projects, point to several common issues of using collocations for dictionary purposes. One of the main ones is the abundance of data, both good and bad. While examining (long) lists of collocation candidates, the lexicographers need to identify the good ones, discard the bad ones, and then also often make a further selection among the good ones. This is far from straightforward; while some bad collocation candidates can be immediately identified, others can be confirmed as bad only after examining corpus examples. Similarly, there are levels of good collocation candidates; cut-off points need to be made not only in terms of how much data the lexicographers need to analyse but also how many collocations one wishes to present to the users. In this respect, it is also crucial to have the criteria for what constitutes a collocation, and what is its relation to other multi-word units, clearly delineated from the onset. The approach we used is described in Kosem et al. (2019) and Gantar et al. (2019).

A related issue is the origin of corpus data and the quality of annotation, which affects the quality of collocation candidates. The origin of bad collocation candidates can often be attributed to the problematic contents of the corpus (e.g., machine-translated texts from the web, Koppel et al., 2019) or errors in lemmatisation, part-of-speech tagging or parsing (Koppel et al., 2019; Pori and Kosem, 2021).

Another challenge is the data model, i.e., where and how is the collocational data stored, which lexicographic decisions are stored (only good candidates or also bad), how are the latest changes in the language monitored and incorporated into the existing data etc. The approach of editing data directly in relational databases where the data can be shared across headwords (i.e., lexical items) is being used by an increasing number of institutions, however editing the dictionary in the XML format still seems to dominate (Tiberius et al., 2022: 9).

Even after addressing all these issues and publishing the dictionary, there is one other aspect to consider, namely the dictionary user. In the next section, we present the findings of the studies conducted among the users of the Collocations Dictionary of

³ <https://rjecnik.hr/mreznik/>

Modern Slovene, as well as other relevant research on the use and consultation of collocations.

2.1 User studies

The most influential for the development of the second version of the Collocations Dictionary was the study by Pori et al. (2020; 2021), which investigated the attitudes of four different groups of users (teachers of Slovene as L1, teachers of Slovene as L2, proofreaders and translators, lexicographers) towards the Collocations Dictionary, and the way in which they used the dictionary. Using the evaluation interview based on the guided think-aloud method, the users were asked to conduct random searches of their own choice, conduct pre-determined searches, and comment on the general usefulness of the dictionary and its look. The most important findings can be summarized as follows:

- the attitudes towards the inclusion of automatic collocations were overwhelmingly positive, under the condition that the users are provided with corpus examples for context and a clear warning about the nature of such data (this being particularly stressed by language teachers).
- the pyramid icon indicating the level of entry completeness was considered by many to be not noticeable enough, the information it conveys should have been presented more clearly.
- the dictionary interface was evaluated as very good, all the features were found to be very useful and easy to use. An often-mentioned suggestion was the use of clear headlines or descriptions instead of icons, or at least adding descriptions of icons.
- while initially showing a selection of the top four most salient collocations of each syntactic structure, and having all the collocations in the structure available on a click was considered useful by participants, there were some doubts over whether most of the users ever get to the additional content. This can be considered problematic given that corpus examples are only provided at the stage of seeing all the collocations.
- the links to the corpus were considered very important, crucial even.
- some users wanted additional information on collocations, for example the information on frequency or saliency.
- the crowdsourcing part was considered useful by some participants, especially proof-readers and translators, although they usually lack time to contribute. On the other hand, teachers expressed concerns about the usefulness of the feature if used by less advanced language users.

Another relevant study was conducted by Arhar Holdt (2021) who looked at the preferences (and expectations) of 415 users of the Collocations Dictionary on the ordering of collocations in the dictionary interface. The questionnaire consisted of asking the participants to: list by memory three collocations of a given headword; select the top three syntactic structures they would like to see in the entry; select five collocations among the ones offered for a given headword and order them according to the perceived importance; provide the criteria used for ordering; provide other comments. The findings showed that the user expectations in terms of preferred syntactic structures more or less matched the order of structures provided in the dictionary. On the other hand, the users clearly preferred, and expected, the collocations to be ordered by frequency rather than by saliency; this is in contrast to how the collocations were ordered in the interface of Collocations Dictionary 1.0. Interestingly, other dictionaries are also not unified in this approach: the Estonian Collocations Dictionary orders collocations by frequency, and the Dutch *Woordcombinaties*, *Mrežnik* and the Macmillan Collocations Dictionary by alphabetical order.

Relevant to the crowdsourcing aspects of the Collocations Dictionary was the study by Pori and Kosem (2021), which included an experiment with six linguists who voted on the suitability of collocation candidates based on the collocation and its randomly selected example. The possible answers to the question of whether a candidate is a collocation were Yes, No, I don't know. While the main aim was to evaluate the reliability of the automatic extraction method, the study also revealed that one needs to have a clear definition of collocation to be able to decide on its relevance/suitability. Furthermore, in the pilot study, the participants often pointed out that many collocations seem perfectly fine and only a highly skilled person who knows what to look for can spot issues such as collocation not matching the syntactic structure (e.g., “*angažirati izvedenca*”, eng. to hire an expert, found in the syntactic structure verb + noun in genitive whereas it is in fact verb + noun in accusative). One other finding was that often more than one example was needed to be able to validate the collocation.

Valuable experience for crowdsourcing collocations was gained when developing the Game of Words (Arhar Holdt et al., 2021). Testing various game modes showed that for crowdsourcing collocations an implicit, gamification method is much more appropriate than an explicit method. In other words, much better and more reliable results are obtained if the users (players) are not aware they are providing collocational information, for example by listing collocates or distributing them to relevant headwords, as opposed to being asked directly whether something is a collocation or not. Relatedly, we also conducted an experiment where a group of students was asked to assign examples of collocations to relevant senses of selected headwords; the findings proved such a task to be extremely reliable (there was 100 % annotator agreement in over 80 % of cases) for various purposes: determining the understandability of indicators and sense division, indicating whether examples have enough context and indirectly

determining their quality/suitability for dictionary purposes, and to some extent confirming the relevance of the collocation (even though this was not the primary goal).

The findings of all these studies provided a point of departure in our planning of the second version of the Collocations Dictionary of Modern Slovene.

3. Collocations Dictionary of Modern Slovene 2.0

The second version of the Collocations Dictionary of Modern Slovene (Kosem et al., 2022)⁴ contains 81,445 headwords, nearly 4.5 million collocations, and more than 17 million examples. In comparison with version 1.0, there are more than twice as many headwords (35,989 in version 1.0), but 40% fewer collocations and nearly 50% fewer examples. This is a direct consequence of newly introduced extraction parameters, which is only one of the many changes introduced in version 2.0.

3.1 Data extraction – a new methodology

One of the important methodological differences from the first version of the Collocations Dictionary is the method of automatic extraction, of both collocations and examples. Collocations are entirely new, i.e., they were extracted from syntactically parsed corpus data (Krek et al., 2022; Krek et al., 2021), as opposed to an extraction based on POS-tagged data which was used for the first version. A new formalism defines dependency syntactic relation within a collocation, and also defines “constraints on any level of annotation, from morphology (parts-of-speech and their properties), syntactic dependency relations, concrete lexical items, and any other types of annotation that can be used for other purposes, e. g. semantic roles, semantic types, word senses, etc.” (Krek et al., 2022: 241). These constraints can be also used to specify the form of each component found in the corpus to be used in a specific collocation, an option that is very important for storing the collocation in the database as well as its presentation to the users. With a new formalism, we were able to separate verbal structures in terms of negation and reflexiveness, adding more syntactic structures to the list. The total number of syntactic structures is currently 82, and they include collocators belonging to four word classes: nouns, verbs, adjectives and adverbs.

With the new method giving more reliable results, combined with the fact that certain structures excluded from version 1.0 proved to be very important for certain headwords (e.g., the first version did not include ‘subject + verb’ due to many bad collocation candidates), we decided to include all 82 syntactic structures in the second version. It is important to note that on the one hand, headwords only contain structures which include the headword’s part of speech *s* (e.g., ‘noun + preposition + noun in accusative’ is found only for nouns), and on the other hand, the number of structures is even higher

⁴ The dictionary is available at <https://viri.cjvt.si/kolokacije/eng/>.

if we take the position into account (e.g., noun headword can be found in the aforementioned structure in the initial or final position). However, this in return meant reducing/limiting the number of collocations per structure to avoid information overload for the users. While the maximum number of collocations per syntactic structure in version 2.0 is 10, more collocations (up to 25) are offered for the structures that proved more collocationally-productive in the research studies (e.g. verb + noun in the accusative, adjective + noun, noun + noun in the genitive).

As far as headwords are concerned, the decision was made to extract collocations for all the nouns (excluding proper nouns), adjectives, adverbs and verbs in the Slovene Digital Dictionary Database (see the next section). The only other parameter used was a minimum frequency of 4 for collocations. Out of 138,032 candidate headwords, 81,445 met this condition; most of the headwords were single words, only 128 were compounds.⁵ For the automatic extraction, we imposed the aforementioned limits per syntactic structure, except for the 1,608 headwords that were selected for full manual validation (see the next section).

A new approach was also used in the automatic extraction of corpus examples. For version 1.0, we used different GDEX configurations for different parts of speech, with configurations being optimized for the extraction of good examples for collocations. While this approach produced good results, it took a great deal of processing, plus the GDEX score of a corpus sentence depended on a given headword rather than the sentence as a whole. Consequently, we decided to devise one GDEX configuration for an entire corpus - with the help of the Sketch Engine team, we ran the script on the Gigafida 2.0 corpus and assigned a GDEX score to each sentence in the corpus. Part of the automatic collocation extraction was thus also the extraction of the list of all corpus IDs of the sentences in which each collocation appeared; based on that, we extracted for the Collocations Dictionary up to four examples with the highest GDEX score per each collocation.

4. Storing collocational data: Digital Dictionary Database and a data warehouse

Collocations, along with other types of lexical information, are stored in the Slovene Digital Dictionary Database (Kosem et al., 2021), which aims to become a one-for-all database for the Slovenian language, to be used for both in the compilation of language resources and natural language processing tasks. The plans for the database have been described in detail by Klemenc et al. (2017). This trend of data consolidation can be observed across Europe, with the most noticeable case studies being the attempts for

⁵ There are many more compounds in the Digital Dictionary Database, however for now only 128 have collocations.

Estonian (Tavast et al., 2018), German (Geyken, 2019), Polish (Żmigrodzki, 2018), and Dutch (Colman, 2016).

The first version of the Collocations Dictionary was part of the DDDS from the very beginning. However, due to many changes introduced by the data in the second version (method of collocation extraction, new corpus etc.), we had to first completely remove from the database the automatic collocational data from the first version and then import the new data. While we were preparing for the import of new data, other data had been imported, i.e., synonyms from the Thesaurus of Modern Slovene⁶ (Arhar Holdt et al., 2018), and bilingual data from the Comprehensive Slovenian-Hungarian Dictionary⁷ (Kosem et al., 2021), the latter also containing collocations. It is worth noting that the lexicographic process of the compilation of the Comprehensive Slovenian-Hungarian Dictionary includes a separate step of compiling entries from scratch for various purposes, which means that much more information (especially collocations and examples) is included than is needed, and ends up, in a bilingual dictionary.

Another relevant resource for the import of collocations was a data warehouse, which served as a storage for all the collocation candidates extracted from the corpus (over 63 million collocation candidates in total). The Digital Dictionary Database thus contains a subset of collocations from the data warehouse. In the data warehouse, we keep additional information such as IDs of corpus sentences in which the collocation is found, sense(s) under which the collocation belongs, the relevance of the collocation for the Collocations Dictionary for each of its components etc. Using the data warehouse facilitates the analysis of data, statistics, data extraction, and maintaining the link to corpus metadata. Having a record of not only good but also bad collocation candidates is crucial to preventing the duplication of work in the future.

A significant challenge at the import stage of new automatic collocations from the data warehouse proved to be matching the already identified collocations found in the digital dictionary database with newly automatically extracted ones, which had to be done to prevent duplication. Among other things, this also included analysing compounds, which may have received a status of a compound in a bilingual dictionary, but were considered legitimate collocations in a collocations dictionary. This process resulted in two types of entries - the ones with fully automatic collocations only, and others with a combination of manually inspected and automatically extracted collocations.

For a selection of 1,608 headwords,⁸ we compiled fully manually validated entries. For these headwords, we did not use the same limitations in terms of a number of automatic

⁶ <https://viri.cjvt.si/sopomenke/eng/>

⁷ <https://viri.cjvt.si/slovensko-madzarski/eng/>

⁸ The initial number was 2,000 but we ended up with fewer entries due to time constraints and work being needed on the matching of automatic collocations with existing manually

collocations per syntactic structure but rather exported all the collocations with the frequency of 4 and above. We, therefore, aimed to inspect all the collocations of a headword, however for frequent headwords with a great number of collocations (over a thousand) we set a minimum value $\logDice \geq 4.0$ for analysis. This roughly meant that whenever this threshold was applied, we ended up analysing under 300 collocations. We had three types of decisions: is a collocation, is a collocation but not relevant for the collocations dictionary, is not a collocation. The collocations in the first group ended up in the Collocations Dictionary, and the collocations in the second group ended up in the Digital Dictionary Database but not in the Collocations Dictionary.

A thorough analysis of collocations for 1,608 entries also served as an evaluation of the quality of automatic data in each syntactic structure. The results show high relevance of many structures (i.e. many structures contain many good collocation candidates) but also very poor results in certain structures. Table 1 and 2 show the top five syntactic structures with the highest percentage of good collocation candidates, and the top five syntactic structures with the highest percentage of bad collocation candidates, respectively.⁹

structure	percentage of good collocation candidates	number of examined collocations
adjective + preposition + noun in instrumental	90.91	396
adjective + noun	90.85	33271
verb + noun in accusative	87.72	6783
reflexive verb + noun in accusative	85.67	317
adjective + noun in dative	84.76	105

Table 1: Top five syntactic structures with the highest percentage of good collocation candidates.

validated collocations in the database.

⁹ Syntactic structures with fewer than 100 collocations were excluded from these lists.

structure	percentage of bad collocation candidates	number of examined collocations
adjective + <i>and/or</i> + adjective	85.62	1210
noun + negative verb	81.17	154
noun + noun in dative	84.13	252
noun in nominative + verb in 3rd person	84.03	4722
noun + <i>and/or</i> + noun	76.56	9789

Table 2: Top five syntactic structures with the highest percentage of bad collocation candidates.

4.1 Interface and data presentation

The interface of the Collocations Dictionary has undergone some significant changes, on account of the harmonization with the interface of other language resources of the Centre for Language Resources and Technologies at the University of Ljubljana (CJVT UL), and, more importantly, of the findings of the studies with the users. The former changes were widening the page layout (to reduce scrolling and show more content initially), changing the font (to a more online-friendly one which supports many different characters and languages), and moving the menu box (with sense menu and structure filter) from left-hand column position to the top line above the content (see Figure 1b). The Collocations Dictionary 2.0 has also adopted the entry layout from other CJVT UL dictionaries (and according to the approach observed in foreign collocations dictionaries), abandoning the previous approach where the collocations were never clearly distributed under senses in the main window (the user had to use the sense filter to get the information of which collocations belonged to each sense) - the comparison is provided in Figures 1a and 1b.

The layout change is already quite noticeable, but even more noteworthy and relevant for the users are some other changes, which were informed by user studies. For example, there is now less clicking in general: all the collocations are offered immediately, with various data manipulation options available on the click of a button. These options include: limiting the view to a selection of most frequent collocations (Less/More icon); ordering collocates by frequency (the default option), alphabetical order, reversed alphabetical order, and length; filtering collocates to only 4768 lemmas on the Reference

List of Slovene Frequent Common Words (Pollak et al., 2020; Arhar et al. 2020); and showing or hiding the headword in the collocation (the headword is shown by default). With the exception of the Less/More option, all the options are part of the Settings row and are thus used for all subsequent searches once set.

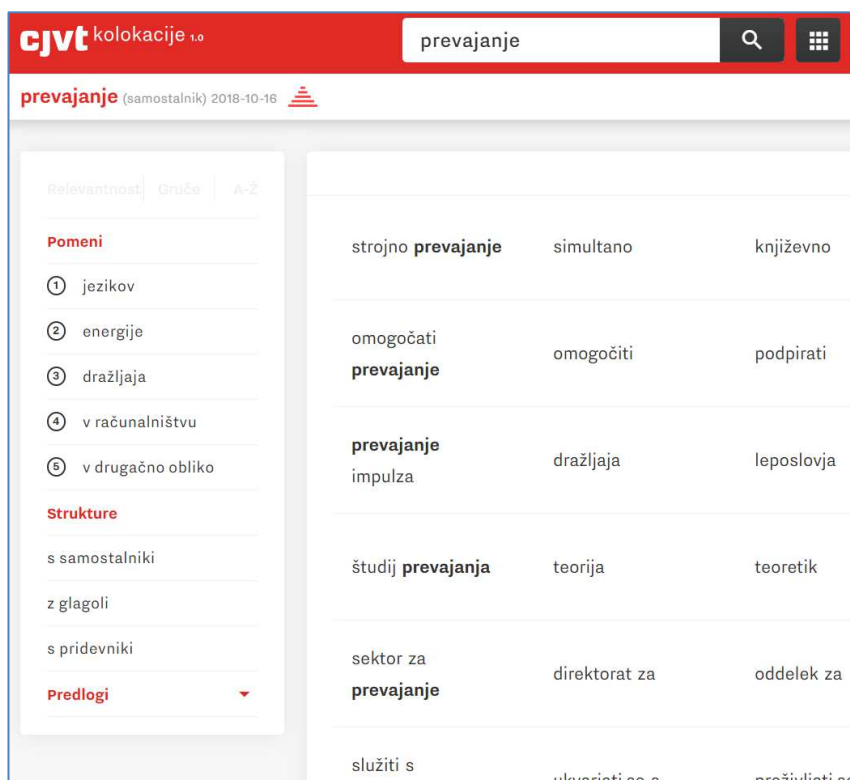


Figure 1a. Entry layout in version 1.0.



Figure 1b. Entry layout in version 2.0.

A lot of thought and effort has been put in improving the clarity of presentation in the interface. The phase pyramid has been abandoned, and instead we added clear headings for two boxes with different types of collocational information. Collocations that have been manually validated and distributed under senses are found in the “Collocations” box, whereas automatic and not yet inspected collocations are found in the box titled “Automatically extracted collocations.” In this way, we reduced the previous five-stage entry progress shown by the pyramid icon (which was often missed or considered unclear by the users) to a three-type entry status which is immediately apparent and needs no additional status icon. The three types of entries in the Collocations Dictionary of Modern Slovene are:

- entries with sense division and only manually validated collocations. These entries have only the “Collocations” box.
- entries with sense division and manually validated collocations in these senses, but also with automatically extracted collocations without an assigned sense. These entries contain both the “Collocations” and “Automatically extraction collocations” boxes.

- entries with only automatically extracted collocations. These entries contain only the “Automatically extraction collocations” box.

We also changed the presentation of syntactic structure titles, as now they are clearly presented as titles under which collocations are grouped (in version 1.0, the structure name was only made available on mouseover). The presentation of examples remained the same; they can be viewed by clicking on a collocation. The link to the corpus showing all the examples of a particular collocation is also available at that point.

Another more significant change, which is related to the user experience, is the enhancement of the crowdsourcing aspect of the dictionary. In the first version of the dictionary, the only crowdsourcing feature was the option to mark collocations as good or bad (using upvote and downvote) on the page of each structure. The feature was rarely used, and as shown by research, such a task is far too demanding for an average user. In the second version, we opted to introduce crowdsourcing at an example level; the users can now not only confirm the validity of the collocation in each example provided but also select the relevant sense (if sense division for a particular headword has already been made). This is in line with our findings that examples rather than collocations are much more suitable for direct crowdsourcing.

5. Conclusions and future plans

The Collocation Dictionary of Modern Slovene, version 2.0, has introduced many changes to both the collocational data it contains, and to the way the data is presented to the user. The changes took into account the latest developments in automatic collocation extraction from corpora, and the findings of various user studies. The dictionary has reaped the benefits of storing the data in the Digital Dictionary Database and in a data warehouse, not only because of avoiding the duplication of work but also because we were able to utilize the lexical data produced in other dictionary projects.

Short-term plans include the preparation of the dictionary database in the XML format and its upload to the CLARIN.SI repository. In line with the policy at the CJVT UL, the database will be available under the CC BY-SA 4.0 license (Creative Commons - Attribution-ShareAlike 4.0 International). Moreover, we are currently working on making the user voting information immediately available next to each collocation; the idea is to show the sense number(s), or the tick or cross icon next to the collocation as soon as the user vote is cast.¹⁰

Long-term, we would like to add other types of grouping of the collocations, for example by questions such as *Mrežnik* and *eleziko*, and/or by semantic properties (e.g., using semantic types). There are also plans to conduct further user studies to identify further

¹⁰ The hold up is mainly technical as we are solving some performance issues.

improvements to the interface. Based on the evaluation of the data of 1,608 manually completed entries, improvements to the automatic extraction method will be made.

An important development expected in the next months will be the introduction of an editor for the Digital Dictionary Database which will facilitate entry compilation and publication, enabling us to make updates to the Collocations Dictionary on a more regular basis.

6. Acknowledgements

The authors acknowledge that the project Empirical foundations for digitally-supported development of writing skills (J7-3159) and the programmes Language Resources and Technologies for Slovene (P6-0411) and Slovene Language – Basic, Contrastive, and Applied Studies (P6-0215) were financially supported by the Slovenian Research Agency. The project Upgrading fundamental dictionary resources and databases of CJVT UL was funded by the Ministry of Culture of the Republic of Slovenia in the period 2021–2022.

7. References

- Arhar Holdt, Š., Pollak, S., Robnik Šikonja, M. & Krek, S. (2020). *Referenčni seznam pogostih splošnih besed za slovenščino*. Proceedings of the Conference on Language Technologies and Digital Humanities, pp. 10-15.
- Arhar Holdt, Š., Čibej, J., Dobrovoljc, K., Gantar, P., Gorjanc, V., Klemenc, B., Kosem, I., Krek, S., Laskowski, C., Robnik-Šikonja, M. (2018). Thesaurus of Modern Slovene: By the Community for the Community. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (eds.) *Proceedings of the XVIII EURALEX International Congress: Lexicography in global contexts*, pp. 401–410. Ljubljana: Znanstvena založba Filozofske fakultete. Available at: <https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/118/211/3000-1>.
- Arhar Holdt, Š. (2021) Razvrstitev kolokacij v slovarskem vmesniku: uporabniške prioritete. In I. Kosem, Iztok (ed.) *Kolokacije v slovenščini*. Ljubljana: Znanstvena založba Filozofske fakultete, pp. 125-157. Available at: <https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/318/465/6974-1>.
- Arhar Holdt, Š., Logar, N., Pori, E., Kosem, I. (2021). Game of words: play the game, clean the database. In Z. Gavriilidou, L. Mitits, S. Kiosses (eds.) *Lexicography for inclusion: EURALEX XIX: Congress of the European Association for Lexicography: 7-9 September 2021*, Vol. 2. Komotini: Democritus University of Thrace, pp. 41-49, Available at: <https://euralex.org/publications/game-of-words-play-the-game-clean-the-database/>.
- Colman, L. & Tiberius, C. (2018). A good match: a Dutch collocation, idiom and pattern dictionary combined. *Proceedings of the XVIII EURALEX International Congress*, pp. 233-246.
- Colman, Lut. (2016). Sustainable lexicography: where to go from here with the ANW

- (Algemeen Nederlands Woordenboek, an online general language dictionary of contemporary Dutch)? *International Journal of Lexicography*, 29/2, pp. 139-155.
- Firth, John. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*, 1957, pp. 10-32.
- Gantar, P., Kosem, I. & Krek, S. (2016): Discovering automated lexicography: the case of Slovene lexical database. *International journal of lexicography*, 29 (2), pp. 200–225.
- Gantar, P., Krek, S. & Kosem, I. (2021). Opredelitev kolokacij v digitalnih slovarskih virih za slovenščino. In I. Kosem (ed.). *Kolokacije v slovenščini*. Ljubljana: Znanstvena založba Filozofske fakultete, pp. 15-41. Available at: <https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/318/465/6969-1>.
- Geyken, A. (2019). The Centre for Digital Lexicography of the German Language: New Perspectives for Smart Lexicography. I. Kosem & T. Zingano Kuhn (eds.) *Electronic lexicography in the 21st century (eLex 2019): Smart lexicography. Book of abstracts*. Lexical Computing CZ s.r.o., Brno, Czech Republic.
- Hanks, P. (2004). Corpus pattern analysis. *Proceedings of the XI EURALEX International Congress*, pp. 87-98.
- Haß, U. (ed.). (2005). Grundfragen der elektronischen Lexikographie. *elexiko – das Online-Informationssystem zum deutschen Wortschatz*. (Schriften des Instituts für Deutsche Sprache). Berlin/New York: de Gruyter.
- Hudeček, L. & Mihaljević, M. (2020a). The Croatian Web Dictionary – Mrežnik Project – Goals And Achievements. *Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje* 46/2, pp. 645–667.
- Hudeček, L. & Mihaljević, M. (2020b). Collocations in the Croatian Web Dictionary – Mrežnik. *Slovenščina 2.0*. 8/2, pp. 78–111.
- Kallas, J., Kilgarriff, A., Koppel, K., Kudritski, E., Langemets, M., Michelfeit, J., Tuulik, M., & Viks, Ü. (2015). Automatic generation of the Estonian Collocations Dictionary database. In *Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference*, 11-13 August 2015, Herstmonceux Castle, United Kingdom Ljubljana/Brighton: Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd., pp. 1–20.
- 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*. Ljubljana: Ljubljana University Press, Faculty of Arts, pp. 10-22.
- Klosa, A. (2015). Wortgruppenartikel in elexiko: Ein neuer Artikeltyp im Onlinewörterbuch. *Sprachreport Jg*, 31(4), pp. 34–41.
- Koppel, K., Kallas, J., Khokhlova, M., Suchomel, V., Baisa, V. & Michelfeit, J. (2019). SkELL corpora as a part of the language portal Sõnaveeb: problems and perspectives. 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*. 1-3 October 2019, Sintra, Portugal. Brno: Lexical Computing CZ, s.r.o., pp. 763-782.

- Kosem, I., Arhar Holdt, Š., Krek, S., Gantar, P., Pori, E., Čibej, J., Klemenc, B., Laskowski, C., Dobrovoljc, K., Gorjanc, V. & Ljubešić, N. (2022). *Kolokacijski slovar sodobne slovenščine 2.0*. Znanstvena založba Filozofske fakultete Univerze v Ljubljani.
- 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://eknjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/118/211/3000-1>
- Kosem, I., Krek, S. & Gantar, P. (2020). Defining collocation for Slovenian lexical resources. *Slovenščina 2.0*, 2, pp. 1-27. DOI: 10.4312/slo2.0.2020.2.1-27.
- Kosem, I., Bálint Čeh, J., Ponikvar, P., Zaranšek, P., Kamenšek, U., Koša, P., Gróf, A., Böröcz, N., Harmat Császár, J., Szijártó, I., Šantak, B., Gantar, P., Krek, S., Roblek, R., Zgaga, K., Logar, U., Pori, E., Arhar Holdt, Š., Gorjanc, V. (2021). *Comprehensive Slovenian-Hungarian Dictionary 1.0*, Slovenian language resource repository, CLARIN.SI, <http://hdl.handle.net/11356/1453>.
- Kosem, I., Krek, S. & Gantar, P. (2021a). Semantic data should no longer exist in isolation: the Digital Dictionary Database of Slovenian. In Z. Gavriilidou, L. Mitits, S. Kiosses (eds.), *EURALEX XIX, Congress of the European Association for Lexicography, Lexicography for inclusion*. Komotini: Democritus University of Thrace, pp. 81–83. Available at: https://euralex2020.gr/wp-content/uploads/2021/09/EURALEX2020_BookOfAbstracts-Preview-1.pdf
- Krek, S., Gantar, P., Kosem, I. (2022). Extraction of collocations from the Gigafida 2.1 corpus of Slovene. In A. Klosa (ed.). *EURALEX 2022, Proceedings of the XX EURALEX International Congress, 12-16 July 2022, Mannheim, Germany*. [S. l.]: IDS-Verlag, pp. 240-252. https://euralex.org/wp-content/themes/euralex/proceedings/Euralex%202022/EURALEX2022_Pr_p230-239_Kosem.pdf.
- Krek, S., Gantar, P., Kosem, I. & Dobrovoljc, K. (2021). Opis modela za pridobivanje in strukturiranje kolokacijskih podatkov iz korpusa. In Š. Arhar Holdt (ed.) *Nova slovnica sodobne standardne slovenščine*. Ljubljana: Znanstvena založba Filozofske fakultete, pp. 160–194.
- Pollak, S., Arhar Holdt, Š., Krek, S. & Robnik Šikonja, M. (2020). Reference List of Slovene Frequent Common Words, Slovenian language resource repository CLARIN.SI, <http://hdl.handle.net/11356/1346>.
- Pori, E. & Kosem, I. (2021). Evalvacija avtomatskega luščenja kolokacijskih podatkov iz besednih skic v orodju Sketch Engine. In I. Kosem (ed.). *Kolokacije v slovenščini*. Ljubljana: Znanstvena založba Filozofske fakultete, pp. 43-77. <https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/318/465/6974-1>.
- Pori, E., Čibej, J., Kosem, I. and Arhar Holdt, Š. (2020): The attitude of dictionary users towards automatically extracted collocation data: a user study. *Slovenščina 2.0*, 8(2): 168–201. DOI: <https://doi.org/10.4312/slo2.0.2020.2.168-201>

- Pori, E., Kosem, I., Čibej, J. & Arhar Holdt, Š. (2021). Evalvacija uporabniškega vmesnika Kolokacijskega slovarja sodobne slovenščine. In I. Kosem (ed.). *Kolokacije v slovenščini*. Ljubljana: Znanstvena založba Filozofske fakultete, pp. 235-268.
- Storjohann, P. (2005). elexiko: A Corpus-Based Monolingual German Dictionary. *Hermes, Journal of Linguistics*, 34, pp. 55–82.
- 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: Lexicography in Global Contexts*. Ljubljana: Znanstvena založba Filozofske fakultete, pp. 749-761.
- Tiberius, C., Munda, T., Repar, A. and Krek, S. (2022). Lexicographic data in ELEXIS. Deliverable of the ELEXIS project. https://elex.is/wp-content/uploads/ELEXIS_D1_6_Lexicographic_data_in_ELEXIS.pdf
- Żmigrodzki, P. (2018). Methodological issues of the compilation of the Polish Academy of Sciences Great Dictionary of Polish. In J. Čibej, V. Gorjanc, I. Kosem, S. Krek (eds.). *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*. Ljubljana: Znanstvena založba Filozofske fakultete, pp. 209-219.

The end of lexicography? Can ChatGPT outperform current tools for post-editing lexicography?

Miloš Jakubíček^{1,2}, Michael Rundell¹

¹Lexical Computing, Czechia & United Kingdom

² Faculty of Informatics, Masaryk University, Brno, Czechia

E-mail: milos.jakubicek@sketchengine.eu, michael.rundell@gmail.com

Abstract

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

1. Introduction

Lexicographic tasks have been subject to automation efforts since the inception of corpus-based lexicography (see Rundell & Kilgarriff, 2011; Rundell et al., 2020). Methods and tools for automatic production of word lists, example sentences or collocations were developed, alongside of large corpora (Jakubíček et al., 2013). Those tools were typically task-specific and were applied individually or collectively to draft a complete dictionary entry and thereby streamline the process of dictionary-making. In this paper we elaborate on the use of ChatGPT, a chatbot based on a very large language model (LLM), which may be perceived as a system that – seemingly – combines all the tools so far produced for lexicography into one, and, when prompted with a simple natural language query such as “Can you give me a dictionary entry for the word table?”, answers with a natural language response mimicking a typical entry structure (with arbitrary components).

In this paper we discuss the advantages and disadvantages of using such a system for lexicographic tasks, based on our observations and on an experiment we carried out on a small set of very heterogeneous English headwords. We introduce the system’s principal properties and their implications as well as contemporary features that might or might not change in the near future. While our experiment was carried out for English only, we address the multilingualism of ChatGPT right at the beginning.

The purpose of this paper is in the first place educative and speculative, rather than recommending or judgmental. The evaluation we carried out was done mainly for illustrative purposes and is of very limited reproducibility. As with any new technology, or rather in this case, an emerging technology (large language models) used in a new context (lexicography), it is of the utmost importance for the lexicographic community to be aware of all the issues around LLMs, the principal caveats and practical questions to ask, before any decision to apply the technology in their work.

2. ChatGPT and OpenAI’s GPT-based models

While we will not attempt a technical description of the system from the NLP point of view, it is necessary to introduce it at a broad level to be able to discuss some of its properties. ChatGPT (Ouyang et al., 2022) is a chatbot based on the GPT-3 language model (Brown et al., 2020) launched by OpenAI in November 2022. GPT stands for Generative Pre-trained Transformer, a type of neural network that is trained on a large unannotated corpus (i.e. plain text), yielding a language model, i.e. a probabilistic distribution over words given prior words. Such a model makes it possible to carry out what is formally called decoding or inference, and in practical terms generates the most likely word sequence given a prompt.

The level of details we can give on how exactly the model was trained and how exactly the inference works is limited. ChatGPT is a closed-source proprietary product of OpenAI, a Microsoft-co-owned company¹. The aforementioned academic publications discuss many aspects of transformer-based neural network training and usage, yet it is unclear to what extent they describe the actual product. This uncertainty extends to the training corpus data. To understand its level, it is just enough to read page 12 of the very comprehensive report on GPT training data provided by Thompson in March 2022 (Thompson, 2022). All we know is that it was trained on a filtered version of the Common Crawl², two unspecified book corpora, one unspecified web corpus and Wikipedia making about 500 billion tokens all together. Unlike the model traditionally used in corpus linguistics, tokens follow the so called subword tokenization – one word typically consists of multiple tokens (or rather, multiple characters form a single token) which – among other benefits – makes it possible for the model to handle morphology. Compared to a corpus-linguistic approach to tokenization, which for English boils down to white-space tokenization where 500 billion tokens would amount to some 450 billion word forms (arbitrarily defined, of course), the subword tokenization approach entails a much smaller word set – the authors estimate two or three times smaller, of the order of “only” about 150–200 billion words.

The tokenizer is one of the very few open components of the system and is available as an online tool³ as well as a Python package. Different models use slightly different tokenizers though (so that, the GPT-4 tokenizer is not the same as the one used in GPT-3) and OpenAI currently claims that 1 token approximates to 4 characters in English⁴. Perhaps more explanatory is the overall size of the vocabulary used by the various GPT models, i.e. number of different tokens, which was only 50,257 for both GPT-2 and GPT-3.

Both training and use of such a model is extremely hardware intensive. It is not easy to quantify this accurately, but it is safe to say that both are far from what is affordable as commodity hardware (broadly defined as any platform costing less than \$100,000), the training being of course more demanding than the inference which is more influenced by the particular workload scenario (number of users, size of the prompts and responses etc.).

Multiple new versions of ChatGPT have been released based on the same language model (GPT-3). First, an updated model (GPT-3.5) appeared, and then in March 2023 a new language model (and new version of ChatGPT using it) labeled GPT-4 was released

¹ See <https://blogs.microsoft.com/blog/2023/01/23/microsoftandopenaiextendpartnership/>

² <https://commoncrawl.org/>

³ At <https://platform.openai.com/tokenizer>

⁴ <https://help.openai.com/en/articles/4936856-what-are-tokens-and-how-to-count-them>

too. While OpenAI provided an API to the actual GPT models, there was no API for ChatGPT initially. Our first experiments therefore used a very tedious workaround based on simulating browser access⁵ and were later replicated using the official API for ChatGPT based on the GPT-3.5 model. The newly released GPT-4-based ChatGPT is not yet generally available through the API and is extremely rate-limited through browser access⁶, thus it was not used for generating the dictionary described in this paper.

Certainly! Here is a dictionary entry for the word "table":

Noun:

1. A piece of furniture with a flat top and one or more legs, used as a surface for working at, eating from, or on which to place things.
2. A set of facts or figures systematically displayed, especially in columns.
3. A list of contents or topics covered in a publication or document.
4. A group of people sitting at a table for a meal or meeting.
5. A level area, typically at the top or bottom of a slope.

Verb:

1. To postpone consideration of (a motion, bill, or other matter).
2. To enter (a proposition) in a table or list.
3. To present formally for discussion or consideration at a meeting.
4. (British) To offer (someone) a meal or entertainment.

Example sentences:

- I'm going to set the dishes on the table.
- The data is presented in a table for easy comparison.
- The table of contents lists all the chapters in the book.
- We sat around the table and discussed our plans for the future.
- We need to table this motion until we have more information.
- Let's table that idea and come back to it later.
- I'd like to table a motion for the next meeting.
- We'd be happy to table you for dinner tonight.

Figure 1: Sample output of GPT-3.5 for the prompt “Can you give me a dictionary entry for the word table?”

Sample output for the prompt “Can you give me a dictionary entry for the word table?” generated by ChatGPT based on GPT3.5 and GPT4 is given in Figures 1 and 2. It is important to realize that because of the non-deterministic nature of the system (which means that the same question may generate a different answer), no reliable conclusions can be drawn from the comparison. But regardless of these uncertainties and of ongoing product development (which invalidates findings particular to certain versions of the system), we briefly enumerate here what appear to be durable properties shared across all current ChatGPT versions:

1. language modelling – not language reasoning

Perhaps the most important point to recognise about how large language models work is that the actual usage (inference) of the model strictly follows the probabilistic nature of the system as learned from the training data. The model does not perform any *logical* inference in terms of formal reasoning, nor does it build any kind of

⁵ Using <https://github.com/mmabrouk/chatgpt-wrapper>

⁶ By the time of writing this paper this was 25 messages every 3 hours only.

table (noun)

1. A piece of furniture consisting of a flat, horizontal surface supported by one or more legs, used for eating, writing, working, or displaying objects.
Example: We sat down at the table to have dinner together.
2. A systematic arrangement of data or information, usually in rows and columns, that presents facts or figures for easy reference and analysis.
Example: The table in the report shows the monthly sales figures for the past year.
3. A list or register of topics or items, especially in a book or document, for reference or information.
Example: The periodic table organizes chemical elements based on their atomic number and properties.
4. In computing, a structured set of data elements, often consisting of rows and columns, used to store and manage information in a database.
Example: The customer table in the database contains information about all the clients.

table (verb)

1. To present (a proposal, resolution, or other matter) for consideration or discussion.
Example: The senator tabled a bill to increase funding for public schools.
2. In British English, to postpone or suspend the consideration of (a matter) for a later time.
Example: The committee decided to table the discussion until next month's meeting.

Figure 2: Sample output of GPT-4 for the prompt “Can you give me a dictionary entry for the word table?”

knowledge base of facts. Many recent controversies⁷ around the GPT models (both in and outside of the academic community) result from ignoring, misunderstanding or simply not being aware of this basic principle.

In the context of generating dictionary entries, it is important to emphasize that the overall structure of the entries is also completely learned from the training data. There is no explicit information the system can use to determine which entry components to generate, how to typeset an entry, how to visualize homographs or polysemous entries or that they should be presented in a numbered list. All of that comes through seeing existing entries of existing dictionaries that were part of the training data.

2. non-deterministic learning and inference

As with many other neural networks, training of the transformer language model is non-deterministic, mostly because some model parameters are initialized at random. This means that repeated training on the same training data creates a (possibly substantially) different model.

Moreover, the inference carried out by ChatGPT through the GPT models is by default non-deterministic too, i.e. it yields different answers for repeated prompts. This results from the fact that finding the optimal answer for a given prompt (or, in other words, the most probable sequence in the model) is not tractable in a model of this size. Different inference heuristics are being applied⁸ to mitigate this issue, and it is not absolutely clear which one is used by ChatGPT⁹. In the API,

⁷ Such as <https://www.bbc.com/news/technology-65202597>

⁸ See <https://huggingface.co/blog/how-to-generate> for a very reader-friendly introduction to this topic.

⁹ although based on the API parameters it is likely a variant of nucleus sampling

the so called temperature parameter may be used to tune the greediness of the inference, and by setting it to 0 one gets deterministic outputs – at the cost of a (possibly substantially) worse output quality, obviously, because a greedy search is rarely the optimal one.

3. **static model**

Once the model is trained, it is static and it is in principle not possible to make any incremental updates easily. This is why the system frequently informs users that the training data are not newer than 2021. Including newer data means retraining the model completely (which in principle always has unforeseen implications). ChatGPT plugins introduced recently do not change the model in any way, but are rather used as part of the prompting mechanism.¹⁰

4. **no source data reference**

In its current forms, the model does not keep references to training sources and it is not straightforward to implement this¹¹. Obviously, the model is unable to “cite” whole sentences verbatim as this is not the way it operates. If the output contains whole sentences that occur in the training data, it is a random (and unintentional) artifact of the training process.

5. **limited prompt and response length**

ChatGPT’s input (prompt) and output (response) length is limited because the model inference is hardware-consuming too. Depending on the model, it varies between 2,048 and 4,096 tokens (so, approximately 500 and 1,000 words). For many applications this is not a problem, but generating long responses that might sometimes be needed in lexicography could be affected.

6. **translation through multi-lingualism**

The vast majority (92%) of GPT-3 training data is English only.¹² The ability to carry out high-quality translations without parallel data (which are hard to obtain for many low resource languages, with no prospect of this changing in the near future) is certainly one of the biggest achievements of the transformer architecture but it also has certain disadvantages, particularly for lexicographic use cases. Translation using ChatGPT also works through language modelling (like anything else) and is achieved by being able to match language-specific contexts cross-lingually. This means that prompts formulated in other languages than English are answered using also (and mainly) English data.

This gets particularly manifested when prompting the system with a cross-lingually polysemous word. To give a simple example, we asked ChatGPT (based on GPT-3.5 and GPT4) in Czech to generate a dictionary entry the Czech noun *stát* which means *state* in the sense of a *country*. The result is given in Figures 3 and 4. Both versions responded with an entry that includes a non-existing sense coming from the translation into the English *state* and one of its English (but not Czech) senses being *status*. Moreover, the example sentence in Czech says *Our company is in a good financial status and is not in an endangered country.*, demonstrating that another translation error occurred in the other direction by translating the English *status* back as a *country*.

¹⁰ <https://openai.com/blog/chatgpt-plugins>

¹¹ Although there are similar attempts, see <https://www.longshot.ai/features/longshot-fact-gpt> or perplexity.ai

¹² An overview by OpenAI is provided at https://github.com/openai/gpt-3/blob/master/dataset_statistics/languages_by_character_count.csv

Určitě! Zde je slovníkové heslo pro podstatné jméno "stát":

stát (mužský rod)

Významy:

1. Politická a teritoriální jednotka řízená vládou a tvořící svrchovanou moc na určitém území.
2. Země, území, region, oblast nebo teritorium.
3. Mocnost, suverénní autorita nebo řídicí orgán, který má kontrolu nad určitým územím nebo skupinou lidí.
4. Stav nebo podmínka, ve které se něco nachází.

Příklady:

- Česká republika je státem ve střední Evropě.
- Ve státě Texas v USA je mnoho ropných vrtných plošin.
- V takovémto státě by nebyla svoboda projevu.
- Naše firma je v dobrém finančním stavu a není v ohrožení státu.

Poznámka: Slovníkové heslo může být upraveno a doplněno v závislosti na slovním spojení nebo kontextu, ve kterém se slovo vyskytuje.

Figure 3: Sample output of GPT-3.5-based ChatGPT for the prompt “Můžeš mi vygenerovat slovníkové heslo pro podstatné jméno stát?” (Can you give me a dictionary entry for the noun country?). The erroneous sense and example sentences are the last (fourth) ones.

Stát (podstatné jméno, rod mužský)

1. Územně a politicky organizovaný celek, který má svou vládu, území, obyvatelstvo a zákony; země, republika, království.
Příklad: Česká republika je stát ve střední Evropě.
2. Instituce a orgány, které vykonávají moc a správu v daném území; veřejná moc, vláda.
Příklad: Stát zajišťuje základní veřejné služby pro občany.
3. Obecně, stav nebo poloha, ve kterém se něco nebo někdo nachází; kondice, situace.
Příklad: Přítel mi pověděl o svém zdravotním stavu.

Synonyma: země, republika, království, veřejná moc, vláda, území, situace, kondice

Související pojmy: vláda, ústava, politika, zákony, suverenita, území, obyvatelstvo, státní správa

Figure 4: Sample output of GPT-4-based ChatGPT for the prompt “Můžeš mi vygenerovat slovníkové heslo pro podstatné jméno stát?” (Can you give me a dictionary entry for the noun country?). The erroneous sense and example sentences are the last (third) ones. Correspondingly the list of synonyms contains erroneous items too (*situace, kondice*).

7. **prompting** The very nature of the system consists of natural language prompts and natural language responses. While this is a big advantage for everyday usage, it represents an obstacle for more formal approaches, as the optimal design of the prompts becomes crucial for high quality output. Moreover the best prompt design for a particular task may change unnoticed between versions or languages. All we know is that finding the 'right' prompt is critical to getting the best response, but there is no reliable way of knowing how to find those best prompts. On top of this, the non-deterministic nature makes it hard to evaluate even a single prompt.

3. Making an English dictionary using ChatGPT

We generated two monolingual English mini-dictionaries: one using the January 9, 2023 version of ChatGPT (based on the GPT3.5 model) by the time of submitting the extended abstract of this paper; and one using the March 23rd, 2023 version of ChatGPT (based on the same model) by the time we were preparing the full paper. Both dictionaries are publicly available with the Lexonomy platform (Měchura et al., 2017) at <https://lexonomy.eu/chatgpt> and <https://www.lexonomy.eu/chatgpt35>. The former was, for reasons explained earlier, done by simulating browser access in the user interface, the latter through the official API that became available meanwhile.

The entries of these two dictionaries were generated for 99 English single- and multi-word headwords which are listed in full as Appendix A. Because the limited availability of the system made it impossible to create a bigger dictionary sample while preparing this paper, we wanted the dataset to be very diverse and therefore adapted a sample headword list used in the preparation of the DANTE lexical database for English (Convery et al., 2010).

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

1. What does the word H mean?
2. Generate a dictionary entry for H .
3. Generate a dictionary entry for H including possible word forms, word senses, pronunciation, collocations, synonyms, antonyms and examples of usage.

These three questions were asked in this particular order in one conversation. As the inference of the system is generally not deterministic, we repeated this whole conversation three times independently in a new ChatGPT context, so that there would be no influence between the three runs. Altogether we thus obtained 297 entries consisting of verbatim answers to the three questions composing each conversation. In Lexonomy, entry names bear the .a1, .a2 and .a3 suffixes for the first, second and third run, respectively.

4. Investigating the mini-dictionary: a lexicographic evaluation

The simplest way to evaluate ChatGPT's responses in this task is to see how well it handles each of the principal components in a dictionary entry. We will therefore consider

its performance across the following elements: word-sense disambiguation, definitions, grammatical information, 'marked' items (such as words which are formal, archaic, or offensive), and example sentences. In each case, we compare ChatGPT's output with equivalent entries in two high-quality 'human-produced' dictionaries: the *Oxford Dictionary of English* (ODE), which is now the default source for a Google search on the lines of 'define X'; and the *Macmillan English Dictionary* (MED). We refer to these as our 'reference dictionaries'.

4.1 word-senses

The challenges here are well known. Establishing a set of word senses for a given headword is generally considered the hardest task in lexicography – not least because meaning is so contextually-determined that 'it makes sense to ask whether words do in fact have meaning at all' (Hanks, 2013: p. 65). The discrete numbered senses in dictionaries are in reality a lexicographic construct, and in many cases no two dictionaries will present the same inventory of senses for a polysemous word. Nevertheless, within this conventional paradigm, we can still judge whether a given dictionary's analysis of a word's meanings is a fair – and practically useful – reflection of the way the word is used in real communicative situations.

Even allowing for the inherent difficulty of the task, ChatGPT does not perform well in this area. Furthermore, our sample did not include any headwords of great complexity (words with, say, six or more senses in a traditional dictionary), so we can assume that – in its current form, at least – it would be defeated by any highly polysemous headword.

A recurring problem is what we might call 'false polysemy', where the system enumerates multiple senses, with different definitions, in cases where there is really only one. A standout example is its treatment of the word *climate*. In both our reference dictionaries, *climate* has two main senses: the weather-related one, and a metaphorical use encoded in expressions like 'in the current economic climate' or 'a climate of fear'. ChatGPT (in this case response a1) gives the following senses:

1. The long-term patterns of temperature, humidity, wind, and precipitation in a particular region.
2. The overall weather conditions of a place over a period of time, typically 30 years or more.
3. The typical or average weather conditions of a place.
4. The general set of weather conditions of a planet or region.
5. The state of the atmosphere in a region in terms of temperature, humidity, wind patterns and precipitation.
6. The average of weather conditions over a period of time, typically 30 years or more.

This goes far beyond what is known in the lexicographic trade as 'splitting' (as opposed to 'lumping'): there is essentially just a single meaning here, explained in six different ways. What is worse, the system fails to take account of the second, very common, metaphorical use identified in our reference dictionaries.

While this is the most egregious instance of false polysemy, there are very few cases in the sample where the system performs adequately (*empty* is probably the best of a bad lot).

At *command*, we find a similar tendency to split one sense unnecessarily while completely missing another common meaning. The IT-related noun use ('the "insert block" command is executed') is correctly identified, but the system posits an equivalent verb use (with the implausible example 'To shut down the computer, you need to command it to shut down'), for which there is little evidence. In response a3, the core sense of 'giving an order' is needlessly split to cover the case of pets: 'To control or direct a pet, animal, or machine through the use of specific commands.' At the same time, other frequent usages are overlooked, with nothing to account for sentences such as 'truffles command a high price' or 'the fort commands a panoramic view of the coast' – all well covered in our two reference dictionaries.

Even simple concrete nouns do not escape these problems, with the word *potato* given no fewer than five 'senses' in response a2:

1. (Botany) A starchy, tuberous crop from the perennial nightshade *Solanum tuberosum*, native to the Andes in South America.
2. (Food) A staple food in many parts of the world, often boiled, baked, or fried.
3. (Industry) Used in the production of various food products, such as potato chips and French fries.
4. (Alcohol) Also used as an ingredient in the production of alcohol, such as vodka.
5. (Variety) Can come in various varieties with different colors, shapes, and textures.

These are all legitimate things to say about potatoes and their use, but this treatment suggests that the system does not really understand what humans mean by a 'dictionary word sense'.

Identifying word senses is rarely straightforward, but when even a simple word like *ameliorate* ends up with three senses, it is clear that ChatGPT is not up to the task.

4.2 definitions

Here the news is more promising, and definitions are in general one of ChatGPT's stronger points. Definitions such as 'An order or instruction given by a person in authority' (*command*, noun use), or 'Capable of producing desired results with a minimum of effort or energy' (*efficient*) give the right information in an accessible form, and compare favourably with those in the reference dictionaries. Some say too much and end up being longer than is desirable: the entry for *bargain* (response a1) includes 'an agreement between two or more parties in which each party agrees to certain terms, often used to refer to a transaction where goods or services are exchanged for an agreed-upon price that is typically lower than the market value.' A tweak to the prompt question might resolve this, specifying a maximum word count (as some dictionary styleguides do).

Other definitions employ familiar lexicographic formulae: *closure* (response a3) has 'The act or process of closing or the state of being closed', and one version of *slavish* (the others are better) includes 'Resembling or characteristic of a slave'. Styles like this, which are unhelpful for users, were widespread in older publications but are less often found in good contemporary dictionaries. Occasionally a definition will fail to include a key meaning component: thus *garden* ('A piece of land used for growing plants, flowers, or vegetables.')

does not mention that gardens are typically attached to houses; similarly, one version of *beach* describes it as 'a place of recreation or relaxation, where people go to swim, sunbathe, and engage in other outdoor activities', without noting its adjacency to the sea or a lake. But there is plenty that is positive. The system seems to perform especially well when defining technical terms. All versions of *carbon cycle*, for example, are well (and clearly) defined (if sometimes over-long), and duly mention the key related terms photosynthesis and respiration. This is important because, of all the components in a dictionary entry, definitions have so far proved the least tractable in terms of automation. ChatGPT may be at least part of the answer.

4.3 grammatical information

In other experiments we have specifically prompted ChatGPT to identify the syntax patterns that typically follow a given word – in the way that pedagogical dictionaries usually do. (Results have been patchy.) This was not done in the case of the mini-dictionary, so our focus here is on the way grammatical features are dealt with at a general level. Transitivity is not always handled well. Thus the entry for *empty* (verb) fails to cover intransitive uses like this (from MED): 'the stadium began to empty'.

More worryingly, some words are wrongly categorised in terms of word class. In one version of *aside* (a3), a sense which is explained as 'to one side: He pushed the plate aside' is labelled as a preposition. In other cases, the form of a definition does not match the word class, as in sense 2 of the verb *haunt* (response a3), defined as if it was both an adjective and a noun:

1. Visit frequently, or reside in as a ghost or spirit.
2. Constantly present in one's mind; an obsession.
3. To frequent a place or places frequently.

Problems like these are pervasive, and significantly compromise the value of ChatGPT's output.

4.4 'marked' items

Most lexical items are 'unmarked', but some are specialised in terms of their distribution across text-types. Dictionaries typically use 'labels' (such as *formal*, *offensive*, or *old-fashioned*) to draw users' attention to these features, though other strategies are sometimes employed too. Some of the words in our sample list were specifically included in order to see how well ChatGPT coped with this aspect of language.

In general, the system performed well on this topic. Its response to the word *half-caste* (once a common word for a person of mixed race, but now universally regarded as offensive) was exemplary. In its response, the explanation of meaning was preceded by a warning that it is 'considered to be a derogatory term used to describe a person of mixed racial heritage'. And this definition is followed by further advice: 'It is now considered offensive and outdated and it is better to use terms such as "mixed race" or "multiracial" instead.' It would be difficult to improve on this. Similarly, *betimes* was correctly identified as an

archaic word whose 'usage is rare in modern English'. Unsurprisingly, it failed to recognise *bockety*, an Irish-English word meaning unstable or rickety. Though this does appear in ODE (but not MED) its frequency in a general English corpus is very low. Its response to *ameliorate* was somewhat disappointing. European cognates of this word (such as French *améliorer*) are typically unmarked, but in English it is a rare and rather formal word, and is marked as such in MED. However, it carries no label in ODE or Merriam-Webster, so it would be unfair to criticise ChatGPT for this omission.

4.5 example sentences

As prompted, all of our sample entries included at least one example sentence for every word and sense covered. An unexpected feature of these examples – given that the system is based on such a large corpus – is that they often look as if they have been made up by a rather unimaginative human editor. A persistent and very noticeable issue, identified in every experiment we have made with ChatGPT – is that examples predominantly follow the formula '3rd person subject with simple past verb', typically opening with a definite article. One of the entries for *aside* (a2) ends with this example set:

- “She put aside her book and listened to the music.”
- “The judge set aside the verdict and ordered a new trial.”
- “He whispered something aside to his friend before he began to speak.”
- “The actor broke character for a moment and delivered an aside to the audience.”
- “The singer added an aside to the melody, making the song more interesting.”
- “The author inserted an aside in the text to comment on the society of his time.”

In pre-corpus times, this pattern was a reliable predictor of an invented example – to the point that lexicographers working on the MED were explicitly warned to avoid using this formula in examples, unless corpus data showed the pattern to be typical of a word's behaviour. ChatGPT's examples are for the most part unconvincing, and when there is a set of examples, they exhibit far too little diversity in terms of structures and even subject matter. (This is something that skilled lexicographers pay a lot of attention to.) One of the worst instances (at *command*), 'The commander commanded his troops to march forward', looks like something invented by a not very good apprentice lexicographer without access to a corpus. In the current state-of-the-art, lexicographers are offered candidate examples filtered by the GDEX software (Kilgarriff et al., 2008), and in most cases it is easy to find a suitable example, which can either be used as is or with minimal tweaking. At the moment, there is probably more mileage in further refining the GDEX algorithm than in trying to get ChatGPT to produce more natural-sounding examples.

5. Conclusions

5.1 Arguments in favour

Some of the arguments in favour of using ChatGPT in a lexicographic context are very straightforward – ChatGPT is:

- extremely simple to use,

- very likely going to be continuously developed and improved,
- very affordable as for pricing¹³,
- naturally multilingual,
- equipped with an API that is easy to integrate into a dictionary writing system.

Moreover the chat environment allows the lexicographer to challenge the response which gets frequently improved in the second round.

But there are also less obvious advantages of ChatGPT. When it comes to generating particular entry components, the system is often on a par with state-of-the-art solutions (though this is very hard to evaluate of course) for some of them, but there are also some where it likely exceeds the state-of-the-art. In our experiments this was most striking on the task of differentiation between synonyms and antonyms. There are many solutions for generating a distributional thesaurus, but discriminating between synonyms and antonyms (and possibly other relations) still remains a challenge. Initial observations however indicate a very low error rate for this task: asking *Can you classify the following list of words into synonyms and antonyms: word1, word2, ...* generates decent (though not perfect) responses (an example of such a prompt is given in Figure 5). This is not surprising though: modelling contexts of words is the very job of a language model!

Similarly, generating dictionary definitions is something that ChatGPT seems to be good at, even though their mapping to a particular sense distribution is problematic (mainly subject to repetitions). A problematic part of the generated definitions might be the inspiration by old dictionaries (using “the act of” definition style, or incorrectly referring to particular dictionaries such as “according to Merriam-Webster X is defined as” and then using an arbitrary generated definition), but this could likely be avoided with better prompts that would instruct the system to avoid such styles.

PROMPT:

The following is a list of similar words for the word "fast": slow, quick, efficient, powerful, expensive, cheap, reliable, simple, complex, small. Can you classify this list into synonyms and antonyms?

RESPONSE:

Sure! I can classify the list into synonyms and antonyms for the word "fast":

Synonyms: quick, efficient, powerful

Antonyms: slow, simple, complex, small

The remaining words (expensive, cheap, reliable) don't have a direct relationship with the word "fast," so I haven't classified them as either synonyms or antonyms.

Figure 5: Sample conversation with GPT-4-based ChatGPT targeted at synonym/antonym classification.

5.2 Arguments against

Reasons not to use ChatGPT are perhaps less obvious but no less important for that, since some of them are quite fundamental.

¹³ See <https://openai.com/pricing>

1. **Googleology is a bad science. And GPTology too.**

We borrow the title of Kilgarriff’s paper (Kilgarriff, 2007) where he argued against using Google search as a corpus search system. Many of the then-used arguments are valid now as well. ChatGPT is using unknown data sources, with non-deterministic (and very likely soon-to-be-personalized) responses, very limited stability and reproducibility. Using it as a general purpose search system in a scientific context inevitably suffers from all the issues a Google search-based approach does.

2. **Vicious data circle**

We explained that GPT knows what an entry looks like from existing dictionaries online that formed part of the training data. This represents a challenge: in all likelihood, it is not the best and most up-to-date dictionaries which were freely available for mass download (though CommonCrawl or similar) and which the system learned from. It is notably easy to trigger the kinds of ‘lexicographese’ (‘the act or state of X’, ‘characterized by Y’, etc.) which were once pervasive in dictionaries but are now (thankfully) being abandoned.

Lexicography has undergone some radical changes in the past 20 years: the arrival of big corpora, NLP analytics, the migration from print to digital dictionaries. All of these have had massive implications on the way lexicographers work and on the range and quality of information that has been uncovered. And these developments are ongoing. Using a system whose training data often pre-dates those changes is somewhat problematic from this point of view.

3. **Evidence generating or evidence observing?**

Last but not least, a dictionary-making process which relies entirely on the use of tools like GPT implies the abandonment of the lexicographer’s current role of scrutinising and verifying the evidence suggested by an analytic system. Most NLP tools for lexicography interrogate a corpus, perform some (often very complex) analysis but track back to corpus evidence in the form of concordance lines, so that the lexicographer can determine whether the automatic results match what is in the corpus (and check the corpus content, metadata, annotation etc.). In the present state-of-the-art, we see this stage as an essential part of the process, and we have significant misgivings about the removal of human actors from the data generation chain. ChatGPT and GPT-like models do not make back-linking evidence possible at the moment, and it is questionable whether this would ever be possible.

This issue also relates to the whole notion of corpus-driven lexicography. In the case of dictionary examples, for instance, it is generally accepted that they should reflect what the data shows us to be the contexts and patterns in which a word most typically occurs: examples shouldn’t be made up, but should be found in the corpus (and shortened or lightly edited if needed). Example sentences generated by ChatGPT cannot be found anywhere. There is no guarantee that they were ever produced by a human writer or speaker, nor (as we have seen above) that their typicality matches what lexicographers would choose.

5.3 Summary

The introduction of ChatGPT has gained huge attention worldwide, often generating excitable or hyperbolic reactions, both positive and negative (see e.g. Beckett, 2023). This paper attempts a more sober-minded evaluation of the potential of this emerging technology, and is cautious about claims that ChatGPT can – to paraphrase a recent talk

by de Schryver – handle (almost) all of the lexicographer’s tasks (or make us believe it can), with successful results ¹⁴.

Our various experiments with ChatGPT (notably but not only with the mini-dictionary described in this paper) have convinced us that it cannot (yet) replace the involvement of lexicographers in the dictionary-making process, and moreover that for some of the requisite tasks (such as sense discrimination and example-writing) its performance is significantly worse than what established technologies can do.

But this certainly does not mean that lexicographers should ignore ChatGPT. For over two decades, we have been adapting lexicographic workflows to emerging technology trends, always with the goal of producing better dictionaries at a lower cost in time and resources. We now need to consider what ChatGPT can contribute to these goals, taking account of the caveats raised in this paper but also of its positive outcomes in some areas. ChatGPT is a general purpose solution and we argue that lexicography needs custom solutions (e.g. through fine tuning of these large language models for particular lexicographic tasks) to mitigate some of the issues discussed in this paper. What these custom solutions may learn from GPT models are all the relevant technological lessons, such as successful application of neural networks as a machine learning computational model and the absolutely crucial role of big datasets. GPT models at the moment represent a highlight of a trend (which has been developing for at least a decade) of using large unannotated datasets for machine-learning purposes. It is up to anyone working in computational lexicography to follow up on this with practical solutions which do not compromise on fundamental principles (above all the idea of a data-driven approach) which have been established over time, since large corpora first became available. And this needs to happen in a workflow model, such as post-editing lexicography, that does not leave the lexicographers sand-blinded as ChatGPT does.

6. References

- Beckett, C. (2023). GPT-4 has brought a storm of hype and fright – is it marketing froth, or is this a revolution? *The Guardian*. URL <https://www.theguardian.com/commentisfree/2023/mar/17/gpt-4-ai-tools-fashion-architecture>.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33, pp. 1877–1901.
- Convery, C., Mianáin, P., Raghallaigh, M.Ó., Atkins, S., Kilgarrieff, A. & Rundell, M. (2010). The DANTE Database (Database of ANalysed Texts of English). In *Proceedings of the XIV EURALEX International Conference*. pp. 293–5.
- Hanks, P. (2013). *Lexical Analysis: Norms and Exploitations*. The MIT Press. URL <https://doi.org/10.7551/mitpress/9780262018579.001.0001>.
- Jakubíček, M., Kilgarrieff, A., Kovář, V., Rychlý, P. & Suchomel, V. (2013). The TenTen corpus family. In *7th International Corpus Linguistics Conference CL2013*. Lancaster University, pp. 125–127.
- Kilgarrieff, A. (2007). Last words: Googleology is bad science. *Computational linguistics*, 33(1), pp. 147–151.

¹⁴ See <https://www.youtube.com/watch?v=mEorw0yefAs&list=PLXmFdQASofcdnRRs0PM1kCzpuoyRTFLmm&index=5&t=566s>

- Kilgarriff, A., Husák, M., McAdam, K., Rundell, M. & Rychlý, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus. In *Proceedings of the XIII EURALEX international congress*, volume 1. Universitat Pompeu Fabra Barcelona, pp. 425–432.
- 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.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A. et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, pp. 27730–27744.
- Rundell, M., Jakubíček, M. & Kovář, V. (2020). Technology and English Dictionaries.
- 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.
- Thompson, A.D. (2022). What’s in my ai. *A comprehensive analysis of datasets used to train GPT-1, GPT-2, GPT-3, GPT-NeoX-20B, Megatron-11B, MT-NLG, and Gopher. LifeArchitect. ai Report*. URL <https://lifearchitect.ai/whats-in-my-ai/>.

A. DANTE sample headword list

command	asleep	mackerel
echo	azure	potato
empty	Belleek	Protestant
haunt	betimes	suitable
leaf	bockety	wake
stomach	Canada	how are you
Amazon	Canada goose	after
Amazonian	carbon	however
beach	carbon cycle	might
DJ	climate	this
echoing	climate change	altogether
efficient	climate control	aside
emptiness	cookie	hereinafter
empty-handed	couch potato	might
grave	DNA	moreover
grave	fart	notwithstanding
gravely	half-caste	nowhere
haunted	Leaving Certificate	provided
haunting	moralize	towards
hauntingly	moralizing	somewhat
leafy	ouch	AIDS
bargain	slag	anti
butter	snowboarding	can't
camp	wed	chug
camper	Wed.	-een
camping	wireless	gutter
slave	also	gutters
slavery	always	Shaw
slavish	anyhow	Shavian
slavishly	anyway	meander
spite	busy	speck
spiteful	careful	swathe
spitefully	closure	Czech
ameliorate	garden	

From a dictionary towards the Hungarian Constructicon

Bálint Sass

Hungarian Research Centre for Linguistics, Institute for Lexicology

E-mail: sass.balint@nytud.hu

Abstract

We present the Hungarian Constructicon, a lexical resource which is an inventory of Hungarian constructions. It was derived mostly automatically from a dictionary. Main step of the processing was to identify constructions in the dictionary and lift them out and create individual entries for them. The tool is supplemented by a sophisticated online frontend which applies a so called dynamic toolbox to the constructicon database in order to be able to give an answer to any one-word or multiword query. Elements of this toolbox are analysed search, dynamic referencing and virtual entries which contains cross-references to elements of cxns present in the constructicon. In this way, the constructicon can handle inflected and derived forms in the query providing all plausible interpretations without needing to know a specific query formalism. This also covers the cases where a word can be interpreted as a regular form and an irregular form as well (cf. *'leaves'*). The Hungarian Constructicon combines the advantages of dictionaries and ccns and is equipped with an intuitive user interface.

Keywords: construction; constructicon; analysed search; dynamic referencing; virtual entry

1. Introduction

The term *constructicon* (Lyngfelt et al., 2018b: 97) (Fillmore, 2008: 49) (Jurafsky, 1991: 18) stands for the inventory of constructions of a language – by analogy to the term *lexicon*¹. Accepting the position of Construction Grammar that utterances are not put together from words, but by combining cxns, it is quite straightforward that the basic unit of a lexical resource should be the cxn: a form–function pairing possibly spanning across linguistic levels. There is a considerable interest in developing (Lyngfelt et al., 2018a) and investigating (Dunn, 2023) ccns nowadays.

A ccn is not a list-like structure, but rather a network of cxns, employing different kinds of part-whole relations. Accordingly, a sophisticated cross-reference system is an important feature of ccns: from a cxn to its parts and vice versa. While in traditional dictionaries phrasemes, collocations and the like are often treated only incidentally (Fellbaum, 2016), ccns treat all kinds of meaning-bearing building units with equal care in a unified way as cxns, regardless of how complex they are. It is common to develop ccns by importing lexical information from existing lexical resources, especially from FrameNets.

In this paper, we present a Hungarian Constructicon (ccn-hu). The architecture of the ccn-hu consists of two components. The static part is the actual XML database of the

¹ Following Lyngfelt (2018) we will use the abbreviation *ccn* for constructicon and *cxn* (plural: *cxns*) for construction throughout this paper. We introduce the term *head-construction* (abbreviated as *hcxn*) denoting an entry in a ccn by analogy to the term *headword*. While *cxn* is a construction in general, *hcxn* is a concrete *cxn* having an entry in a ccn.

Hungarian cxns containing a structured entry for each cxn. The dynamic part is a set of tools which processes user input to find out which cxn (or cxns) are to be shown to the user and how. After discussing the status of single-element entities (Section 2), we compare the ccn-hu to other lexical resources (Section 3). Then, we elaborate on the static and the dynamic part in Section 4 and in Sections 5–7 respectively.

2. How many elements does a cxn contain?

An important point of our approach is that single-element units *are* cxns. Words are cxns, morphemes are cxns as well. Though this is not a novel idea (Goldberg, 2006: 5), we emphasize this here. In general, the cxn used to be a multi-element entity while the word (or the morpheme) is a single-element entity. Therefore, a common criterion for a cxn is to consist of at least two elements. Our idea is to include all units of meaning (Teubert, 2005) in the ccn regardless the number of their sub-elements. Namely, single-element entities are also treated as cxns, albeit a somewhat special case of them. The fundamental goal of ccns is to be able to handle all linguistic units in a unified way and allow the user to find all kinds of meaning-bearing units in them. By analogy, it is like to treat 1 (one) just like other positive integers, despite it is quite special being neither prime nor composite. This decision makes it easier to come up with simpler general descriptions of numbers or simpler methods which can handle all of them. We anticipate that the similar decision in lexicology, i.e. the inclusion of single-element entities into the notion of cxn, will have similar advantages.

Covering single-element entities, ccns can integrate dictionaries into themselves. In this way, a ccn can grasp the complete network of the language and show the connections between linguistic units.

3. Comparison to other lexical resources

In this section, we compare the Hungarian Constructicon to other ccns and then to some online dictionaries. As the ccn-hu contains single-element entities (e.g. simple words) as well (see Section 2), it can be considered a dictionary from a certain point of view, accordingly, the latter comparison is also relevant.

3.1 Comparison to ccns: Swedish and Russian

Many ccns have been created in recent years. We compare the approach and features of ours (ccn-hu) to the Swedish (Lyngfelt et al., 2018b) (ccn-sw) and to the Russian Constructicon (Janda et al., 2020; Bast et al., 2021) (ccn-ru) as well.

Size. Ccn-sw contains 393 cxns, ccn-ru contains 2277 cxns while ccn-hu contains more than 13000 cxns in the current version. It should be noted that the Hungarian cxns are less abstract for the most part. E.g. ccn-sw contains many high-level cxns like ‘*artikel*’ (‘article’) or ‘*passiv*’ (‘passive’), but contains several more concrete cxns at the same time, like ‘*antingen X eller Y*’ (‘either X or Y’) or ‘*dra OBJ*’ (‘drag OBJ’) and non-mutable fixed ones like ‘*bla bla bla*’ (‘and so on’). The latter are also characteristic of ccn-hu.

Formalizedness. In our view, ccns should be formalized to the most possible extent rendering the database as machine readable as possible. We consider the level of formal-

ization to be low for both examined ccns. They seem to be just human readable, not inherently machine readable. This is future work for ccn-hu as well.

Connection to FrameNet. As the original English Constructicon (Fillmore, 2008) used FrameNet as starting point, it is somewhat surprising that none of the three ccns in question is directly connected to the corresponding FrameNet.

Single-element units. Differently from ccn-hu, neither ccn-sw nor cnn-ru contains single-element units.

Availability. Data of ccn-ru is freely available, it is in fact a simple `.csv` table. For ccn-sw, the whole cxn-list can be copied from the website. Concerning ccn-hu, the query interface will be freely usable for personal and research use, it is not decided yet whether the software and the data itself will be freely available or not.

3.2 Comparison to online dictionaries: DWDS and OALD

Here, we compare the ccn-hu to online dictionaries: DWDS (BBAW, 2023) and OALD (Oxford University Press, 2023) according to various aspects.

Multiword input. DWDS can not handle simple multiword input like *‘das Buch’*. While *‘zur Verfügung stellen’* is not handled, *‘zur Verfügung haben’* is. OALD does not respond to *‘black dog’*. Creating ccn-hu, one of our important aims is to be able to handle multiword input, even to give an answer to any possible query.

Irregular inflection. On the one hand, it is common that the irregular forms are included in dictionaries. DWDS is at least not totally complete in this sense as to the query *‘Bücher’* it responds with the entry of *‘der Bucher’* (booker) which is misleading. Similarly, *‘hast’* takes the user to *‘die Hast’*. However *‘ziehst’* works well. These kind of redirecting is solved in ccn-hu by analysed search (see Section 5), cf. *‘lovat’* (*‘ló’* (horse) in accusative case).

On the other hand, it is also common that the regular forms are not included in dictionaries. Maybe they are considered out-of-scope and set aside as being part of “grammar”. In OALD *‘books’* silently redirects to *‘book’*, it does not tell the user any information about the connection between the query and the resulting entry. For a language learner, this connection can be important. It holds especially for morphologically richer languages, it seems to be a good behaviour for a Finnish dictionary/ccn to respond with *‘talo’* (house) + *‘-ssa/-ssä’* (in) to the query *‘talossa’*. This kind of redirecting is also solved in ccn-hu, cf. *‘asztalt’* (*‘asztal’* (table) in accusative case).

The most problematic case combines the above too: some word forms represent an irregular form of a word and a regular form of another word at the same time. In these cases dictionaries tend to present the irregular solution and tend to hide the regular, which can be misleading. Consider the English word form *‘leaves’* and enter it to OALD. The irregular plural of *‘leaf’* will be provided but the third person singular of *‘leave’* does not. This solved in ccn-hu as well, cf. *‘terem’*.

Use what you have. We take the position that it is better to have an incomplete entry for a cxn than nothing. DWDS responds to query *‘Nagellackentferner’* with a partial entry

which contains cross-references to ‘*Nagellack*’ and ‘*Entferner*’ and some corpus examples, but no definition (cf. Janssen, 2008). This is clearly an automatically generated entry, but a very useful one: it helps the user understand the queried word. Virtual entries (see Section 7) implement this feature in ccn-hu.

Down-references. Cross-references from a cxn to its parts can be called *down-references*. DWDS do have down-references under ‘*Wortzerlegung*’ word decomposition. OALD has it as well, it is accessible by double-clicking elements of cxns (see e.g. ‘*red herring*’). The ccn-hu has down references for every unit which has elements.

Formalizedness. Dictionaries are generally optimized for human-readability, so they tend to be less formalized compared to ccns. For example, DWDS still uses old-fashioned textual abbreviations like ‘*etw. jmd.*’.

We note that ccn-hu could be compared to a machine translation system too. The big picture is that such a systems usually work as a dictionary for one-word queries and as a translator for multiword queries, the ccn-hu works like a dictionary for multiword queries as well.

4. Lifting out cxns from a dictionary

In this section, we cover the static part of the ccn-hu, i.e. how we created its XML database.

In absence of a Hungarian FrameNet, we started from a monolingual dictionary and derive the ccn to a great extent automatically. Our initial dictionary was (Pusztai, 2003) which is a common reference work for Hungarian and contains more than 73000 entries. The automatic ccn-creation process was carried out as described in the following.

Firstly, we carried out some basic XML preprocessing: fixed UTF-8 character encoding, normalized whitespaces and lowercased the whole dictionary. Then we made the initial XML a bit more data-centric converting some text nodes to attribute nodes. For example, the homonymy indexes were converted from `<hom>1</hom>` to `<hom value="1"/>`. This was a simple *vertical operation*, i.e. a transformation which affects the dictionary as a whole at once.

After that we identified cxns in the “collocation” part of the dictionary entries (marked by the `<coll>` element in the initial XML), we lifted out the XML subtree representing the cxn and created a new individual entry for it on its own. The lemma of the new entry becomes the textual form of the cxn and part-of-speech is set to “cxn” simply. Then we created cross-references from the original place of the cxn to the newly created entry. (These links are colored green in the user interface, see Section 9.) Thus, an additional 14000 entries were added to the ccn being prepared.

An online lexical resource does not encounter any size limits, so we resolved common abbreviations and the tilde (˜) headword placeholder. The latter was not a trivial task as due to traditional practice in Hungarian lexicography in some cases the headword had to be altered before replacing the tilde.

In the final step we converted the ccn into a HTML form which is suitable for displaying and easily queryable using XSLT at the same time and added the entry–query links (see Section 8). Then, the finished material was put behind a Flask frontend for online use.

Goldberg (2006: 5) presents a long list of cxn types. Many different types appear in our final ccn database: bound morphemes, simple words, compound words, filled idioms. All these are non-mutable, continuous cxns. Handling more complex, mutable, non-continuous or partially filled cxns remains a future work (see Section 10).

5. Analysed search

Interacting with a ccn, you should have the opportunity to search for cxns not just words. We will introduce a new type of search called *analysed search* – suitable for ccns – to eliminate the need for users (e.g. language learners) to learn a formal language or a specific search tool (Sato, 2012). The user is allowed to enter free text in a plain search box, then we apply automatic morphological analysis to the text, and direct the user to the appropriate identified cxn(s). This process is applied to the ccn database described in Section 4 and performed for every type of cxn from simple or compound words to e.g. preverb-verb combinations.

Hungarian is a morphologically rich language (Megyesi, 1998) with an extensive inflectional and derivational system. Additionally, compounding is also happens inside the word, i.e. compounds are written together as one word. We use the *e-magyar* system (Indig et al., 2019) for processing user input. The *emMorph* (Novák et al., 2016) morphological analyser module can break down words into morphemes, for example ‘*gyerekeket*’ is broken down to these elements: ‘*gyerek*’ (‘child’) + ‘*-k*’ (‘plural’) + ‘*-t*’ (‘accusative case’), or ‘*hatalmán*’ to these: ‘*hatalom*’ (‘power’) + ‘*-a*’ (‘possessive suffix’) + ‘*-n*’ (‘on’). This is exactly what is needed because the basic elements of cxns are morphemes.

The algorithm of analysed search is the following for a one-word input:

1. if the input is a hcxn on its own, take it into account;
2. perform the morphological analysis;
3. consider all analyses and take one segmentation from each: choose the segmentation with the longest left side part which is a hcxn;
4. we omit possible duplications collecting results into a set.

If there are several alternative results at the end, all of them are considered and presented one after another.

There has been a long-standing debate about whether a certain Hungarian word is compound or not, what is the lemma (the base form) of a certain Hungarian word, i.e. which derivational suffix should be removed and which one should not. Our approach allows us to put this debate aside. Taking the ccn itself as an oracle, we say that if a compound or a derived form is present as a hcxn, then it is accepted as is. An example concerning compounds: ‘*rendőr*’ (‘order guard’ = policeman) will be presented as a cxn as it is a hcxn, while ‘*kapuőr*’ (‘gate guard’ = gatekeeper) will be presented as a compound of two words ‘*kapu*’ (gate) and ‘*őr*’ (guard). The 3rd point of the above algorithm implements this mode of operation.

We note that analysed search is one of the rare cases where a classic low-level natural language processing tool, i.e the morphological analyser, can be used not only for solving a subtle subtask but also directly to meet the needs of end users.

We also note that applying analysed search we make heavy use of the fact that the Hungarian Constructicon is an inherently online tool. It would be hard to include e.g. all compound words future users may ever think of in a printed dictionary.

6. Identification of cxns

In Section 5 we discussed the case of one-word input only. It is important that the input can be multiword naturally, in fact it can be any linguistic element or combination: a morpheme, a word, a phrase or even a short text. A major task is to be able to identify (possible multiword) cxns in multiword input. While a complete solution – handling e.g. complex non-continuous verbal cxns – remains future work (see Section 10), there is already a partial solution of this task handling two easier cases.

On the one hand, the system recognizes non-mutable continuous cxns on their own or even as a part of a query. The algorithm matches the input text greedily to the hcxs and gives the longest one as a result. For example, as ‘*ad hoc*’ and ‘*ad hoc bizottság*’ (ad hoc committee) are both hcxs the queries presented in Table 1 will provide the cxns shown.

query	identified cxns
(a) ‘ <i>ad hoc dolog</i> ’	‘ <i>ad hoc</i> ’ + ‘ <i>dolog</i> ’ (thing)
(b) ‘ <i>ad hoc bizottság dönt</i> ’	‘ <i>ad hoc bizottság</i> ’ (ad hoc committee) + ‘ <i>dönt</i> ’ (decides)

Table 1: An illustration of the operation of the greedy cxn-identification algorithm. If there is a choice (see (b)), the longer cxn will be identified.

On the other hand, the system recognizes a kind of non-continuous cxns as well, namely the preverb-verb combinations. While the **emMorph** morphological analyser module (see Section 5) does all kinds of analyses inside tokens, **emPreverb** (Pethő et al., 2022) module adds the functionality of connecting separated preverb tokens to their verbs. In Hungarian the preverb (or verbal prefix) is written together with the verb in certain cases, but it constitutes an independent token in others, placed possibly several words away from the verb (cf. Megyesi, 1998: 9). The algorithm loops over the tokens of the input. Processing a verb, the algorithm picks up the corresponding preverb (if there is one), connects it to the verb and reanalyses the resulting connected form, and when it comes to a preverb which is already connected, the algorithm simply skips it. Table 2 shows an example of this feature using ‘*bejön*’ (come in) in which ‘*be*’ (in) is the preverb and ‘*jön*’ (come) is the verb.

query	identified cxns
(a) ‘ <i>bejön</i> ’	‘ <i>bejön</i> ’ (come in)
(b) ‘ <i>most jön be</i> ’	‘ <i>most</i> ’ (now) + ‘ <i>bejön</i> ’ (come in)

Table 2: An illustration of preverb-verb cxn identification. The separated preverb in query (b) is handled properly.

7. Dynamic referencing and virtual entries

Analysed search (Section 5) is supplemented by a novel cross-referencing process called *dynamic referencing*. If the search query does not have a matching cxn, but its parts do, a so called *virtual entry* is created on-the-fly automatically: containing nothing but references to the parts. For example, *‘almafa’* (apple tree) is a hcxn, so the user will get its entry immediately, but *‘grépfrútfa’* (grapefruit tree) is not, so the virtual entry created will contain a link to *‘grépfrút’* (grapefruit) and another to *‘fa’* (tree) beyond the information that the original query is a compound construction.

Perhaps it is not surprising that an overwhelming majority of possible queries will result in a virtual entry. Let us review the following cases from the simplest to the most complex using different cxns, all containing the morpheme *‘asztal’* (table).

1. **Simple word.** The query for a simple word, e.g. *‘asztal’* will simply provide its original real entry from (Pusztai, 2003). Words that do have an inner structure but present in the ccn as a hcxn on their own will behave the same way, see e.g. *‘asztalos’* (‘table + -s suffix’ = carpenter). Results for all the other query types below will be virtual entries.
2. **Suffixed word.** For example, *‘asztalra’* (‘table+onto’ = onto table) not being a hcxn on its own, will be analysed and its two parts will be shown in a virtual entry as *‘asztal’* (table) + *‘-ra/-re’* (onto). Fortunately, case markers and other suffixes like *‘-ra/-re’* have a real entry in the initial dictionary already, so they can be presented using a hand-crafted mapping between emMorph codes and them.
3. **Compound word.** Compounds are treated similarly as they are cxns containing more than one morphemes just like suffixed words. For example, *‘faasztal’* (wooden table) will result in a virtual entry containing *‘fa’* (wooden) + *‘asztal’* (table).
4. **Sequence of words.** Word sequences are firstly tokenized using the emToken tokenizer module (Mittelholcz, 2017) and then treated according to point 2 token by token. The result will be a sequence of (virtual) entries, for example *‘három’* (three) and *‘asztal’* (table) for the query *‘három asztal’* (three table).
5. **Non-mutable continuous cxn.** Fixed continuous cxns are identified inside query text (see Section 6), so *‘nem az ő asztala’* (‘not his table’ = none of his business) will be found and its original entry will be presented.
6. **Non-continuous preverb-verb cxn.** These cxns are also identified (see Section 6), and will be presented as a real or a virtual entry.
7. Handling more complex cxns remains future work, see Section 10.

What if the meaning is more than the meaning of the parts presented in the virtual entry? This is a matter of completeness of the ccn. If a cxn is not present in the ccn, we can not do anything but show information about the parts of it. Obviously, in the present version of the Hungarian Constructicon we can only work with those cxns that were included in the initial dictionary.

We do not think that our ccn is complete in any sense, it just contains quite a large amount of cxns. Instead of trying to make the ccn complete at all costs, we focus on making it easy to expand. Clearly, any expansion will influence dynamic referencing as it will decrease the need for virtual entries. If a brand new entry for *‘grépfrútfa’* (grapefruit

tree) will be added in the future to the ccn, its own real entry will be presented for this query and virtual entry creation will no longer be needed thenceforth. This behaviour was successfully tested in the system.

We can refer to analysed search, dynamic referencing and virtual entries together as *the dynamic toolbox*. The point of the dynamic toolbox is that it allows the ccn to give an answer *always* to *any* queries to the best of its ability. If the ccn itself improves, the responses will improve as well.

8. Entry–query links

Ccn-hu will also offer a feature called *entry–query* links, which adds to its overall convenience. This is a kind of cross-referencing system from a lexicographic perspective and a user-friendly feature from user experience point of view.

It means that every word in the text of real or virtual entries functions as a link to start a query that looks up the word itself in the ccn. Unsurprisingly, the entry for ‘*sárga*’ (yellow) contains the word ‘*citrom*’ (lemon) in the definition part. Just click on *citrom* to reach the entry of this very word. The whole dynamic toolbox machinery described in the previous sections will start working as if it would be a query entered directly by the user. This allows us to add entry–query links to every word appearing in the entries.

This feature can help investigating the ccn itself as a subject of lexicological research. We can examine lexical loops (cf. Levary et al., 2012), or the question whether members of the definition vocabulary are themselves defined (cf. Atkins & Rundell, 2008: 448).

9. Availability

The Hungarian Constructicon is available for the scientific community and the general public as well at <http://ccn.nytud.hu/intro>. Please authenticate (username: **eLex2023** password: **letssee**) and feel free to try all examples typesetted like ‘*példa*’ presented in this paper.

The user interface consists of a simple search box and a short description of the system. There are some clickable examples in the description text. A small icon to the right of the *Search* button gives some information about what is going on in the background: ✓ means that the result is a real entry (cf. point 1 in Section 7); ✗ means that no result can be provided; and the magic wand which appears in other cases means that some elements of the dynamic toolbox was applied.

The implementation is based on python3, Flask, XML, lxml and XSLT technologies. Recognizing non-mutable continuous cxns (Section 6) uses a simple hash for finding cxns.

10. Future work

There are many directions in which our works can be further developed. Some of them are listed below from easy ones to difficult ones.

- Create *up-references*, i.e cross-reference every cxn from the entries of its elements (cf. *down-references* in Section 3.2).

- Test the Hungarian Constructicon with end users, collect and investigate real life queries, and shape further development along the learned lessons.
- Integrate other lexical resources which can be used as a cxn source (e.g. Sass & Pajzs, 2010).
- To support the tasks below, develop a formal representation of cxns, or use an existing one, if possible.
- Handle inflected form of multiword cxns. Can be considered as a special case of the next one.
- Handle complex non-continuous verbal cxns with or without free slots (cf. point 7 in Section 7). The difficulty of this task lies in the fact that elements (words and bound morphemes) of this kind of cxns can appear in several different order with possible intervening words. The representation is to be worked out as well as the algorithm which can efficiently use it. Dependency parsing may have a role in the solution.
- Refer to the appropriate meaning of any cxn and “grey out” the others on the user interface. Seems to be a very hard problem.

11. Summary

In this paper, we presented the current version of the Hungarian Constructicon (ccn), a lexical resource which is an inventory of Hungarian constructions (cxns). The ccn was derived mostly automatically from a dictionary. To be able to handle all kinds of linguistic units in a uniform way we included morphemes and words into the category of cxns. The main step of the processing was to identify cxns in the dictionary and lift them out creating individual entries for them. The number of entries was increased by about 20 percent in this way.

The ccn is supplemented by a sophisticated online frontend which applies a so called dynamic toolbox to the ccn database in order to be able to give an answer to any one-word or multiword query. Elements of this toolbox are analysed search which provides an analysed version of the input query, dynamic referencing which creates virtual entries containing cross-references to elements of cxns which are not present in the ccn.

In this way, the ccn can handle inflected and derived forms in the query providing all plausible interpretations without needing to know a specific query formalism. This also covers the cases where a word can be interpreted as a regular form and an irregular form as well like in case of the English example ‘*leaves*’ or Hungarian example ‘*terem*’.

Combining the advantages of dictionaries and ccns we consider our methodology a step towards creating a general purpose “ultimate” lexical resource.

12. References

- Atkins, B.T.S. & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford University Press.
- Bast, R., Endresen, A., Janda, L.A., Lund, M., Lyashevskaya, O., Mordashova, D., Nessel, T., Rakhilina, E., Tyers, F.M. & Zhukova, V. (2021). The Russian Constructicon. An electronic database of the Russian grammatical constructions. URL <https://constructicon.github.io/russian>.

- BBAW (2023). DWDS – Digitales Wörterbuch der deutschen Sprache. Das Wortauskunftssystem zur deutschen Sprache in Geschichte und Gegenwart. URL <https://www.dwds.de>.
- Dunn, J. (2023). Exploring the Constructicon: Linguistic Analysis of a Computational CxG. In *Proceedings of the Workshop on CxGs and NLP / SyntaxFest*. Association for Computational Linguistics.
- Fellbaum, C. (2016). The Treatment of Multi-word Units in Lexicography. In P. Durkin (ed.) *The Oxford Handbook of Lexicography*. Oxford: Oxford University Press, pp. 411–424.
- Fillmore, C.J. (2008). Border Conflicts: FrameNet Meets Construction Grammar. In E. Bernal & J. DeCesaris (eds.) *Proceedings of the XIII EURALEX International Congress*. Barcelona: Universitat Pompeu Fabra, pp. 49–68.
- Goldberg, A.E. (2006). *Constructions at work: The nature of generalization in language*. Oxford University Press.
- Indig, B., Sass, B., Simon, E., Mittelholcz, I., Vadász, N. & Makrai, M. (2019). One format to rule them all – The **emtsv** pipeline for Hungarian. In *Proceedings of the 13th Linguistic Annotation Workshop*. Florence, Italy: Association for Computational Linguistics, pp. 155–165. URL <https://www.aclweb.org/anthology/W19-4018>.
- Janda, L., Endresen, A., Zhukova, V., Mordashova, D. & Rakhilina, E. (2020). How to build a constructicon in five years: The Russian example. *Belgian Journal of Linguistics*, 34, pp. 162–175.
- Janssen, M. (2008). Meaningless Dictionaries. In E. Bernal & J. DeCesaris (eds.) *Proceedings of the XIII. EURALEX International Congress*. Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, Barcelona, pp. 409–420.
- Jurafsky, D. (1991). *An On-line Computational Model of Human Sentence Interpretation: A Theory of the Representation and Use of Linguistic Knowledge*. Ph.D. thesis, Department of Electrical engineering and computer sciences, University of California, Berkeley.
- Levary, D., Eckmann, J.P., Moses, E., & Thusty, T. (2012). Loops and Self-Reference in the Construction of Dictionaries. *Phys. Rev.*, X(2), p. 031018.
- Lyngfelt, B. (2018). Introduction: Constructicons and constructicography. In Lyngfelt et al. (2018a), pp. 1–18.
- Lyngfelt, B., Borin, L., Ohara, K. & Torrent, T.T. (eds.) (2018a). *Constructicography: Constructicon development across languages*. Amsterdam: John Benjamins.
- Lyngfelt, B., Bäckström, L., Borin, L., Ehrlemark, A. & Rydstedt, R. (2018b). Constructicography at work: Theory meets practice in the Swedish Constructicon. In Lyngfelt et al. (2018a), pp. 41–106.
- Megyesi, B. (1998). *The Hungarian Language: A Short Descriptive Grammar*.
- Mittelholcz, I. (2017). **emToken**: Unicode-képes tokenizáló magyar nyelvre. [**emToken**: a Unicode-capable tokenizer for Hungarian.]. In V. Vincze (ed.) *MSZNY2017*. Szegedi Tudományegyetem, Informatikai Tanszék csoport, pp. 70–78.
- Novák, A., Siklósi, B. & Oravecz, Cs. (2016). A New Integrated Open-source Morphological Analyzer for Hungarian. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk & S. Piperidis (eds.) *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Paris, France: European Language Resources Association (ELRA).
- Oxford University Press (2023). *Oxford Advanced Learner’s Dictionary*. URL <https://www.oxfordlearnersdictionaries.com>.

- Pethő, G., Sass, B., Kalivoda, Á., Simon, L. & Lipp, V. (2022). Igekötő-kapcsolás [Connecting perverbs to verbs]. In G. Berend, G. Gosztolya & V. Vincze (eds.) *MSZNY 2022*. Szegedi Tudományegyetem, Informatikai Intézet, pp. 77–91.
- Pusztai, F. (ed.) (2003). *Magyar Értelmező Kéziszótár [Hungarian Monolingual Explanatory Dictionary]*. Akadémiai Kiadó.
- Sass, B. & Pajzs, J. (2010). FDVC – Creating a Corpus-driven Frequency Dictionary of Verb Phrase Constructions for Hungarian. In S. Granger & M. Paquot (eds.) *Proceedings of eLex 2009*. Louvain-la-Neuve: Presses universitaires de Louvain, pp. 263–272.
- Sato, H. (2012). A Search Tool for FrameNet Constructicon. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA), pp. 1655–1658.
- Teubert, W. (2005). My version of corpus linguistics. *International Journal of Corpus Linguistics*, 10(1), pp. 1–13.

Probing visualizations of neural word embeddings for lexicographic use

Ágoston Tóth¹, Esra Abdelzaher²

¹ University of Debrecen, Faculty of Humanities, Department of English Linguistics

² University of Debrecen, Institute of English and American Studies

University of Debrecen, Doctoral School of Linguistics

E-mail: toth.agoston@arts.unideb.hu, esra.abdelzaher@gmail.com

Abstract

Our study explores the possibility of using the distributional characteristics of headwords as exemplified in the online Oxford Learner's Dictionaries, captured by contextualized word embeddings and displayed in two dimensions to help lexicographers find sense categories, detect variations across senses and select potential example sentences. In addition to the dictionary examples, we added British National Corpus data that contained the headwords. BERT word embeddings were extracted for all occurrences of the headword, then two-dimensional representations of the resulting high-dimensional BERT embedding vectors were created using 4 algorithms: MDS, Isomap, Spectral and t-SNE. Clustering was assisted by k -means clustering and Silhouette scoring for different k values. Our investigation showed that Silhouette scores for k -means increased after dimension reduction; furthermore, spectral and t-SNE visualizations were associated with the most cohesive clusters. The highest Silhouette scores recommended a number of clusters different from the number of dictionary senses, but semantic and syntactic patterns were detectable across the recommended clusters.

Keywords: sense delineation; word embedding visualization; BERT

1. Introduction

Lexicography is open to incorporating advances in information technology, especially when a large amount of manual labour can be substituted. Consider how quickly concordancing became computerized, also the swift adaptation of database management systems to store lexicographic data, or the introduction of methods for quantitative corpus analysis, including those for detecting potential collocations via scoring first-order (syntagmatic) word co-occurrence patterns using t-score, MI-score, etc.

The idea that word distribution can be directly exploited for capturing meaning was pointed out by Firth (1957), who famously argued that the meaning of a word is distributed over the neighbouring words, or the company that words keep. Words may be distributionally similar (therefore, they appear in paradigmatic relations in their second-order co-occurrence patterns) for semantic and structural reasons; the presence of the semantic component is now being actively exploited in Natural Language Processing and Artificial Intelligence research. In what follows, we will refer to this area

of interest as Distributional Semantics (DS; cf. Lenci, 2008).

In the 2010s, the quick spread of connectionist language modelling and the eventual introduction of Large Language Models (LLMs) changed Distributional Semantics in its implementation, and expanded the range of applications in Natural Language Processing. Machine learning algorithms based on artificial neural networks get distributional data from large amounts of text while learning to solve distribution-related tasks (such as masked-word prediction, next-word prediction and context prediction). While doing so, they internally characterize the tokens of the text that they are processing; we call these internal characterizations *word embeddings*. The latest generation of LLMs, which includes the ELMo model (Peters et al., 2018), BERT (Devlin et al., 2019) and GPT (Radford et al., 2018), are designed to dynamically associate actual uses of tokens with their distributional features, giving us *contextualized* embeddings. It is reasonable to evaluate whether contextualized word embeddings can be used for identifying senses for lexicographic use, too.

Sense delineation presents a significant challenge to practicing lexicographers, given the complexity and fuzziness of meaning categories. Explaining the meaning of a simple word such as *dog* requires knowledge about multiple semantic fields including shape, movement and sound. Linguists have the means to discuss the complexity of the meaning of words and how they may overlap when sharing the same conceptual base or schematic structure (e.g. Langacker, 1999; Lakoff, 1987 and Fillmore & Atkins, 1992). Lexicographers, however, need to represent word meaning as a finite list of senses. In this regard, deducing word senses from corpus uses is very challenging. Using the target word as part of a name or sublanguage is likewise problematic for lexicographers. Lexicographers have to decide whether this is a different unpredictable sense that should be recorded in a dictionary or not. Moreover, non-standard word use always depends on deviation from the known use. However, the new use is not always salient for users, specifically if triggered by a combination of words rather than a single target word (Kilgarriff, 2007).

In this paper, we explore the possibility of employing BERT word embeddings as tools for identifying senses of words as they appear in dictionary examples and also in additional corpus sentences. Section 2 of this paper discusses related work in the literature. Section 3 presents the methodology of the current research from data collection, through producing 2-dimensional visualizations that may assist lexicographic work, to the examination of the clusters. Section 4 has the qualitative analysis of the visualizations for the four words that we have selected for this analysis. Our concluding remarks are presented in Section 5, where we also discuss the limitations of our research.

2. Related work

Rychlý & Kilgarriff (2007) offered a DS method for building distributional thesauri. They used a corpus of lemmatized and parsed language to gather information about

how words are used in context, including the grammatical relations between a target word and other (context) words in sentences. The method then identifies other words that share similar contexts. This function is also available in the Sketch Engine, where “Sketch differences” rely on lexical collocates and grammatical relations in the contexts to show how (dis)similar two words are (Kilgarriff et al., 2014). This type of information has been useful in unveiling word senses that are not present in dictionaries (see, for instance, Abdelzaher & Tóth, 2020). The “Sketch differences” tool does not use contextualized word embeddings.

Jatowta, Tahmasebi & Borin (2021) give a review of the literature that tracks meaning change in a diachronic setting using distributional data of words, and tackle the question of visualization, too. The paper illustrates that even static embeddings can help us compare different states of the language if we generate snapshots for the states under scrutiny, generate static embeddings for them and compare these embeddings. Unfortunately, static embeddings contain a mix of all senses, all usages of the given word, so they cannot directly help the sense delineation process. The possibility of using contextualized word embeddings is pointed out by the authors as a possible future direction.

Montes & Heylen (2022) visualize distributional semantic data for testing different word embedding parameter sets (which is common practice with static “count-type” embeddings) and also for checking the distributional properties of the word under scrutiny – the Dutch word *heffen* with 2 senses. Their study is presented in the context of cognitive linguistics. In our present paper, we utilize a single, pre-trained distributional model that implements a modern contextualized word embedding type designed to collect token-level distributional information in a context-sensitive way; the parameters that we test are related to the visualization step rather than distribution modelling, and our focus is on sense delineation within the context of lexicography.

In our work, we use BERT word embeddings (Bidirectional Encoder Representations from Transformers; Devlin et al., 2019), which is a well-established contextualized embedding type in Natural Language Processing. BERT is based on the Transformer architecture (Vaswani et al., 2017). The model learns to predict a masked word in a sentence and to decide if two sentences appeared sequentially in the training corpus. As a contextualized model, BERT captures the distributional properties of actual uses of words (more precisely, those of tokens in its vocabulary) in given contexts. Outside of the field of lexicography, contextualized word embeddings have been proven to form distinct clusters corresponding to different word senses in Wiedemann et al. (2019).

3. Methods

3.1 Data collection

In our analysis, we created two-dimensional (2D) visualizations of BERT embeddings for instances of four headwords: *full*, *mouth*, *risk* and *sound*, as exemplified in dictionary example sentences and found in the British National Corpus.

The professionally selected and edited *dictionary* examples were taken from the online *Oxford Learner's Dictionaries* at <http://www.oxfordlearnersdictionaries.com> (OD). We took all examples (including the “Extra Examples”) of the selected headwords in all senses, but we had to discard those examples that contained an inflected form of the headword, as inflected forms are treated as different BERT tokens (which may get related in their representations, but the analysis of the relation between the embeddings of headwords and inflected forms is beyond the scope of this paper) or, in some cases, sequences of tokens. Hornby's *Idiomatic and Syntactic English Dictionary* (Hornby, 1948), which is known for its inclusion of syntactic information and its focus on word complementation, is part of OD's heritage, which may be reflected in the example sentences OD provides for each word sense. For this reason, different syntactic patterns corresponding to different senses are expected to stand out in the visualized representations.

The additional *corpus* sentences (1000 for each headword) were taken from the British National Corpus (BNC) available via <http://www.sketchengine.eu>. We used the sentence concordancer option, looked up the word, shuffled the output and exported the data. We did not filter for part of speech. While BNC may not be the most extensive or most up-to-date corpus of English, it is a balanced representation of British English (Leech, 1992). We collected examples that contained the exact headword.

3.2 Creating BERT embeddings

We produced contextualized word embeddings for the headwords in the dictionary example sentences and corpus examples. The embeddings were created using the Huggingface BERT libraries (<https://huggingface.co>). We relied on a pre-trained BERT model (*bert-large-uncased*, <https://huggingface.co/bert-large-uncased>) and the corresponding *bert-large-uncased* tokenizer from Huggingface. The BERT-large model contains 336 million trained parameters with 24 layers and 16 attention heads. We did not fine-tune the network, as we wanted to visualize pure distributional data acquired for the standard BERT learning goals. The resulting word embeddings were vectors that contained 1024 floating point numbers for each use of the given headword in the dictionary examples and corpus sentences; we used the embedding developed in the last layer of BERT in the position of the target word. According to the distributional

hypothesis, more similar uses of the target words are in closer proximity to one another when we visualize distributional feature vectors in the resulting 1024-dimensional space.

3.3 Dimension reduction

We used manifold learning algorithms for dimension reduction from 1024 to 2 dimensions as they are capable of preserving the underlying structure of the data.

We employed four algorithms: Multidimensional Scaling (MDS), Isomap, Spectral and t-SNE. MDS is a linear method, which is computationally efficient, while the three non-linear methods should be able to learn more complex relationships between the data dimensions.

MDS creates a low-dimensional representation by minimizing the difference between distances of data point pairs in the high-dimensional space and pairwise distances in the low-dimensional space. The main contributions to the field of MDS are reviewed in Groenen & Borg (2014).

Isomap (Tenenbaum, de Silva & Langford, 2000) is based on graph theory. It uses geodesic distance, which is a path between two points on a surface – rather than along a straight line. The Isomap graph is created by connecting neighbouring points and computing the geodesic distance between each pair of points. The algorithm uses MDS to embed the data into a low-dimensional space preserving the pairwise geodesic distances.

Spectral clustering employs the graph Laplacian to encode the similarity between data points. The top eigenvectors of the Laplacian matrix are considered to capture the global structure of the data. Spectral embedding is known to be able to capture non-linear structures and different types of relationships. For details, see Ng, Jordan & Weiss (2002).

Finally, t-SNE (van der Maaten & Hinton, 2008) is a non-linear method that constructs a probability distribution over pairs of high-dimensional data points and a similar distribution over pairs of low-dimensional points, and it minimizes the difference between these two distributions using gradient descent in an iterative fashion. t-SNE is considered very effective at preserving the local structure of data at the expense of non-local structure.

t-SNE is often used in current Natural Language Processing research for dimension reduction. It is the infrequent use of the remaining three methods that led us to test the possibility of utilizing them for the task at hand. We suppose that lexicographers carrying out the manual evaluation of corpus data, and looking for – otherwise hidden – second-order co-occurrence patterns, would benefit from getting access to multiple methods to work with. Compare it to the range of tools we can use for detecting

potential collocates (and, in general, first-order co-occurrence patterns): t-core, MI-score, etc.

We used a free tool, the Orange Data Mining toolkit (Demsar et al., 2013; <https://orangedatamining.com>) for converting the 1024D token embeddings to 2D using the above manifold learning algorithms, and also for visualization of the 2D outputs as scatterplots. Figures 2, 3, 5, 6, 7, 8 and 9 of this paper were prepared using this program. The interactive scatterplots that you have access to while using the toolkit also offer zoom functionality and can show or hide sentences as data labels. These interactive services, which are not shown in this study, made an important contribution to our work. Note, however, that the Orange toolkit is not designed to be a “lexicographer’s workbench”.

3.4 *k*-means analysis of the clusters using Silhouette scores

In addition to visual observation of the low-dimensional representations, we also studied the original high-dimensional feature space and its 2D representations using *k*-means clustering with additional Silhouette scoring for selecting *k*.

K-means clustering is commonly used for grouping data points into clusters automatically, based on their similarity to each other. In our case, *k* centroids are initially selected using the *k*-means++ algorithm (Ostrovsky et al. 2006). Then data points are assigned to the closest centroids based on squared Euclidean distances. After this assignment step, an update step is carried out, which recalculates the centroids to optimize the overall result of the clustering. In our experiment, we allowed for a maximum of 5000 iterations over the assignment and update steps. The algorithm is sensitive to the initial selection of the centroids (even with the *k*-means++ initial centroids); therefore, 20 reruns were performed, and the run with the lowest within-cluster error (lowest sum of squares) was kept.

The selection of the number of the clusters is of special importance in our case. It runs parallel to the lexicographic task of sense delineation, which involves drawing borderlines between senses, polysemous and homonymous, where polysemous senses are related in their meaning by definition. The lexicographical task of splitting and lumping senses is known to be challenging, and it is not automatized. In our exploratory research, we took OD’s senses as reference points, but we also wanted to know the number of clusters that BERT data (raw and 2D-converted) naturally exhibited. Therefore, we used Silhouette scoring (Rousseeuw, 1987) of different *k* values in *k*-means analysis. Silhouette scoring is a measure of how well data points fit into their clusters, and it “shows which objects lie well within their cluster, and which ones are merely somewhere in between clusters” (ibid.). A higher score indicates better clustering.

We carried out *k*-means clustering and calculated the Silhouette scores using the Orange Data Mining toolkit. We did not perform added quantitative evaluation of the clusters

(using Rand index or V-measure, for instance) in addition to what we have access to in the toolkit. Quantitative and qualitative analyses of the resulting plots are provided in the next section.

4. Results

4.1 Silhouette scores and k -means clusters before and after dimension reduction

Silhouette scores increased for all words after dimension reduction. In most cases, the number of clusters (C) was similar before and after dimension reduction and for the different visualization methods. However, for *risk*, the number of the suggested best clusters based on the 1024D distributional representations differed considerably from that recommended after t-SNE visualization. Figure 1 shows the Silhouette scores for different k -means clusters before and after the dimension reduction of the distributional representations of *risk*.

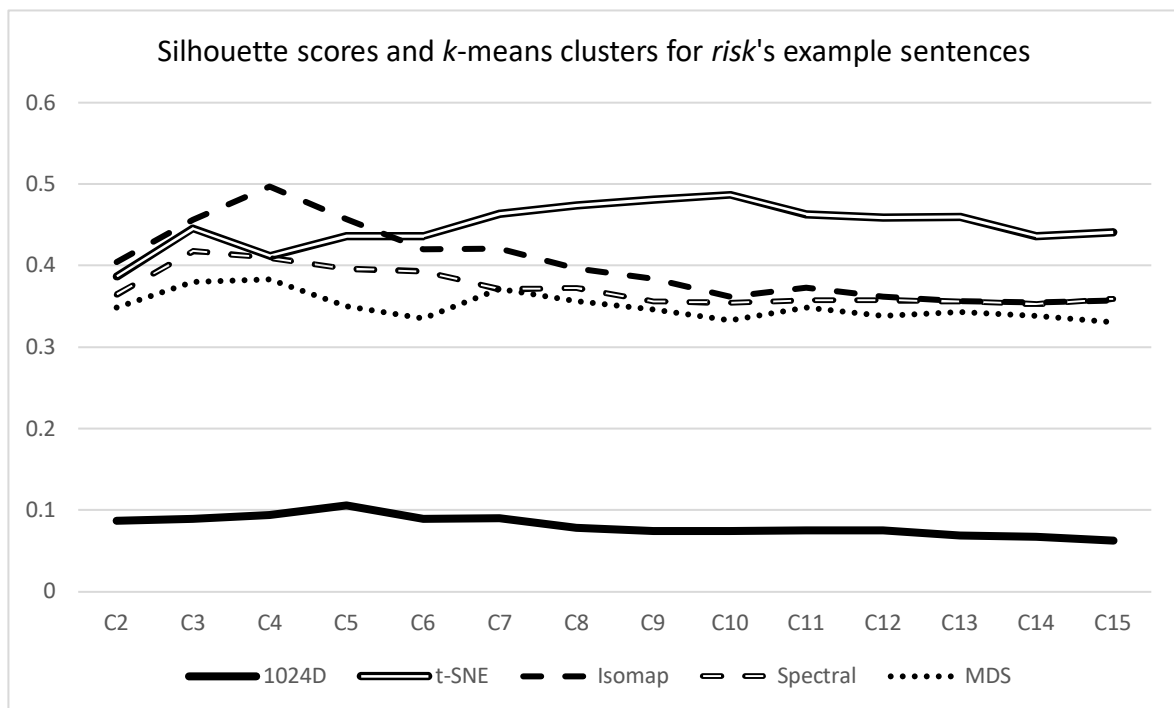


Figure 1: Silhouette scores for 2-15 clusters of *risk* before and after dimension reduction

The Silhouette scores for 2-15 clusters based on the 1024 dimensions represent an almost linear line on the chart without any significant peaks, at a consistently low value. On the contrary, for the t-SNE visualization, there is an increase in the Silhouette score for cluster three (0.456) and cluster ten (0.487). The best Silhouette score is associated with four clusters based on the Isomap visualization (0.497).

Before dimension reduction, the suggested five clusters hardly reflected any patterns. Figure 2 visualizes the box plot of the *k*-means clusters and a sample of the sentences in each cluster based on the 1024D representation of *risk*. Whereas the BNC sentences were distributed across the five clusters, the verbal senses of *risk* clustered together in C5. However, the same cluster usually contained heterogeneous sets of the uses of *risk*. C5 included the verbal senses of *risk* as recorded in the OD sentences and also had some of the nominal senses. C1 included only the nominal uses, but several contexts were present in the cluster. Medical risk was dominant in C1, but instances of *risk* in statistical and economic contexts appeared towards the end of the cluster. C2 was mostly associated with financial risks but also included several health-related risks towards the end of the cluster. Sentences in C3 referred to social, environmental, economic and medical risks. Sentences in C4 generally referred to risky situations without specification (at the top of the cluster) and associated *risk* with business loss and body injuries, among others.

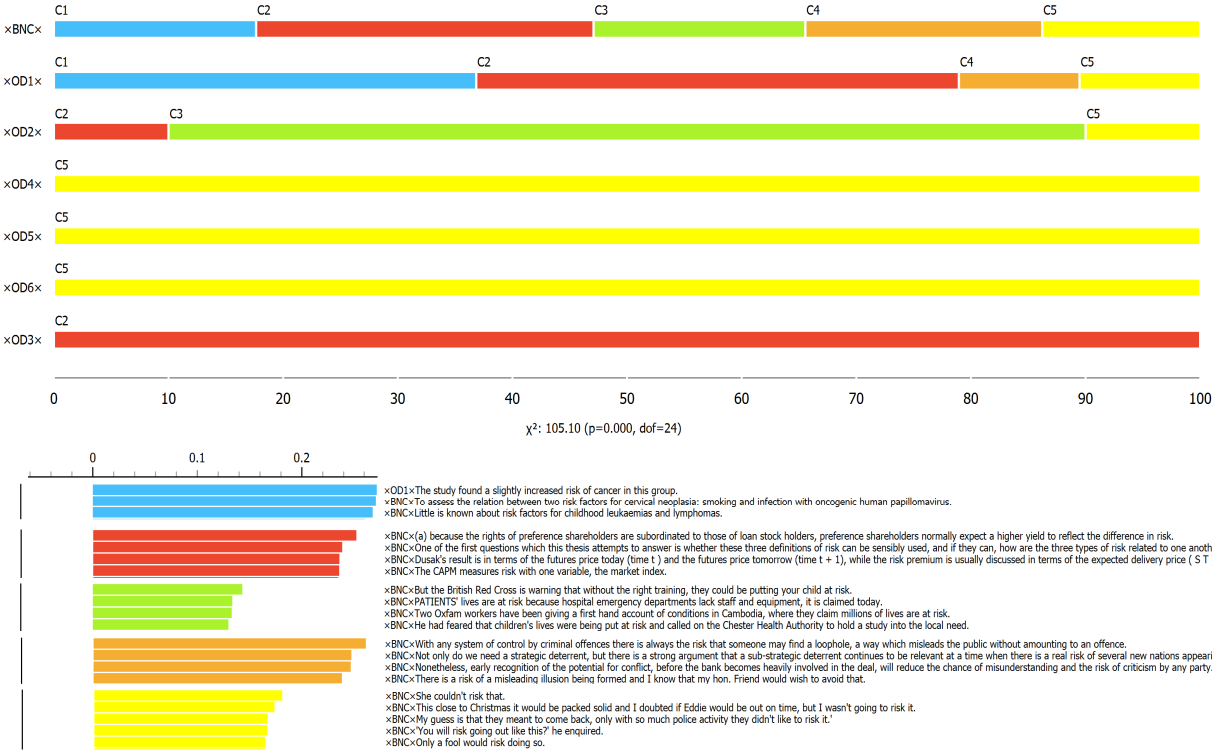


Figure 2: *k*-means clusters and sample sentences for *risk* in 1024D space

The increased Silhouette scores after the dimension reduction were reflected in the sentences grouped in each cluster. The suggested ten clusters based on the t-SNE visualization showed semantic and syntactic patterns shared among most of the sentences in a cluster. First, the verbal senses of *risk* clustered in C3 with verbal uses from the BNC, without nominal senses from OD in the cluster. Second, patterns, such as $V_{be} risk to NP$ in C1, $increase(d)/reduce(d)/ high/ low risk of NP$ in C2, $risk (of)+ing$ and $risk+that+clause$ in C4, started to appear in the clusters frequently. Third, compounds such as $adj+risk+N$ were most frequent in C6, whereas collocates such as

at risk distinguished the sentences in C7. Fourth, sentences referring to health-related risks were primarily placed in C2, whereas business and financial risks dominated C5. Figure 3 displays the box plot of the k -means clusters and a sample of the sentences with *risk* after t-SNE visualization.



Figure 3: t-SNE-based k -means clusters and sample sentences for *risk*

Unlike the case of *risk*, the differences in the k -means clusters were minor for *mouth*. Figure 4 shows the Silhouette scores for *mouth* before and after dimension reduction. The Silhouette scores for different k values for the MDS visualization are almost similar, and they are considerably low. The best Silhouette score was 0.112 for two clusters before dimension reduction. After dimension reduction, the four visualization methods suggested three clusters as the best categorization of the five OD senses of *mouth* (i.e., part of the face, a person needing food, of a river, opening or entrance and way of speaking). The Silhouette score was best for the Spectral-based clusters (0.577), followed by Isomap (0.559), t-SNE (0.481) and MDS (0.375).

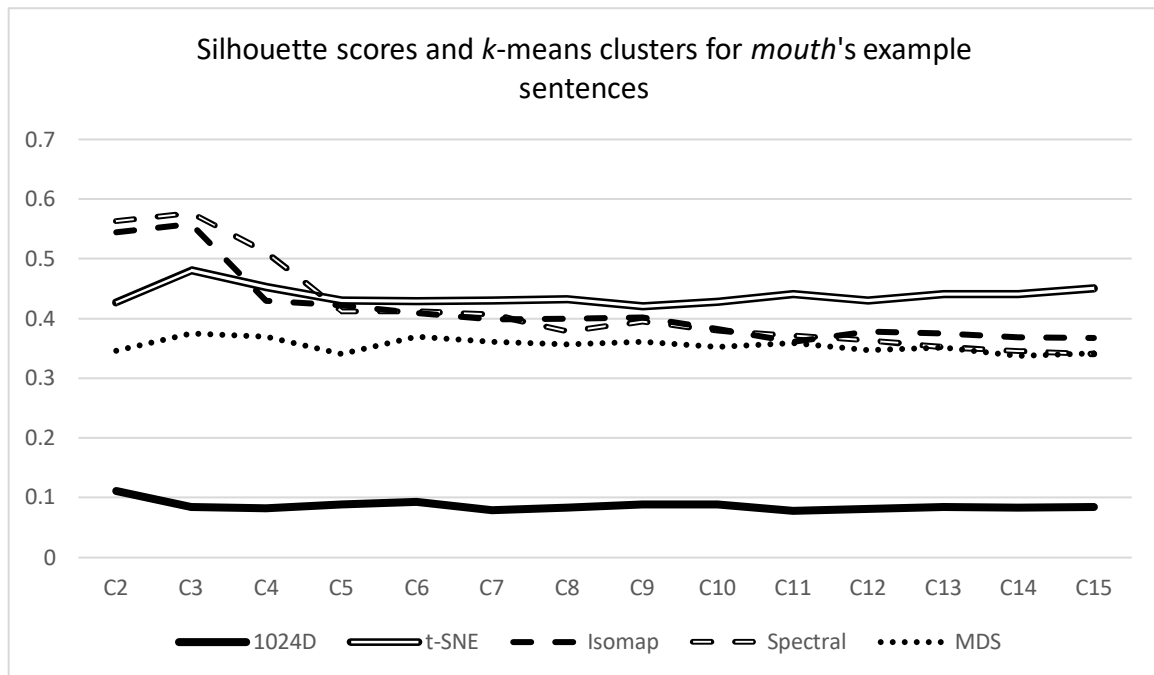


Figure 4: Silhouette scores for 2-15 k -means clusters for *mouth*'s example sentences

The remaining part of this section explores the sentences in the suggested two clusters based on the 1024D distributional representations and in the three clusters suggested based on the Spectral representation. Figure 5 shows the box plot of the k -means clusters for *mouth* in 1024D and the Silhouette plot of a sample of the sentences in the two clusters. As visualized, all OD senses are clustered in a single category, whereas a group of BNC sentences form a distinctive cluster. The first cluster contained a diaspora of heterogeneous sentences, and the second cluster mostly had sentences in which *mouth* was used in a romantic fiction genre. The literal sense of *mouth* (part of face), the metaphoric sense (opening) and the metonymic sense (way of speaking) appear in the same cluster.

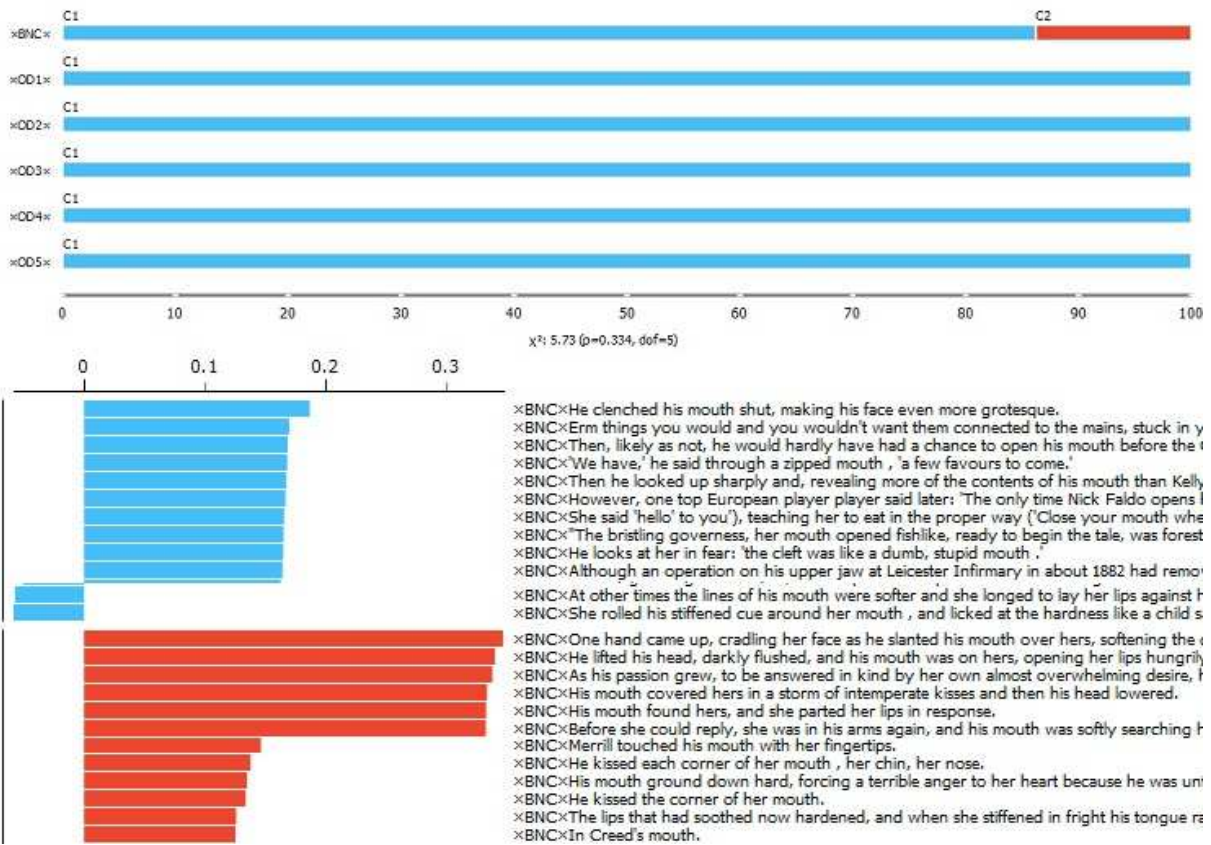


Figure 5: *k*-means clusters and sample sentences of *mouth* before dimension reductions

After dimension reduction, Spectral visualization showed the best Silhouette scores (0.577) for three clusters. The first cluster contained most of the senses of *mouth* (senses 2, 3, 4 and 5 in OD and some sentences from sense 1) and most BNC examples. Cluster two included the same romance-related uses of *mouth*, which clustered likewise before dimension reduction. However, a new category appeared and separated the uses of *mouth* to make facial expressions from other senses. The newly introduced cluster grouped sentences from OD’s sense 1 and BNC examples.

4.2 Silhouette scores and *k*-means clusters: two perspectives

This section compares the best *k*-means clusters recommended by the Silhouette scoring to *k*-means clusters with *k* set to the number of dictionary senses. For *mouth*, the recommended clusters after using the four visualization methods were three as mentioned in the previous section (C3: making facial expressions, C2: romance-related sense, and C1: all other senses). We had five OD dictionary senses for *mouth*. Preselecting the number of clusters to five slightly improved the sub-clusters of the sentences, but it did not correspond to the dictionary senses. The three categories of *mouth* in romantic contexts, speaking and making facial expressions stood out again, although the literal use of the mouth to speak and the metonymic use as a way of speaking overlapped in clusters 1 and 5. The two added clusters contained a diaspora

of uses. For instance, cluster 1 included sentences referring to *mouth* in a medical context, as a way of speaking and with reference to eating and drinking. Cluster 5 grouped the metaphoric uses of *mouth* as ‘mouth of a river’ or ‘entrance of a cave’ with the literal uses of *mouth* in speaking. Figure 6 shows some of the similarity patterns in the sentences based on Spectral visualization of 5 *k*-means clusters.

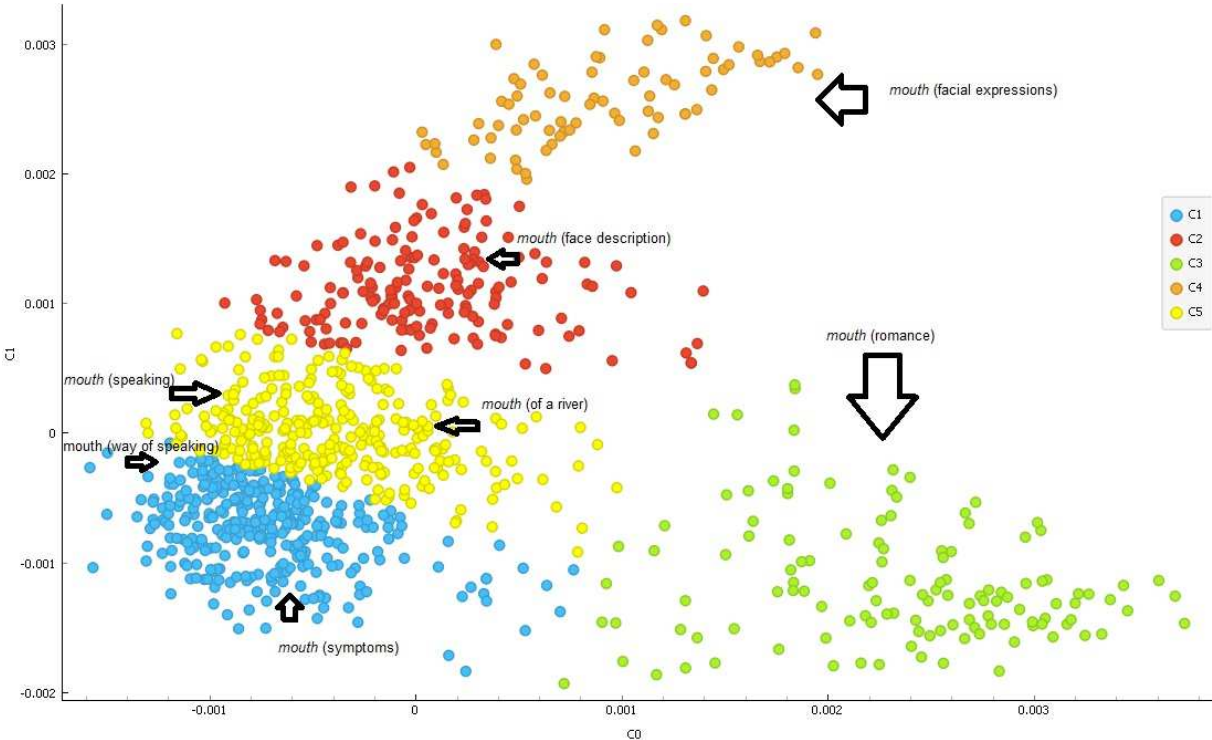


Figure 6: Scatter plot of 5 *k*-means cluster based on Spectral visualization of *mouth*'s sentences (colours indicate different *k*-means clusters as shown in the chart legend)

The same applies to *full*, which has 11 dictionary senses in the current study. However, before and after dimension reduction, the best Silhouette scores recommend two or three clusters for all the sentences of *full*. After manually setting the *k*-means clusters to 11, sentences in the clusters did not reflect the dictionary sense delineation. On the contrary, the same cluster contained semantically and syntactically dissimilar sentences whereas similar sentences overlapped in different clusters. As illustrated in figure 7, sentences expressing the literal and metaphoric senses of *full* as ‘having a lot’ appeared in four neighbouring clusters with no explicit patterns separating or joining them. In addition, the pattern *full* + noun which denotes ‘complete’ was frequent in two different categories.

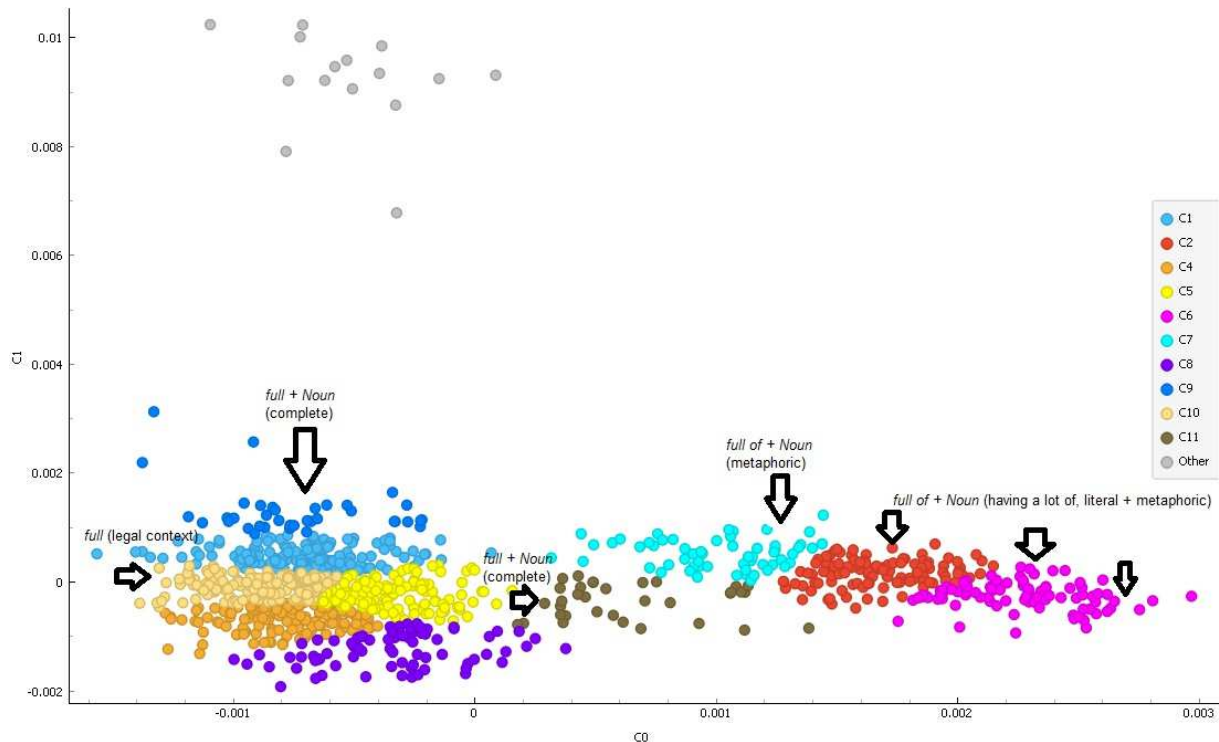


Figure 7: Spectral-based scatter plot of sentences with *full* in 11 pre-set *k*-means clusters

It is evident from the four case studies, investigated in this research, that pre-setting the number of clusters to match dictionary senses will not be helpful. However, depending on the automatically calculated highest Silhouette scores may be a better reflection of the patterns of use and, accordingly, of word senses, too, in or outside lexicographical contexts.

4.3 Comparing different visualization methods

Spectral, t-SNE and Isomap showed the best Silhouette scores for all words, unlike MDS. Figure 8 shows the four visualizations of the sentences of *sound* in a 2D space. Sentences are sporadically distributed all over the space with MDS, even if they instantiate the same sense. On the contrary, the visualized spaces created by Spectral, t-SNE and Isomap cluster the sentences closer to each other in major classes based on the part of speech. Sense categories are more salient in the t-SNE visualization of the examples of *sound*. First, the different parts of speech formed distinctive clusters all over the 2D space. Second, dissimilar senses belonging to the same POS appeared in different clusters. For instance, the nominal sense of *sound* as a passage of water appeared in a distinctive cluster other than the phonetics-, music- and television-related senses. Also, the verbal senses of *sound* as ‘give impression’ versus ‘make a sound’ appeared in two clusters with considerable distance between them. The similar nominal and verbal senses of *sound* as ‘an impression’ and ‘give impression’ formed close, but separate clusters.

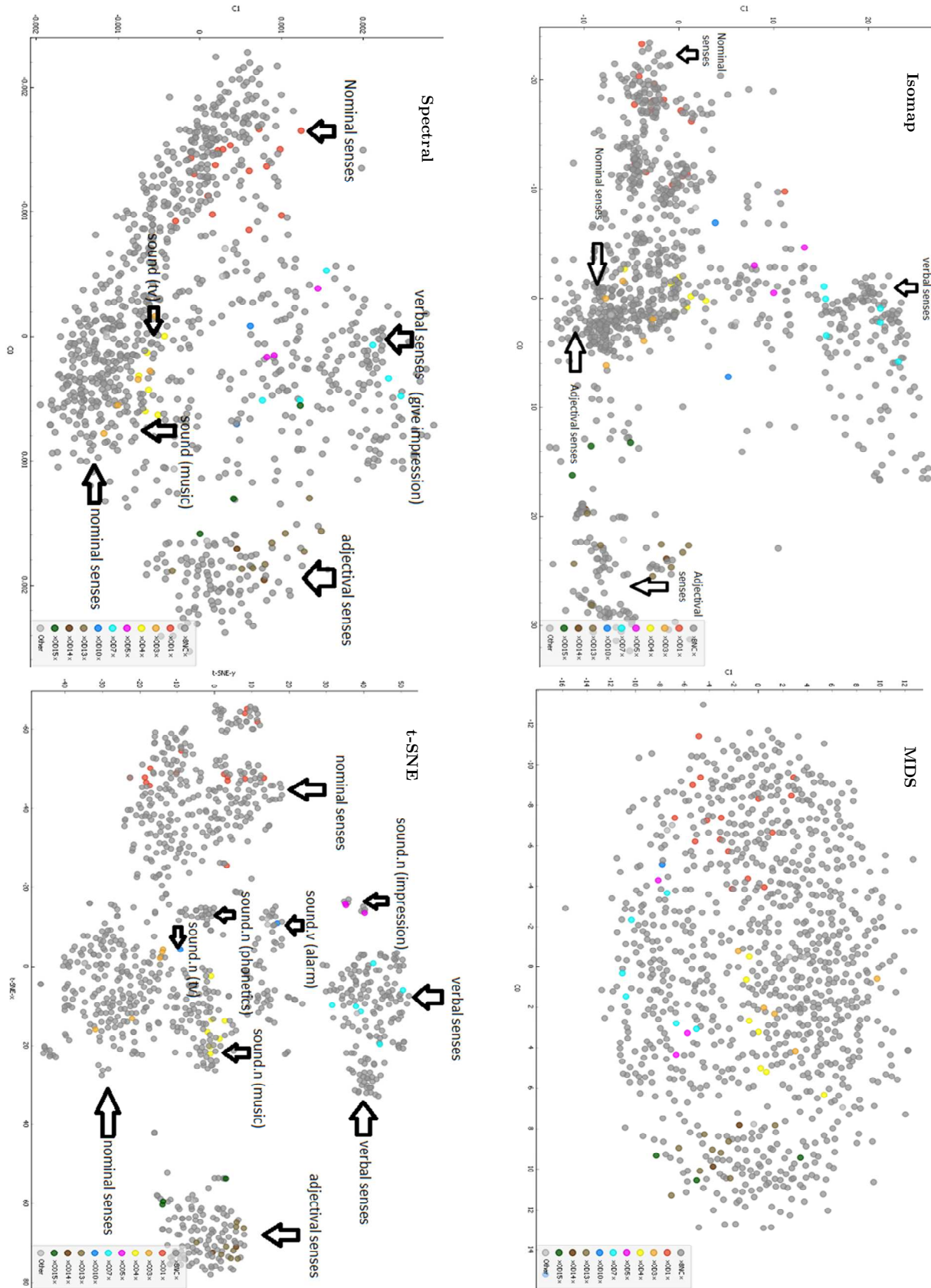


Figure 8: Four 2D visualizations of *sound's* example sentences

In the initial phase of our experiments, manual parameter tuning was carried out based on Silhouette scores and also on the qualitative features of the resulting clusters, typically with one or two words. The parameter sets that we settled with for the different dimension reduction methods are shown in Table 1.

Dimension reduction	Settings
t-SNE	perplexity = 20 distance = Euclidian initialization = PCA max. iterations = 3000 learning rate = 200
MDS	initialization = PCA max. iterations = 5000
Isomap	neighbours = 20
Spectral	affinity = RBF kernel

Table 1: Parameter choices for the dimension reduction methods

We do not argue, however, that a single parameter set will cover all usage scenarios, all words of interest, all corpus sizes, etc. Instead, we recommend that the user should be given choices and the opportunity to find the most useful methods and settings. The t-SNE algorithm, for instance, is notoriously sensitive to the perplexity parameter, which balances the effect of local vs. distant neighbours on the resulting low-dimensional representation. We tried different values, and, in addition, we also explored different distance metrics, including Euclidean, Manhattan and Chebychev. Whereas the number of recommended clusters remained almost the same for all words, the Silhouette scores changed slightly. The best scores were mainly associated with the Euclidean metric and perplexity set as 20. Table 2 shows the suggested cluster numbers for *sound* corresponding to several t-SNE settings.

Distance metric	Perplexity	Clusters	Silhouette Scores
Euclidean	10	4	0.574
Euclidean	20	4	0.591
Euclidean	30	4	0.589
Manhattan	10	4	0.572
Manhattan	20	4	0.591
Manhattan	30	4	0.582
Chebychev	10	4	0.552
Chebychev	20	4	0.557
Chebychev	30	4	0.546

Table 2: The suggested clusters and Silhouette scores for *sound* in different t-SNE settings

Importantly, changing the parameters did not influence the inclusion of the OD sentences in the clusters or their overall position in the charts. The adjectival senses remained in the same cluster (C1) and appeared together on the t-SNE charts. Also,

the verbal and nominal senses of *sound* as ‘to give an impression’ and ‘the idea or impression’ were close to each other on the charts and formed a single cluster (C3). The nominal senses of *sound* with reference to phonetics, as a ‘passage of water’ and as ‘audible signals’ formed sub-clusters in cluster two (C2). The fourth cluster contained the verbal and nominal senses of *sound* as ‘something you hear’ and ‘produce a sound’.

Changing the affinity measures for the Spectral algorithm had a considerable influence on the results. For *mouth*, *risk* and *sound*, the nearest neighbour affinity retrieved better results than RBF kernel. It was the opposite for the word *full*, however. Table 3 depicts the suggested clusters for all words using RBF kernel and nearest neighbour in the Spectral algorithm.

Word	Affinity	Clusters	Silhouette score
Full	RBF kernel	2	0.838
Full	Nearest neighbour	3	0.601
Mouth	RBF kernel	3	0.577
Mouth	Nearest neighbour	3	0.775
Risk	RBF kernel	3	0.418
Risk	Nearest neighbour	4	0.517
Sound	RBF kernel	4	0.529
Sound	Nearest neighbour	3	0.730

Table 3: The suggested clusters and Silhouette scores based on Spectral’s affinity measures

Let us point out, however, that while the Silhouette scores increased with the nearest neighbour affinity, the homogeneity of the classes decreased in most cases. Figure 9 shows the distribution of the sentences with *mouth* over the Spectral space using the nearest neighbour measure. The cohesion of the clusters is evident, and the distance between some uses (e.g. ‘using the mouth to make facial expressions’ and ‘reference to the mouth in face description’) is noticeable. However, the overlap between the example sentences shows the heterogeneity of the sentences that form cohesive clusters. The figurative use of *mouth* as ‘an opening of a hole or cave’, the collocation *mouth open* with reference to surprise and *mouth* in relation to the medical field overlapped in the same cluster. Also, a mixture of literal and metaphoric uses of *mouth* and *open* were merged in the same cluster.

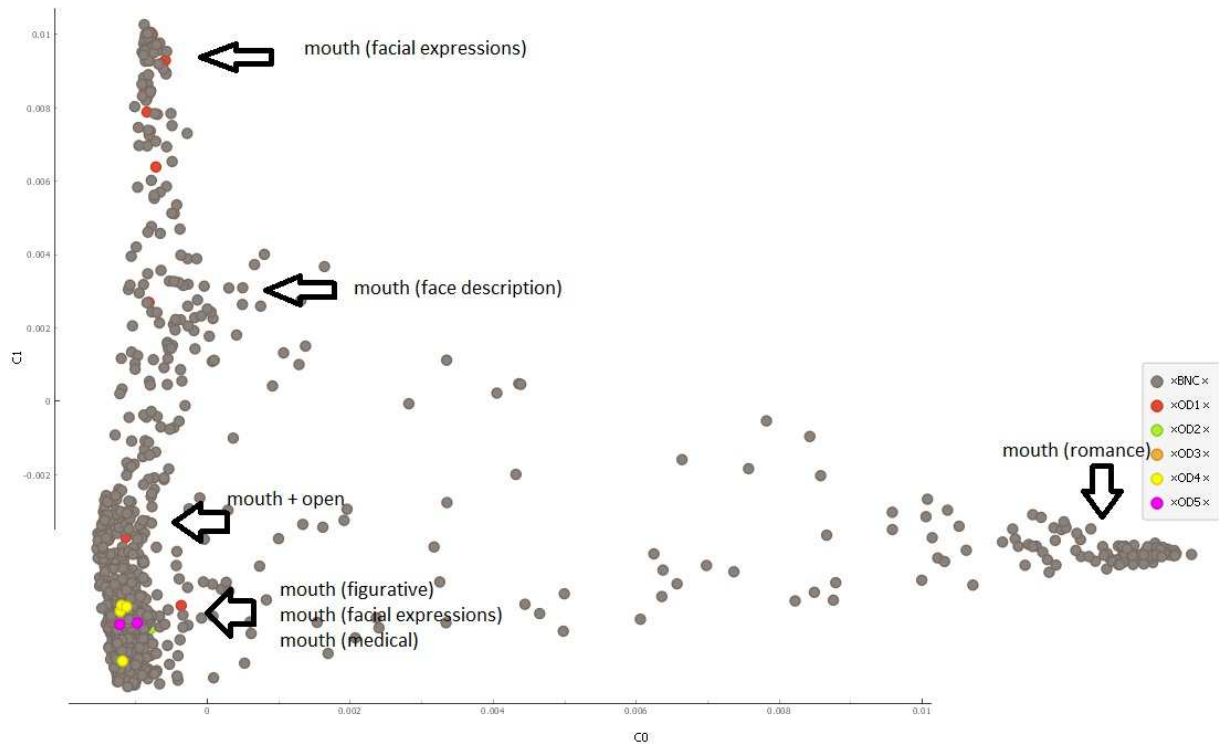


Figure 9: Spectral visualization using nearest neighbour measure for *mouth* sentences

The use of RBF kernel decreased the Silhouette score of the clusters, but the homogeneity of the clusters and sub-clusters within improved. Figure 6 has already illustrated the distribution of *mouth* sentences using RBF kernel in the Spectral algorithm. It showed the separation between the metaphoric, metonymic and literal senses of *mouth* in the clusters and the closeness between face-related senses in clusters 2 and 4 and speaking-related senses in clusters 1 and 5.

Regardless of the parameters, the cohesion of the clusters increased after dimension reduction. Figure 10 summarizes the Silhouette scores of the *k*-means before and after using different 2D visualization methods for the four words examined in this study. It is evident that the cohesion of clusters considerably increased after the dimension reduction for all words. Also, the suggested best number of clusters differed across words and visualization methods. The highest Silhouette score was 0.838 for Spectral visualization of the sentences of *full*. For the same word, the Silhouette score for the MDS visualization was the lowest (0.392), although the two visualization methods recommended the same number of clusters. The visualization created by Spectral clustered the sentences closer to each other in two major classes. Most sentences following the pattern *full*+noun formed a cluster different from sentences following the pattern noun+*V_{be}*+*full* of+noun. Some sentences were sporadically distributed over the two clusters. However, they also showed some patterns, such as the collocations *full up* and *full to* and the pattern noun+*V_{be}*+*full*. Although the original senses of *full* in OD are 12, the Spectral visualization did not show sensitivity to the semantic differences between the sentences corresponding to the 12 senses. For instance, the metaphoric

senses of *full* (e.g., full of pain or joy) and the literal ones (e.g., full of books, clothes) are clustered in one category.

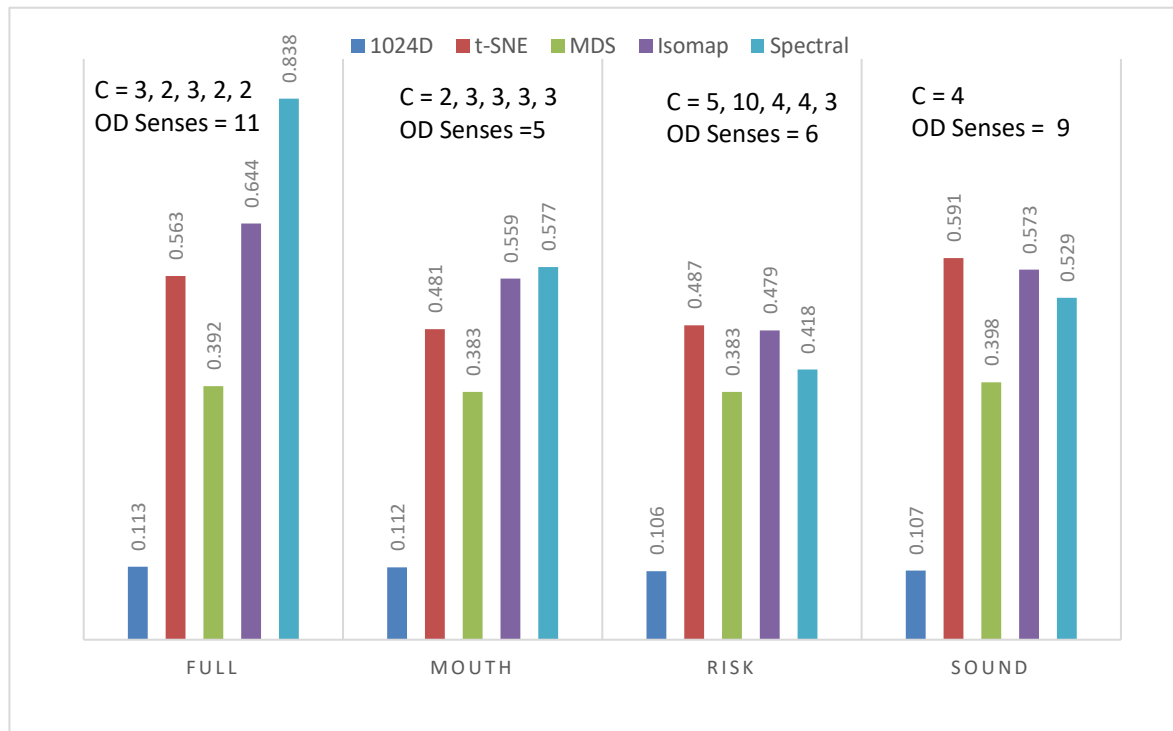


Figure 10: The highest Silhouette scores for the four studied words before and after dimension reduction

Several theoretical and computational approaches have been implemented in the literature to cluster dictionary senses into new categories. The clusters differed qualitatively and quantitatively according to the adopted approach. Whereas some studies depended on extensive qualitative analysis of dictionary data to improve the representation of senses for human users (e.g., Geeraerts, 2001; Lewandowska-Tomaszczyk, 2007; Molina, 2008), others aimed at improving the automatic performance of NLP tasks (for instance, Buitelaar, 1998, 2010; Ide & Wilks, 2007). Therefore, the number and members of the suggested clusters differed considerably.

Theory-based studies in lexicography highlighted the necessity of finding meaning relations among word senses (e.g., metaphoric and metonymic extensions of the literal senses), identifying the core literal meaning or meanings from which other meanings descend and organize word senses in homogenous categories that have always differed from those in the dictionaries. Although our study depended on distributional, rather than cognitive linguistic, approaches, the separation between the metaphoric, metonymic and literal senses of words such as *mouth* and *sound* was done automatically based on the distributional features of the word uses. Also, the uses of words with relevance to specific semantic fields (e.g., *risk* in financial domains, *mouth* to make

facial expressions, *full* with relevance to emotions) stood out in the automatically generated clusters.

The automatically generated clusters lumped several dictionary senses in the same cluster. It was most evident in the case of *full*, which had 11 fine-granular dictionary senses in our study. Yet, the different algorithms suggested 2 or 3 clusters only. Although the sub-clusters separated the metaphoric and the literal uses which were lumped in the dictionary, they also lumped the different levels of fullness which were split in the dictionary.

In almost all cases, the four algorithms reduced the number of OD’s sense categories. Some dictionary distinctions were preserved within the sub-clusters (e.g., *sound* of music vs. *sound* of TV and radio), but others were lost (e.g. the four verbal senses of *risk*). Reducing the number of dictionary senses has been proposed in some NLP initiatives that prioritize the improvement of the quantitative indicators (the accuracy of word sense disambiguation). They, however, sometimes opt for solutions that are incompatible with the lexicographic practice, such as maintaining only meaning distinctions at the highest ontological levels, as discussed by Ide and Wilks (2007).

Our study aimed at combining extrinsic assessment of the clusters with qualitative analysis of their homogeneity so that the experiments can be relevant to both lexicographers and NLP scholars interested in sense-related tasks.

5. Conclusion

This study explored the possible use of 2D visualizations of contextualized word embeddings in lexicographic context, specifically sense delineation and example selection. It presented case studies for lexicographers to test the applicability of employing the suggested visualization methods in lexicographic investigations. Although the distributionally-created clusters did not correspond to the number of dictionary senses, they showed BERT’s sensitivity to semantic and syntactic similarities between word uses.

Before dimension reduction, Silhouette scores of the k -means clusters were low, and so was the qualitative cohesion between the sentences in the cluster. Accordingly, providing lexicographers with distributionally-recommended clusters based on the original high-dimensional word embeddings are not helpful.

Visualizing BERT representations in 2-dimensional spaces using Spectral, t-SNE, Isomap and MDS algorithms showed quantitative and qualitative improvements that can be beneficial to lexicographers. For instance, not only the Silhouette scores of the k -means clusters increased, but also semantic and syntactic similarities appeared in the clusters and the manually identified sub-clusters within them.

Although the scope of the present study is limited to four words, to four dimension-

reduction methods and a single contextualized word embedding type (albeit a powerful one), we find these results novel and useful. The visualization of contextualized word embeddings of neologisms can help lexicographers identify their collocational patterns, POS usages and semantic preferences. Such patterns consistently appeared in the four case studies. Also, these visualizations can be helpful in enriching dictionary entries with additional, corpus-based examples; the closest BNC sentences to the OD examples mostly reflected very similar semantic and syntactic patterns in the four cases. In our charts, we also saw thematically-motivated clusters of BNC sentences that were ignored during exemplification of the OD headword (consider the uses of the word *mouth* in romantic literature), a situation which – when a representative corpus is used for the analysis – indicates a hiatus in the entry, which is not readily observable in concordances.

By taking advantage of the power of contextualized word embeddings and dimension reduction algorithms, we should be able to provide methods for lexicographers to explore and better understand the complex relationships between words and their meaning. These methods – enabled by current advances in Natural Language Processing – do not replace any subtask of the human “art and craft” of dictionary compilation, but they contribute to computer-assisted lexicography.

6. Acknowledgements

This publication was supported by the University of Debrecen Faculty of Humanities Scholarly Fund.

7. References

- Abdelzaher, E. & Tóth, Á. (2020). Defining Crime: A multifaceted approach based on Lexicographic Relevance and Distributional Semantics. *Argumentum*, 16, pp. 44–63, <https://doi.org/10.34103/ARGUMENTUM/2020/4>.
- Buitelaar, P. (1998). *CORELEX: Systematic polysemy and underspecification*. PhD thesis. Waltham, Massachusetts: Brandeis University.
- Buitelaar, P. (2010). Ontology-based semantic lexicons: Mapping between terms and object descriptions. In C. Huang, N. Calzolari, A. Gangemi, A. Lenci, A. Oltramari & L. Prevot (eds.) *Ontology and the lexicon: A natural language processing perspective*. Cambridge: Cambridge University Press, pp. 212–223.
- Demsar J., Curk, T., Erjavec, A., Gorup, C., Hocevar, T., Milutinovic, M., Mozina, M., Polajnar, M., Toplak, M., Staric, A., Stajdohar, M., Umek, L., Zagar, L., Zbonitar, J., Zitnik, M., Zupan, B. (2013.) Orange: Data Mining Toolbox in Python. *Journal of Machine Learning Research*, 14, pp. 2349–2353.
- Devlin, J., Chang, M., Lee, K. & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186.

- Fillmore, C. & Atkins, S. (1992). Towards a frame-based organization of the lexicon: the semantics of RISK and its neighbors. In A. Lehrer and E. Kittay (eds.) *Frames, fields, and contrasts: New essays in semantic and lexical organization*, pp. 75–102.
- Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. In J. R. Firth (ed.): *Studies in linguistic analysis*. Oxford: Basil Blackwell, pp. 1–32.
- Geeraerts, D. (2001). The definitional practice of dictionaries and the cognitive semantic conception of polysemy. *Lexicographica*, 17, pp. 6–21.
- Groenen, P. J. F. & Borg, I. (2014). The Past, Present, and Future of Multidimensional Scaling. In J. Blasius & M. Greenacre (eds.) *Computer Science and Data Analysis Series: Visualization and Verbalization of Data*. CRC Press: Boca Raton, FL, USA; Taylor & Francis Group: Didcot Melton Park/London/Abingdon, UK, pp. 95–117.
- Hornby, A. S. (1948). *Idiomatic and Syntactic English Dictionary*. Institute for Research in Language Teaching. Tokyo: Kaitakusha.
- Ide, N. & Wilks, Y. (2007). Making sense about sense. In E. Agirre & P. Edmonds (eds.) *Word sense disambiguation*. Dordrecht: Springer, pp. 47–73.
- Jatowta, A., Tahmasebib, N. & Borinb, L. (2021). Computational approaches to lexical semantic change: Visualization systems and novel applications. *Computational approaches to semantic change*, 6(311).
- Kilgarriff, A. (2007). Googleology is Bad Science. *Computational Linguistics*, 33(1), pp. 147–151, <http://doi.org/10.1162/coli.2007.33.1.147>.
- 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(1), pp. 7–36, <http://doi.org/10.1007/s40607-014-0009-9>.
- Lakoff, G. (1987). *Women, fire, and dangerous things*. Chicago: University of Chicago Press.
- Leech, G. (1992). 100 million words of English: The British National Corpus (BNC). *Language research*, 28(1), pp. 1–13.
- Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *Italian Journal of Linguistics*, 20(1), pp. 1–31.
- Lewandowska-Tomaszczyk, B. (2007). Polysemy, prototypes, and radial categories. D. Geeraerts & H. Cuyckens (eds.) *The Oxford handbook of cognitive linguistics*. Oxford: Oxford University Press, pp. 139–169.
- Molina, C. (2008). Historical dictionary definitions revisited from a prototype theoretical standpoint. *Annual Review of Cognitive Linguistics* 6(1), pp. 1–22.
- Montes, M. & Heylen, K. (2022). Visualizing distributional semantics. In D. Tay & M. X. Pan (eds.) *Data Analytics in Cognitive Linguistics: Methods and Insights*. Berlin, Boston: De Gruyter Mouton, pp. 103–137.
- Ng, A. Y., Jordan, M. I. & Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems (NIPS)*, 14, pp. 849–856.
- OD: Oxford Learner’s Dictionaries, online version, accessed at:

- <https://www.oxfordlearnersdictionaries.com> (7 February 2023).
- Ostrovsky, R., Rabani, Y., Schulman, L. J. & Swamy, C. (2006). The Effectiveness of Lloyd-Type Methods for the k-Means Problem. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*. IEEE., pp. 165–174.
- Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. & Zettlemoyer, L. (2018). Deep Contextualized Word Representations. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 2227-2237.
- Radford, A, Narasimhan, K, Salimans, T. & Sutskever, I. (2018). Improving language understanding by generative pre-training. Available at https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf
- Rousseeuw, P. J. (1987). Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics*, 20, pp. 53–65.
- Rychlý, P. & Kilgarriff, A. (2007). An efficient algorithm for building a distributional thesaurus (and other Sketch Engine developments). In S. Ananiadou (ed.) *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*. Stroudsburg, PA: Association for Computational Linguistics, pp. 41–44, <http://doi.org/10.3115/1557769.1557783>.
- Tenenbaum, J. B., de Silva, V. & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500), pp. 2319–2323.
- van der Maaten, L. & Hinton, G. (2008). Visualizing Data Using t-SNE. *Journal of Machine Learning Research*, 1, pp. 1–48.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u. & Polosukhin, I. (2017). Attention is all you need. In: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan & R. Garnett (eds.) *Advances in Neural Information Processing Systems*, 30, pp. 5998–6008.
- Wiedemann, G., Remus, S., Chawla, A. & Biemann, C. (2019). Does BERT Make Any Sense? Interpretable Word Sense Disambiguation with Contextualized Embeddings. In *Proceedings of the 15th Conference on Natural Language Processing, KONVENS 2019, Erlangen, Germany, October 9-11, 2019*.

The *SERBOVERB* Language Resource and Its Multifunctionality

Saša Marjanović

University of Belgrade – Faculty of Philology,
Studentski trg 3, Belgrade 11 000, Serbia
sasa.marjanovic@fil.bg.ac.rs

Abstract

Serbian verb inflection is known for its complexity and unpredictability, posing a challenge for L2 Serbian speakers. Existing dictionaries are not well-suited to address the needs of L2 speakers. To overcome these challenges, the author presents *Serbo Verb*, an electronic resource and application that offers a dynamic approach to processing Serbian verb inflection. *Serbo Verb* includes a conjugation, dictionary, and gamification module, and offers paradigms for more than 34,000 verbs. The resource has been developed through a research project between the University of Toulouse Jean Jaurès (France) and the University of Belgrade, Faculty of Philology (Serbia). The author describes the structure and multifunctionality of *Serbo Verb*, highlighting its potential to provide a more accessible and user-friendly resource for L2 Serbian speakers seeking to resolve their communication problems and improve their language skills. By offering a multifunctional and comprehensive approach to Serbian verb inflection, *Serbo Verb* represents a significant step forward in electronic lexicography.

Keywords: *Serbo Verb*; verb inflection; Serbian; resource; dictionary

1. Introduction

Serbian verb inflection is quite complex. The paradigm of the average Serbian verb in the active voice includes hundreds of inflected forms (*cf.* Krstev, 1997; Tošović, 2012). The relationship between these inflected forms and their basic (lemma) form — which is conventionally used to represent the entire verb paradigm — is only predictable in a small number of inflectional classes (Jelaska, 2005; Marjanović, 2016b). Hence, mastering Serbian verb inflection can be quite challenging for average L2 Serbian speakers (*cf.* Krajišnik, 2011; Babić, 2021). The task is rendered even more difficult by the fact that some inflected forms are hard to match to their lemma form. The existing Serbian dictionaries, both mono- and bilingual, where L2 speakers might search for an inflection information, are not well tailored to the needs of average L2 speakers: they list verbs generally only in the lemma form, while the forms relevant for establishing the entire paradigm (*cf.* Marković, 2014) are very often lacking (Marjanović, 2016a). Although there are different ways to process Serbian verb inflection in printed dictionaries to satisfy all the prototypical communication-related and cognitive needs of target users (see Marjanović, 2016a & 2016b), we believe that the most appropriate and up-to-date solution is found in electronic lexicography, in the form of an electronic

conjugator.

This paper provides an overview and evaluation of currently available Serbian language conjugators in Section 2. Since these resources have some limitations and a new one is needed, Section 3 examines the existing inflection lexicons developed for Serbian language processing that could serve as a starting point for a new conjugator. Section 4 introduces *SerboVerb*, an innovative linguistic resource and its application, designed for Serbian L2 speakers. Developed as part of a research collaboration between the University of Toulouse – Jean Jaurès, France, and the University of Belgrade, Serbia, *SerboVerb* processes Serbian verb inflection dynamically, eliminating the limitations of static paper-based resources. The resource is accessible for free via a website and mobile app for Android and iOS. The paper emphasizes the potential of electronic lexicography to overcome traditional resource limitations and better meet the needs of L2 speakers. In addition, this paper details the structure of *SerboVerb*, highlighting its exhaustiveness, simplicity, and availability in processing verb inflection. Section 5 delves into its various functionalities. The paper concludes by outlining future plans in Section 6 and providing closing remarks in Section 7. Overall, the paper aims to showcase the multifunctionality of *SerboVerb* as a valuable language resource for learners of Serbian.

2. Previous Resources

SerboVerb is not the sole Serbian conjugator intended for human use, nor is it the first. To the best of our knowledge, several such electronic tools have been developed since the 1990s. Section 2.1 of this paper provides a chronological review of existing conjugators, while Section 2.2 offers a comparative evaluation of their strengths and weaknesses.

2.1 Existing Conjugators

The first Serbian conjugator was developed by a private company *Lexicom* (<https://lexicom.rs>) based in Belgrade (Serbia). However, there is no accompanying technical or scientific documentation related to this resource, so it is unclear how extensive the resource is and how many verbs it processes. The resource was freely searchable through the company’s website (*cf.* Marjanović, 2016a), but is no longer accessible. The verb paradigm was presented in a tabular format. It is worth noting that, while Serbian can be written using both the Cyrillic and Latin alphabets, the verb lemma search and display in this particular case were exclusively limited to the Latin script.

The *Grammatical Dictionary of Serbian* is the second conjugator, a linguistic resource created by the private company *Srbosoft* from Obrenovac (Serbia), which offers a range of Serbian language lexicographic resources, mostly retro-digitized from previous

paper editions. The resource is available on the company's website (<http://srpskijezik.com>). It has been available online since the end of 2017 and can be searched with an annual subscription. While there is no documentation for the resource either, it contains approximately 117,296 lemmas, which would include around 20,000 verb paradigms. The database can be searched by lemma using both Cyrillic and Latin alphabets, but the paradigm output is exclusively in the Cyrillic alphabet. The output is presented in plain text format, showing one tense at a time. To access other tenses, users need to click on the corresponding tab. However, it's important to note that the paradigm display presents inflected forms in a tabular format, numbered from 1 to 6. This means that the third person plural is listed as the sixth person. This sequencing might lead to potential confusion among users. The resource also provides accent markings for all inflected forms, allowing the user to obtain information about the pronunciation of each form.

The *Verbix* conjugator (<https://www.verbix.com>) is the third conjugator available for Serbian and provides access to conjugators for over one hundred languages. Users can search the verb database by entering any form of the verb without creating an account. However, unlike the previous two resources, *Verbix* can only be searched in Cyrillic script, and the output of the verb paradigm is also only in Cyrillic. The resource includes both simple and compound forms, but does not provide verb participles nor verb adverbs. There are typographical and encoding errors, as well as frequent instances of uncorrected inflected forms, which may compromise its reliability. However, the advantage of this conjugator is its more accessible paradigm layout. Additionally, 20 randomly selected verbs belonging to the same inflectional class are listed in the lemma form, prompting users to consider the similarities and differences between the paradigms of related verbs.

In addition to the three conjugators for the Serbian language, a Croatian conjugator called *Croatian Morphological Lexicon* (hereafter referred to as *CML*) (<http://hml.ffzg.hr>) has also been available since 2005 (Tadić & Fulgosi, 2003; Tadić, 2005; cf. Ljubešić et al., 2016). The relevance of Croatian conjugators to this paper lies in the fact that Croatian and Serbian are standardized micro-languages that are part of the same macro-language system. They share the same inflectional patterns and have a significant overlap in their lexical systems. Access to this *CML* conjugator requires an account approved by the author. However, it is not possible to reliably present the resource as access to it was not obtained at the time of writing. Based on literature (Tadić & Fulgosi, 2003), the resource contains about 36,000 lemmas, of which 7,735 are verbs, with two types of searches possible: by lemma and by any inflected form. The results of the searches are not hyperlinked, meaning that the user cannot access the complete paradigm of the selected lemma from an inflected form without conducting a new search. The first version of the conjugator listed the inflected forms alphabetically (Tadić, 2003), while the second version grouped them into traditionally organized paradigms (Tadić & Fulgosi, 2003). Additionally, in both versions, the inflected forms were tagged with a morphosyntactic code.

Finally, the *Croatian Language Portal* (hereafter referred to as *CLP*) (<https://hjp.znanje.hr>), a combination of retro-digitalized previously published Croatian monolingual dictionaries, includes a conjugator that provides users with the complete conjugation of 12,011 out of 15,699 Croatian verbs. However, the paradigms generated by the Portal's conjugator do not always match the data provided in the traditional morphological blocks of the dictionary entries, resulting in inconsistencies in the data presented to users. It is worth noting that the morphological block, which forms an essential component of the *CLP* dictionary entry, includes only the relevant inflection data necessary to establish the complete paradigm of a verb. In some instances, the conjugator offers only a single paradigm for a verb, disregarding the possibility of multiple potential paradigms as indicated by the data in the morphological block. Furthermore, the paradigms provided by the conjugator do not indicate any competing forms within corresponding tenses, further eroding the overall credibility of the *CLP* as a reliable language resource.

2.2 Comparative Evaluation

While the conjugators mentioned earlier can be useful for L2 Serbian speakers, each of them has its own limitations that electronic lexicographic resources should strive to overcome (*cf.* Tarp, 2008; Tarp, 2012; Lew, 2012; Grønvik & Smith Ore, 2013; Simonsen, 2014; Simonsen, 2015). These limitations concern the following eight points: *availability, access, content, scope, reliability, updating, searchability, and display.*

Regarding the first point, it can be concluded that all conjugators are available except for the first one (*Lexicom*), which, to the best of our knowledge, cannot be accessed for unknown reasons. Previously, *Lexicom* was open and available for free search without an account, much like *Verbix* and *CLP*. However, to search the *Srbosoft* conjugator and *CML*, users need to create an account, which is then verified by administrators before use. Unlike the others, *Srbosoft* conjugator access is not free and requires an annual subscription. Therefore, only two conjugators (*Verbix* and *CLP*) are currently available for completely open access.

In terms of content, Serbian and Croatian conjugators can be divided as follows: the first group (*Lexicom*, *Srbosoft*, and *Verbix*), exclusively includes verbs pertaining to the Ekavian variety. Conversely, the second group (*CML* and *CLP*), only contains Ijekavian verbs. These variations are a result of the different diatopic reflexes of the Old Slavonic sound *yat*. Consequently, a single verb that previously had the *yat* sound can now have two standard variants: the Ekavian — *e* — variant (e.g., *deliti*) and the Ijekavian — *ije* or *je* — variant (e.g., *dijeliti*), which are marked by areal distinctions. While it is expected for the Standard Croatian to include only Ijekavian forms, it is not justified for Serbian conjugators, as the Standard Serbian encompasses both Ekavian and Ijekavian variants. As a result, Serbian conjugators may not be helpful to users in need of inflection data on Ijekavian verb forms. Additionally, although

Serbian can be written in both Cyrillic and Latin scripts, all Serbian conjugators are available in only one script, with *Lexicom* and *Verbix* in Latin and *Srbosoft* in Cyrillic. Croatian conjugators use the Latin script, as it is the only script of the Croatian standard. This can pose a challenge for users who are not proficient in both alphabets.

When it comes to the scope of these resources, there are noticeable differences. Regarding the number of lemma, *Srbosoft* has the highest number of verbs (around 20,000), followed by *CLP* (12,011) and *CML* (7,735). Data on the number of verbs for *Lexicom* and *Verbix* is not available, but a random search of fifty verbs on *Verbix* reveals that even the most common verbs are missing. As only *CLP* and *Verbix* are freely accessible, it can be inferred that *CLP* has the most comprehensive coverage, but as it is a Croatian resource, it does not include verbs unique to the Serbian standard. Nonetheless, *CLP* is also the most inclusive concerning the number of inflected forms it encompasses, incorporating all simple and compound inflected forms. Conversely, *Verbix* excludes non-finite verb forms, while *Srbosoft* does not provide a paradigm for compound forms. There is no information on the data for *Lexicom* and *CML*, but as their resources were primarily created for *NLP* purposes, it is likely that these conjugators exclude compound forms.

Furthermore, *Verbix* contains many spelling, encoding, and material errors, while the other conjugators are reliable. However, this would not be a problem if the *Verbix* database were regularly updated and errors corrected. Unfortunately, this is probably not the case. It is also not clear whether any of these resources are regularly expanded with new verbs.

When it comes to searchability, all conjugators allow searching by entering the corresponding verb in the alphabet in which the verbs are stored in the database. Only the *Srbosoft* conjugator enables alphabet-insensitive search, which means that the user does not have to use Cyrillic script in the search field, but the search results will still be in Cyrillic. This can be convenient for users who do not have Cyrillic keyboards. *Verbix* and *CLP* offer the most flexibility regarding the linguistic form that can be entered in the search field. Users of these conjugators can enter any form of the verb in the search field, not just the lemma form, as is the case when searching with *Lexicom* and *Srbosoft* conjugators. On the other hand, *CML* is somewhere in between: this conjugator allows searching both by lemma and by inflected forms, but in the case of searching for an inflected form, the user is informed of its morphosyntactic description and directed to its lemma, which they need to search again if they require the entire verb paradigm.

It should be noted that while morphosyntactic identification of the searched form in *CML* is very useful, the MULTEXT-East format in which this description is encoded may be difficult for average users to decode. On the other hand, *Verbix* and *CLP* do not provide identification of the searched inflected form, but only display the full paradigm. It should also be pointed out that *Verbix* offers an autocomplete option

when entering the verb in the search field, which saves time needed for typing the rest of the word.

Regarding paradigm display, most conjugators list inflected forms in a row, one after the other, with each tense being named. The *Lexicom*, *Srbosoft*, *CML*, and *CLP* conjugators provide numbered forms for persons. With all, except *Srbosoft*, the numbering follows the traditional description of three singular and three plural persons, meaning that singular and plural are numbered separately. In the *Srbosoft* conjugator, however, all persons are numbered continuously, with plural forms being numbered 4–6. This may be confusing for users accustomed to the traditional didactic description of forms and verb paradigm presentation.

On the other hand, unlike the aforementioned conjugators, the forms in *Verbix* are not numbered at all, which reduces the paradigm's readability. *Srbosoft* compensates for this shortcoming with a better paradigm view: only one verb tense is shown at a time, while others are selected by clicking on a tab above the verb forms. In other conjugators, the paradigm view is uninterrupted, and the user must scroll down to find data not immediately visible.

Finally, it can be said that conjugators do not provide information on the meaning of verbs. This is expected given that conjugators only offer data on verb inflection. However, in the era of linked resources, it is regrettable that the presented conjugators cannot be used with other tools. The only exception is *CLP*, which provides, for each entry, a description from integrated Croatian monolingual dictionaries, but it should be noticed that its data does not always match the data provided by morphological blocks. Therefore, there is a need for a new conjugator that would address all the shortcomings mentioned in this evaluation.

3. Related Resources

The starting point in the development of a new conjugator can be the use of the outcomes of Serbian and Croatian language processing. The first results date back to the 1990s, but they were not available for long. Tadić's conjugator, mentioned above, is based on the author's *Croatian Morphological Lexicon*, which has been available through META-share since 2012. It consists of entries in triples format: first, the inflected form is listed, followed by the lemma, and finally, the morphosyntactic description encoded according to the MULTTEXT-East recommendations (Tadić, 2003; Tadić & Fulgosi, 2003). However, this lexicon is based on entries from a medium-size one-volume Croatian dictionary, which limits its coverage of Serbian standard vocabulary due to its focus on the most frequent Croatian words.

At the same time, a more extensive resource called the *Serbian Morphological Dictionaries (SrpMD)* was developed in the DELA format, relying on UNITEX systems (Krstev, 1997; 2008). This resource consists of several text files, including one

containing simple-word lemmas (DELAS), one for multi-word lemmas (DELAC), and two files for inflected forms of simple and complex words, respectively (DELAF and DELACF), generated automatically (see Stanković et al., 2018). The lemma lexicon includes entries in lemma form, their corresponding POS category, and the label of a finite-state transducer, which allows for the unambiguous production of all inflected forms and their morphosyntactic properties. The inflected form lexicons include entries in inflected form, their lemma, and their morphosyntactic properties. The lemma lexicon also often includes a series of markers that indicate features of the entry or indicate the type of feature and specify its value. The resource includes both Ekavian and Ijekavian word forms of the Serbian language and is encoded in ASCII to neutralize the difference between Cyrillic and Latin characters. The number of entries in *SrpMD* is constantly increasing, and according to the literature, its size has grown significantly over the years. The initial version of the simple word lexicon DELAS comprised 6,569 lemmas, with 1,884 of them being verbs (Krstev, 1997). Ten years later, the lexicon expanded to include 84,607 lemmas, of which 15,907 were verbs (Krstev, 2008). Presently, the lexicon contains a total of 205,003 lemmas, with 21,159 of them classified as verbs (Ružević, 2022: 32). Development of this resource was initially carried out through the *WS4LR* application interface, which was later upgraded and renamed to *LeXimir* (Stanković et al., 2018). Although this resource is indexed on Meta-Share, it is only available to a limited group of users upon request, and other researchers — unfortunately — cannot use or distribute it for either commercial or non-commercial purposes (see Ljubešić et al., 2016; Miletic, 2017 & 2018).

Another noteworthy lexicon for Serbian language processing is the accentual-morphological lexicon developed for the *AlphaNum* speech synthesizer (Sečujski & Delić, 2011). This lexicon contains entries with information about the lemma, encoded accentual configurations, and morphosyntactic properties. As of 2011, it contained around 100,000 lemmas, with ongoing additions facilitated by the *ARecnik* user interface. The interface enables manual entry of new words or automatic input from connected text files. Based on the entered data, the program generates inflected forms, morphosyntactic properties, and accentual configurations. However, this lexicon is not available for download.

According to published references (Tošović, 2012 & 2014), significant efforts were made between 2008 and 2015 to carry out morphological annotation of inflected and uninflected words in Serbian, Croatian, and Bosnian. The project aimed to establish the minimum number of rules required to generate the maximum and complete system of inflected forms using the *MorfoGenerator* system. The project covered 30,030 verbs out of 112,000 words, using 378 out of 822 rules to generate inflected forms for each verb. The resulting lexicon, *Gralis-MorfoGenerator*, was used for morphosyntactic annotation of texts in the multilingual *Gralis* corpus. Regrettably, the manually verified inflected form paradigms and the *MorfoGenerator* tool, which were intended

to be publicly available, are not currently accessible for search or download from any repository. Furthermore, the webpage cited in the papers is no longer reachable.

The first freely available morphological lexicon of the Serbian language, *Wikimorph-sr*, was derived by parsing the pages of the Serbo-Croatian version of *Wiktionary* based on a dump from October 2, 2015 (Miletic, 2017). The primary purpose of the lexicon was to enable multilayered annotation of Serbian texts in the multilingual parallel corpus *ParCoLab* (Miletic et al., 2017). It was supplemented with a list of entries extracted from a previous manually POS-tagged Serbian texts. The lexicon is in triples format, in accordance with MULTEXT-East recommendations, and contains 117,445 lemmas, including 11,299 verbs. Its coverage was tested on three contemporary Serbian novels, consisting of around 150,000 tokens, or 28,980 unique word forms, of which over 50% appear only once. The lexicon was found to cover 72% of word forms in these novels, which increases to around 80% for words that appear more than 10 times. The author notes that this result may be higher if a larger sample of texts were tested, but also suggests that the lexicon should be manually supplemented.

SrLex (Ljubešić et al., 2016) is another open-source lexicon that was created alongside the Croatian lexicon *hrLex*. These lexicons were built by expanding a publicly available lexicon from the *Apertium* machine translation system, which contained 10,183 lemmas assigned to 413 inflectional patterns. To identify missing words, the *hrWaC* and *srWaC* corpora were searched by frequency. A team of six linguists then used a graphical interface to review the missing Croatian words. They could either accept one of the automatically predicted lemma and inflectional pattern candidates or flag the word as not belonging to any of the predicted candidates. The process was repeated six times to improve coverage. The Serbian data was processed in just two rounds due to the significant lexical overlap with Croatian. As a result of the expansion, the Serbian lexicon (*srLex*) contains 105,358 lemmas, with an increase in the number of verb patterns from 167 in the original *Apertium* lexicon to 568 in *srLex*. The lexicon is freely available in both MULTEXT-East and Universal Dependencies formats.

In a study by Miletic (2018), the last two lexicons were mutually compared. It was shown that *Wikimorph-sr* contains only 21% of the entries found in *srLex*, while *srLex* contains 41% of the entries from *Wikimorph-sr*. Although the first finding is not surprising, the latter is less expected. Therefore, these resources were integrated into a single resource called *ParCoLex*, to assess whether their combined use could provide better coverage of *ParCoLab* text samples. The assessment used a sample of 16,389 tokens, corresponding to 6,301 unique inflected forms. The results showed that *srLex* provided better coverage than *Wikimorph-sr*, with 94% coverage of tokens compared to 73% for *Wikimorph-sr*, and 93% coverage of unique inflected forms compared to 63% for *Wikimorph-sr*. However, the newly integrated *ParCoLex* outperformed both resources, achieving 98% coverage for all tokens and 95% coverage for unique inflected forms. With its largest number of lemmas (157,886, including 14,562 verbs), *ParCoLex*

can serve as a valuable resource for researchers and developers working on Serbian language-related projects, such as *SerboVerb* (presented in next sections), since it offers a comprehensive and relatively reliable source of morphological information.

4. The *SerboVerb* Language Resource

In response to the limitations of existing conjugators for Serbian (and Croatian), as discussed in Section 2, a project was launched in 2017 at the University of Toulouse - Jean Jaurès (France) to develop a new, comprehensive, and multifunctional conjugator for Serbian, which was named *SerboVerb*. The project aimed to create an electronic resource that could be easily searched through a user-friendly application, taking into account the availability of an extensive morphological lexicon for non-commercial use (as discussed in Section 3).

The development of the *SerboVerb* application was funded by the Research Valorization Unit of the University of Toulouse – Jean Jaurès (France) and Toulouse Tech Transfert, a French company dedicated to promoting local research results through technology transfer. The development of the *SerboVerb* resource began in 2018 and has been ongoing since then. It is being carried out by an expert group consisting of linguists, lexicographers, and NLP researchers from the University Toulouse – Jean Jaurès and the University of Belgrade, Faculty of Philology (Serbia). This expert group had already established an intensive collaborative relationship in the field of NLP (*cf.* Miletic et al., 2017). External collaborators were also involved, including volunteers from both universities.

The entire resource is hosted on servers provided by Huma-Num, the French digital infrastructure supported by the CNRS (the French National Center for Scientific Research). It can be accessed for free via the website (<https://serboverb.com>), as well as through a mobile app available for Android and iOS operating systems, which can be downloaded from the Google Play Store and the App Store, respectively. The web application also serves as a resource management system. Figure 1 shows the homepage of the web application.

In order to enhance the overall functionality of the resource, a complex verb database, including their inflection paradigms and foreign languages equivalents, was implemented into the application, along with additional external educational materials. Consequently, the *SerboVerb* application now comprises three modules: a conjugation module, a dictionary module, and a gamification module, which will be presented in the following subsections.

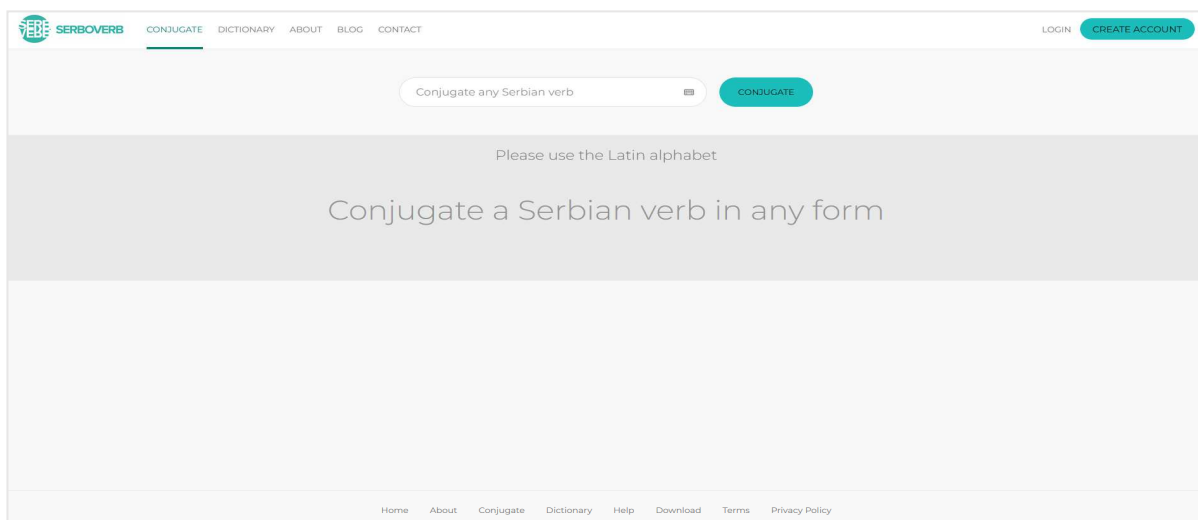


Figure 1: Homepage of the web-based *SerboVerb* application

4.1 Conjugation Module

The Conjugation module is a part of the application that enables users to search the verb database and display the inflectional paradigms of the searched verbs. The *SerboVerb* database was created based on the *ParCoLex* morphological lexicon (see Section 3). The lexicon was converted from a text file in MULTEXT-East format to XML using the P5 schema of the Text Encoding Initiative (TEI). Additionally, since the *ParCoLex* lexicon exclusively stored simple verb forms such as present, imperative, synthetic future, aorist, imperfect, active and passive participle forms, as well as present and past participle forms, active compound forms — including perfect tense, analytical future, future II, conditional, and pluperfect forms — were automatically generated. Passive compound forms were not included. When generating these compound forms, special attention was given to include forms that usually occur in context with a subject, as well as forms when the subject is omitted. However, it was noticed that some relatively common verbs were missing from this extensive inflection database of 14,562 verbs, so work on expanding the *SerboVerb* database began immediately.

The expansion work was carried out in four phases. In the first phase, all the verbs in their lemma form were extracted from the *ParCoLex* lexicon and arranged in tabular form. Then, all the verbs were extracted from the *Reverse Dictionary of the Serbian Language* (Nikolić, 2000) and compared with the list of verbs in the lexicon. Any verbs missing were retained, and merged with the first list. Since reflexive verbs in *ParCoLex* do not contain the reflexive particle *se*, while the *Reverse Dictionary* includes reflexive verbs in their lemma form, merging these two lists enabled the identification of existing reflexive verbs in the *SerboVerb* database. Additionally, a small number of verbs were manually added from other specialized paper lexicographic resources, primarily slang dictionaries, dictionaries of neologisms and anglicisms.

Finally, as all Serbian lexicographic sources are based on relatively outdated material, all missing verbs found in the *srWac* and *hrWac* corpora (Ljubešić & Klubička, 2014) were automatically extracted. The resulting list included 34,049 verbs. In the second phase, the verbs were annotated. Two annotators worked on this task, which lasted for six months. Firstly, based on the existing linguistic descriptions, 121 inflectional patterns were identified. Then, for each verb, a manual tag was assigned to indicate its membership to one of these patterns. In cases where a verb could also have a paradigm according to another inflectional pattern, an additional tag was assigned. However, the patterns did not include imperfect tense forms. For each verb in the database, a verb aspect was also indicated to mark the absence of certain verb forms (e.g. imperfect tense forms for perfective verbs, aorist tense forms for imperfective verbs). Each verb associated with either the Ekavian or Ijekavian variety was annotated with a distinct tag, and its corresponding counterpart in the opposite variety was added. Similarly, a subset of approximately 16,000 most frequent verbs and a subset of 1,844 core Serbian verbs (*cf.* Section 4.2) were specifically tagged. The lists of these most frequent and core verbs were published in the form of a paper-based conjugation dictionary for the needs of Serbian L2 speakers (Marjanović & Radosavljević, 2019). However, the entire *SerboVerb* database has not been made available for distribution.

In the third phase, rules for generating verb paradigms belonging to the most frequent and productive inflectional patterns were developed, as the inflectional patterns were designed to allow for the creation of rules for unambiguous generation of the complete verb paradigm. Simple and compound inflected forms were generated for all verbs that follow productive inflectional classes, which were missing from the database generated based on the *ParCoLex* lexicon. The imperfect tense forms were generated using a separate set of rules. The newly generated forms were added to the *SerboVerb* database at the end of 2018. The source element in the XML structure of the *SerboVerb* database provided clear indication of the *ParCoLex* paradigms and the newly added verbs, as well as their generated inflected forms.

During the fourth and final phase, the manual verification of newly generated inflected forms began in the spring of 2019. The verb paradigms formed on the basis of data from the *ParCoLex* lexicon were immediately published and have since been accessible to end-users. Initially, these forms were not subject to verification, as the creation of the *srLex* resource, which formed the basis of *ParCoLex*, involved linguists who verified the verb lemmas and their predicted paradigms (*cf.* Ljubešić et al., 2016; see Section 3). However, within the *SerboVerb* application, these verbs are internally labelled as unchecked. This label does not imply that the paradigms of these verbs are entirely accurate, nor does it mean that they will remain unchecked. The decision was made to prioritize the verification of the newly generated forms to speed up the process of verifying the entire database. As a result, the verification of the paradigms of these verbs will be conducted after the verification of the newly generated forms. Additionally, special attention is given to verifying the imperfect tense and passive participle forms of these verbs, as the imperfect tense forms of some verbs were not

generated simultaneously with the other inflected forms. Furthermore, the transitivity of some verbs was not marked in the manually annotated database, necessitating thorough verification of the resulting paradigms.

The accuracy verification of the forms is carried out in rounds, which are organized once a year. Each round covers 4,000 verbs and is conducted in two stages, with each stage lasting four months. In the first stage, a group of 10 trained and experienced native speakers of Serbian receive a batch of verbs and, following detailed instructions provided by the *SerboVerb* team, verify, correct, and supplement their paradigms. If there are no errors in the generated verb paradigm, the collaborator marks the verb with an appropriate flag. If a collaborator encounters a problem or has a doubt about a particular inflected form, they flag it for further review. In the second stage, the *SerboVerb* team coordinators provide additional verification. They publish verified verbs that are ready for publication and simultaneously review, correct, and supplement verb paradigms for which collaborators had doubts. At the time of writing this paper, 20,158 verbs have been reviewed. The remaining verbs will be reviewed in the following rounds.

4.2 Dictionary Module

The Dictionary module is a component of the application used to search and display the multilingual dictionary database of the *SerboVerb* language resource. The database is also structured in XML format according to TEI Guidelines, since it is merged with the *SerboVerb* inflection database. It can be searched in the same way as the conjugation module database (see Section 4.1). In the dictionary module, users can enter a verb lemma or any inflected form of the verb, and receive a bilingual dictionary description of the desired Serbian verb in one of the 36 available languages (both European and non-European).

The dictionary description contains one or more senses introduced by a gloss, marked with one or more labels, followed by one or more equivalents, each of which may also contain one or more labels, and finally, one or more translated examples. Therefore, this is a dictionary description in which Serbian is the source language, and other languages are the target languages (TL). Users can choose the TL they need for the first dictionary look-up, and that language will remain as the default language for subsequent searches in the dictionary module.

The development of the multilingual database started in autumn 2022. During the first phase, basic equivalents were added for a list of 1,844 core Serbian verbs (previously mentioned in Section 4.1), extracted from the annotated *SerboVerb* database. These verbs are representative enough for most L2 speakers up to level B2 (Upper Intermediate level) according to the Common European Framework of Reference for Languages. The selection criteria for these verbs are not discussed in this paper. Currently, the entries for core verbs have basic equivalents in Albanian,

English, French, German, Portuguese, Russian, Spanish, and Ukrainian. However, equivalents for Czech, Danish, Italian, Norwegian, Polish, Slovak, Swedish, and Turkish are still being added. Insertion of equivalents in Bulgarian, Greek, Hungarian, Macedonian, Romanian, and Slovene started in April 2023. Equivalents for other languages such as Arabic, Chinese, Dutch, Estonian, Farsi, Finnish, Hebrew, Japanese, Korean, Latvian, Lithuanian, Romani, Rusyn, and Swahili are being prepared for autumn 2023. The insertion of equivalents is carried out by a team of collaborators who possess a minimum proficiency level of C1 (Advanced level) in the respective languages. Each group comprises one to four members, and their work lasts for up to four months. Once the equivalents have been entered for all languages, the coordinators of the *SerboVerb* team plan to conduct a manual cross-check of all entries to ensure that the dictionary module is consistent across all languages.

4.3 Gamification Module

The Gamification module is designed to provide an interactive way for L2 Serbian speakers to learn, practice, and improve their verb inflection skills. Development of the module began in autumn 2022 and is currently ongoing. The initial content was created by the *SerboVerb* team, and external collaborators with expertise in teaching Serbian as an L2 or heritage language have been engaged to prepare additional education material. This material is expected to be added to the module in the near future, further enhancing its value as a learning tool.

The educational material in the gamification module is presented as a series of learning games, with various types available (see Mihaljević & Hudeček, 2022), such as quizzes, drag-and-drop exercises, fill-in-the-blanks, find-the-match, puzzles, crosswords, memory games, and hangman games. External collaborators may also contribute unique games. All games contain at least two gamification elements, such as levels, scoring, leaderboard, and time limit. The educational material is classified according to the required language competencies in Serbian as an L2 needed to solve them and is marked accordingly. Users are provided with a score of their performance to boost motivation. Based on their performance, they are ranked against other users who have completed the same game. Additionally, some games have a time limit.

All of the educational material is prepared using open-access gamification platforms that are freely accessible. As a result, this module is the least consistent in terms of content and presentation. However, this is not a problem, as the involvement of different and numerous collaborators ensures a variety of approaches and a wider reach in the use of *SerboVerb* app and its resources.

5. Multifunctionality of *SerboVerb*

The differences among the Serbian conjugators discussed in Section 2 can significantly influence the user's experience. Hence, it was crucial to take these aspects into account when creating the *SerboVerb* application as they can greatly impact the efficiency and effectiveness of the end product. Moreover, comprehending the advantages and limitations of each conjugator could help the *SerboVerb* team develop an application that cater for user's specific requirements better.

As previously demonstrated in the literature (Tarp, 2008), according to Function Theory, users for whom a particular language is a foreign language (in this case, Serbian) may have a primary or secondary need for inflection information, which can be satisfied by seeking help from a dictionary in all extralexigraphic situations, including communicative (receptive and productive) and cognitive ones. The following subsections illustrate how the *SerboVerb* application provides data based on which appropriate information can be derived in all three mentioned situations.

5.1 Receptive Functions

If an L2 Serbian speaker is not familiar with or unable to recognize a certain inflected form of a verb, they can search for it in the *SerboVerb* web-based or mobile application without creating an account and completely free of charge. Within the Conjugation module, the user can enter the unrecognized form in the search field (see Figure 2a). The searched form can be in its lemma or non-lemma form. Through the autocomplete feature (see Figure 2b), the application will suggest one or multiple possible results, along with a brief morphosyntactic identification of the form. This feature assists the user in identifying the tense in which the searched verb form is located within the written or spoken extralexigraphic context where they first encountered the verb, and provides the corresponding result. By clicking on the appropriate form, the user can access the paradigm of the selected result (see Figure 2c).

The result page consists of two components: a shaded identification block (see Figure 2c) and a brighter paradigm block (see Figure 2c & 3a). The first block provides the user with more reception-relevant data: firstly, it identifies the searched inflected form by placing it in a specific tense from the verb paradigm; secondly, it indicates whether the verb is limited to Ekavian or Ijekavian areas or can be used in all varieties of Serbian standard language. If the usage is limited to a specific area, a cross-reference to the counterpart form is provided to the user. Then, the aspectual value of the verb is presented to the user. Finally, the identification block also provides basic equivalents for 1844 core Serbian verbs, which provide the lexical meaning of the searched verb and facilitate its reception. If the user needs a language that is not provided by default, they can select the appropriate language from the drop-down menu list (Figure 3b). If the user requires additional information (e.g., on the usage of

the verb in context) to further understand its lexical meaning, they can click on the icon that opens the Dictionary module, which offers more data from the dictionary database. The second element in this result page provides the complete paradigm of the searched verb. By scrolling down, the user can locate the searched form within the full inflectional paradigm.

5.2 Productive Functions

In situations where an L2 Serbian speaker is not familiar with the inflectional paradigm of a certain verb, or is unsure about it, but needs it for text production purposes, they can search for the verb's inflectional paradigm in the Conjugation module. As in receptive situations (*cf.* Section 5.1), the search can be performed based on the form of the verb that the user first recalls. This can be either the lemma form or any inflected form. The search result page displays a shaded identification block and a brighter paradigm block. Unlike in receptive needs, where the identification block carries more informative weight, in productive needs, the primary importance of the data is in the paradigm block. In this block (see Figure 3a), the user scrolls down to search for the verb tenses that they require in the production situation. The verb tenses are arranged so that the most frequent ones in contemporary Serbian, and the ones that are first learned in Serbian L2 courses (present, imperative, perfect, and future tense), come first. Regarding the data in the paradigm, it should be noted that the user can also obtain information about all the compound tenses, as well as the paradigm of reflexive verbs, where forms have different word order depending on whether the subject is present or not. Furthermore, the graphical interface is designed to enable the user to quickly scroll through the paradigm, both up and down and left and right (especially when displaying forms for the appropriate gender). Moreover, in the identification block, the user can check whether the searched verb is used in the appropriate Ekavian or Ijekavian area and what aspectual value it carries. Then, if they need information about the use of the verb in context, they can switch to the Dictionary module (see Figure 3c).

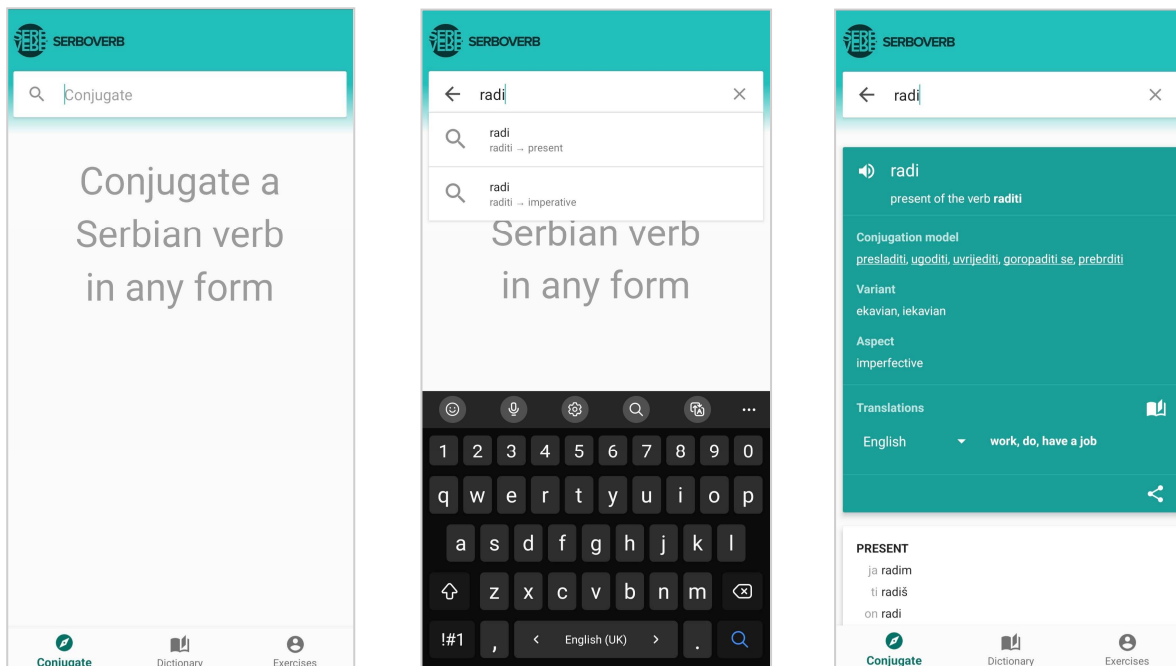


Figure 2: The conjugation module in the Android app version:
 a) the homepage, b) a search action, c) the conjugation result page

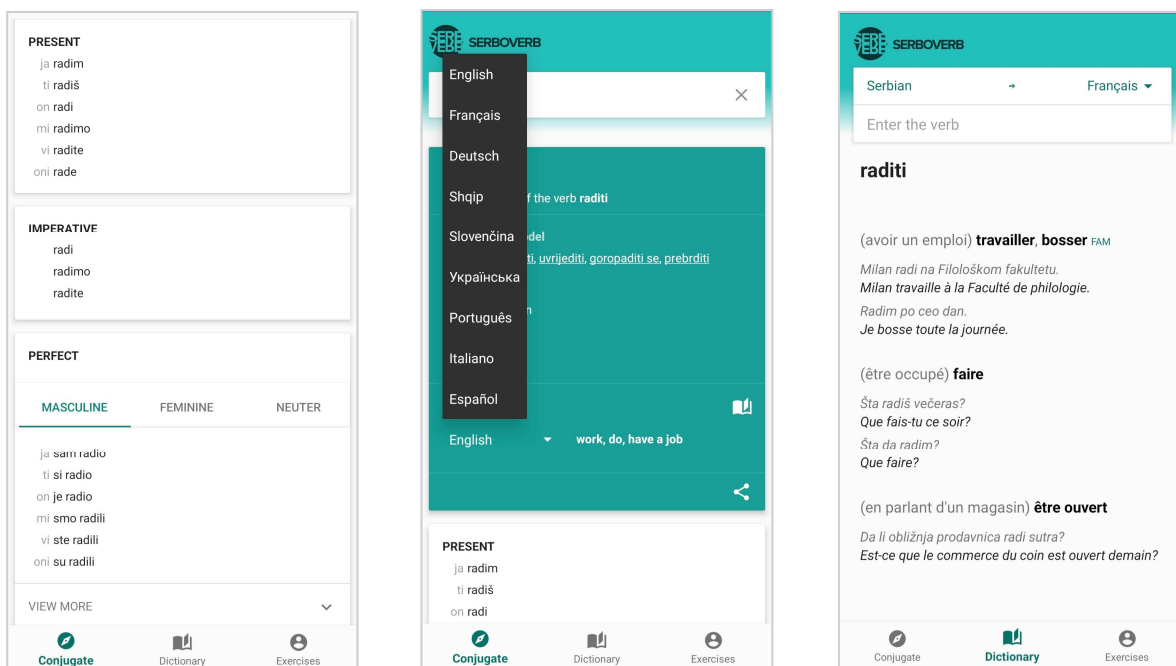


Figure 3: The conjugation and the dictionary modules in the Android app version:
 a) the conjugation result page, b) switching a language, c) dictionary view

5.3 Cognitive Functions

L2 Serbian speakers can use the *SerboVerb* application not only when they need to solve a communication-based problem but also in cognitive situations, where they want to independently confirm or acquire knowledge about the paradigms or inflected forms of certain verbs they are uncertain about. In such cases, the *SerboVerb* resource in the application can be searched in the same way as described in previous subsections (see Sections 5.1 & 5.2). An additional feature that is not relevant to the previous two functions is the cross-reference to five randomly selected verbs from the *SerboVerb* inflection database that belong to the same inflectional pattern. By comparing the paradigms of similar verbs, users can acquire and expand their knowledge of the conjugation properties of individual inflectional classes. Additionally, cognitive functions are satisfied through the use of a gamification module, described in Section 4.3.

6. Future Development

As stated in the previous sections, *SerboVerb* is an application and a language resource that are still in development. Further development is based on user needs, gathered through log file analysis and direct communication with users. So far, several needs have been identified for which both short-term and long-term plans have been made.

Despite the fact that the search field indicates that the verb database should be searched using Latin characters, it has been noticed that users occasionally search for verbs using Cyrillic alphabet. As a result, the short-term plan involves introducing an algorithm in the search field that instantly transliterates Cyrillic letters into Latin characters, enabling users to input forms in their preferred alphabet. Furthermore, the short-term plan entails conducting further verification of the inflection database to ensure that all verbs become available to users in the near future and that the data is as reliable as possible.

In the long term, the plan is to expand the multilingual database by adding examples for core verbs in their basic meanings, expressed in lexically simple and concise syntactic patterns, and translated into available languages. The gamification module will also receive regular updates with new content to cater for different types of users. Finally, a pronunciation module will be developed that enables users to hear the correct pronunciation of the searched form and other forms in the verb paradigm.

7. Conclusion

This paper introduced an innovative language resource called *SerboVerb* and its accompanying application, which enable L2 Serbian speakers to effectively and dynamically meet all their needs related to verb inflection in various communicative and cognitive situations. As demonstrated, the application was designed to be freely

and openly accessible, with a comprehensive database of verbs and their inflected forms, continuously updated and expanded, with flexible search capabilities, and an effective and highly readable graphical interface that presents a large amount of data in a clear manner. Additionally, the main inflection database is linked with other resources, such as dictionaries and educational content, further enhancing its utility. By relying on this more trustworthy tool than on previous conjugators, L2 Serbian speakers now have access to a valuable resource that includes Serbian, a language often considered low-resourced, thus enriching the electronic lexicographic landscape.

8. Acknowledgements

The paper was produced as part of the Scientific Research Program under a contract no. 200167, signed by the University of Belgrade, Faculty of Philology, and the Ministry of Science, Technological Development and Innovations of the Republic of Serbia.

9. References

- Babić, B. (2021). *Unutarjezičke greške u nastavi srpskog jezika kao stranog*. Novi Sad: Filozofski fakultet.
- Grønvik, O. & Smith Ore, Ch-E. (2013). What should the electronic dictionary do for you – and how? In I. Kosem et al. (eds.) *2013. Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia*. Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut, pp. 243–260.
- Jelaska, Z. (2005). Glagolske vrste. In Z. Jelaska et. al. (eds.) *Hrvatski kao drugi i strani jezik*. Zagreb: Hrvatska sveučilišna naklada, pp. 170–185.
- Krajišnik, V. (2011). Rječnik u nastavi srpskog kao stranog jezika. *Analiz Filološkog fakulteta*, 23(2), pp. 245–258.
- Krstev, C. (1997). *Jedan prilaz informatičkom modeliranju teksta i algoritmi njegove transformacije*. PhD thesis. Beograd: Matematički fakultet.
- Krstev, C. (2008). *Processing of Serbian: Automata, Texts and Electronic Dictionaries*. Belgrade: Faculty of Philology.
- Lew, R. (2012). How can we make electronic dictionaries more effective? In S. Granger & M. Paquot (eds.) *Electronic lexicography*. Oxford: University Press, pp. 343–362.
- Ljubešić, N. & Klubička, F. (2014). {bs,hr,sr}WaC – web corpora of Bosnian, Croatian and Serbian. In *Proceedings of the 9th Web as Corpus Workshop (WaC 9)*. Gothenburg: Association for Computational Linguistics, pp. 29–35.
- Ljubešić, N., Klubička, F., Agić, Ž. & Jazbec, I.-P. (2016). New Inflectional Lexicons and Training Corpora for Improved Morphosyntactic Annotation of Croatian and Serbian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož: European Language Resources

- Association (ELRA), pp. 4264–4270.
- Marjanović, S. & Radosavljević, N. (2019). *Srpski glagoli: Konjugacijski rečnik glagola srpskoga jezika*. Beograd: Klett.
- Marjanović, S. (2016a). Glagolska fleksija u rečnicima. O recepcijskim i produkcijskim potrebama stranih korisnika. In V. Krajišnik et al. (eds.) *Srpski kao strani jezik u teoriji i praksi III*. Beograd: Filološki fakultet, pp. 261–277.
- Marjanović, S. (2016b). Glagolska fleksija u dvojezičnom rečniku sa srpskim kao ciljnim jezikom. *Zbornik Matice srpske za filologiju i lingvistiku*, 59(2), pp. 109–128.
- Marković, A. (2014). Gramatika u srpskim rečnicima. In R. Dragičević (ed.) *Savremena srpska leksikografija u teoriji i praksi*. Beograd: Filološki fakultet, pp. 69–91.
- Mihaljević, J. & Hudeček, L. (2022). Model for developing educational games based on data from dictionary structure. *Studia lexicographica*, 16(30), pp. 111–133.
- Miletic, A. (2017). Building a morphosyntactic lexicon for Serbian using Wiktionary. In *6e édition des Journées d'étude toulousaines : Les interfaces en Sciences du Langage. Actes des Journées d'études toulousaines 18 et 19 mai 2017*. Toulouse: Université Toulouse Jean Jaurès, pp. 30–34.
- Miletic, A. (2018). *Un treebank pour le serbe : constitution et exploitations*. PhD thesis. Toulouse: Université Toulouse – Jean Jaurès.
- Miletic, A., Stosic, D. & Marjanović, S. (2017). ParCoLab: A Parallel Corpus for Serbian, French and English. In K. Ekštejn & V. Matoušek (eds.) *Text, Speech, and Dialogue. 20th International Conference, TSD 2017, Prague, Czech Republic, August 27–31, 2017 Proceedings. Lecture Notes in Artificial Intelligence* 10415. Cham: Springer, pp. 156–164.
- Nikolić, M. (2000). *Obratni rečnik srpskoga jezika*. Novi Sad – Beograd: Matica srpska, Institut za srpski jezik SANU, Palčić.
- Rujević, B. (2022). *Rečnici u digitalnom dobu – informatička podrška za srpski jezik*. PhD thesis. Beograd: Filološki fakultet.
- Sečujski, M. & Delić, V. (2011). *Automatska konverzija tekstualnih informacija u govor*. Beograd: Vojnotehnički institut.
- Simonsen, H. K. (2014). Mobile Lexicography: A Survey of the Mobile User Situation. In A. Abel, C. Vettori & N. Ralli (eds.) *Proceedings of the XVI EURALEX International Congress: The User in Focus. 15-19 July 2014, Bolzano/Bozen*, pp. 249–261.
- Simonsen, H. K. (2015). Mobile Lexicography: Let's Do it Right This Time! In I. Kosem et al. (eds.) *2015. 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. 84–104.
- Stanković, R., Krstev, C., Lazić, B. & Škorić, M. (2018). Electronic Dictionaries – from File System to *lemon* Based Lexical Database. In *Proceedings of the 11th*

- International Conference on Language Resources and Evaluation (LREC 2018) - W23 6th Workshop on Linked Data in Linguistics : Towards Linguistic Data Science (LDL-2018)*, Miyazaki, Japan, May 7-12, 2018.
- Tadić, M. & Fulgosi, S. (2003). Building the Croatian Morphological Lexicon. In *Proceedings of the 2003 EAACL Workshop on Morphological Processing of Slavic Languages*. Budapest: Association for Computational Linguistics, pp. 41–45.
- Tadić, M. (2003). *Jezične tehnologije i hrvatski jezik*. Zagreb: Ex Libris.
- Tarp, S (2008). *Lexicography in the Borderland between Knowledge and the Non-Knowledge*. Tübingen: Niemeyer.
- Tarp, S. (2012). Theoretical challenges in the transition from lexicographical p-works to e-tools. In S. Granger & M. Paquot (eds.) *Electronic lexicography*. Oxford: University Press, pp. 107–118.
- Tošović, B. (2012). Morfogeneratorska tipologija glagola (na korpusu Rečnika srpskog jezika). *Slavistika*, 16, pp. 135–142.
- Tošović, B. (2014). Automatsko kodiranje pomoću Morfogeneratora. *Slavistika*, 18, pp. 207–214.

Operationalising and Representing Conceptual Variation for a Corpus-driven Encyclopaedia

Santiago Chambó^{1, 2}, Pilar León-Araúz²

¹ Doctoral School in Humanities Social Sciences and Law,
University of Granada, Avenida Madrid 13, 18071, Granada, Spain

² Department of Translation and Interpreting,
University of Granada, Buensuceso 11, 18071 Granada, Spain

E-mail: santiagochambo@ugr.es, pleon@ugr.es

Abstract

Detecting conceptual variation among humanitarian actors in textual sources is one of the challenging objectives of the Humanitarian Encyclopedia. This article proposes a method to operationalise and represent conceptual variation. Conceptual variation is a phenomenon whereby individuals and organisations show different understandings of the intensions and extensions of concepts. Despite the existence of a shared vocabulary, humanitarian concepts are presupposed to be affected by conceptual variation due to the recent professionalisation and diversity of the sector. In a pilot study, the four humanitarian principles (i.e., HUMANITY, IMPARTIALITY, NEUTRALITY, and INDEPENDENCE) were analysed with a hybrid methodology that combines Frame-based Terminology and Content Analysis. Definitions were extracted from a corpus of humanitarian documents, coded inductively to unveil definitional elements, and consolidated with corpus metadata to associate them with specific types of humanitarian organisations. Finally, a conceptual profile for each concept was represented by plotting its definitional elements and the number of occurrences on radar charts. Occurrences were subsequently disaggregated by organisation type to reveal differences between humanitarian actors. Several cases of conceptual variation were preliminarily detected. Minor cases of semantic overlap were also identified. Our preliminary results suggest that this method can detect and represent conceptual variation satisfactorily.

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

1. Introduction

The humanitarian domain is a multidisciplinary and recently professionalised field that comprises numerous specialised organisations ran by people with different professional and cultural backgrounds (Eberwein and Saurugger 2013). This diversity plays a role in how humanitarians conceptualise their domain (Stroup 2012; Sezgin and Dijkzeul 2015), giving rise to highly unstable concepts such as RESILIENCE (Béné et al. 2012), EVIDENCE (Knox Clarke and Ramalingan 2014) and LOCAL ORGANISATION (Khan and Kontinen 2022). In this context, the Humanitarian Encyclopedia (HE; humanitarianencyclopedia.org) has entered the stage as a descriptive reference work of the humanitarian domain. The objective of the HE is to describe humanitarian concepts

by combining expert knowledge and corpus-driven conceptual analyses provided by a team of linguists. This is meant to minimise biases and content gaps that can arise due to diverse backgrounds of entry authors (Humanitarian Encyclopedia 2021b).

Given this context of diversity, the HE's mission statement is to foster a shared understanding of humanitarian notions by describing 129 key humanitarian concepts. The HE requested its team of linguists to conduct conceptual analyses on the four humanitarian principles (i.e., HUMANITY, IMPARTIALITY, NEUTRALITY, and INDEPENDENCE) with the objective of identifying whether humanitarian organisations show different understandings (i.e., conceptual variation). These four principles are key domain concepts that, according to Hansen (2008, 125), are conceptualised solidly and are well understood in non-Western cultures. However, other works like Abu-Sada (2012) claim that both humanitarian organisations and affected populations do not have a shared understanding of these notions, which leads to disappointed expectations and miscommunication around the role of humanitarian practitioners.

This article describes a pilot study conducted by a team of linguists at the HE. The objectives of the study are (1) to determine the meaning of the four humanitarian principles by elucidating conceptual characteristics from lexical data obtained from a corpus of humanitarian documents, (2) to associate conceptual characteristics to humanitarian actors, and (3) to compare the distribution of conceptual characteristics across humanitarian actors by means of data visualisations to detect whether they display divergent understandings. The rest of this article is structured as follows. Section 2 briefly examines the phenomenon of conceptual variation and methodological considerations to approach its study. Section 3 details the materials and methods used in this study. Section 4 presents the results of each conceptual analysis and examines detected cases of conceptual variation. Lastly, Section 5 draws a conclusion and delineates future research lines.

2. Operationalising Conceptual Variation

Conceptual variation refers to the diversity of understandings among people about the intensions and extensions of concepts. There is evidence of variation in how individuals conceptualise notions (Hampton 2020), resulting in fuzzy, highly diverse, and multidimensional conceptualisations (León-Araúz 2017, 215). Therefore, it is reasonable to assume that human collectives like humanitarian organisations may also be subject to conceptual variation.

Studying conceptual variation from textual sources requires a method for conceptual analysis driven by lexical data to determine the meaning of concepts. Multiple methods for conceptual analysis have been devised in several disciplines. With methodological differences, they are similar in that they aim to elucidate conceptual characteristics by deriving them from textual evidence. Concretely, terminological methods for conceptual analysis are recognised as the most sophisticated thanks to their metaconceptual and

detailed description tools (Nuopponen 2010). These are used to build concept systems with universal top-level categories (Gil-Berrozpe, León-Araúz, and Faber 2019) and catalogues of semantic relations (Nuopponen 2022).

In 2020, the HE adopted a Frame-based Terminology (Faber 2015; 2022) approach to conceptual analysis through the systematic extraction and curation of lexical data from corpora. This is done by querying corpora with textual markers, such as knowledge patterns (KPs), that help linguists extract definitions and other knowledge-rich contexts (KRCs). KRCs contain useful data to describe the meaning of concepts (Meyer 2001; Marshman 2022) and are therefore used to substantiate conceptual characteristics.

FBT is well-equipped to elucidate characteristics by focusing on compact single-word and multi-word expressions (Faber 2022, 366), which constitute a defining feature of well-established specialised languages, especially those describing the physical world, such as the medical and environmental domains. These nominal expressions are used to designate the concept nodes in conceptual systems. However, useful KRCs may also “include entire clauses that are difficult to merge into a single concept” (León-Araúz and Reimerink 2019, 128). This applies extensively to humanitarian KRCs, which contain a high level of lexical heterogeneity, making it difficult to elucidate characteristics. When faced by this type of KRC, FBT is not as well-equipped.

When conceptual characteristics are designated by a diverse range of expressions, it is necessary to classify them into manageable categories. Conceptual Content Analysis (Bengtsson 2016; Lindgren 2016) provides inductive categorisation techniques of lexical data to derive themes, categories and detect the presence of concepts in corpora (Kyngäs 2020, 14) from open data observation. This method enables conceptual analysts to generate compact designations for conceptual characteristics by coding sections of text in KRCs and subsuming them into quantifiable categories. These categories can be then linked to the corpus metadata of their KRCs and modelled into datasets.

Combining corpus linguistics and quantitative analysis of conceptual characteristics is one of the main challenges of analysing complex social concepts (Kantner and Overbeck 2020, 186) like the four humanitarian principles. In this study, we combine (1) targeted extraction of KRCs through corpus linguistics techniques provided by FBT with (2) coding and inductive categorisation techniques provided by Content Analysis. By doing so, we generated datasets that link conceptual characteristics to the metadata of documents published by different types of humanitarian organisations. If KRCs are associated to corpus metadata, conceptual characteristics inherit their attributes, which generates useful data to describe and compare them. This enables the disaggregation of characteristics by organisation type, thus operationalising conceptual variation.

To interpret the data, we once attempted to detect cases of conceptual variation by producing data visualisations with a software package designed for business purposes in Chambó and León-Araúz (2021), but to no avail. We concluded that a more powerful

and flexible solution was necessary to disaggregate conceptual characteristics, represent their quantitative dimension and enable comparison of results. Section 3 describes the materials and methods used to detect conceptual variation from textual sources.

3. Materials and Methods

This section examines the materials and the methods used to collect, model, and visualise lexical data for each humanitarian principle and establish their conceptual profiles by generating radar charts.

3.1 Materials

3.1.1 The HE Corpus

In 2019, the HE compiled a corpus of 4,824 humanitarian documents, published between 2004 and 2019. The HE Corpus (Humanitarian Encyclopedia 2021a) amounts to a total of 84,926,707 tokens and 71,201,157 words. Corpus metadata include a taxonomy of organisation types and subtypes, region of publication, and year of publication, among others. In this study, data disaggregation was limited to organisation types. Table 1 details the codes for each organisation type, their description, and the number of documents in the corpus.

Code	Description	Documents
NGO	Non-governmental organisation, e.g., ACTED	2,128
NGO_Fed	Federations of NGOs, e.g., ActionAid	878
IGO	Inter-governmental organisations, e.g., AESAN	453
RC	National Red Cross and Red Crescent Societies and international organisations, e.g., ICRC.	375
Net	Sector-wide networks of humanitarian agents, e.g., ALNAP	339
Found	Foundations, e.g., the Breteau Foundation	240
State	Governments and state agencies, e.g., USAID	157

RE	Religious entities, e.g., Australian Lutheran World Service	146
C/B	Corporate, business and think-tank organisations, e.g., the Overseas Development Institute (ODI)	72
Project	Specific time-bound projects, e.g., The Sphere Project	22

Table 1: Organisation type metadata in the corpus

3.1.2 Sketch Engine

The HE Corpus was uploaded onto Sketch Engine (Kilgarriff et al. 2014), a browser-based corpus management and query software. The HE Corpus was queried systematically for definitions of the four humanitarian principles using the Concordance tool, which queries the corpus with Corpus Query Language (CQL) expressions and displays matches in a key word in context (KWIC) concordance view. Queries and results can be further processed with additional functionalities. In this study, we used the Lemma context filter functionality to limit the extraction of concordances with a selection of definitional KPs within a window of -5, 5 tokens.

3.1.3 Definitional KPs

Definitions are considered the starting point in semantic analysis (Sierra et al. 2010, 76) as well as high-density units of analysis (León-Araúz and Reimerink 2019). For this reason, the conceptual profiles of each humanitarian principle were built based on data obtained from definitions. Definitions were extracted combining CQL expressions and a set of definitional KPs. These include definitional verbal patterns (e.g., ‘defined as’, ‘understood as’, ‘means’) as in Sierra et al. (2008; 2010), and paralinguistic patterns such as colons and round brackets as in Dorantes et al. (2017). In a manner similar to Kovář, Močiariková, and Rychlý (2016), definitions were extracted by designing simple CQL queries and creating macros by including a catalogue of definitional KPs in the Lemma context filter functionality (Table 2).

Strategy	KPs
Is_a CQL query	be a, be not, be one, be the
KPs with Lemma filter	call, categorise, concept, conceptualise, conceptualize, consider, deem, define, definition, entail, idea, imply, involve, mean, meaning, notion, refer, require, requirement, term, understand, word, : (colon), ((opening round bracket).

Table 2: Definitional KPs

3.1.4 Taguette

The extracted definitions from Sketch Engine were then imported into Taguette (Rampin and Rampin 2021), a free and open-source qualitative data analysis software, which enables the user to annotate documents by highlighting sections of text and assigning tags inductively or deductively. Each tag is intended to represent a characteristic of each concept by subsuming semantically similar sections of texts from definitions, with an approach similar to Lindgren (2016). Annotations can be subsequently exported and combined with corpus metadata obtained from Sketch Engine to create a dataset containing conceptual characteristics, their originating definitions, and the organisation type of the document where they were found.

3.1.5 The fmsb Package for R

Once each conceptual characteristic is quantified and associated to organisation types, datasets were processed with the R programming language and the fmsb package (Nakazawa 2023) to produce data visualisations. The fmsb package contains a function to generate radar charts. These can be used to visualise conceptual characteristics as vertices of a polygon, which form the axes of a chart. The number of occurrences for each characteristic are also represented as vertices of another polygon that is placed over the axes of the chart. Multiple radar charts were generated for the total number of occurrences and each organisation type. These radar charts are presented in Section 4.

3.2 Methods

3.2.1 Extraction of Definitions: Corpus Querying and Manual Curation

Definitions for the four humanitarian principles were extracted from the HE Corpus with Sketch Engine following the two querying strategies detailed in Section 3.1.3. Firstly, the corpus was queried with the following CQL expression:

```
[lemma_lc="X"][] {0,3}[lemma_lc="be"][lemma_lc="not"]?[lemma_lc="a|the|one"]
```

where X corresponds to the designation of each humanitarian principle, i.e., ‘humanity’, ‘neutrality’, ‘impartiality’ and ‘independence’. Secondly, the corpus was again queried for each humanitarian principle with a macro built using the Lemma filter functionality and the catalogue of definitional KPs detailed in Table 2 (Section 3.1.3). Finally, candidates were manually curated and exported into a dataset containing definitions and corpus metadata.

Concept	Occurrences	Strategy	Candidates	Selection
HUMANITY	7041	Is_a CQL query	89	1
		KPs with Lemma filter	932	40
IMPARTIALITY	1423	Is_a CQL query	13	0
		KPs with Lemma filter	234	60
NEUTRALITY	1402	Is_a CQL query	14	0
		KPs with Lemma filter	223	37
INDEPENDENCE	5052	Is_a CQL query	36	1
		KPs with Lemma filter	505	38

Table 3: Definitional candidates and selection for each humanitarian principle

Table 3 compares the total number of occurrences in the HE corpus for each concept with the number of candidates obtained with each extraction strategy. The most productive strategy was, by far, querying the corpus with the Lemma filter

functionality. This method extracted a total of 177 definitions, with 41 for HUMANITY, 60 for IMPARTIALITY, 37 for NEUTRALITY, and 39 for INDEPENDENCE.

3.2.2 Elucidation of Conceptual Characteristics: Inductive Coding

Definitions were imported into Taguette (Section 3.1.4) and coded inductively by decomposing each definition into textual fragments and categorising them semantically. By doing so, the definitional elements of each humanitarian principle were elucidated and associated with textual evidence from each definition. In total, 39 tags were created, with an average of 1.72 tags per definition.

Definitions for the concept of HUMANITY generated 13 tags, with an average of 2.37 tags per definition and the highest number of diverse tags among the four humanitarian principles. In total, definitions for HUMANITY were tagged 98 times, with the most productive definition containing 5 tags and the least productive, only 1 tag. The most prominent definitional elements describe HUMANITY as a principle whereby humanitarian assistance should be delivered wherever it is needed (24 occurrences), with the goal to alleviate human suffering (19), prevent it (10), simply address it (9) or save human lives (10). Other eight marginal definitional elements were identified. Table 4 describes all tags created, details their number of occurrences, and provides examples from our sample of definitions.

HUMANITY			
Tag	Cases	Description	Example
Anywhere	24	Humanitarian assistance should be delivered wherever needed.	Human suffering must be addressed <u>wherever it is found</u> .
Alleviate human suffering	19	Humanitarian assistance should aim at alleviating human suffering.	...humanity (meaning the centrality of saving lives and <u>alleviating suffering</u> wherever it is found)...
Prevent human suffering	10	Humanitarian assistance should aim at preventing human suffering.	To <u>prevent</u> and alleviate <u>human suffering</u> wherever it may be found.
Save human lives	10	Humanitarian assistance should aim at saving human lives.	...humanity, meaning the centrality of <u>saving human lives</u> and alleviating suffering wherever it is found.
Address human suffering	9	Humanitarian assistance should address human suffering.	...in which the principle of humanity (i.e., <u>responding only to human suffering</u>) is the highest principle...

Right to dignity	7	Humanity requires acknowledging the right of all human beings to dignity.	Humanity: <u>people's right to a life in dignity</u> takes precedence over politics and principles.
Focus on most vulnerable populations	4	Humanitarian assistance should focus on the most vulnerable.	Humanity: human suffering must be addressed wherever it is found, <u>with particular attention (paid) to the most vulnerable in the population.</u>
Non-discrimination	3	Humanitarian assistance should be delivered without discrimination on any grounds.	Humanity: The International Red Cross and Red Crescent Movement, born of a desire to <u>bring assistance without discrimination</u> to the wounded on the battlefield...
Needs-based assistance	3	Humanitarian assistance should be delivered based on the needs of affected populations.	Humanity: <u>allocation of aid solely in proportion to needs</u> , as part of the overall aim of preventing and alleviating human suffering.
Human freedom	2	Humanity requires acknowledging that all human beings are born free.	...principle of humanity: that <u>all human beings are born free</u> and equal in dignity and rights.
Human equality	2	Humanity requires acknowledging that all human beings are equal.	...principle of humanity: that all human beings are born free and <u>equal in dignity and rights.</u>
Care for people	2	Humanity is caring for people.	Humanity: <u>people caring for people.</u>
Shared decency	1	Humanity is decency shared by all human beings.	It called to our collective humanity, to our <u>shared decency.</u>

Table 4: Characteristics of HUMANITY as coded in Taguette

For the concept of IMPARTIALITY, a set of 11 tags was created, with an average of 1.77 tags per definition and a total of 108 tags, being the highest number of tags generated among the four humanitarian principles. Definitional productivity ranges between 3 and 1 definitional elements. Semantically, the principle of IMPARTIALITY displays a solid core, whereby humanitarian assistance should be delivered without discriminating against recipients on the grounds of nationality, race, sex, class, or other distinctions (44 occurrences) and strictly be provided according to the needs of affected populations (33). Other nine less prominent definitional elements were identified, which consider that humanitarian assistance, when driven by this principle, should focus on targeting the most vulnerable (10), prioritise the most urgent cases (9) and deliver aid in

proportion to the needs of affected people (6). All the tags obtained for IMPARTIALITY are detailed in Table 5.

IMPARTIALITY			
Tag	Cases	Description	Example
Non-discrimination	44	Humanitarian assistance should be delivered without discrimination on any grounds.	Impartiality requires humanitarian actors to <u>make no distinctions on the basis of nationality, race, gender, religious beliefs, class or political opinions in their operations...</u>
Need-based assistance	33	Humanitarian assistance should be delivered based on the needs of affected populations.	Impartiality: we provide our assistance to those who are suffering, <u>according to need.</u>
Target the most vulnerable	10	Humanitarian assistance should target on the most vulnerable.	...impartiality of assistance, requires us to <u>provide aid to those who need it most</u> , wherever they may live.
Urgency prioritisation	9	Humanitarian assistance should prioritise the most urgent cases.	Impartiality requires humanitarian actors to make no discrimination..., <u>giving priority to the most urgent cases of distress.</u>
Proportionality	6	Humanitarian assistance should be proportional to the needs of affected people.	...the principle of impartiality, which requires that it be provided solely on the basis of need and <u>in proportion to need.</u>
Alleviate human suffering	2	Humanitarian assistance should aim at alleviating human suffering.	Impartiality requires humanitarian actors to make no discrimination...in their operations and <u>to relieve suffering</u> , giving priority to the most urgent cases of distress.
Deliver services close to the frontline	1	Impartiality implies delivering services to affected people close to the frontline	...the principle of impartiality, implies that they <u>should deliver their services as close to the frontline as possible.</u>
Gender equality	1	Humanitarians should pay attention to achieving fairness between women and men.	The humanitarian aims of proportionality and impartiality mean that <u>attention must be paid to achieving fairness between women and men and ensuring equality of outcome.</u>

Fair and transparent contracting	1	Impartiality implies conducting fair and transparent contracting procedures.	Impartiality: <u>Fair and transparent contracting procedures</u> are essential to avoid suspicion of favouritism or corruption.
Anywhere	1	Humanitarian assistance should be delivered wherever needed.	...impartiality of assistance, requires us to provide aid to those who need it most, <u>wherever they may live.</u>
Non-partisanship	1	Humanitarians should not take sides.	Impartiality: LPI conducts its work in an inclusive and <u>non-partisan way...</u>

Table 5: Characteristics of IMPARTIALITY as coded in Taguette

The principle of NEUTRALITY generated 11 distinct tags and a total of 68 tags distributed across 37 definitions. With an average of 1.62 tags per definition, definitional productivity ranges between 3 and 1 definitional elements. The semantic core of NEUTRALITY comprises three prominent definitional elements, compelling humanitarians not to take sides in conflicts (20 occurrences), avoiding engaging in controversies of ideological nature (15) and refraining from favouring conflict parties (12). Other eight additional less prominent definitional elements were identified and are also detailed in Table 6.

NEUTRALITY			
Tag	Cases	Description	Example
No side-taking in conflicts	20	Humanitarians should not take sides in conflict.	The principle of neutrality means that <u>in a situation of conflict, no one takes sides with one of the parties involved.</u>
No engagement in controversies	15	Humanitarians should not engage in political, religious, or ideological controversies.	Neutrality: Humanitarian actors must not take sides in hostilities or <u>engaging controversies of a political, racial, religious or ideological nature.</u>
No favouring conflict party	12	Humanitarians should not favour parties to a conflict.	The provision of humanitarian assistance of the Czech Republic is governed by...neutrality (<u>the humanitarian actors do not favour any part of a given conflict</u>)...

Free from political or religious affiliation	3	Humanitarians should not be affiliated to religious or political causes.	Neutrality – <u>we are not affiliated to any political or religious constituency.</u>
No commercial gain	2	Humanitarians should not seek commercial gains.	Neutrality: provision of assistance <u>without seeking</u> to further a particular political or religious standpoint or <u>to obtain commercial gain...</u>
Abide by national and international law	2	Humanitarian assistance should take place in line with national and international law.	Neutrality: provision of assistance... <u>abiding by applicable national and international law...</u>
Provides trust	2	Independence generates trusts in humanitarian actors.	Neutrality: <u>humanitarian initiatives need trust.</u>
Perception	1	To be perceived as neutral.	Neutrality requires humanitarian organisations...that their action does not provide support to either side of the conflict, or <u>is perceived as doing so.</u>
No engagement with States	1	Humanitarians should not engage with governments.	...neutrality requires <u>avoiding engagement with state structures...</u>
Non-discrimination	1	Humanitarian assistance should be delivered without discrimination on any grounds	Neutrality: Slovenia provides humanitarian aid independently of the sides to a conflict, <u>whereby the aid is offered under the same conditions...</u>
Needs-based assistance	1	Humanitarian assistance should be delivered based on the needs of affected populations.	Neutrality: Slovenia provides humanitarian aid independently of the sides to a conflict... <u>based on the current needs of the affected population.</u>

Table 6: Characteristics of NEUTRALITY as coded in Taguette

Lastly, the concept of NEUTRALITY generated the fewest number of distinct tags, with a total of 4 definitional elements and a definitional productivity of 1 tag. However, NEUTRALITY displays the most well-defined semantic core with a single outstanding definitional element: Autonomy. The principle of NEUTRALITY compels humanitarians to act autonomously from the objectives of political, military, economic or other actors like donors (28 occurrences). Marginally, some definitions also consider that being neutral requires humanitarian actors not to be affiliated to religious or political causes

(6), which contrasts with the existence of religious entities in the sector, as seen in Table 1 (Section 3.1.1.). Details for all 4 tags are presented in Table 7.

INDEPENDENCE			
Tag	Cases	Description	Example
Autonomy	28	Humanitarian action should be autonomous from the objectives of political, economic, military and other actors.	Operational Independence: our humanitarian actions are <u>autonomous of any political, economic, military or other objectives of its donors or other actors</u>
Free from political or religious affiliation	6	Humanitarians should not be affiliated to religious or political causes.	Independence: <u>from any religious or party-political affiliation.</u>
Holistic approach to services for the most vulnerable	2	Independence requires services to the most vulnerable be delivered in a holistic way.	Sustaining independence requires a <u>holistic approach which incorporates other key local services such as housing, education, health and social protection for those who are most vulnerable.</u>
Transfer responsibility to locals	1	Independence requires transferring the responsibility over infrastructure to local actors.	Independence: <u>transfer the infrastructure to local responsibility.</u>

Table 7: Characteristics of INDEPENDENCE as coded in Taguette

3.2.3 Uncovering Conceptual Variation: Data Consolidation and Visualisation

Once definitions have been decomposed and quantified through inductive coding, definitional elements were associated to organisation types by combining the datasets obtained for each concept in Taguette with the corpus metadata obtained from Sketch Engine. By doing so, definitional elements can be disaggregated by organisation type. If definitional elements are distributed markedly unevenly across organisation types, it can therefore be argued that a concept may display conceptual variation.

Our datasets for each concept were loaded onto R and visualised with the `fmsb` package with radar charts (Section 3.1.5.). Radar charts represent definitional elements with the vertices of a polygon, forming the axes of a chart. The occurrences of definitional elements are represented by colour polygons, whose vertices represent the occurrences

of each definitional element by comparing their position against the chart’s axes. Figure 1 illustrates how concepts can be represented with radar charts.

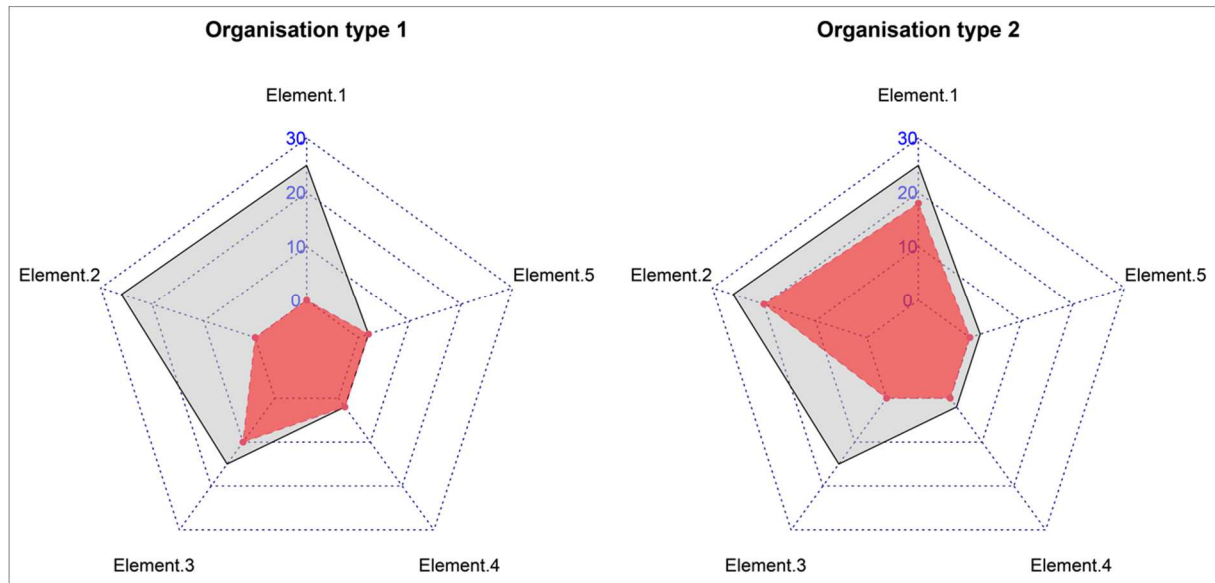


Figure 1: Example of conceptual variation with radar charts

The grey polygon represents all definitional elements and the total number of occurrences. This provides a graphic representation of a concept’s semantic core. In the example of Figure 1, definitional elements 1, 2 and 3 are the most prominent and therefore constitute the semantic core. Thanks to data disaggregation, occurrences by organisation type are represented with an overlapping red polygon, which can be compared against the semantic core. In addition, the shape of red polygons can also be compared by juxtaposing radar charts, which is useful to reveal possible cases of conceptual variation. In Figure 1, organisation type 1 generates most instances of definitional element 3, whereas organisation type 2 produces most definitional elements 1 and 2. This reveals a stark contrast between the two organisation types, suggesting a case of conceptual variation.

A total of 36 radar chart visualisations were produced for each humanitarian principle. There were two organisation types that did not generate data for any concept, namely RE and Found. Additionally, NEUTRALITY did not produce definitional elements from IGO and Project documents, nor did the concept of INDEPENDENCE, which lacks data from Project documents too. In Section 4, we present and discuss the results by interpreting these data visualisations.

4. Results

Representing the quantitative dimension of definitional elements helps build conceptual profiles, enabling comparison within and between concepts. This Section interprets the data visualisations obtained with the method described in Section 3. It presents the

conceptual profiles of HUMANITY, IMPARTIALITY, NEUTRALITY, and INDEPENDENCE in Section 4.1. Each concept is then further analysed by comparing the contributions from each organisation type against the four conceptual profiles in Section 4.2.

4.1 Conceptual Profiles

As described in Section 3.2.2., concepts have a semantic core, i.e., the set of the most quantitatively prominent characteristics. They may also present marginal characteristics with low numbers of occurrences as well as a limited or wide range of characteristics. The more definitional elements are found in a concept, it is safe to assume that it will be more likely to be subject to conceptual variation.

Figure 2 displays the conceptual profiles for the four humanitarian principles by representing with a grey polygon the quantitative dimension of their definitional elements. The semantic core of each concept is therefore represented by the most protruding sides of their polygons. The cases of HUMANITY, IMPARTIALITY and NEUTRALITY show a wide range of features. HUMANITY has a semantic core formed by a dominant and a less dominant module. The former consists of two prominent definitional elements (Anywhere and Alleviate_human_suffering), while the latter comprises four less prominent but comparatively more relevant (Address_human_suffering, Save_human_lives, Prevent_human_suffering and Right_to_dignity) than the rest of marginal features.

Similarly, IMPARTIALITY presents a two-module semantic core with a markedly dominant one (Need_based_assistance and Non_discrimination) and a less prominent module (Urgency_prioritisation, Target_the_most_vulnerable and Proportionality). In contrast, the concepts of NEUTRALITY and INDEPENDENCE show more compact semantic cores. NEUTRALITY presents a well-defined three-pronged core (No_side_taking_in_conflicts, No_engagement_in_controversies and No_favouring_conflict_parties), while INDEPENDENCE stands out for its semantic core formed by one definitional element (Autonomy).

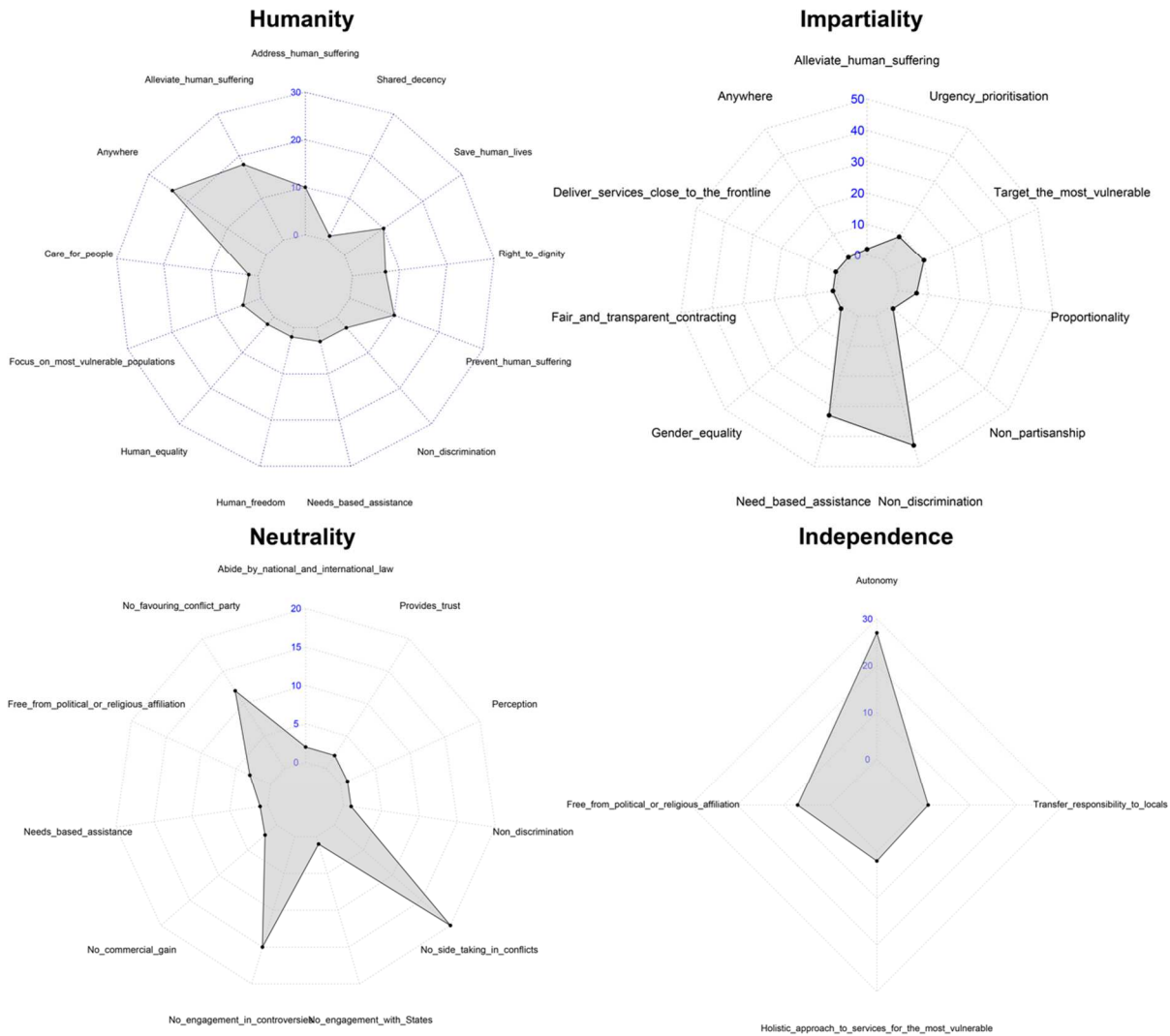


Figure 2: Conceptual profiles for HUMANITY, IMPARTIALITY, NEUTRALITY, and INDEPENDENCE.

In general, Figure 2 suggests that the four humanitarian concepts are well-defined, as shown by their relatively clear-cut semantic cores. However, HUMANITY and IMPARTIALITY present notable secondary definitional elements that contrast with the compact semantic cores of NEUTRALITY and INDEPENDENCE. Additionally, there are multiple definitional elements found across concepts, with key features in one concept constituting marginal ones in another. For example, Anywhere in HUMANITY is prominent, but it is clearly marginal in IMPARTIALITY.

Another example is Non_discrimination, which is found as a marginal feature in NEUTRALITY, although it is part of the semantic core of IMPARTIALITY. This phenomenon may indicate peripheral cases of confusion between humanitarian principles. The most striking case is found between NEUTRALITY and INDEPENDENCE with respect to Free_from_political_or_religious_affiliations. Despite its clearly

defined semantic core, INDEPENDENCE contains a non-negligible number of occurrences of this definitional element, which is also found in NEUTRALITY.

4.2 Detecting Conceptual Variation

Several cases of differences in distributions of definitional elements were detected among organisation types. These suggest that humanitarian organisations may have slightly different understandings of the four humanitarian principles or attach more importance to some characteristics over others. Figures 3 to 6 contain radar charts for each concept and organisation type.

Firstly, the understanding of HUMANITY is distributed unevenly, with organisations presenting similar distributions and others constituting clear outliers (Figure 3). C/B and State show similar profiles, with preference for the definitional elements of Anywhere, Alleviate_human_suffering, Save_human_lives and Focus_on_vulnerable_populations. Concretely, C/B definitions present a slightly higher number of occurrences for Save_human_lives, while State definitions appear to highlight Focus_on_vulnerable_populations more. Similarly, Net and NGO_Fed also present similar profiles, as they both coincide on Anywhere and Address_human_suffering.

By contrast, RC, NGO, and Project generate completely dissimilar profiles. The most striking difference is found in RC definitions with regards to Prevent_human_suffering, containing all its occurrences. With a relatively smaller number of occurrences, Project definitions appear to emphasise more the acknowledgement of human rights, as evidenced by their preference for Human_equality, Human_freedom and Right_to_dignity. IGO contributes with little definitional data, rendering its profile negligible.

Secondly, IMPARTIALITY presents divergent distributions of definitional elements in its semantic core, with the most notable differences being those found between RC, Project and NGO_Fed compared to the rest of organisations. (Figure 4). C/B, State and NGO documents mostly coincide on Need_based_assistance and Non_discrimination, which constitute the semantic core of this concept. C/B definitions differ in that they also consider Target_the_most_vulnerable as a definitional element of the concept, while State definitions do not.

Conversely, RC definitions describe IMPARTIALITY almost exclusively in terms of Non_discrimination, whereas NGO_Fed and Project focus solely on Need_based_assistance. In terms of marginal definitional elements, Project definitions uniquely highlight Proportionality, while NGO_Fed diverges slightly by emphasising Urgency_prioritisation.

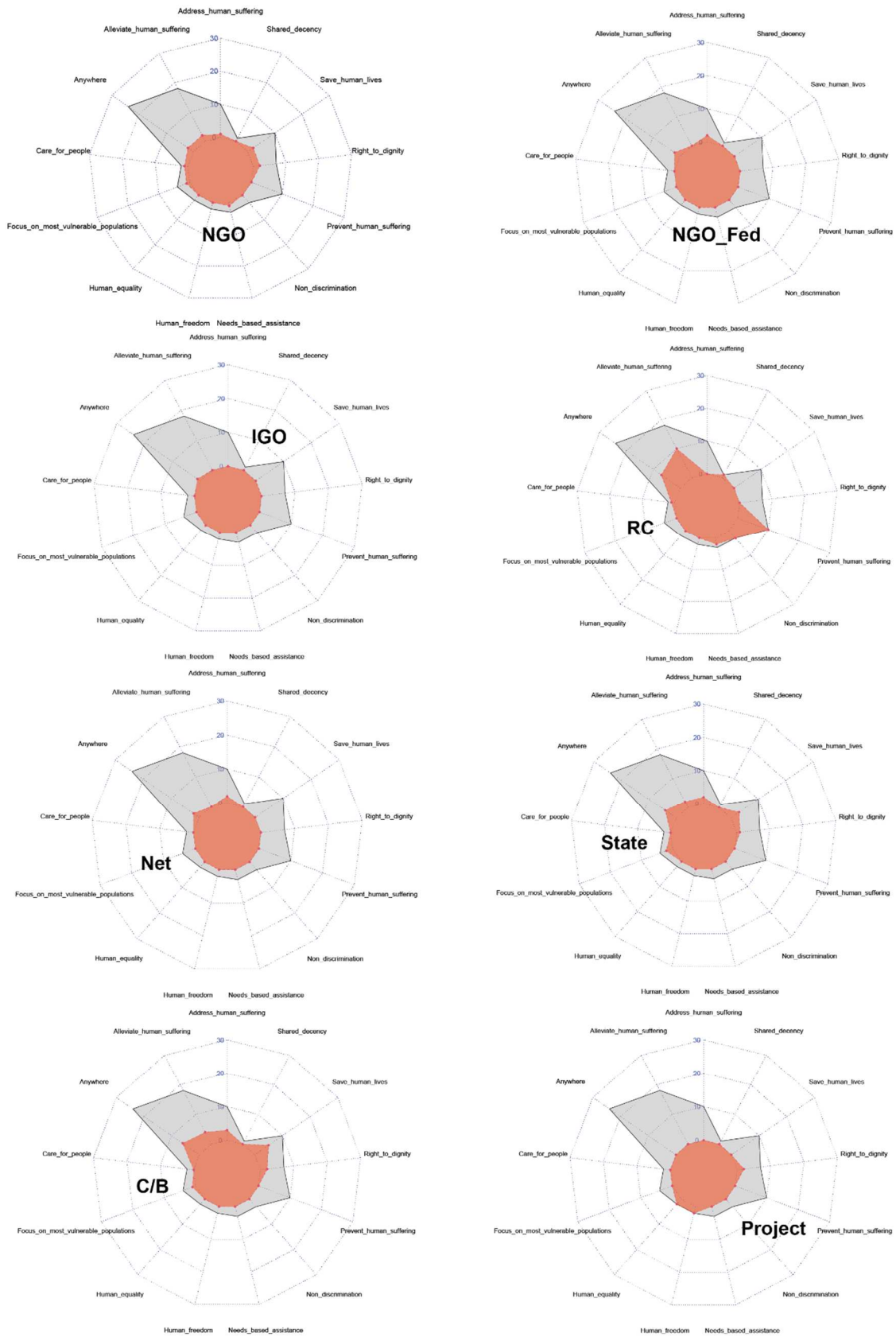


Figure 3: Distribution of definitional elements by organisation type for HUMANITY

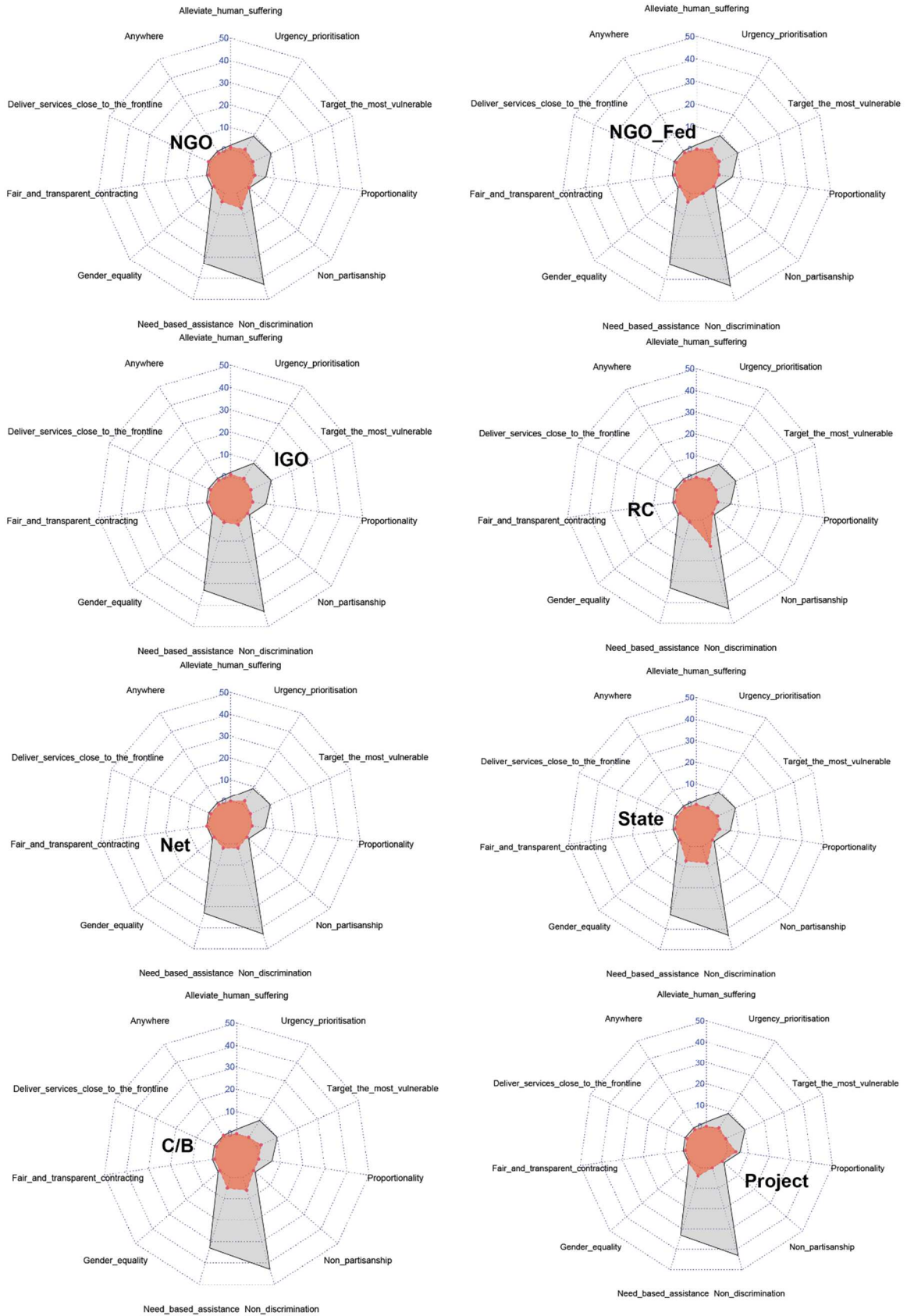


Figure 4: Distribution of definitional elements by organisation type for IMPARTIALITY

Thirdly, several cases of variation were also found for NEUTRALITY (Figure 5). NGO, NGO_Fed and Net definitions show similar profiles, with preference for No_engagement_in_controversies and No_side_taking_in_conflicts. The starkest difference in distribution was, once again, found between C/B and State, and RC. RC definitions provide most occurrences of No_engagement_in_controversies and No_side_taking_in_conflicts, while uniquely describing the concept in terms of Abide_by_national_and_international_law and Provides_trust.

However, State definitions focus on No_favouring_conflict_party and completely disregard RC definitional elements. Marginally, some State definitions also appear to confuse NEUTRALITY with IMPARTIALITY, as they are the only ones containing all the occurrences of Needs_based_assistance, a prominent definitional element in the semantic core of IMPARTIALITY. C/B definitions also prefer No_favouring_conflict_party, but they coincide slightly with RC definitions on No_engagement_in_controversies and No_side_taking_in_conflicts. IGO and Project documents provided no definitional data for NEUTRALITY.

Lastly, a subtle case of conceptual variation was detected in INDEPENDENCE (Figure 6) between NGO_Fed definitions and the rest of organisation types. NGO_Fed definitions account for all the occurrences of marginal elements, with Free_from_political_or_religious_affiliation being the most prominent. This definitional element is also found in NEUTRALITY, with a relative low number of occurrences compared to its semantic core. This suggests a slight semantic overlap between NEUTRALITY and INDEPENDENCE. In fact, NGO_Fed definitions show a higher number of occurrences of Free_from_political_or_religious_affiliation than Autonomy, which is the dominant definitional element in INDEPENDENCE. This indicates that NGO_Fed definitions may conceptualise INDEPENDENCE in a different manner.

In summary, the four humanitarian principles show sufficiently well-defined semantic cores, with primary and secondary sets of 1 to 3 prominent definitional elements. Organisational differences in the distribution of definitional elements were detected, especially between pairs of organisations with similar distributions (i.e., NGO_Fed and Net, State and C/B) and RC. Definitions from RC documents deviate the most from the rest of organisation types for the concepts of HUMANITY, IMPARTIALITY and NEUTRALITY, making RC a clear outlier. Project and NGO definitions also deviate notably with respect to the dominant organisation types. In HUMANITY, both organisations display unique profiles. In IMPARTIALITY, NGO aligns with C/B and State, whereas Project exhibit a unique distribution that overlaps partially with NGO_Fed. In NEUTRALITY, NGO appears to coincide partially with NGO_Fed and Net. As for INDEPENDENCE, all organisation types, save for NGO_Fed, appear to agree on a shared understanding of the concept.

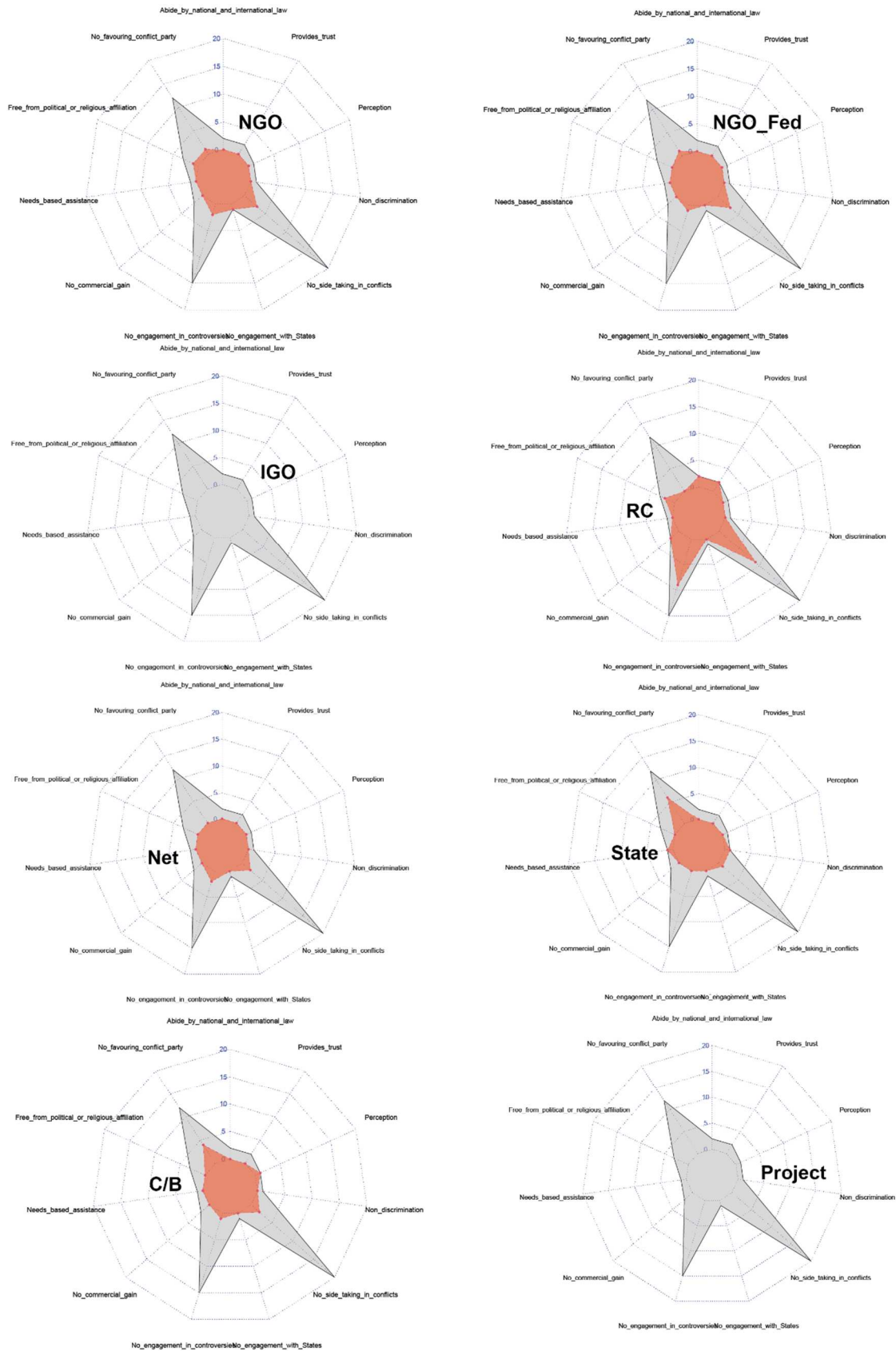


Figure 5: Distribution of definitional elements by organisation type for NEUTRALITY

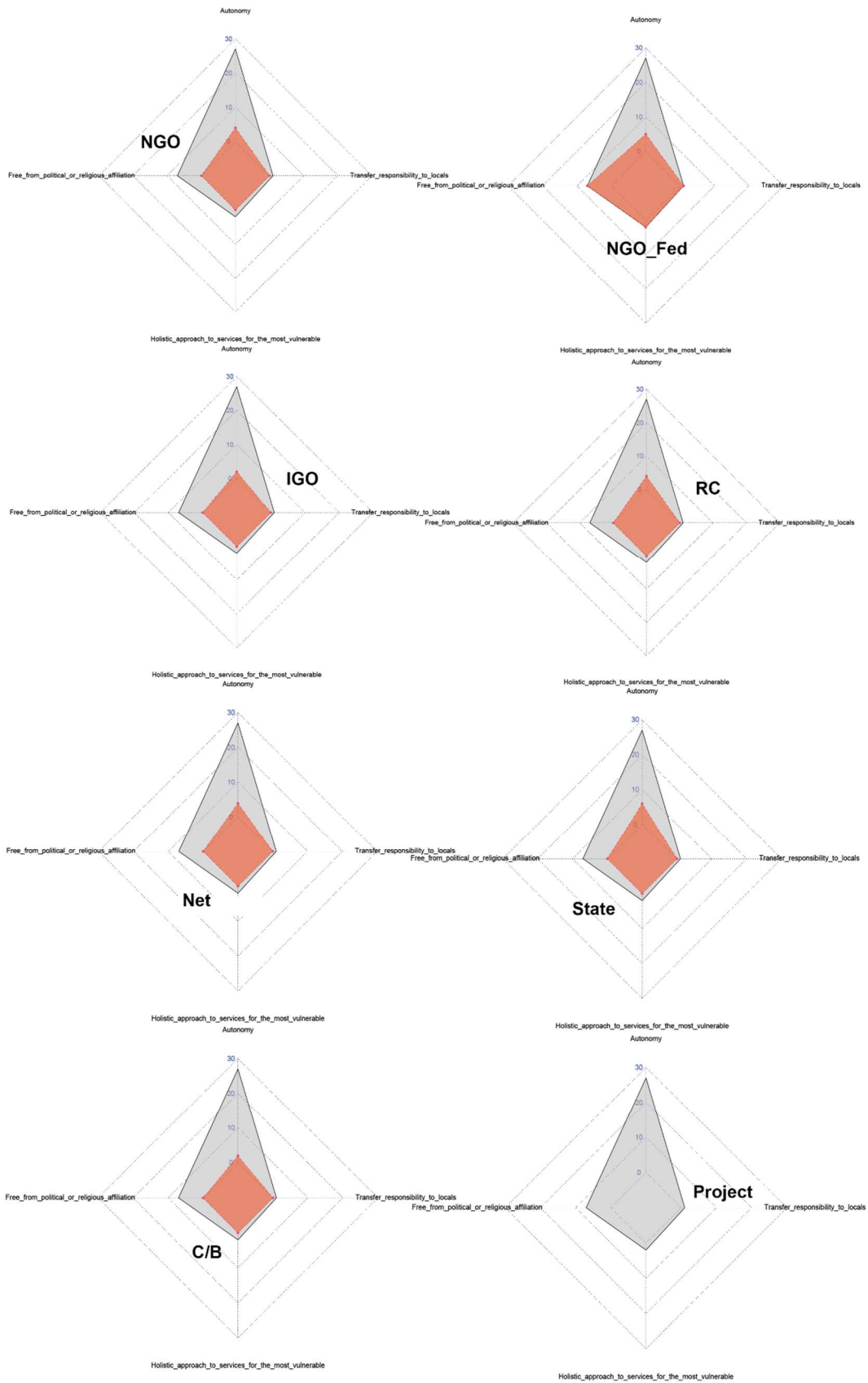


Figure 6: Distribution of definitional elements by organisation type for INDEPENDENCE

5. Conclusion

This pilot study combined a Frame-based Terminology approach, inductive content analysis and data visualisation to determine whether four key humanitarian concepts (i.e., the principles of HUMANITY, IMPARTIALITY, NEUTRALITY, and INDEPENDENCE) are affected by conceptual variation. To study conceptual variation, we proposed a method to create datasets of definitional elements linked to different types of humanitarian organisations. We extracted definitions from a corpus of humanitarian documents, coded said definitions inductively and consolidated the quantification of definitional elements with corpus metadata for each definition. For data interpretation, we then represented each humanitarian concept by plotting their definitional elements, together with their respective occurrences, on radar charts. We also disaggregated definitional elements by organisation type to detect distributional differences. This study demonstrated that radar charts are an effective way to both represent the semantic core of a concept and detect possible cases of conceptual variation among subsets.

Future studies will be conducted with more data obtained from knowledge-rich contexts beyond definitions to represent variation in hierarchical and non-hierarchical relations. Additional efforts will be required for two main purposes. Firstly, new workflows will be designed to produce interactive visualisations automatically, which will enable us to represent more data, accelerate its interpretation and facilitate the detection of more compelling cases of conceptual variation. Secondly, more sophisticated metrics, as well as data weighting methods, will be considered to account for the different sizes of subcorpora and produce more rigorous statistical representations of concepts in humanitarian discourse. In a future project, we plan to include another language (i.e., Spanish) in the study of conceptual variation. This will require additional efforts to design a suitable methodology that will help establish equivalences between definitional elements and represent data in way that can be easily interpreted.

6. Acknowledgements

This research was carried out as part of project VariTerminHum: Analysis and Representation of Terminological Variation in the Humanitarian Domain (PROYEXCEL_00369) funded by the regional government of Andalusia, Spain. We would also like to thank the anonymous reviewers, whose critical reading and suggestions helped improve this paper.

7. References

- Abu-Sada, Caroline, ed. 2012. *In the Eyes of Others: How People in Crises Perceive Humanitarian Aid*. USA: MSF-USA.
- Béné, Christophe, Rachel Godfrey Wood, Andrew Newsham, and Mark Davies. 2012. 'Resilience: New Utopia or New Tyranny? Reflection about the Potentials and

- Limits of the Concept of Resilience in Relation to Vulnerability Reduction Programmes’. *IDS Working Papers* 2012 (405): 1–61.
<https://doi.org/10.1111/j.2040-0209.2012.00405.x>.
- Bengtsson, Mariette. 2016. ‘How to Plan and Perform a Qualitative Study Using Content Analysis’. *NursingPlus Open* 2 (January): 8–14.
<https://doi.org/10.1016/j.npls.2016.01.001>.
- Chambó, Santiago, and Pilar León-Araúz. 2021. ‘Visualising Lexical Data for a Corpus-Driven Encyclopaedia’. In *Proceedings of ELex 2021*, 29–55. Brno, Czech Republic: Lexical Computing CZ s.r.o.
- Dorantes, Miguel Alejandro, Alejandro Pimentel, Gerardo Sierra, Gemma Bel-Enguix, and Claudio Molina. 2017. ‘Extracción automática de definiciones analíticas y relaciones semánticas de hiponimia-hiperonimia con un sistema basado en patrones lingüísticos’. *Linguamática* 9 (2): 33–44.
<https://doi.org/10.21814/lm.9.2.257>.
- Eberwein, Wolf-Dieter, and Sabine Saurugger. 2013. ‘The Professionalization of International Non-Governmental Organizations’. In *Routledge Handbook of International Organization*, edited by Bob Reinalda, 257–69. Abingdon-on-Thames, England: Routledge.
- Faber, Pamela. 2015. ‘Frames as a Framework for Terminology’. In *Handbook of Terminology: Volume 1*, edited by Hendrik J. Kockaert and Frieda Steurs, 14–33. Amsterdam, The Netherlands: John Benjamins Publishing Company.
<https://benjamins.com/catalog/hot.1.fra1>.
- . 2022. ‘Chapter 16. Frame-Based Terminology’. In *Theoretical Perspectives on Terminology: Explaining Terms, Concepts and Specialized Knowledge*, edited by Pamela Faber and Marie-Claude L’Homme, 23:353–76. Terminology and Lexicography Research and Practice. Amsterdam, The Netherlands: John Benjamins Publishing Company. <https://doi.org/10.1075/tlrp.23.16fab>.
- Gil-Berrozpe, Juan Carlos, Pilar León-Araúz, and Pamela Faber. 2019. ‘Ontological Knowledge Enhancement in EcoLexicon’. In *Proceedings of the ELex 2019*, 177–97. Sintra, Portugal: Lexical Computing.
<https://dialnet.unirioja.es/servlet/articulo?codigo=7302779>.
- Hampton, J. A. 2020. ‘Investigating Differences in People’s: Concept Representations’. In , edited by T. Marques and A. Wikforss, 67–82. Oxford, UK: Oxford University Press.
<https://doi.org/10.1093/oso/9780198803331.003.0005>.
- Hansen, Greg. 2008. ‘The Ethos-Practice Gap: Perceptions of Humanitarianism in Iraq’. *International Review of the Red Cross* 90 (899).
- Humanitarian Encyclopedia. 2021a. ‘Corpus’. Humanitarian Encyclopedia. 2021.
<https://humanitarianencyclopedia.org/corpus>.
- . 2021b. ‘Methodology’. Humanitarian Encyclopedia. 2021.
<https://humanitarianencyclopedia.org/methodology>.
- Kantner, Cathleen, and Maximilian Overbeck. 2020. ‘Exploring Soft Concepts with Hard Corpus-Analytic Methods’. In *Reflektierte Algorithmische Textanalyse*,

- edited by Niels Reiter, Axel Pichler, and Jonas Kuhn, 169–90. Berlin, Germany: De Gruyter. <https://doi.org/10.1515/9783110693973-008>.
- Khan, Abdul Kadir, and Tiina Kontinen. 2022. ‘Impediments to Localization Agenda: Humanitarian Space in the Rohingya Response in Bangladesh’. *Journal of International Humanitarian Action* 7 (1): 14. <https://doi.org/10.1186/s41018-022-00122-1>.
- Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. ‘The Sketch Engine: Ten Years On’. *Lexicography* 1 (1): 7–36. <https://doi.org/10.1007/s40607-014-0009-9>.
- Knox Clarke, P., and B. Ramalingan. 2014. ‘Meeting the Urban Challenge: Adapting Humanitarian Efforts to an Urban World’. London: ALNAP/ODI. <https://www.alnap.org/help-library/meeting-the-urban-challenge-adapting-humanitarian-efforts-to-an-urban-world>.
- Kovář, Vojtěch, Monika Močiariková, and Pavel Rychlý. 2016. ‘Finding Definitions in Large Corpora with Sketch Engine’. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, 391–94. Portorož, Slovenia: European Language Resources Association (ELRA). <https://aclanthology.org/L16-1061>.
- Kyngäs, Helvi. 2020. ‘Inductive Content Analysis’. In *The Application of Content Analysis in Nursing Science Research*, edited by Helvi Kyngäs, Kristina Mikkonen, and Maria Kääriäinen, 13–21. Cham, Switzerland: Springer International Publishing. https://doi.org/10.1007/978-3-030-30199-6_2.
- León-Araúz, Pilar. 2017. ‘Term and Concept Variation in Specialized Knowledge Dynamics’. In *Multiple Perspectives on Terminological Variation*, edited by Patrick Drouin, Aline Francoeur, John Humbley, and Aurélie Picton. Amsterdam, The Netherlands: John Benjamins Publishing Company.
- León-Araúz, Pilar, and Arianne Reimerink. 2019. ‘High-Density Knowledge Rich Contexts’. *Argentinian Journal of Applied Linguistics* (1): 109–30.
- Lindgren, Simon. 2016. ‘Introducing Connected Concept Analysis: A Network Approach to Big Text Datasets’. *Text & Talk* 36 (3): 341–62. <https://doi.org/10.1515/text-2016-0016>.
- Marshman, Elizabeth. 2022. ‘Chapter 13. Knowledge Patterns in Corpora’. In *Theoretical Perspectives on Terminology: Explaining Terms, Concepts and Specialized Knowledge*, edited by Pamela Faber and Marie-Claude L’Homme, 23:291–310. Terminology and Lexicography Research and Practice. Amsterdam, The Netherlands: John Benjamins Publishing Company. <https://doi.org/10.1075/tlrp.23.13mar>.
- Meyer, Ingrid. 2001. ‘Extracting Knowledge-Rich Contexts for Terminography’. In *Recent Advances in Computational Terminology*, edited by Didier Bourigault, Christian Jacquemin, and Marie-Claude L’Homme, 279–302. Amsterdam, The Netherlands: John Benjamins Publishing Company. <https://benjamins.com/catalog/nlp.2>.

- Nakazawa, Minato. 2023. 'Fmsb: Functions for Medical Statistics Book with Some Demographic Data'. <https://CRAN.R-project.org/package=fmsb>.
- Nuopponen, Anita. 2010. 'Methods of Concept Analysis-a Comparative Study Part 1 of 3'. *LSP Journal* 1 (1).
- . 2022. 'Chapter 3. Conceptual Relations: From the General Theory of Terminology to Knowledge Bases'. In *Theoretical Perspectives on Terminology: Explaining Terms, Concepts and Specialized Knowledge*, edited by Pamela Faber and Marie-Claude L'Homme, 23:63–86. Terminology and Lexicography Research and Practice. Amsterdam, The Netherlands: John Benjamins Publishing Company. <https://doi.org/10.1075/tlrp.23.03nuo>.
- Rampin, Rémi, and Vicky Rampin. 2021. 'Taguette: Open-Source Qualitative Data Analysis'. *Journal of Open Source Software* 6 (68): 3522. <https://doi.org/10.21105/joss.03522>.
- Sezgin, Zeynep, and Dennis Dijkzeul, eds. 2015. *The New Humanitarians in International Practice: Emerging Actors and Contested Principles*. London, UK: Routledge. <https://doi.org/10.4324/9781315737621>.
- Sierra, Gerardo, Rodrigo Alarcón, César Aguilar, and Carme Bach. 2008. 'Definitional Verbal Patterns for Semantic Relation Extraction'. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication* 14 (1): 74–98. <https://doi.org/10.1075/term.14.1.05sie>.
- . 2010. 'Definitional Verb Patterns for Semantic Relation Extraction'. In *Probing Semantic Relations: Exploration and Identification in Specialized Texts*, edited by Alain Auger and Caroline Barrière, 74–96. Benjamins Current Topics. Amsterdam, The Netherlands: John Benjamins Publishing Company. <http://gen.lib.rus.ec/book/index.php?md5=9c2603bdd3a75152d427b43b8068911f>.
- Stroup, Sarah S. 2012. *Borders among Activists: International NGOs in the United States, Britain, and France*. Borders among Activists. Ithaca, NY: Cornell University Press. <https://doi.org/10.7591/9780801464256>.

Rapid Ukrainian-English Dictionary Creation

Using Post-Edited Corpus Data

Marek Blahuš¹, Michal Cukr¹, Ondřej Herman^{1,2}, Miloš Jakubíček^{1,2}, Vojtěch Kovář^{1,2}, Jan Kraus¹, Marek Medved^{1,2}, Vlasta Ohlídalová^{1,2}, Vít Suchomel^{1,2}

¹Lexical Computing, Brno, Czechia

² Faculty of Informatics, Masaryk University, Brno, Czechia
E-mail: firstname.lastname@sketchengine.eu

Abstract

This paper describes the development of a new corpus-based Ukrainian-English dictionary. The dictionary was built from scratch, we used no pre-existing dictionary data. A rapid dictionary development method was used which consists of generating dictionary parts directly from a large corpus, and of post-editing the automatically generated data by native speakers of Ukrainian (not professional lexicographers). The method builds on Baisa et al. (2019) which was improved and updated, and we used a different data management model. As the data source, a 3-billion-word Ukrainian web corpus from the TenTen series (Jakubíček et al., 2013) was used.

The paper briefly describes the corpus, then we thoroughly explain the individual steps of the *automatic generation—post-editing* workflow, including the volume of the manual work needed for the particular phases in terms of person-days. We also present details about the newly created dictionary and discuss directions for its further development.

Keywords: Ukrainian; post-editing; dictionary; lexicography

1. Introduction

For decades, language corpora have served as source data for dictionary building. In the last years, corpora were also used for automatic generation of various dictionary parts (Rundell & Kilgariff, 2011; Kosem et al., 2018; Gantar et al., 2016; Kallas et al., 2015). These automatic outputs were then post-edited by professional lexicographers to ensure the data quality in the resulting dictionary.

With the advancement of technology, it is now possible to create whole dictionaries using this scenario of automatic generation and post-editing by native speakers (not necessarily professional lexicographers). The methodology was used before (Baisa et al., 2019); we have improved the process and used it in a new project aimed at creating a Ukrainian–English dictionary using a 3-billion-word Ukrainian corpus.

This paper covers all our work on this particular project. We describe building, cleaning and tagging the new multi-billion web corpus of Ukrainian. Then, we discuss the rapid dictionary creation method and our particular implementation which is different from (Baisa et al., 2019) especially in the data management approach.

In the last part, we describe the resulting dictionary that contains more than 55,000 verified headwords but due to time and budget constraints, we were able to fully complete only 10,000 entries, so there is still large space for improvements.

2. New Ukrainian Web Corpus

We were able to identify three Ukrainian corpora the new dictionary could be based on:

- General Regionally Annotated Corpus of Ukrainian (GRAC) (Shvedova, 2020; Starko, 2021)
- UberText Ukrainian corpus by Lang-uk¹, a web corpus of 665 million tokens
- ukTenTen14 web corpus from 2014, consisting of 2.73 billion tokens

Of these corpora, the first one is not available for download. The second one is a rather small, topic-specific corpus (mostly news). It is distributed in the form of shuffled sentences, which prevents the selection of headwords by document frequency. For our dictionary work, we took the third one, enlarged it and updated it into a new Ukrainian web corpus. In this stage we followed the methodology of the TenTen corpora family (Jakubíček et al., 2013).

The crawler (Suchomel & Pomikálek, 2012) was instructed to download from Ukrainian top-level domains `.ua` and `.укр` and generic domains such as `.com`, `.org`, or `.net`. A character trigram based model trained on a 200 kB sample of manually checked Ukrainian plaintext was used to stop crawling websites that did not contain text in Ukrainian.

The crawl was initialized by nearly 6 million unique seed URLs:

- 194 manually identified news sites
- 94,000 websites from web directories
- 336,000 URLs of web pages found by search engine Bing by searching Ukrainian words
- 5,410,000 URLs found in ukTenTen14

Table 1: Number of documents by TLD in the final merged and cleaned data from 2014 and 2022

TLD	documents	tokens	% corpus tokens
ua	4 640 585	2 122 675 553	65
com	1 099 646	591 327 114	18
org	1 089 027	397 328 162	12
net	318 197	143 994 060	4.4
eu	16 046	8 759 810	0.27

Data obtained by the crawler were converted to UTF-8 with the help of the Chared tool (Pomikálek & Suchomel, 2011) and cleaned by jusText (Pomikálek, 2011). The

¹ <https://lang.org.ua/en/corpora/>, accessed in April 2023.

Table 2: Websites contributing the most tokens to the final merged and cleaned data from 2014 and 2022

Website	description	documents	tokens	% corpus tokens
uk.wikipedia.org	encyclopedia	791 134	243 194 981	7.4 %
uapatents.com	government, patents	36 829	36 339 611	1.1 %
pulib.if.ua	tech encyclopedia	11 669	26 054 618	0.79 %
techtrend.com.ua	tech encyclopedia	18 746	22 706 445	0.69 %
litopys.org.ua	text library	4 592	22 501 121	0.69 %
ligazakon.ua	legal	17 622	22 334 382	0.68 %
uad.exdat.com	(site down in 2023)	8 220	15 928 398	0.49 %
alls.in.ua	(site down in 2023)	14 022	15 292 017	0.47 %
maidan.org.ua	politics, news	11 791	14 826 687	0.45 %
ua.textreferat.com	essays, schoolwork	18 536	14 614 928	0.45 %
economy.nayka.com.ua	economic news	6 025	14 418 873	0.44 %
uatxt.ensayoes.com	(site down in 2023)	7 401	13 575 562	0.41 %
gazeta.dt.ua	news	9 306	13 047 965	0.40 %
uadocs.exdat.com	(site down in 2023)	7 221	12 810 126	0.39 %
zakon-ua.com	(legal, down in 2023)	6 385	12 249 178	0.37 %

result was merged with the old ukTenTen14 and with 1,040,000 articles from Ukrainian Wikipedia downloaded by the Wiki2corpus tool.² Duplicate paragraphs were removed by Onion (Pomikálek, 2011) and manual cleaning was performed according to Suchomel & Kraus (2021).

The final size of the merged Ukrainian corpus is 3,280 million tokens and 2,593 million words in 7.2 million documents with 52% texts downloaded in 2014 and 48% texts downloaded in 2020. Sizes of parts of the corpus coming from selected TLDs and websites are in Table 1 and Table 2, respectively. As can be seen there, the most contributing sites are encyclopedias, technology sites, news sites and legal related sites. Distribution of genres and topics assigned using the method described in Suchomel & Kraus (2022) can be found in Table 3 and in Table 4, respectively.

The corpus was then tagged using RFTagger (Schmid & Laws, 2008) and lemmatized using CST lemmatiser (Jongejan & Dalianis, 2009). The RFTagger model was trained on the Universal Dependencies corpus for Ukrainian³ and the Brown corpus of the Ukrainian language (Starko & Rysin, 2023). Training was also supplemented by an additional morphological database generated from the Ukrainian Brown dictionary (Starko & Rysin, 2020). The model for the CST lemmatiser was trained on Ukrainian Brown dictionary using Affixtrain.⁴ As the last step, heuristic postprocessing of the tagging and lemmatization was applied based on manual inspection of the corpus data.

3. Rapid Dictionary Development by Post-editing

The post-editing methodology we are building on assumes that all lexicographic content is automatically generated from an annotated corpus, and step-by-step post-edited, re-

² <https://corpus.tools/wiki/wiki2corpus>

³ https://github.com/UniversalDependencies/UD_Ukrainian-IU

⁴ <https://github.com/kuhumcst/affixtrain>

Table 4: Subcorpus sizes by topic

Topic	documents	tokens	% corpus
society	484 425	244 264 763	7.4 %
business	100 807	76 791 974	2.3 %
science	76 214	61 546 682	1.9 %
arts	86 876	51 562 726	1.6 %
health	60 184	39 859 674	1.2 %
home	81 442	39 208 234	1.2 %
recreation	23 241	14 499 974	0.44 %
games	18 548	11 291 503	0.34 %
sports	23 357	7 331 632	0.22 %
technology	5 561	1 622 874	0.049 %

Table 3: Subcorpus sizes by genre

Genre	documents	tokens	% corpus
news	1 507 101	584 037 607	18 %
encyclopedia	1 080 862	510 102 047	16 %
legal	165 224	87 684 930	2.7 %
blog	57 846	36 407 663	1.1 %
discussion	24 547	17 370 881	0.53 %

informing the corpus to maximize the mutual completion between the data and the editors, thereby minimizing the editorial effort. Central to this process are two databases: the corpus and the dictionary draft which get mutually updated. The entry components are generated separately according to their dependencies, as illustrated in Figure 1.

After an entry component is generated and post-edited by human editors, the edited data are incorporated into the corpus annotation and used for generating further entry components. For example, having word sense post-edited leads to the introduction of sense identifiers in the corpus, which in turn yields sense-based analysis for a distributional thesaurus or example sentences (which would not be very reliable otherwise).

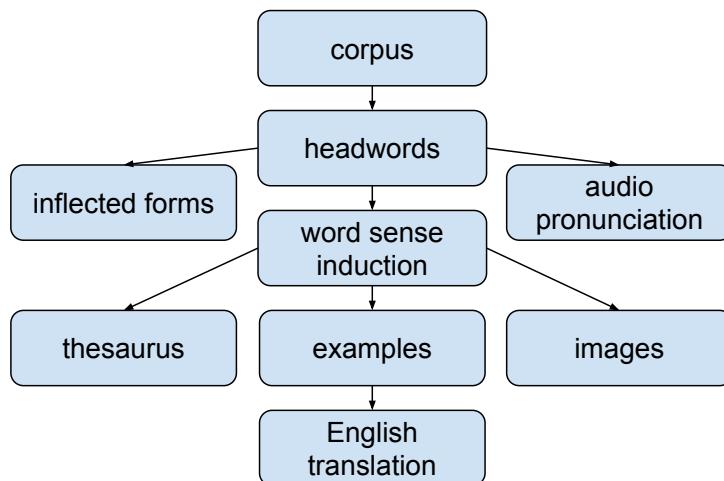


Figure 1: A high-level workflow overview of the post-editing process

In the next sections we explain in detail how we developed a large-scale dictionary with a fraction of human effort compared to the standard setting in which the lexicographers themselves interrogate the corpus. We show the method can rely on existing (imperfect) NLP tools but requires a radical change to the typical lexicographic workflow and a robust data management process between the corpus, the dictionary and the editors.

3.1 Training the Native Speakers

Annotators should be native speakers of the source language, but they are not expected to have any previous lexicographic training. For tasks that involve translation, written capacity in the target language (English) is required. English was also the prevailing language of instruction.

Good training helps annotators understand their tasks well and leads to high-quality output. Each step in the dictionary creation process needs to be clearly explained—containing all relevant information; giving illustrative examples; describing potential conflicts or marginal cases; mentioning the recommended amount of time per entry in each particular task.

Therefore, the training for each task consists of three parts:

1. e-learning describing the task in general, providing English examples, explaining the underlying linguistic concepts, including test questions to verify that the annotator understood the essence of the task
2. half-day face-to-face training where we explain the whole task with real Ukrainian examples and language-specific issues
3. manual of 2-3 pages with the necessary instructions

Most of the time, annotators work using the Lexonomy on-line dictionary building tool (Měchura, 2017; Jakubíček et al., 2018). We have developed a dedicated user interface (customized entry editor) in Lexonomy for each task.

3.2 Headwords

The annotator sees a list of headword candidates (i.e. combinations of lemma and part of speech) and their task is to assign a flag to each according to its perceived correctness. Flagging can be performed with the mouse, but using keyboard shortcuts is preferred. Available flags are given short English names and color codes. The key to attributing flags to headword candidates, reproduced here as Figure 2, is shown to the editor all the time.

After familiarizing themselves with the concepts of *lemma* and *part of speech* and having learned about specifics of handling them in Ukrainian and in the applied tagger, annotators train by using the key to flag headword candidates.

In this project, a total of 119,615 headword candidates were evaluated, 87% of which received at least two annotations and 24% were annotated at least three times. Multiple annotations are taken to create a margin for detecting errors and conflicts of opinion. Eight annotators took part in the annotation effort, the work was split into 289 batches and in total 285,177 annotations were made.

The most frequently assigned flag was “ok” (38.4%), followed by “not a lemma” (25.9%) and “wrong POS” (21.2%), then came “proper name” (5.1%) and “I don’t know” (5.0%), later “non-standard (register or spelling)” (2.7%), and at last “not Ukrainian” (1.6%).

Total time annotators spent on this task was 2114 hours, i.e. one annotation took on average 27 seconds. Speed varied greatly between annotators, ranging from 12 seconds

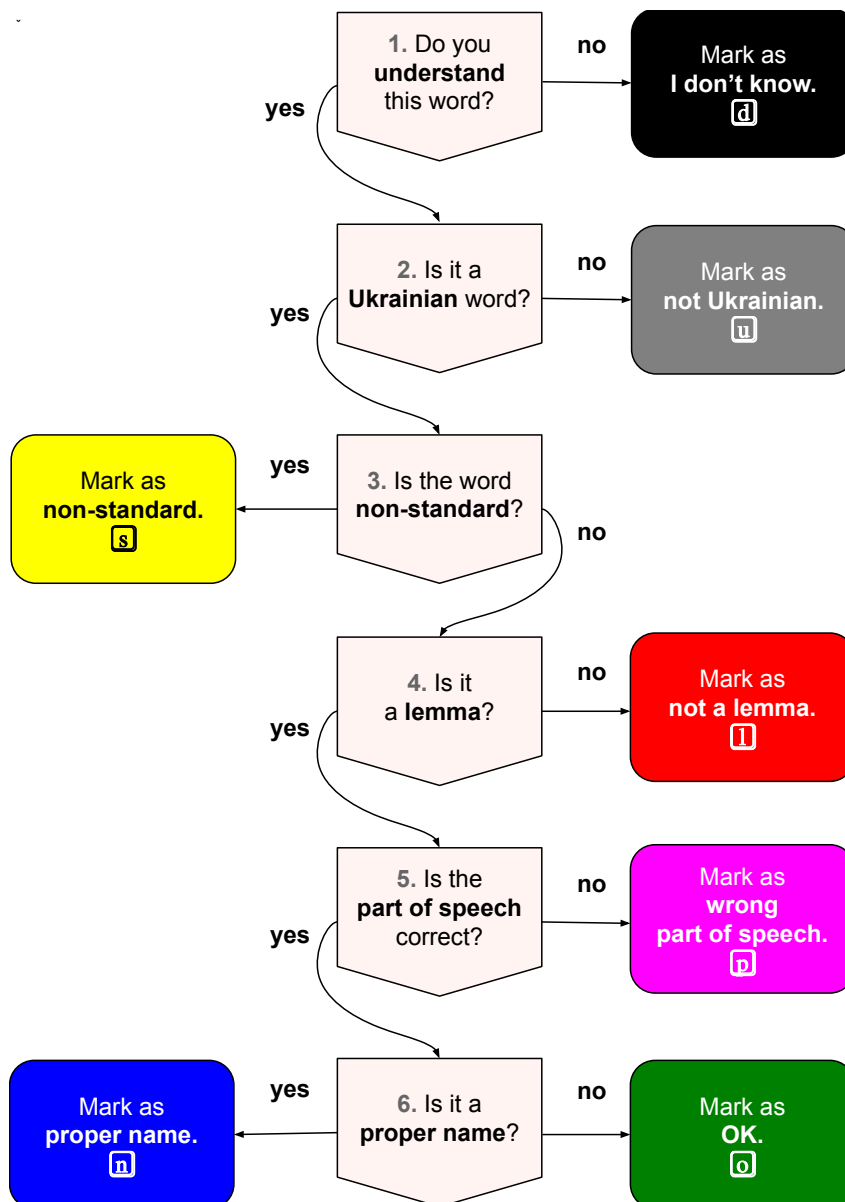


Figure 2: Key to attributing flags to headword candidates, color-coded and with keyboard shortcuts

to 64 seconds per annotation, influenced by factors such as annotator’s self-confidence, computer skills (use of clicking vs. pressing keys), reliance on external resources, work habits or tiredness.

Out of the presented headword candidates, 49,131 (41%) were eventually accepted as correct headwords into the final dictionary. Major contributor of noise in the input data was inter-POS homonymy, produced by early versions of the used tagger from before we managed to reduce it by integrating a larger morphological database. If only lemmas are counted, 66% of the candidates made it into the dictionary. The lempos to lemma ratio has decreased from 1.45 among the headword candidates to only 1.02 among the accepted headwords. Low homonymy between parts of speech was expected since it is a strong property of Slavic languages.

3.3 Headword Revision


In Headword Revision, annotators get the chance to review headword candidates that were rejected in the Headwords task but could be turned into correct headwords. For each such rejected headword, Lexonomy displays a form in the right-hand pane (see Figure 3), whose exact content varies depending on what is signalled to be an issue with the headword (e.g. not a lemma, wrong part of speech, non-standard spelling).

For instance, if only part of speech is believed to be wrong, then the lemma field is pre-filled and the annotator is asked to select a different part of speech from a dropdown. However, they still have the option to modify the lemma as well, at their own discretion. For cases of ambiguity, it is possible to enter multiple revisions for a headword. The annotator can also decide that the headword be ignored (without revision), or accepted as is (call it correct). Due to the decisive character of this task, it should be commissioned with priority to annotators with high proficiency in the language and good performance in the Headwords task.

ПОЛІКЛІНІК PoS: noun

CORRECT HEADWORD SHOULD BE:

lemma: PoS: proper name?



I DO NOT UNDERSTAND THIS WORD

THIS IS NOT A UKRAINIAN WORD

THIS HEADWORD IS CORRECT

EXAMPLES

1. Хоча керівництво **поліклініки** й надалі переконує – внески добровільні.
2. Зусиллями міської влади і депутатів басейн повернули на баланс **поліклініки** .
3. Його буде встановлено на першому поверсі хірургічного корпусу **поліклініки** .
4. Один випадок був зафіксований й на території однієї з **поліклінік** міста .
5. Проведення медоглядів у **поліклініках** у присутності батьків є логічним.

Figure 3: Interface for the Headword Revision task

In this project, 54,503 headword candidates were sent for revision. Some of them eventually underwent revision more than once (in order to explore inter-annotator agreement), what resulted in 5,820 duplicate entries (though with possibly differing annotations). Four annotators contributed to this task, which was split into 66 batches.

To make an annotation, the user clicks a radio button. If the headword is to be corrected, then they also enter the correct lemma, pick the correct part of speech and indicate whether it is a proper name.

In 94.9% of cases, a revision was resolved by providing an alternative headword. In 3.2%, annotators said that the displayed headword was in fact correct. The remaining 1.9% were cases of unrecognized words or words considered non-Ukrainian. In the typical situation when correct headwords were provided to replace an incorrect headword, in 91.6% there was just one replacement headword, in 7.4% there were two and in 1.0% three or more (up to six).

Total time annotators spent on this task was 722 hours, i.e. annotating one entry took on average 43 seconds. Speed fluctuated a lot across annotators in this task too, with the fastest person taking just 28 seconds per entry and the slowest one needing 77 seconds.

3.4 Word Forms

The Forms task is concerned with inflection. Ukrainian is an inflected language and we want to collect as many inflected forms of each headword (lemma) in the corpus as possible. Annotators are first trained to distinguish inflection from derivation. Then, in Lexonomy, their task is to tell apart correct and incorrect items in a list of possible inflected forms for each headword. A link to concordance is available for case of doubt, but in practice, most items are resolved swiftly. “Correct” is the default, so the annotator needs to act only in case of incorrect forms. This task has a threshold only slightly higher than the Headwords task, it can be introduced quite early in the process and no other later tasks depend on it, which makes it a universal task for times of delay etc.

держати (verb)  I DON'T KNOW

Inflected forms:























	form	correct?	
1.	держати	=headword	
2.	держити	 	
3.	держало	 	
4.	держ	 	
5.	держимо	 	
6.	держатиме	 	
7.	держатимуть	 	
8.	Держать	 	

Figure 4: Interface for the Forms task

In this project, word forms were sought for 42,694 headwords, for which there was a total of 578,327 form candidates (i.e. an average of 13.5 form candidates per headword). Among the form candidates, nearly all (99.2%) only appeared with a single headword. This means that the task was not as much about checking the form-lemma relationship, rather than about checking the correctness and acceptability of the form itself (the used tagger is permissive and accepts even some archaic and corrupted word forms). All seven annotators available at the time were made to work on this task. The work was divided into 43 batches.

The observed ratio of reported incorrect forms was 21.6%. In almost four out of five such cases (79.4%), the rejected form candidates started with a capital letter – and, for lemmas which start with a small letter themselves, such word forms differing in letter case are highly unlikely in Ukrainian. In fact, 77.4% of all capitalized word forms ended up marked incorrect.

Annotators spent 1269 hours checking the word forms, which means that they took on average 107 seconds per headword, or 8 seconds per word form. The fastest annotator needed only 2.5 seconds per word form (or 35 seconds per headword), while the slowest required 22 seconds per word form (or 328 seconds per headword). Explanation of these inter-annotator differences must be looked for in the same factors as mentioned with the Headwords task.

We did not make any automatic judgments on the correctness of words forms, but we benefited from the large corpus to extract a rather satisfying list of them – both in terms of precision (we have shown above that majority of the presented candidates were correct) and recall (although we did not attempted to quantify it, as we explicitly did not aim at acquiring a “full” word form list, whatever it should mean). The average number of unique word forms per lemma was 18.0 for verbs, 13.1 for adjectives, 9.4 for pronouns, 7.3 for nouns, 5.4 for numerals, and below 1.3 for other parts of speech (uninflected). Depending on details of the used processing pipeline, orthographic or phonetically motivated variants of words may have been represented either as “inflected forms” or as separate headwords.

3.5 Audio Recordings

Instead of relying on phonetic transcriptions to indicate pronunciation or on the traditional stress marks to indicate word stress, we make an audio recording of the headword’s pronunciation by a native speaker and store it as a part of the dictionary entry. This is the only part of the entry creation process that is done fully manually, since we want to be in control of the quality of the result and automatic text-to-speech output could not be post-edited. However, apart from having to face a few challenges such as preserving a steady loudness or maintaining a low noise level, it turned out to be also one of the simplest tasks. This is also the only step that does not use Lexonomy, but a specially developed audio-recording software, and the only step which necessitates physical presence of the annotator in dedicated premises (a soundproof audio cabin with high-quality recording hardware) during the whole work time.

In this project, we recorded audio pronunciation for all the 55,632 headwords in the final dictionary. Some of the headwords were recorded multiple times and, due to the recording occurring in parallel with the rest of the dictionary building, we also made recordings of

some headwords which eventually did not make it into the final dictionary. In total, 57,800 audio files were created (i.e. 3.9% overhead). The work was divided into 60 batches, 59 of which were assigned to the same annotator so that same voice is used throughout the dictionary. Only the last batch (1.3% of headwords) was assigned to a different person because the original speaker was not available anymore.

The recording station in the audio booth was controlled with a special small 6-key keyboard (the available keys were marked with pictograms meaning YES, NO, SKIP, DOWN, UP, QUIT, respectively). This was done to save desk space else occupied by a regular keyboard, concentrate all controls in a single location, reduce the chance of typos, limit noise generated by keystrokes and improve user comfort for the annotator. The processing of each headword consists of seeing it displayed on the screen, recording its pronunciation, then listening to the recording to check its quality, and possible re-recording if the quality is not sufficient.

It took the annotators 553 hours, or about 14 weeks (of 40 work hours each), to make the recordings. That means an average of 36 seconds per headword. This time, however, includes regular break time, because it is demanding if not impossible for a non-trained person to stay concentrated in a small booth and keep speaking using a fresh voice of stable strength for the whole day. In fact, in most of the cases when a headword had to be recorded repeatedly, the reason for this was the software stepping in with an automatic low-volume alert.

3.6 Word Senses

Identification of word senses for each headword is an important step in the dictionary building process, because all subsequent tasks are performed on sense level instead of headword level, and therefore dependent on the word-sense distinctions made here. After annotators learn that there is not a single perfect solution for the problem (Kilgarriff, 1997), reaching common ground with regard to granularity of sense distinctions is attempted by means of joint practice and discussions on each other's proposed solutions.

Annotators' invention is effectively limited to automatically induced word sense data (read more in 4.2.5), represented in Lexonomy as *example usages* (i.e. collocations, each including a longest-commonest match (Kilgarriff et al., 2015)) and grouped into clusters, each of which could be considered a word sense candidate. Having reviewed this data, however, the annotator has the freedom to establish a number of senses of their choice, to distribute the collocations among them freely, not to assign a collocation to any particular sense (by marking it either as "mixed sense" or "error") and even come up with a sense not linked to any of the collocations (the latter is allowed so that no important word senses are lost due to possible deficiencies of the word sense induction algorithm). Each sense is also given a disambiguating gloss (in the language of the dictionary), one or more English translations, and a mark saying whether it is offensive in meaning.

The Word Senses task might be the most difficult task to be properly trained, and the quality of its outcome directly influences the quality of data in all upcoming tasks. In this project, 10,098 post-edited word sense disambiguations were performed in this way, for a total of 10,016 distinct headwords. In each processed entry, there were on average 43 collocations, divided into 9 clusters. Four annotators were chosen and trained for this task and the work was divided into 55 batches.

In terms of part of speech, 45.2% of the annotated headwords were nouns, 24.7% adjectives, 21.6% verbs and 5.8% adverbs; the remaining 2.7% were other parts of speech, for which word sense disambiguation is not always applicable. Figure 5 shows an example of one cluster, with three collocations. Annotators assign collocations to senses by clicking numbered buttons. The available buttons multiply as soon as more senses are declared – which is done by providing a disambiguating gloss, English translation(s) and possibly switching a toggle to mark offensiveness. To reflect real-world conditions, English translations can be shared by multiple senses, again by means of numbered buttons, thus reducing the need for typing. And when a collocate is not self-explaining, the annotator has the option to view a corresponding concordance in the corpus.

Group 1				
Mark all: <input type="button" value="1"/> <input type="button" value="NEW"/> <input type="button" value="MIXED"/> <input type="button" value="ERROR"/>				
example usage	actions	collocate	relation to headword	concordance
<i>бродіння відбуватися</i>	<input checked="" type="button" value="1"/> <input type="button" value="NEW"/> <input type="button" value="MIXED"/> <input type="button" value="ERROR"/>	бродіння NOUN	"відбуватися" молочнокисль ...	🔗
<i>відбувається масаж внутрішніх органів</i>	<input checked="" type="button" value="1"/> <input type="button" value="NEW"/> <input type="button" value="MIXED"/> <input type="button" value="ERROR"/>	орган NOUN	"відбуватися" масаж ...	🔗
<i>заміщення відбуватися</i>	<input checked="" type="button" value="1"/> <input type="button" value="NEW"/> <input type="button" value="MIXED"/> <input type="button" value="ERROR"/>	заміщення NOUN	"відбуватися" поступовий ...	🔗

Figure 5: Part of the interface for the Word Senses interface

For 60.1% of headwords, a single sense was identified; 18.5% was split into two senses; 10.5% into three senses; 5.0% into four senses; the remaining 5.8% into five senses or more. Overall average number of senses identified for the processed headwords was 1.84. Among annotators, the average number of identified senses reached from 1.38 to 2.31. The highest number of senses (not necessarily an ideal) was routinely found by an annotator who happened to have some formal education in the field of linguistics. The same annotator was also the only one who would, exceptionally, go to great lengths by establishing more than 10 senses for a headword.

Annotators spent a total of 1203 hours on the Word Senses task, i.e. about 7 minutes per headword. Three of the annotators had very close averages (from 7.3 to 9.1 minutes per headword), only the fourth annotator (the one with linguistic education) differed substantially when she took a much lower average of 4.5 minutes per headword.

Of the listed collocations, only 1.8% were declared incorrect or incomprehensible, and 1.5% could not be conclusively attributed to particular sense (when there were more senses to choose from). Remaining collocations were either all attributed to a single sense, or distributed among two senses (in the average ratio of 79:21) or three senses (67:23:10). Even with four senses, the least-frequent one still corresponded to approximately three collocations, which indicates that even in highly competitive situations, all senses were solidly backed up by corpus data (in contrast with senses defined without any corpus evidence, which are disregarded in this computation).

Annotators entered a total of 26,715 English translations (usually single words, but sometimes multi-word expressions, and exceptionally even descriptions of concepts that lack a direct English translation), which means an average of 2.65 translations per headword. This is close to the average number of pre-generated machine translations of the headword into English, which was 2.45.

Only a tiny fraction (25, 0.1%) of the identified senses has been marked offensive, although the annotators were aware of this possibility and each of them used it at least once. We believe that more of the headwords could be used in an offensive or derogatory way and suspect that the annotators may have under-annotated them under the influence of the previous tasks, in which we had to repeatedly stress that also bad words are to be included in the dictionary and that they should be treated *as any other word*.

3.7 Thesaurus

In the Thesaurus task, annotators are trained to evaluate thesaurus candidates (i.e. selected related headwords, read more in 4.2.6) for a given headword *in one of its senses* (this subdivision into senses is maintained across the rest of the dictionary building). Each thesaurus item can be put into one of three categories: Synonym, Antonym and Similar word (i.e. not a synonym or antonym, but still somehow related). A fourth option, named Other, is the default choice and results in the candidate being discarded.

океан NOUN I DON'T KNOW

translations: **ocean**

thesaurus candidates:

	candidate	type			
1.	море <small>NOUN</small>	synonym	antonym	similar	other
2.	затока <small>NOUN</small>	synonym	antonym	similar	other
3.	озеро <small>NOUN</small>	synonym	antonym	similar	other
4.	річка <small>NOUN</small>	synonym	antonym	similar	other
5.	пустеля <small>NOUN</small>	synonym	antonym	similar	other
6.	ріка <small>NOUN</small>	synonym	antonym	similar	other
7.	гора <small>NOUN</small>	synonym	antonym	similar	other

Figure 6: Interface for the Thesaurus task

In this project, two annotators were assigned to the Thesaurus task and, at the time of writing, they had processed in total 10,377 entries (headwords in individual senses), divided into 12 batches. Each entry contained exactly 20 thesaurus candidates.

Out of all thesaurus candidates, three fourths (75.5%) were discarded (marked as Other), while 15.0% were accepted as Similar, 8.2% were classified as Synonyms and 1.3% as Antonyms. In the training phase, we realized that one annotator had developed a preference towards marking many *related* words as Similar, while the other preferred Synonym in these cases. During data inspection, we found out that Similar:Synonym ratio was 83:17

for the first annotator and 61:39 for the second one. We could, however, not find solid grounds on which we could convince one or the other to change their preference. The percentage of identified antonyms was consistently low with both annotators.

Work on the Thesaurus took 364 hours, with one of the annotators being significantly slower (527 seconds per entry) than the other (74 seconds).⁵

On average, 4.9 thesaurus candidates were accepted for each headword. Since the candidates were scored by Sketch Engine and shown in that order, we would expect that items higher in the list have a higher chance of being accepted as thesaurus items. And indeed, the probability that a candidate item had been accepted was found to be inversely proportional to the item's rank; it was 48.9% for the first item in the list, 38.1% for the second one, 32.3% for the third one; 19.3% for position ten, 15.6% (minimum) for position fifteen. Positions 16–20 were exceptions, because they had been reserved for top-scored thesaurus candidates for the headword, regardless of sense. These items had a higher chance of being accepted (21.2–27.3%), comparable to that of the (sense-specific) positions 5–9.

3.8 Usage Examples

Choosing a good, easy to understand, illustrative dictionary example for a headword (in one of its senses) is a challenging task. So although GDEX (Rychlý et al., 2008) is used to pre-select candidate sentences (read more in 4.2.7), annotators need to be well trained to choose the best one of the five pre-selected sentences and redact them when necessary (shorten them or remove controversial information). In rare cases, annotators may even be forced to come up with an example sentence of their own (for this purpose, they have on hand a link to the first one hundred GDEX-scored collocation lines from the corpus as source of inspiration), although writing example sentences anew is strongly discouraged for reasons of time expense and authenticity.

In the user interface, the annotator selects their preferred sentence by clicking on a button next to it. Clicking directly on the sentence activates a text input field in which its text can be modified as needed. After an example sentence is selected, it changes color from red to green and another text field opens below it, pre-filled with machine translation of the original sentence into English. It is the annotator's responsibility to check and fix the English translation as needed and to make sure that the sentences in the two languages stay in sync.

Four annotators were trained in this activity and 20 batches were finished at the time of writing this paper. In those batches, the annotators processed a total of 14,474 entries, each containing five pre-selected and pre-translated example sentences. The work took them 693 hours, which averages to 2.9 minutes per entry. The average time spent on an entry varied greatly across the annotators (0.8 minutes, 4.3 minutes, 6.0 minutes, 14.7 minutes). The differences are likely to have been caused by each annotator's differently strong criteria for a good example. Slower annotators edited their chosen examples more heavily, often fully rewriting them because they thought it necessary.

It seems that the position of the five offered sentences in the list (they were order by decreasing GDEX score) correctly reflected their quality, or at least that sentences closer

⁵ Due to the charitable dimension of the project, the work with annotators had defects which would not be tolerated in a fully commercial setup.

examples:

1.

Термін дії проміжних нарядів не повинен перевищувати терміну дії загального наряду.

NO YES
2.

Це гарантує високу міцність і тривалий термін служби.

NO YES

translation
3.

Термін дії візи буде точно відповідати тривалості навчання.

NO YES

Figure 7: Interface for the Examples task

to the top attracted annotators' attention more and would be more probably chosen in case if multiple comparably good candidates were present. The chance of the sentence in position one to be chosen was 34.7%; position two 18.9%; position three 15.0%; position four 12.6%; position five 12.3%.

The average length of the chosen example in its original (from the corpus) and accepted (possibly modified) form was 63.1 and 56.5 characters (8.7 and 7.8 words), respectively, which suggests a welcome tendency of the annotators to produce shorter examples. The same tendency was found also with regard to the length of the sentence's English translation (decrease from 67.6 to 60.8 characters; from 11.3 to 10.2 words). Evaluation of Levenshtein distance (minimum number of insertions, deletions, and substitutions) between the generated and post-edited Ukrainian sentences reveals that 67.3% of the 13,449 studied sentences did not need any modifications at all, and the average edit distance was 12.6 on the whole set (and 38.5 just on those sentences which needed modification).

The pre-generated machine translations of the original Ukrainian sentences into English and their final forms (often updated both for language and for linguistic deficiencies in the Ukrainian originals) differed more, as expected, but not substantially: the edit distance was 15.9 (and 34.6 on just the modified sentences, which is even a decrease). Also, surprisingly, 54.1% of the machine-translated sentences were considered good enough by the annotators to be left intact. This seems to suggest that the machine translation is reliable and saves time during annotation. Indeed, in cases when the Ukrainian sentence was left unmodified, 76.0% machine translations were also not modified; and other 7.7% only required up to 5 edits (insertions, deletions or substitutions) to be performed in order to fix the English sentence. The average edit distance of English sentences in these cases of unmodified Ukrainian sentences was 2.6 (or 10.6 just on the modified sentences).

3.9 Images

The Images task was not yet administrated at the time of writing this paper, but we foresee using an interface similar to the one depicted in (Baisa et al., 2019). Freely licensed images relevant to the headwords will be identified and a top list will be offered to annotators to choose from.

3.10 Final Review

Final Review is the last phase of the dictionary building process. In it, a complete dictionary entry is composed out of the collected components (see entry structure in 5.1) and visualized for the first time. The annotator’s task is to fix any typos and mistakes and to check the overall coherence of the entry.

For instance, senses (however well-defined) are perceived differently across annotators, who may produce translations, usage examples and images that are not fully compatible with each other. In Final Review, a skilled annotator has the last say and can modify or delete entry components to achieve coherence. Addition of information, however, is discouraged at this step. Final-reviewed entries have got their definite form (in terms of data management, not visualization), in which they will appear in the final dictionary.

4. Data Management

Baisa et al. (2019) reported on issues with data management. Although the paper itself is not very specific about this issue, we have learned from the authors that the issues were connected to the fact that the XML annotations from all the phases described above were exported from and imported into one centralized database. Once an annotation was imported into the database, it could not be easily changed and re-imported, because the entries could have been changed by following imports. The approach would be probably working fine if all the annotations and import/export processes were perfect and consistent, however, that was not the case. Every inconsistency in annotation and all the small bugs in the automatic import/export procedures, propagated and resulted in a decent amount of entries containing inconsistent information which must have been manually corrected, generating delays and additional costs. Moreover, as new versions of the source corpus were produced (e.g. due to improvements of lemmatization and tagging), some parts of the data became inconsistent with the corpus.

Therefore, our approach to the data management is different. We take the source corpus and the native speaker annotations as source data for fully automated procedure that creates respective dictionary parts, merges them into the complete dictionary and generates new data for annotation. The procedure is implemented as a Makefile which makes it easy to define dependencies among the individual components, and is illustrated in Figure 8. In case of any change (new annotations available, new version of the source corpus, ...) all the data are re-processed, new versions of partial dictionaries and derived corpora are created, and a new version of the dictionary is automatically generated. Also, new data for annotation, if needed, are created.

This approach gives us the flexibility to fix any problem or inconsistency in the source data, or in the manual annotations, that previously passed unnoticed, and re-generate

the whole dictionary easily. The fully automated procedure therefore enforces consistency across all the pieces of data involved in the process. Also, it can be used instantly for a new corpus and a new language.

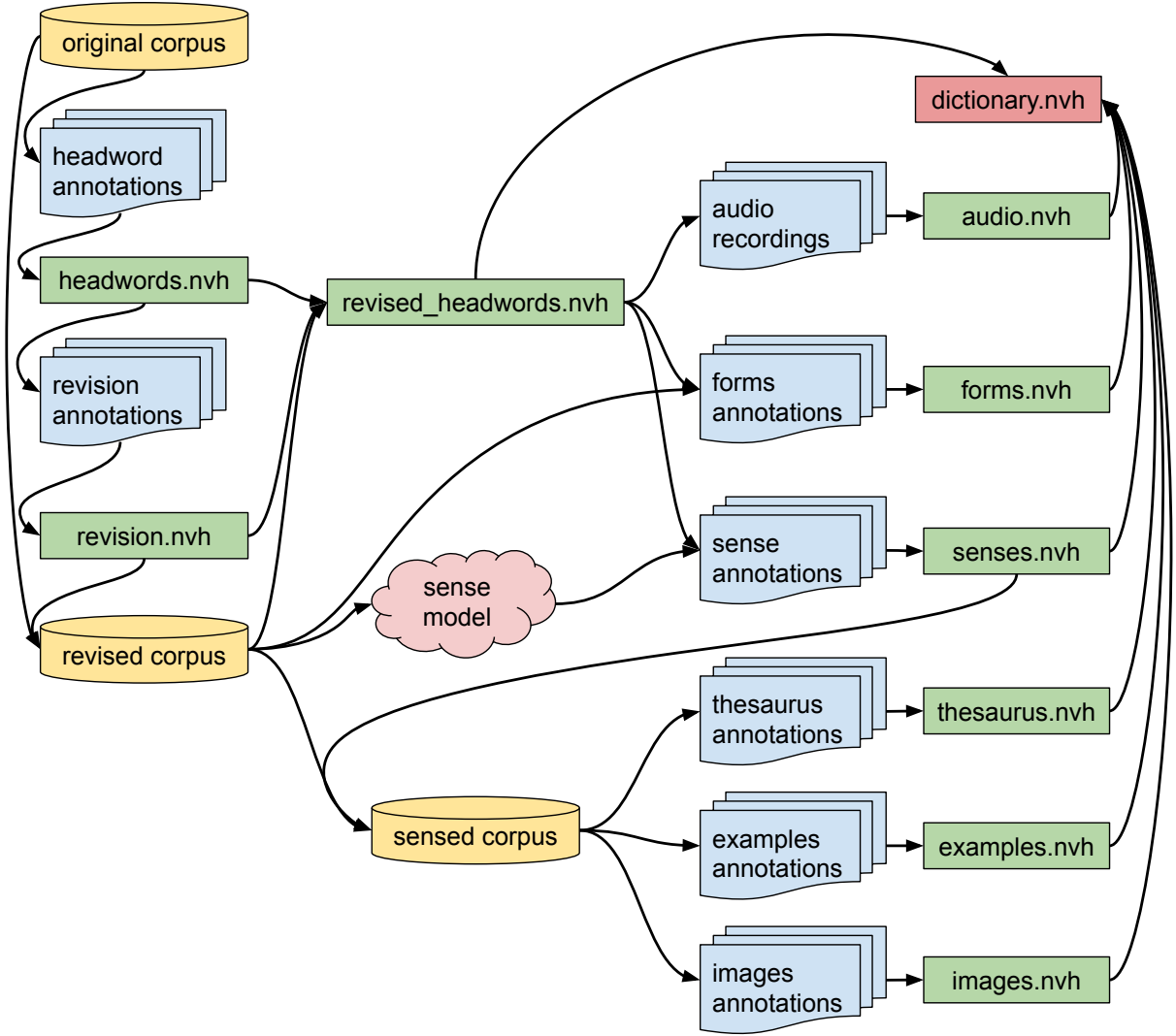


Figure 8: Illustration of the data management process

4.1 Formats

For the partial dictionaries (green rectangles in Figure 8) as well as for the resulting dictionary, we used the NVH – name-value hierarchy – format (Jakubíček et al., 2022), a text format easily readable for both humans and simple automatic text processing tools, which is suitable for dictionary data and significantly less complex than XML.

For the manual annotations (blue “documents”), XML was used as the internal format of the Lexonomy software where the annotators worked.

4.2 Generating the Dictionary

In this part we describe the automatic procedure more in detail. In Figure 8, every shape represents a target in a Makefile, and the arrows represent the dependencies among the particular targets. Typically, there is one Python script (or a few calls of the standard UNIX tools) for each of the targets, which generates the target contents from its dependencies.⁶

For clarity, we have split the description into parts, but please bear in mind that all the content of this section is one fully automatic process that runs as a whole over partial data, and can be repeated as many times as needed.

4.2.1 Headwords

At the very beginning, there is a source corpus, tagged and lemmatized automatically, using available software tools. The first step of the procedure takes the word list of lemmoses (lemmas with a one-letter part-of-speech suffix) from the corpus and generates annotation batches (“headword annotations” in Figure 8) for the N most frequent words (by document frequency). In this project, a total of 102,323 lemmoses received 2 annotations from different annotators, and if they were not in agreement, further annotations were collected until there was at least 50% agreement.⁷

From the headword annotations, a partial dictionary `headwords.nvh` is generated, containing lemmos, its annotations, final decision and the percentage of agreement, for each of the headwords.

4.2.2 Headword Revision

If the final decision about a headword was *wrong_lemma*, *wrong_pos* or *non_standard*, a *revision* annotation was generated – a next step whose purpose was to fix mistakes of the automatic lemmatizer and tagger and find correct (or standard, respectively) lemmas and parts of speech of the words. Most of the items sent to revision were revised to a word that we already had in the dictionary, but we also obtained 6,177 words that we had not seen before and the dictionary would miss them if the *revision* step was not incorporated.⁸

The outputs of the revision annotations are merged into a partial dictionary `revision.nvh` which records the corrections. This dictionary is then used in two ways:

- Using the recorded revisions and the original corpus, we create a *revised corpus* that contains correct lemmas and parts of speech, and is used as a base for further

⁶ The Figure 8 is slightly simplified. In the real Makefile, there are few more targets of rather technical nature that would split some of the arrows into two. However, they are not important for understanding the principle of the procedure, so we don't discuss them here for the sake of clarity.

⁷ However, this is something we may want to change in the future because collecting too many annotations slightly complicated the task and led to a delay. For the next projects, we would recommend collecting 2 annotations only, and continuing directly to the *revision* step in case of disagreement.

⁸ There is still a theoretical possibility of missing an important word: if the lemma of a frequent word form was ambiguous and the tagger always returned one of the options and never the other; however, we did not encounter such a situation in the project. Also, this problem would be present with any approach based on a list of lemmas from a corpus.

processing, namely word senses. (If we did not take this step, the word sense model would not contain the 6,177 new words at all, and the data for other words would be incomplete.)

- We merge it with `headwords.nvh` and create `revised_headwords.nvh` which contains a final list of headwords for the dictionary together with frequencies and frequency ranks generated from the revised corpus. The next phases do not add more words into the dictionary, they just add more information to words that are already present.

4.2.3 Word Forms

For each of the valid words in `revised_headwords.nvh`, we generate a list of word forms present in the revised corpus into the word form annotation batches. The annotators mark them as correct or wrong and the correct word forms are then exported into `forms.nvh` which is later merged into the final dictionary.

4.2.4 Audio Recordings

Audio batches for recording are generated for all the valid words in `revised_headwords.nvh`. After recording, the audio files are kept separately and the metadata containing information about the location of the particular audio file, are compiled into `audio.nvh` which is then merged into the final dictionary.

4.2.5 Word Senses

From the revised corpus, we generate an automatic model of word senses for all words in the corpus. At first, we used traditional collocation-based approach described in Herman et al. (2019), but the result would frequently miss high frequency senses. The overall quality of the result was not sufficient and significant post-editing effort was necessary to extract useful information. For this reason, we switched to a word sense induction model based on Bartunov et al. (2016), which represents the senses of a word as word embeddings. Then, we map the senses from the model onto (some of) the collocations from word sketch, clustering the collocations. Each cluster of collocations is then considered a candidate sense. From these clusters of collocations, we generate sense annotation batches and ask the annotators to name, fix and translate the automatically identified senses, as discussed in 3.6.

These annotations are then processed into another partial dictionary `senses.nvh` that records the division of each word into senses, the collocations assigned to the particular senses, and the names and translations of the senses. Apart from being an input for the final dictionary, this partial dictionary is used to generate a *sensed corpus* from the revised corpus, where the basic unit of analysis is not a lemma (lempos) anymore, but a *sense*. In our particular implementation, a sense is a lempos concatenated with the sense name, e.g. *bank-n#river* vs. *bank-n#money*, but the exact string is not important, it could as well be *bank-n#1* and *bank-n#2*. The important moment is that now we can work with separate senses instead of lemmas (lemposes)—namely compile word sketches and thesaurus for senses so that word sketch and thesaurus for *bank-n#river* is different from word sketch and thesaurus for *bank-n#money*.

4.2.6 Thesaurus

For each *sense* recorded in `senses.nvh`, a list of similar words (and similar *senses*) is pulled from the *sensed corpus* using Sketch Engine’s thesaurus function (Rychlý & Kilgarriff, 2007) and put into thesaurus annotation batches. Because not all of the occurrences are clustered into senses, we merge thesaurus for the sense with the thesaurus of the (more general) lemma to get more quality data. The results of the annotation are again compiled into a partial dictionary `thesaurus.nvh`.

4.2.7 Usage Examples

For each sense recorded in `senses.nvh`, we generate a set of 5 best candidate example sentences from the corpus with the GDEX tool (Rychlý et al., 2008). For this purpose, a new Ukrainian-specific GDEX configuration was created. The candidate sentences are then automatically translated into English by the DeepL API⁹ and annotation batches are created from the extracted sentences and their automatic translations. The annotators are then asked to read all the sentences, select one best example, edit it (but only if needed) and check and edit (again, only if needed) its automatic translation into English.

The annotations are then processed into a partial dictionary `examples.nvh` which is then merged into the final dictionary.

4.2.8 Images

The images phase of the project is not yet finished at the time of writing this paper, but we intend to implement it in the same frame as the previous phases: automatically search for copyright-free images in several databases, based on English translations for each sense, let the annotators select one best image out of 10, and record the selections in a partial dictionary `images.nvh`.

5. About the Dictionary

So far we discussed the process of compiling the Ukrainian dictionary. This section summarizes some basic information about the resulting dictionary itself.

5.1 Entry structure

The entry structure of the dictionary may be clear from the description of the methods above—however, the following description shows it explicitly:

- **Headword (lemma + part of speech)** is the basic identification of every entry. It is also the primary key of the dictionary in the database sense—we don’t allow multiple entries with the same lemma and part of speech.
- **Flag** specifies the type of the entry: in the final dictionary, we have only *ok*, *name* and *non_standard* but we also keep all the rejected words with the other flags in

⁹ deepl.com

a separate database. *Non_standard* and *name* entries do not contain senses, and *non_standard* also contains a link to the standard form of the headword.

- **Frequency** of the word retrieved automatically from the document frequency in the corpus, i.e. number of the documents the headword occurred in.
- **Rank** of the headword according to the frequency (computed automatically from the frequency).
- **Pronunciation**, or precisely the location of the audio recording with the pronunciation, the output of the audio recording phase.
- **List of word forms**, the output of the word forms post-editing phase.
- **List of senses** identified in the sense annotation phase. Only words marked *ok* have senses and translations. Next, every sense contains:
 - **Sense identifier** or disambiguator which tells the senses apart and may explain them to an extent (but it is neither definition or explanation of the sense). It may be empty if the word is found monosemous (has only a single sense recorded in the dictionary).
 - **One or more translations to English**, as recorded in the sense annotation phase.
 - **Collocations** sorted by grammatical relations, as recorded in the sense annotation phase. Each collocation also contains a short example (typically 3-5 words) automatically extracted from the corpus.
 - **List of synonyms, antonyms and similar words**, as identified in the thesaurus annotation phase.
 - **One usage example** and its translation to English, both results of the example annotation phase.
 - **Image**, if appropriate, selected in the images selection step (not implemented yet). Every image consists of its location, source and license.

The structure is rather shallow, but we believe it contains the most important elements for a decent dictionary entry. Also, the modular nature of the process makes it possible to add further steps easily, such as definitions/explanations or translation into more languages.

In this dictionary, we did not take multi-words into account—but there are already tools available to identify multi-words from corpus n-grams and collocations that would make it relatively easy to enrich the dictionary in this direction.

5.2 Basic statistics

For organizational and budget reasons, we did not complete all the entries all the way through, some of them are “more complete” than others. A relatively long list of valid frequent headwords and word forms is a valuable multi-purpose resource, so we aimed at having a really long list of headwords first, and then continued with the other phases step by step, always starting with the most frequent headwords.

By the time of writing this paper, the project is still not finished, but mainly for budget reasons we slowed it down and now the work continues with only one remaining Ukrainian editor. This means the numbers below are not final but they reflect the state after less than 1 year of intensive work during which **6,918 hours** of manual post-editing work (or approximately 3.5 full-time person-years) were consumed. See Figure 10 for a breakdown by task.

зуб

зуб

зубець

зубка

зубний

зубожілий

зубожіння

зубожіти

зубок

зубопротезування

зубочистка

зубр

зубчастий

зубчатий

зубчик

зуб NOUN ★☆☆☆

rank: 3 776

Inflected forms

зубів, зуби, зубами, зуба, зуб, зубах, зубом, зубам, зубі, зубу, зубові

1 анатомічний

In English

tooth

Synonyms

ікло, моляр

Similar

коронка, протез

Examples

Існує один дуже хороший народний метод відбілювання зубів.
There is one very good folk method of teeth whitening.

2 механічний

In English

tine

Examples

Основні елементи циліндричного зубчастого колеса з прямим зубом.
The main elements of a spur gear with a straight tooth.

3 озброєний до зубів

In English

armed to teeth

Examples

Сюди ми прийшли на катамарані, озброєні до зубів.
We came here on a catamaran armed to the teeth.

Figure 9: Dictionary entry example for the word зуб (tooth). There are 7 senses of the word in total, here we show only the first three of them.

Overall, our database contains 123,574 annotated headword candidates (i.e. all the headwords from the corpus seen by at least one annotator). This figure includes the revised headwords that were originally not present in the corpus—without them, it is 117,397. Of these, 14,141 were only seen by one annotator (better than nothing but not reliable

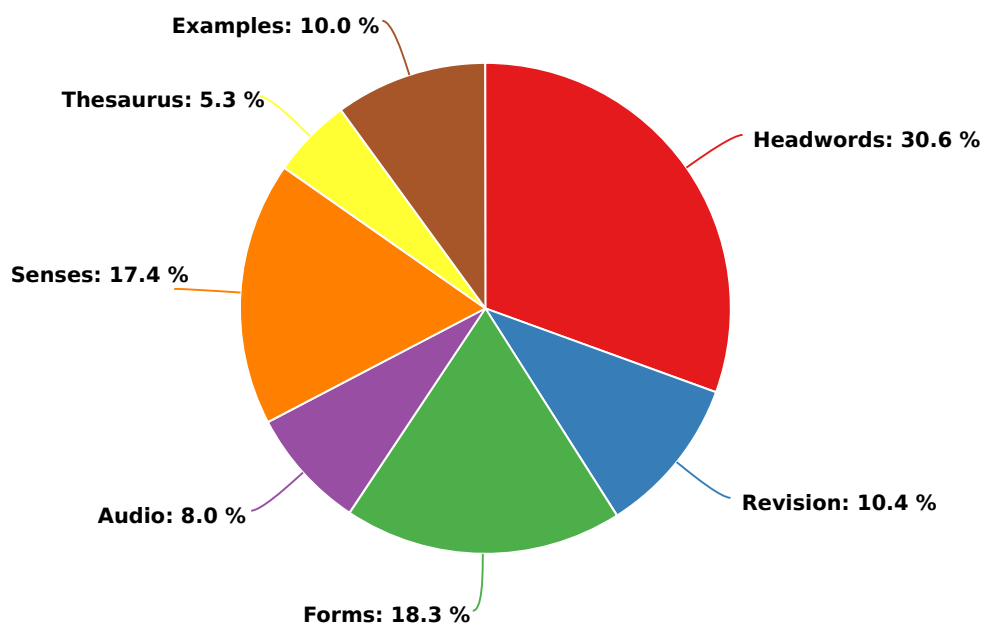


Figure 10: Workload by task (100% = 6,918 hours)

enough for the dictionary) which leaves us with 109,433 headword candidates with reliable annotation.

Of these, 55,632 ended with flags suitable for the final dictionary, namely:

- 46,987 common words (marked *ok*)
- 8,252 proper names
- 393 non-standard words

So we can say that the size of our dictionary is **55,632 entries**. All of these entries contain an audio recording of the pronunciation, as well as frequency and rank derived automatically from the corpus.

42,639 of these headwords contain list of their word forms which is in total 453,010 validated word forms.

The size of the dictionary in terms of complete entries, i.e. entries with verified senses, translations, thesaurus and usage example, is **9,785**. (We still plan to add images in the near future.) Of these, 3,901 entries are polysemous and 5,884 are monosemous. 1,057 words have more than three senses. Total number of senses in the dictionary is 17,973.

In all the process phases, we always proceeded according to document frequency. In other words, we went through the 109,433 most frequent words in the corpus, the dictionary contains the 55,632 most frequent Ukrainian words (according to the corpus) and we have complete entries with senses for the 9,785 most frequent words.

6. Conclusions

We have reported on a rapid corpus-based development of a new Ukrainian-English dictionary using a new process of automatic generating and step-by-step post-editing of

the dictionary. We described building the source corpus, then we went through all phases of the process in detail and explained our approach to dictionary data management during the process.

The resulting dictionary contains ca. 10,000 finished entries; another 45,000 entries for less frequent headwords are partly finished. Overall the process consumed less than 7,000 hours of paid editor's time which is a fraction of both time and money needed to build a similar dictionary in a traditional way with professional lexicographers.

In the future, we will continue working on the dictionary (2–5,000 more finished entries, adding images), and since we made the workflow setup really easy within this project, we are looking forward to running similar projects with new languages soon.

7. Acknowledgements

We cordially thank the Institute for Ukrainian (<https://mova.institute>) for permission to use their manually annotated corpus available through the Universal Dependencies project (https://github.com/UniversalDependencies/UD_Ukrainian-IU). We cordially thank Andriy Rysin, Vasyl Starko and the BrUK team for permission to use the Ukrainian morphological database they developed and made available at https://github.com/brown-uk/dict_uk. This work has been partly supported by the Ministry of Education of CR within the LINDAT-CLARIAH-CZ project LM2023062. This publication was written with the support of the Specific University Research provided by the Ministry of Education, Youth and Sports of the Czech Republic.

8. References

- Baisa, V., Blahuš, M., Cukr, M., Herman, O., Jakubíček, M., Kovář, V., Medved, M., Měchura, M., Rychlý, P. & Suchomel, V. (2019). Automating dictionary production: a Tagalog-English-Korean dictionary from scratch. In *Proceedings of the 6th Biennial Conference on Electronic Lexicography*. Brno, Czech Republic: Lexical Computing CZ s.r.o., pp. 805–818. URL https://elex.link/elex2019/wp-content/uploads/2019/10/eLex-2019_Proceedings.pdf.
- Bartunov, S., Kondrashkin, D., Osokin, A. & Vetrov, D. (2016). Breaking sticks and ambiguities with adaptive skip-gram. In *artificial intelligence and statistics*. PMLR, pp. 130–138.
- 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. URL <https://doi.org/10.1093/ijl/ecw014>. <https://academic.oup.com/ijl/article-pdf/29/2/200/7199846/ecw014.pdf>.
- Herman, O., Jakubíček, M., Rychlý, P. & Kovář, V. (2019). Word Sense Induction Using Word Sketches. In *Statistical Language and Speech Processing: 7th International Conference, SLSP 2019, Ljubljana, Slovenia, October 14–16, 2019, Proceedings 7*. Springer, pp. 83–91.
- Jakubíček, M., Kilgarrieff, 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/>.
- Jakubíček, M., Kovář, V., Měchura, M. & Rambousek, A. (2022). Using NVH as a Backbone Format in the Lexonomy Dictionary Editor. In A.R. Aleš Horák Pavel Rychlý

- (ed.) *Proceedings of the Sixteenth Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2022*. Brno: Tribun EU, pp. 55–61. URL <https://raslan2022.nlp-consulting.net/>.
- Jakubíček, M., Měchura, M., Kovář, V. & Rychlý, P. (2018). Practical Post-editing Lexicography with Lexonomy and Sketch Engine. In *The XVIII EURALEX International Congress*. p. 65.
- Jongejan, B. & Dalianis, H. (2009). Automatic training of lemmatization rules that handle morphological changes in pre-, in- and suffixes alike. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. Suntec, Singapore: Association for Computational Linguistics, pp. 145–153. URL <https://aclanthology.org/P09-1017>.
- Kallas, J., Kilgarriff, A., Koppel, K., Kudritski, E., Langemets, M., Michelfeit, J., Tuulik, M. & Viks, Ü. (2015). Automatic generation of the Estonian Collocations: Dictionary database. In *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*. Trojina, Institute for Applied Slovene Studies, pp. 1–20.
- Kilgarriff, A. (1997). I don't believe in word senses. *Computers and the Humanities*, 31, pp. 91–113.
- Kilgarriff, A., Baisa, V., Rychlý, P. & Jakubíček, M. (2015). Longest–commonest Match. In *Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference*. Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd ..., pp. 11–13.
- Kosem, I., Koppel, K., Zingano Kuhn, T., Michelfeit, J. & Tiberius, C. (2018). Identification and automatic extraction of good dictionary examples: the case(s) of GDEX. *International Journal of Lexicography*, 32(2), pp. 119–137. URL <https://doi.org/10.1093/ijl/ecy014>. <https://academic.oup.com/ijl/article-pdf/32/2/119/28858872/ecy014.pdf>.
- 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*. pp. 19–21.
- Pomikálek, J. (2011). *Removing boilerplate and duplicate content from web corpora*. Ph.D. thesis, Masaryk University.
- Pomikálek, J. & Suchomel, V. (2011). chared: Character Encoding Detection with a Known Language. In *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2011*. pp. 125–129.
- 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., Husák, M., Kilgarriff, A., Rundell, M. & McAdam, K. (2008). GDEX: Automatically finding good dictionary examples in a corpus. In *Proceedings of the XIII EURALEX International Congress*. Barcelona: Institut Universitari de Lingüística Aplicada, pp. 425–432.
- 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)*. Manchester, UK:

- Coling 2008 Organizing Committee, pp. 777–784. URL <https://aclanthology.org/C08-1098>.
- Shvedova, M. (2020). The General Regionally Annotated Corpus of Ukrainian (GRAC, uacorporus.org): Architecture and Functionality. In V. Lytvyn, V. Vysotska, T. Hamon, N. Grabar, N. Sharonova, O. Cherednichenko & O. Kanishcheva (eds.) *Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2020)*. Lviv, Ukraine, April 23-24, 2020, volume 2604 of *CEUR Workshop Proceedings*. CEUR-WS.org, pp. 489–506. URL <https://ceur-ws.org/Vol-2604/paper36.pdf>.
- Starko, V. (2021). Implementing Semantic Annotation in a Ukrainian Corpus. In N. Sharonova et al. (eds.) *Proceedings of the 5th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2021)*. Kharkiv, Ukraine, April 22-23, 2021, volume 2870 of *CEUR Workshop Proceedings*. CEUR-WS.org, pp. 435–447. URL <https://ceur-ws.org/Vol-2870/paper32.pdf>.
- Starko, V. & Rysin, A. (2020). *Velykij elektronnyj slovnyk ukrayins'koyi movy (VESUM) yak zasib NLP dlya ukrayins'koyi movy*. Seriya "Ne vse splyva rikoyu chasu...". Vydavnychyj dim Dmytra Buraho. URL https://www.researchgate.net/profile/Vasyl-Starko/publication/344842033_Velikij_elektronnij_slovník_ukrainiskoi_movi_VESUM_ak_zasib_NLP_dla_ukrainiskoi_movi_Galaktika_Slova_Galini_Makarivni_Gnatuk/links/5fa110cd458515b7cfb5cc97/Velikij-elektronnij-slovník-ukrainiskoi-movi-VESUM-ak-zasib-NLP-dla-ukrainiskoi-movi-Galaktika-Slova-Galini-Makarivni-Gnatuk.pdf.
- Starko, V. & Rysin, A. (2023). Creating a POS Gold Standard Corpus of Modern Ukrainian. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*. Dubrovnik, Croatia: Association for Computational Linguistics, pp. 91–95. URL <https://aclanthology.org/2023.unlp-1.11>.
- Suchomel, V. & Kraus, J. (2021). Website Properties in Relation to the Quality of Text Extracted for Web Corpora. In *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2021*. pp. 167–175. URL <https://nlp.fi.muni.cz/raslan/2021/paper19.pdf>.
- Suchomel, V. & Kraus, J. (2022). Semi-Manual Annotation of Topics and Genres in Web Corpora, The Cheap and Fast Way. In *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2022*. pp. 141–148. URL <https://nlp.fi.muni.cz/raslan/2021/paper22.pdf>.
- Suchomel, V. & Pomikálek, J. (2012). Efficient Web Crawling for Large Text Corpora. In S.S. Adam Kilgarriff (ed.) *Proceedings of the seventh Web as Corpus Workshop (WAC7)*. Lyon, pp. 39–43. URL <http://sigwac.org.uk/raw-attachment/wiki/WAC7/wac7-proc.pdf>.

Adding Information to Multiword Terms in Wiktionary

Thierry Declerck¹, Lenka Bajčetić², Gilles Sérasset³

¹ DFKI GmbH, Multilingual Technologies, Saarland Informatics Campus D3-2, D-66123 Saarbrücken, Germany

² Innovation Center of the School of Electrical Engineering in Belgrade, Bulevar kralja Aleksandra 73, 11000 Belgrade, Serbia

³ Université Grenoble Alpes, CNRS, Grenoble INP*, LIG, 38000 Grenoble, France
E-mail: declerck@dfki.de, lenka.bajcetic@ic.etf.ac.bg.rs, gilles.serasset@imag.fr

Abstract

We describe ongoing work dealing with the potential “auto-enrichment” of “Multiword terms” (MWTs) that are included in the English edition of Wiktionary. The idea is to use and combine information contained in the lexical components of the MWTs and to propagate this extracted and filtered information into the lexical description of the MWTs, as those are typically equipped with less lexical information as it is the case for their lexical components. We started our work with the generation of pronunciation information for such MWTs, on the base of the pronunciation information available for their components. We present in this paper first achievements but also issues we encountered. Addressing those issues lead us to consider additional resources for supporting our approach, like DBnary and WikiPron. This step was ultimately leading to suggestions of adaptations for those additional resources, which, in the case of DBnary, are already implemented. We are currently extending our approach to a morphosyntactic and semantic enrichment of the English MWTs in Wiktionary.

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

1. Introduction

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

We started our work with the generation of pronunciation information, as we noticed that a vast majority of English MWTs in Wiktionary are lacking this type of information. While designing a potential evaluation dataset for the pronunciations generated by our approach, we noticed that only around 3% of MWTs are carrying pronunciation information. We also discovered that other complex lexical constructions (affix + word, or word + affix) are often lacking pronunciation information. We collected for the evaluation dataset 6,768 MWT entries with pronunciation (compared with 252,082 MWT entries that are lacking such information). Our approach for generating pronunciation information for MWTs consisted in combining the pronunciation information included in the lexical description of their components. Results of this work can be integrated in other lexical resources, like the Open English WordNet (McCrae et al., 2020),¹ where pronunciation information has been added for now only for single word entries, as described in (Declerck et al., 2020a).

¹ See also <https://en-word.net/>

A specific issue emerged for the generation of pronunciation information for MWTs that contain (at least) one heteronym.² For this type of lexical entry a specific processing is needed, disambiguating between the different senses of the heteronym for extracting the appropriate pronunciation of this one lexical component to be selected to form the overall pronunciation of the MWT. An example of such a case is given by the Wiktionary entry “acoustic bass”, for which our algorithm has to specify that the pronunciation /beɪs/ (and not /bæs/) has to be selected and combined with /əˈkuːstɪk/. It is important to mention that Wiktionary often lists several pronunciations for various variants of English. In this work we focus on the standard, received pronunciation for English, as encoded by the International Phonetic Alphabet (IPA).³

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

In this paper, we describe first briefly the way multiword terms (MWTs) are introduced in Wiktionary. We summarize then the various approaches we followed for both designing an evaluation dataset and generating pronunciation information, dealing for now with the English edition of Wiktionary. We discuss issues we encountered, and which lead to the consultation of related resources, like DBnary (Sérasset & Tchechmedjiev, 2014; Sérasset, 2015) and WikiPron (Lee et al., 2020). While the cooperation with DBnary has been already established and resulted in improvements of our approach and an adaptation of DBnary itself, which we describe in some details, we are starting with the formulation of suggestions for adaptation for WikiPron. We present our first step towards the enrichment of MWTs with morphosyntactic and semantic information extracted from their components. We close the paper with conclusive remarks and presenting future work.

2. Wiktionary

Wiktionary⁴ is a freely available web-based multilingual dictionary. Like other Wikimedia⁵ supported initiatives, it is a collaborative project. This means that there might be inaccuracies in the resource, but the editing system is helping in mitigating this risk. The coverage and variety of lexical information is much larger than any single curated resource, while Wiktionary is integrating information from expert-based dictionary resources, when their licensing conditions allow it. Nastase and Strapparava (2015) gave some details on the quality (and quantity) of information included in the English Wiktionary edition, also in comparison with WordNet.⁶

² The online Oxford Dictionary gives this definition: “A heteronym is one of two or more words that have the same spelling but different meanings and pronunciation, for example ‘tear’ meaning ‘rip’ and ‘tear’ meaning ‘liquid from the eye’” <https://www.oxfordlearnersdictionaries.com/definition/english/heteronym>, [accessed 20.04.2023.]

³ See <https://www.internationalphoneticalphabet.org/ipa-sounds/ipa-chart-with-sounds/>

⁴ <https://en.wiktionary.org/>

⁵ <https://www.wikimedia.org/>

⁶ See Fellbaum (1998) and <http://wordnetweb.princeton.edu/perl/webwn> for the on-line version of Princeton WordNet.

Wiktionary includes, among others, a thesaurus, a rhyme guide, phrase books, language statistics and extensive appendices. Wiktionary’s information also (partly) includes etymologies, pronunciations, sample quotations, synonyms, antonyms and translations.⁷ Wiktionary developed categorization practices which classify an entry along the lines of linguistics (for example “developed terms by language”) but also topical information (for example “en:Percooid fish”). So that the entry “sea bass” is categorized, among others, both as an instance of “English multiword terms” and of “en:Percooid fish”.⁸

3. Multiword Terms in Wiktionary

The version of the English edition of Wiktionary we worked with is listing 159,169 English multiword terms,⁹ and 75,646 expressions are categorized as “English terms with IPA pronunciation”.¹⁰ This is quite a small number in comparison to the whole English Wiktionary, which has over 8,633,770 pages (among those, 7,387,538 are classified as content pages¹¹). When we analyse these figures, we need to be aware that they are representing the number of pages categorized as a particular category, and a Wiktionary page can often contain several lexical entries, although this is typically not the case for MWTs. Also, it is important to keep in mind that the English Wiktionary contains a lot of terms which are not English. We can see the exact number of Wiktionary pages classified as English lemmas if we look at the category itself.¹² The actual number of 714,732 means that a little over 10% of English lemmas have pronunciation, while approximately 22% of all English lemmas belong in the MWT category. There is clearly a gap that needs to be filled when it comes to pronunciation information in Wiktionary. While introducing pronunciation for the remaining (non MWTs) 90% of lemmas seems like it has to be a manual task (or semi-automatic, using other lexical resources) - we have investigated ways to produce the missing pronunciation for numerous MWTs.

4. A first Approach

We designed a computer program to extract from the Wiktionary XML dumps¹³ the pronunciation information from the component words and to combine them for the corresponding MWT, limiting our work to MWTs with two component words, which are building a majority of the cases, and which are well described in Wiktionary, with clear links to pages containing their component parts, while MWTs having more components are more poorly represented in Wiktionary.

This way, we can straightforwardly create a huge amount of pronunciation information that we can add to English MWTs included in Wiktionary. However, there is this issue concerning the cases in which a MWT is containing at least one heteronym. As the Wiktionary entry of the MWT is pointing back for its lexical components to Wiktionary pages (which often contain more than one lexical entry), but not to the specific entry

⁷ See <https://en.wikipedia.org/wiki/Wiktionary> for more details.

⁸ The categorization system is described at <https://en.wiktionary.org/wiki/Wiktionary:Categorization>

⁹ https://en.wiktionary.org/wiki/Category:English_multiword_terms [accessed 20.04.2023.]

¹⁰ https://en.wiktionary.org/wiki/Category:English_terms_with_IPA_pronunciation [accessed 20.04.2023.]

¹¹ See <https://en.wiktionary.org/wiki/Special:Statistics> [accessed 20.04.2023]

¹² https://en.wiktionary.org/wiki/Category:English_lemmas [accessed 20.04.2023.]

¹³ See <https://dumps.wikimedia.org/> for more details

with the specific sense, we needed to adapt our approach and go into a deeper parsing of Wiktionary, adding complexity to our program. This point let us consider the use of already existing tools that are extracting information from Wiktionary, two of those tools - DBnary and WikiPron - being described in Section 1.2, and in Sections 6 and 7 respectively.

This was particularly relevant for the design of an evaluation dataset, as for this we had to query the category system of Wiktionary, which is not included in the available XML dumps. We had thus to make use of the Wiktionary API, which is a RESTful interface that allows programmers to access the data contained in the Wiktionary dictionary through standard HTTP requests. It may be used to query for definitions, translations, links or categories of a specific Wiktionary page. In our cases, we planned to use it to query each page for its categories. This would be an easy way to go if the size of English edition of Wiktionary was not so massive: more than 8.6 million entries need to be checked, which means 8.6 million requests sent to Wiktionary API. This is quite slow and if not done correctly will lead to being blacklisted from the Wiktionary website. Using this approach, we have extracted over 98% of MWTs from Wiktionary and compiled a list of 153,525 multiword terms without IPA, and a gold standard of 4,979 MWTs with IPA - we can see that only about 3% of MWTs have pronunciation information in Wiktionary. However, this approach was very time-consuming and can only be applied to a specific version of Wiktionary. Hence, as the Wiktionary data is always growing, new MWTs introduced in Wiktionary will not benefit from this work. This is the reason why we tried to reproduce our experiment using the DBnary dataset, which is regularly updated. The move to DBnary offered us some more MWTs with IPA pronunciation included in Wiktionary, resulting in the (current) total number of 6,768 MWT entries with pronunciation.

This work was needed in order to build an evaluation dataset. We aim at an “internal” evaluation of our approach, as a number of MWTs in Wiktionary are in fact equipped with pronunciation information, like “sea bass” (in the IPA encoding /'si:bæs/), so that we can compare our pronunciation extraction applied to “sea” and “bass” and see if it yields the correct pronunciation from the heteronym “bass”. We encountered in this context a number of Wiktionary-related issues . One issue being, that in some cases suprasegmental information is encoded in the IPA transcription of either the component(s) or in the IPA transcription associated with the MWT, so that a proper string matching approach can not be implemented. Another issue being that sometimes syllable boundaries are marked, and sometimes not. And in some cases, the IPA transcription associated with the MWT in Wiktionary is just concatenating the two IPA codes, while in other cases, a blank is introduced between the two IPA codes. We have also some issues related to the regional encodings, as sometimes we have only the US IPA code or the UK IPA code. Last but not least, sometimes two alternative IPA transcriptions are given for a single word entry, while only one is present in the IPA transcription of the corresponding MWT entry. Those issues also lead us to consider for the building of the evaluation dataset the use of the WikiPron resource, which is described in Sections 1.2 and 7.

5. Related Work

Wiktionary is often used as a source for various text-to-speech or speech-to-text models. For instance, the work of Schlippe et al. (2010) developed a system which automatically extracts phonetic notations in IPA from Wiktionary to use for automatic speech recognition. A more

recent example is the work by Peters et al. (2017) which is aimed at improving grapheme-to-phoneme conversion by utilizing Wiktionary. Grapheme-to-phoneme is necessary for text-to-speech and automatic speech recognition systems.

A recent tool is WikiPron (Lee et al., 2020), which is an open-source command-line tool for extracting pronunciation data from Wiktionary. It stores the extracted word/pronunciation pairs in TSV format.¹⁴ We observe that no Wiktionary multiword terms are included in those lists. Also, no (semantic) disambiguation is provided and, for example, the word “lead” is listed twice, with the different pronunciations, but with no sense information, as WikiPron is providing solely word/pronunciation pairs. Results of our work consisting in generating pronunciation information to multiword terms, while taking into consideration heteronyms, could thus be included in WikiPron directly or via Wiktionary updates. But in its actual form, WikiPron can be re-used for our purposes, as it harmonizes phonemic pronunciation data across various Wiktionary language editions, while the pronunciations are segmented, and stress and syllable boundary markers can be on request removed. Especially the latter is relevant for our work, as it will ease future evaluation work (see the issues described in Section 4). This dataset and its relevance for our work, while also discussing some shortcomings, are described in more details in Section 7.

BabelNet (Navigli & Ponzetto, 2010)¹⁵ is one of the resources that is integrating Wiktionary data,¹⁶ with a focus on sense information, in order to support, among others, word sense disambiguation and tasks dealing with word similarity and sense clustering (Camacho-Collados et al., 2016). The result of our work could be relevant for BabelNet, as the audio files displayed by BabelNet are not based on the reading of pronunciation alphabets but on external text-to-speech systems, which are leading to errors, as can be seen in the case of the heteronym “lead”, for which BabelNet offers only one pronunciation.¹⁷

A very relevant resource for our approach is DBnary (Sérasset & Tchechmedjiev, 2014; Sérasset, 2015).¹⁸ DBnary extracts different types of information from Wiktionary (covering 23 languages) and represents it in a structured format, which is compliant to the guidelines of the Linguistic Linked Open Data framework.¹⁹ In the DBnary representation of Wiktionary we find lexical entries (including words, multi word expressions (MWEs) or affixes, but without marking those sub-classes of lexical entries explicitly, an issue that has been fixed in new release of DBnary, as this is requested for continuing our approach in the context of DBnary), their pronunciation (if available in Wiktionary), their sense(s) (definitions in Wiktionary), example sentences and DBnary glosses, which are offering a kind of “topic” for the (disambiguated) entries, but those glosses are not extracted from the category system of Wiktionary. They are taken from available information used to denote the lexical sense of the source of the translation of an entry from English to other languages.

¹⁴ As of today, more than 3 million word/pronunciation pairs from more than 165 languages. Corresponding files are available at <https://github.com/CUNY-CL/wikipron/tree/master/data>.

¹⁵ See also <https://babelnet.org/>.

¹⁶ As far as we are aware of, BabelNet integrates only the English edition of Wiktionary, including all the languages covered by this edition.

¹⁷ See the audio file associated with the two different senses of the entry for “lead”: <https://babelnet.org/synset?id=bn%3A00006915n&orig=lead&lang=EN> and <https://babelnet.org/synset?id=bn%3A00050340n&orig=lead&lang=EN>.

¹⁸ See <http://kaiko.getalp.org/about-dbnary/> for the current state of development of DBnary.

¹⁹ See Declerck et al. (2020b) and <http://www.linguistic-lod.org/>.

DBnary does not include categorial information from Wiktionary, and also did not offer support for dealing with MWTs lacking pronunciation information and that contain (at least) one heteronym. Therefore, we still need(ed) to access and consult Wiktionary directly, using methods that are described in Section 4, also for designing the dataset for evaluating our work (MWTs in Wiktionary that are carrying pronunciation information). Hence, our results can also be integrated in DBnary, directly or via the updated Wiktionary entries. In fact, our work lead to the adaptation of DBnary, as this is briefly described in Section 6.

6. Cooperation with DBnary

DBnary is representing the lexical information extracted from Wiktionary using the Linked Open Data (LOD) principles²⁰ and as such it is using RDF²¹ as its representation model. It is freely available and may be either downloaded or directly queried on the internet. DBnary uses the *OntoLex-Lemon* standard vocabulary (Cimiano et al., 2016),²² displayed in Figure 1 to represent the lexical entries structures, along with *lexvo* (de Melo, 2015) to uniquely identify languages, *lexinfo* (Cimiano et al., 2011)²³ and *Olia* (Chiarcos & Sukhareva, 2015)²⁴ for linguistic data categories.

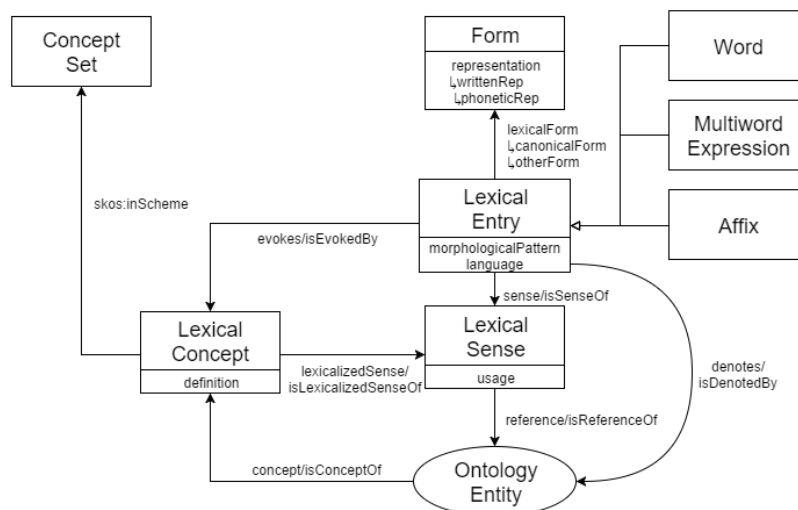


Figure 1: The core module OntoLex-Lemon. Taken from <https://www.w3.org/2016/05/ontolex/#core>

While trying to reproduce with the DBnary engine the work briefly described in Section 4, we noticed that DBnary was lacking some information. First, Wiktionary multiword terms were not marked explicitly. Second, derivation relations between single word lexical entries and MWTs, in which they occur, were not extracted, while this information is

²⁰ See <https://www.w3.org/wiki/LinkedData> for more information on those principles

²¹ The Resource Description Framework (RDF) model is a graph based model for the representation of data and metadata, using URIs to represent resources (nodes) and properties (edges).

²² See also the specification document at <https://www.w3.org/2016/05/ontolex/>.

²³ The latest version of the lexinfo ontology can be downloaded at <https://lexinfo.net/>.

²⁴ The “Ontologies of Linguistic Annotation (OLiA)” is available at <https://acoli-repo.github.io/olia/>.

crucial for the disambiguation of components of MWTs that are heteronyms. The DBnary maintainer²⁵ tuned the extraction program to fix these identified lacks.

These missing elements were added and are now available in versions starting from February 2023. The extraction program now correctly *types* English Wiktionary entries either as `ontolex:Word` or as `ontolex:MultiWordExpression` (for the MWTs). Moreover, derivation relations are now extracted and available in the graph using `dbnary:derivesFrom` transitive property.

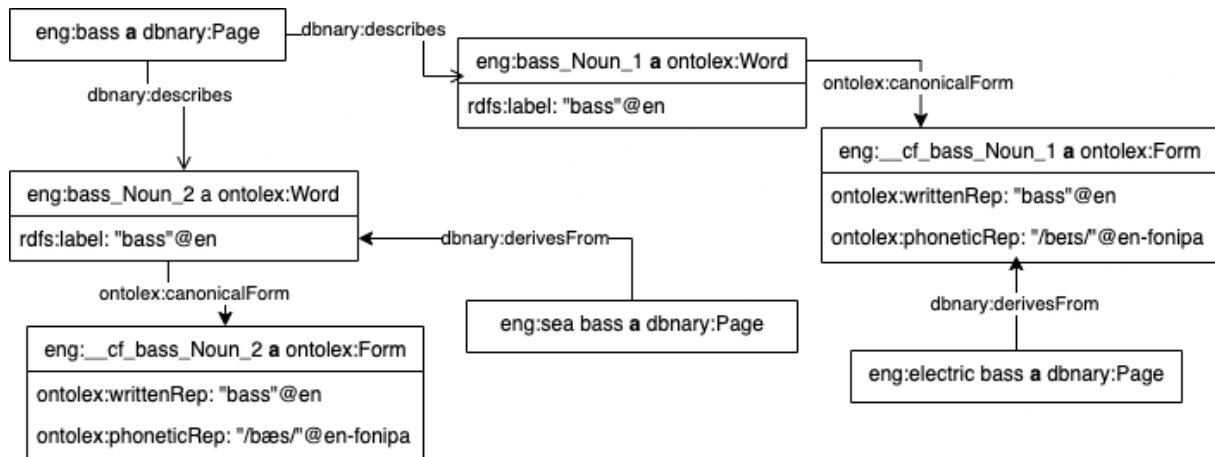


Figure 2: A very small extract of the DBnary graph showing the DBnary page *bass* and 2 of the lexical entries it describes (*bass_Noun_1* [sound, music, instrument] and *bass_Noun_2* [perch, fish]) and their respective canonical forms. The pages *sea bass* and *electric bass* are also represented with their derivation relations.

Figure 2 shows an example of the organisation of two heteronym lexical entries described by the same page, along with their canonical forms (with written and phonetic representations). Figure 2 also shows how the derivation relation is modelled in DBnary, using the transitive `dbnary:derivesFrom` property. It must be noted that in Wiktionary original data, the derivation links point to Wiktionary pages but not to Wiktionary entries, hence, the DBnary modelling reflects this as it is usually difficult to automatically choose which lexical entry (or entries) is (are) the valid target of the derivation relation. But, applying the property in the inverse direction (could be named `dbnary:derivesTo`), the subject/source of the relation is a lexical entry within a Wiktionary page, pointing to a MWT page. As MWT pages consist mainly of only one lexical entry, we can precisely establish a “subterm” relation between a single lexical entry and the MWTs it occurs in, combining if needed both “directions” of use of the property. This point is very important, as it allows projecting all the lexical information of the single lexical entry to the component it builds within a MWT, as this is briefly presented in Section 8.

²⁵ The DBnary extraction programs are open source and available at: <https://gitlab.com/gilles.serasset/dbnary/> where issues can be added to ask for correction or enhancement of the extractors. It is also possible to fix the extractors and create a Merge Request.

7. About Wikipron

We were also confronted with issues with the pronunciation information in various language editions of Wiktionary, as sometimes suprasegmental information or syllables boundaries are present and sometimes not, or the fact that sometimes we have only the phonetic IPA transcription, sometimes only the phonemic transcriptions and sometimes both associated with a Wiktionary page and their entries. Those issues are building an obstacle for the creation of a clean evaluation dataset. Searching for help for this, we looked in more details at the WikiPron resource, as it is providing for a differentiated analysis of the extracted pronunciation information from Wiktionary. WikiPron is also proposing a cleaning of certain pronunciation information. The WikiPron data set is being used for example in an investigation on what phonological information is encoded in character embeddings (Boldsen et al., 2022). But contrary to the authors of this study, we would not call Wikipron a “dictionary”, as we discovered certain issues, that would need to be addressed if the resource should be called a “dictionary”, in a lexicographic sense.

A first issue (already discussed above) is the fact that WikiPron does not consider the extraction of pronunciation information associated with Wiktionary MWTs – although we think that the tool could (and should) extract and deliver the word-IPA pairs for those MWTs. But, as in the case of DBnary, this should be an easy “fix” to implement.

A second issue, more significant, is the fact that entries that have more than one IPA transcription are encoded in the word-IPA codes pairs as two different units. So for example, for UK English:

electric ə 'l ɛ k t r ɪ k
electric ɪ 'l ɛ k t r ɪ k

This can give the impression that we are dealing with 2 different lexical entries, as WikiPron represents in the same way the two different pronunciations for “lead”, which is a heteronym and which should thus be considered as having two different lexical entries with different pronunciations **and** meanings:

lead l ɛ d
lead l iː d
lead l i d

whereas the two last pronunciations are variant for the second meaning (in fact, the last pronunciation corresponds to a misspelling of the verb.²⁶ A better TSV representation for both words would be:

electric ə 'l ɛ k t r ɪ k | ɪ 'l ɛ k t r ɪ k
lead l ɛ d
lead l iː d | l i d

We note that this way of presenting those cases of pronunciation information can be easily represented in OntoLex-Lemon, and could therefore be encoded directly in DBnary, contributing to another adaptation of this linked data compliant resource.

²⁶ See https://en.wiktionary.org/wiki/lead#Etymology_3

8. Extending the Approach to the Addition of morphosyntactic and semantic Information to MWTs

In addition to pronunciation creation and enrichment, our work can lead to another improved description of Wiktionary multiword terms (represented in DBnary as instances of the class `ontolex:MultiWordExpression`), as we can (in a next step) also add the disambiguated morphosyntactic and semantic information associated to hypernyms included in MWTs, taking as a departure point the senses used in Wiktionary itself.

As DBnary is making use of the OntoLex-Lemon model, we can take advantage of the availability of its “Decomposition” module,²⁷ which is graphically displayed in Figure 3.

We can observe that the property `decomp:subterm` of the Decomposition module is equivalent to the property `dbnary:derivesFrom`, recently introduced in DBnary, in order to represent the Wiktionary section “Derived terms” (see Figure 2 for comparison). Therefore, we can just map the `rdf:Object` of `dbnary:derivesFrom` to the `rdf:Object` of `decomp:subterm`, while the `rdf:Subject` of `decomp:subterm` is the MWT itself, as can be seen in Listing 20.1.

As a result, the recent adaptations of DBnary allow not only to generate pronunciation information for MWTs contained in the English edition of Wiktionary, but also to add morphosyntactic and semantic information to the components of such MWTs, and to encode this information in such a way that the new data set can be published on the Linguistic Linked Open Data cloud.

```
1 :electric_bass_lex a
2   ontolex:MultiwordExpression ;
3   decomp:subterm eng:electric_Adjective_1 ;
4   decomp:subterm :eng:bass_Noun_1 .
```

Listing 20.1: The (simplified) representation of “electric bass” using the Decomposition module of OntoLex-Lemon, with links to lexical data encoded in DBnary

Using this module, we can explicitly encode the morphosyntactic, semantic and domain information of the components of MWTs, which are only implicitly present in Wiktionary. For our example, we know yet that “electric” has PoS “adjective” (Wiktionary lists also a nominal use of the word) and “bass” the PoS “noun” (Wiktionary lists also adjectival and verbal uses), while semantically disambiguating the components of the MWT (in the full DBnary representation, the “`ontolex:Word`”: “`eng:bass_Noun_1`” is linked to the corresponding instances of “`ontolex:Sense`”. And in fact, we can then link to a corresponding Wikidata entry for “bass guitar” (<https://www.wikidata.org/wiki/Q46185>) and the one for “electricity” (<https://www.wikidata.org/wiki/Q12725>)

9. Conclusion and future Work

We described in this paper ongoing work on computing lexical information for multiword terms (MWTs) included in Wiktionary. While progressing, we were repeatedly confronted

²⁷ The specification of OntoLex-Lemon describes “Decomposition” in those terms: “Decomposition is the process of indicating which elements constitute a multiword or compound lexical entry. The simplest way to do this is by means of the subterm property, which indicates that a lexical entry is a part of another entry. This property allows us to specify which lexical entries a certain compound lexical entry is composed of.”. Taken from <https://www.w3.org/2016/05/ontolex/#decomposition-decomp>

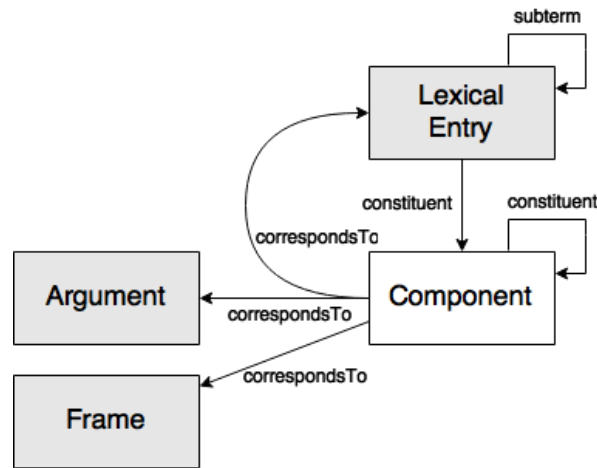


Figure 3: The Decomposition module of OntoLex-Lemon. Taken from <https://www.w3.org/2016/05/ontolex/#decomposition-decomp>

with issues, and we therefore investigated the combined use of other resources resulting from the extraction of information from Wiktionary. We got this way acquainted with the DBnary resource, which is offering a Linked Open Data compliant representation of lexical information extracted from Wiktionary, using at its core the OntoLex-Lemon model and other Semantic Web based vocabularies. As it was immediately clear that using the extraction engine of DBnary is massively easing our work, we teamed with the maintainer of DBnary, who adapted the extraction engine for our needs. Thanks to this cooperation, we discovered also that we can not only generate pronunciation information for MWTs, but that we can also in a straightforward manner extract morphosyntactic and semantic information from the components of MWTs and add those to the lexical description of the MWTs. The enriched information can be encoded in a principled way in OntoLex-Lemon. This will lead to the generation of a new dataset for English MWTs within the Linguistic Linked Data framework. As a result, the DBnary engine is now more than an extractor from Wiktionary and a mapper to an LOD compliant representation, as it generates lexical information that can be used for enriching existing lexical resources.

While confronted with issues related to the precise IPA encoding of pronunciation in Wiktionary, we got acquainted with the WikiPron resource, which is helping us for the building of an evaluation dataset for our pronunciation generation to be associated with MWTs. We also discovered some issues with WikiPron that would need to be addressed, as we want to add elements of this very relevant resource in a lexical framework.

Both DBnary and WikiPron are tools and resources with a large multilingual coverage, a fact that will help us to extend our work to other languages than English.

10. Acknowledgements

The presented work is pursued in the context of the COST Action NexusLinguarum – European network for Web-centered linguistic data science (CA18209), 731015). The DFKI contribution is also pursued in the context of the LT-BRIDGE project, which has received funding from the European Union’s Horizon 2020 Research and Innovation Programme under Grant Agreement No 952194.

We would like to thank the anonymous reviewers for their helpful comments.

11. References

- Boldsen, S., Agirrezabal, M. & Hollenstein, N. (2022). Interpreting Character Embeddings With Perceptual Representations: The Case of Shape, Sound, and Color. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 6819–6836. URL <https://aclanthology.org/2022.acl-long.470>.
- Camacho-Collados, J., Pilehvar, M.T. & Navigli, R. (2016). Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. *Artif. Intell.*, 240, pp. 36–64. URL <https://doi.org/10.1016/j.artint.2016.07.005>.
- Chiarcos, C. & Sukhareva, M. (2015). OLiA – Ontologies of Linguistic Annotation. *Semantic Web*, 6(4), pp. 379–386. Publisher: IOS Press.
- Cimiano, P., Buitelaar, P., McCrae, J. & Sintek, M. (2011). LexInfo: A declarative model for the lexicon-ontology interface. *Journal of Web Semantics*, 9(1), pp. 29–51. URL <https://www.sciencedirect.com/science/article/pii/S1570826810000892>.
- Cimiano, P., McCrae, J. & Buitelaar, P. (2016). Lexicon Model for Ontologies: Community Report, 10 May 2016. Technical report, W3C. URL <https://www.w3.org/2016/05/ontolex/>.
- de Melo, G. (2015). Lexvo.org: Language-Related Information for the Linguistic Linked Data Cloud. *Semantic Web*, 6(4), pp. 393–400.
- Declerck, T., Bajcetic, L. & Siegel, M. (2020a). Adding Pronunciation Information to Wordnets. In *Proceedings of the LREC 2020 Workshop on Multimodal Wordnets (MMW2020)*. Marseille, France: The European Language Resources Association (ELRA), pp. 39–44. URL <https://aclanthology.org/2020.mmw-1.7>.
- Declerck, T., McCrae, J.P., Hartung, M., Gracia, J., Chiarcos, C., Montiel-Ponsoda, E., Cimiano, P., Revenko, A., Saurí, R., Lee, D., Racioppa, S., Abdul Nasir, J., Orlikowski, M., Lanau-Coronas, M., Fäth, C., Rico, M., Elahi, M.F., Khvalchik, M., Gonzalez, M. & Cooney, K. (2020b). Recent Developments for the Linguistic Linked Open Data Infrastructure. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 5660–5667. URL <https://aclanthology.org/2020.lrec-1.695>.
- Fellbaum, C. (ed.) (1998). *WordNet: An Electronic Lexical Database*. Language, Speech, and Communication. Cambridge, MA: MIT Press.
- Lee, J.L., Ashby, L.F., Garza, M.E., Lee-Sikka, Y., Miller, S., Wong, A., McCarthy, A.D. & Gorman, K. (2020). Massively Multilingual Pronunciation Modeling with WikiPron. In *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 4223–4228. URL <https://www.aclweb.org/anthology/2020.lrec-1.521>.
- McCrae, J.P., Rademaker, A., Rudnicka, E. & Bond, F. (2020). English WordNet 2020: Improving and Extending a WordNet for English using an Open-Source Methodology. In *Proceedings of the LREC 2020 Workshop on Multimodal Wordnets (MMW2020)*. Marseille, France: The European Language Resources Association (ELRA), pp. 14–19. URL <https://aclanthology.org/2020.mmw-1.3>.
- Nastase, V. & Strapparava, C. (2015). knoWitiary: A Machine Readable Incarnation of Wiktionary. *Int. J. Comput. Linguistics Appl.*, 6, pp. 61–82.
- Navigli, R. & Ponzetto, S.P. (2010). BabelNet: Building a Very Large Multilingual Semantic Network. In *Proceedings of the 48th Annual Meeting of the Association for*

- Computational Linguistics*. Uppsala, Sweden: Association for Computational Linguistics, pp. 216–225. URL <https://aclanthology.org/P10-1023>.
- Peters, B., Dehdari, J. & van Genabith, J. (2017). Massively Multilingual Neural Grapheme-to-Phoneme Conversion. *CoRR*, abs/1708.01464. URL <http://arxiv.org/abs/1708.01464>.
- Schlippe, T., Ochs, S. & Schultz, T. (2010). Wiktionary as a source for automatic pronunciation extraction. In T. Kobayashi, K. Hirose & S. Nakamura (eds.) *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*. ISCA, pp. 2290–2293. URL http://www.isca-speech.org/archive/interspeech_2010/i10_2290.html.
- Sérasset, G. (2015). DBnary: Wiktionary as a Lemon-Based Multilingual Lexical Resource in RDF. *Semantic Web*, 6, pp. 355–361.
- Sérasset, G. & Tchechmedjiev, A. (2014). Dbnary : Wiktionary as Linked Data for 12 Language Editions with Enhanced Translation Relations. In *3rd Workshop on Linked Data in Linguistics: Multilingual Knowledge Resources and Natural Language Processing*. Reykjavik, Iceland, p. to appear. URL <http://hal.archives-ouvertes.fr/hal-00990876>.

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

Marek Blahuš¹, Michal Cukr¹, Miloš Jakubíček^{1,2},
Vojtěch Kovář^{1,2}, Vít Suchomel^{1,2}

¹Lexical Computing, Brno, Czechia

² Faculty of Informatics, Masaryk University, Brno, Czechia
E-mail: firstname.lastname@sketchengine.eu

Abstract

This paper presents a new generation of terminology extraction grammars for the OneClick Terms system. Unlike previous grammars built using linguistic judgment, the new grammars use rules inspired by patterns frequently observed in existing terminology databases. This evidence-based approach leads to a more accurate coverage of term candidates of lexical structures typical for authentic terms. The internal variety and maximum length of recognized terms have also increased. Due to the use of techniques known from corpus linguistics in their design, the resulting grammars maximize the coverage of terms while keeping a manageable size.

In the paper, we first describe how term grammars are used in OneClick Terms (Baisa et al., 2017) to enable terminology extraction for individual languages. Then we explain the procedure which we use to design next-generation term grammars for seven languages (English, Estonian, French, German, Italian, Portuguese, Spanish). This includes studying the IATE term base (Zorrilla-Agut & Fontenelle, 2019) to derive information on the typical structure of terms in each language. Eventually, we provide figures concerning the new term grammars and their recall of the IATE terms, and we discuss directions for further development.

Keywords: terminology extraction; evidence-based term grammars; OneClick Terms; IATE

1. Introduction

Finding terms in a domain-specific corpus has been a feature of NLP tools for more than a decade (see, e.g., (Aker et al., 2013), (Gojun et al., 2012)). While many of such tools were designed as language-independent, the Sketch Engine corpus management system (Kilgarriff et al., 2014) has pioneered language-aware automatic term extraction for many languages, building on the belief that customization and collaboration with actual speakers of the language lead to higher-quality results (Jakubíček et al., 2014).

Currently, 29 languages are supported in both monolingual or bilingual term extraction. A dedicated web interface called OneClick Terms (Baisa et al., 2017) showcases the essential functionality of Sketch Engine and hides all the background complexity of corpus building, text alignment and the actual term extraction from the eyes of the user. All the user does is upload the document(s) and select the language(s), after which all computation happens seamlessly and the extracted terms are displayed as a result.

For each supported language, OneClick Terms needs a language-specific term grammar. A term grammar is a set of rules which defines the lexical structures, typically noun phrases, which should be included in term extraction. Earliest term grammars for Sketch Engine were prepared for the World Intellectual Property Organisation (WIPO) and typically had the form of a single part-of-speech-based regular expression (e.g. one or more adjectives followed by a noun, for English).

Because these grammars were prepared using linguistic judgment, they could only match term candidates of a limited variety and length. We deem this approach substandard and believe that applying the same principles that are common in corpus linguistics (i.e. statistically exploring large sets of data rather than relying on a linguist’s intrinsic knowledge) would provide higher-quality term grammars. The idea is to observe which lexical structures are common in terminological databases, instead of coming up with a selection of our own. Obviously, not all sequences of tokens of an applicable lexical structure are terms, but the existing term extraction algorithm will take care of distinguishing actual terms from mere term candidates.

In this paper, we describe a new generation of terminology extraction grammars for the OneClick Terms system, which we developed with a strictly empirical approach by studying an existing manually curated terms base, namely the IATE (Interactive Terminology for Europe), created by the Translation Centre for the Bodies of the European Union with terms in 24 languages (Zorrilla-Agut & Fontenelle, 2019). Since the rules are inspired by patterns observed in a terminology database, we call these grammars “evidence-based” term grammars. Our aim is to maximize the term grammar recall of the terms in the terminology database, which serves as the gold standard showing what people actually perceive as terms in the particular language. We have used this new approach to develop evidence-based term grammars for seven languages so far (English, Estonian, French, German, Italian, Portuguese, Spanish) and evaluate them in terms of improvement compared to the existing term grammars and coverage of the terms in IATE.

2. Background

The term extraction in OneClick Terms is based on a corpus-based contrastive technology involving two key steps: (1) applying the rules in the term grammar to a corpus to generate a list of term candidates (2) scoring the candidates by comparing their frequencies in the uploaded document(s) (which form a *focus corpus*) to their frequencies in general language represented by an extensive *reference corpus* (Jakubíček et al., 2014). OneClick Terms uses the corpora of the TenTen Corpus Family (Jakubíček et al. (2013), target size 10^{10} words) as reference. A later extension to the system allows for bilingual terminology extraction from aligned documents (Kovář et al., 2016) based on co-occurrences in aligned segments being ranked using the logDice association score (Rychlý, 2008). Recently, the support for bilingual extraction from non-aligned documents was added (Baisa et al., 2015).

We used Sketch Engine to build a single-purpose *term corpus*, consisting of multi-word terms from the current version (September 2021) of the IATE term base (all domains, all collections, only the “term” term type, any evaluation, any reliability), cleaned by removing any HTML markup (e.g. $\langle i \rangle$), quotation marks, text in brackets, and even full entries if they look like a complex chemical formula (e.g. *6-chloro-N-ethyl-N-(propan-2-yl)-1,3,5-triazine-2,4-diamine*), a list of multiple terms (e.g. *period of driving time, driving*

period) or an incomplete term (e.g. *to inform ... of ...*). Each term is represented as a separate paragraph and the corpus is processed using the standard processing pipeline for the particular language, which includes part-of-speech tagging, lemmatization and morphological annotation.

A report (see Figure 1) is then generated using the Sketch Engine API¹, showing the frequency distribution of part-of-speech combinations (e.g. *adjective + noun*) in the terms (paragraphs) in the term corpus, ordered by descending frequency. For each such combination, a second-level frequency distribution is computed on the morphological tags, revealing that, for instance, in languages with gender agreement, the combination *masculine singular + masculine singular* is much more frequent than *masculine singular + feminine singular*; the latter being either the result of incorrect tagging, or a random grouping of words (if found in a regular running text corpus) which do not form any lexical structure. For each part-of-speech and morphology combination, a list of one hundred random examples of matching terms is displayed to allow for quick inspection during the term grammar design process.

The imposed order within the report makes it possible for the term grammar author to focus on the most frequent patterns and provides hints at probable grammatically incorrect readings and other rare cases unworthy of attention. As a rule of thumb, only items with a relative frequency of at least 0.15% were considered for inclusion in the term grammar. At the same time, collaboration with a speaker of the language makes it easier to understand the observed patterns and generalize them where useful (e.g. enforcing an overall agreement in gender and number instead of running in the risk of omitting some less-frequent cases such as with the plural). On the other hand, some constraints need not be reflected in the rules, such as grammatical case governed by a preposition, because false positives seem to be rare and by not demanding a particular case we allow for possible incorrectly tagged terms to be included and the term grammar to be simpler.

Generalization, compromising and application of linguistic knowledge contribute to shortening the length of the resulting term grammar (i.e. lowering the number of rules), making the internal structure of the grammar easier to understand and also making the computation quicker. It is assumed that a breakdown of the gold-standard terms into part-of-speech and morphological tags is sufficient for the creation of term grammar rules. If, during the rule design or during later evaluation, it is observed that some constituent of a rule should be specified in more detail, it is possible to further limit the accepted words to certain lemmas or word forms (e.g. in most Romance languages, only a limited set of adjectives is permitted to appear *in front* of the noun they relate to), to enforce additional relationships between two constituents of a rule, or to limit the acceptable context (adjacent words) of a valid term candidate within running text.

3. Term Grammars

A term grammar is a carefully crafted set of rules (expressed in CQL, the Corpus Query Language (Jakubíček et al., 2010) describing the lexical structures, typically noun phrases, which should be included in term extraction. Noun phrases are manifested by the presence of a head noun, but their internal morphosyntactic structure is variable and by far not all sequences of words that include a noun are terms.

¹ <https://www.sketchengine.eu/documentation/api-documentation/>

2. adjective + noun (119236 terms, 18.75%)

2.1. JJ NN (109240 terms, 17.18%)

Nuclear housing • active site • aero-medical centre • allelopathic chemical • armed neutrality • back chute • bacterial bed • calcareous grassland • complementary medicine • concurrent liability • critical assembly • dental floss • environmental effectiveness • ever-married survivor • express request • ferrous iron • fragmented mechanization • governmental aid • hedge period • hybrid selection • little plover • louvred fitting • mass effect • medical cannabis • mizzen sail • natural recovery • non-motorized vessel • on-line separation • political instability • poor soil • posterior kidney • preformed joint • private shareholder • public procurement • radiant density • random choice • reverse calf • sealed ampoule • semi-scale brewing • single licence • standard tare • straight lease • synthetic fluid • terminal bar • top performer • two-price system • unobservable variable • up-to-date inventory • variable pad • written assessment...

2.2. JJ NNS (8613 terms, 1.35%)

Introductory Notes • Physical contingencies • administrative courts • adverse consequences • algebraic parentheses • ancillary restrictions • beneficial contracts • calcareous algae • collective arrangements • cumulative grounds • descriptive markings • discouraged people • error-free seconds • essential workers • executive powers • fine seeds • hazardous substances • high-speed data • industrial trucks • interest-induced shifts • journey-related variables • locked points • major effects • mass properties • military mails • minor repairs • missing plants • modal numbers • non-recurring expenses • numeric data • outdated data • peritrichous flagella • photo axes • polar latitudes • preliminary surveys • psychomotor activities • repetitive duties • residual stocks • self-supporting elements • settlable solids • short-horned grasshoppers • social dynamics • speculative damages • structural arrangements • super singles • tamping ties • toothed whales • undercover activities • urban centres • white pages...

2.3. JJ NP (940 terms, 0.15%)

African Eve • Argentine Republic • Bosnian Serbs • Chocolate Point • Dedicated Target • Euro-Mediterranean Partnership • Euromediterranean Bank • European Commission • European Union • Feminist Initiative • Focal Point • Governmental Committee • Honourable Member • Injured Party • Legislative Council • MAb-based ELISA • Molecular Engineering • Neutral Red • Norwegian Trench • Permanent Secretary • Spanish Constitution • Standby Mode • Test-ban Treaty • Transatlantic Forum • Wet Sump • active NFE • anatomical MRI • asynchronous TTY-terminal • climate-neutral Union • competent Court • depletion-mode FET • dideoxy sequencing • east Berlin • far IR • flash EAROM • free-floating e-scooter • glacé ki • helical CT • lazy Susan • multilayer TVS • nationwide ISDN • postgenomic biosciences • prokaryotic promotor • regional programm • serial DAS • soft Brexit • total AMS • underfloor wheel-lathe • visual CAPTCHA • Ħal Qormi...

Figure 1: Part of the report for English IATE data: The *adjective + noun* pattern is the second most common in multi-word terms. Majority of these terms are tagged JJ NN (i.e. adjective followed by a singular or mass noun) in the corpus. In some such terms, the noun is in the plural (JJ NNS). A few terms consist of the adjective followed by a proper noun, what is sometimes the result of inaccuracies in tagging due to the use of title case (e.g. *Governmental Committee* or *Standby Mode*) or due to the fact that acronyms such as *MRI*, *NFE*, or *CT* are tagged as NP. Used tagset is the English TreeTagger PoS tagset with Sketch Engine modifications (see <https://www.sketchengine.eu/tagsets/english-part-of-speech-tagset/>)

It should be noted that the full internal structure of a term candidate is usually not visible in Sketch Engine, because only shallow parsing is performed and the exact dependencies within a complex noun phrase may remain ambiguous. Such cases require our attention, because some isomorphic syntactic structures might erroneously be discarded if rule conditions have been set too tight (with only the prevailing structure in mind). For example, in French noun phrases of the type *noun + preposition + noun + adjective*, imposing gender and number agreement between the last two words (e.g. in *gestion des exploitations agricoles*, i.e. *management of agricultural exploitations*) is wrong, because the adjective can as well link to the noun in the first position (e.g. *pardon des péchés obtenu*, i.e. *obtained forgiveness of sins*).

Besides the CQL query that a sequence of words must match to produce a term candidate, each term grammar rule ensures that the term is represented in its canonical (citation) form. The tradition differs across languages, but it usually includes using the lemma for the head noun and its optional modifiers (Gojun et al., 2012). For many Romance and Slavic languages, the lemma used for adjectives must be gender-respecting (e.g. *nuée ardente* instead of *nuée ardent* in French, see Jakubiček et al. (2014)). The rules are even more complex in German (with its capitalized nouns and adjectives ending in suffixes corresponding to the gender of the related noun).

Full implementation of such rules may rely on special attributes present in the corpus. Examples of attributes that had to be added or modified include: corpus attributes for the comparison of the agreement in number and case, context-based disambiguation of non-conclusive gender and number in the output of the FreeLing tagger², or an extension of gender-respecting lemmas to the past participle (while the past participle behaves like adjectives and appears within terms, its lemma used to be the verb infinitive).

In the formula describing the citation form, individual matched tokens are referred to by their labels (numbers) in the CQL query. For convenience, the numbering of tokens in the query is chosen so as to provide an idea about the syntactic structure of the noun phrase, starting from number 1 for the head noun (with necessary limitations, due to the fact that a single consecutive row of integers is used). In theory, tokens may be present in a different order in the citation form, but we have not found a need for this in any of the languages we have worked with. Sometimes, tokens from the query may be missing in the citation form, usually when they are used only as negative filters, e.g. to ensure that another noun does not follow a matched sequence of nouns, so that *Centro Robert Schuman* is considered a term candidate, but not its substring *Centro Robert*. Such negative restrictions are typically put in place only during the evaluation of a term grammar draft, because the term corpus itself does not contain such incomplete terms.

In most languages, the citation form of terms traditionally uses lower-case letters only. This is convenient in order to reconcile differences in letter case in the word forms (e.g. when a phrase is sometimes spelled in the corpus *in Title Case*) and to cope with the fact that the built-in lemmatization for some languages returns lower-case output only. Another peculiarity is that term grammar rules currently cannot enforce use of the plural for the headword of a citation form, although this is customary in some contexts. As such cases are difficult to recognize, this difference is disregarded and all terms' headwords are rendered in the singular, in turn producing occasional incorrect citation forms (e.g. *united*

² <https://nlp.lsi.upc.edu/freeling/>

state of america). We believe that a future addition to the OneClick Terms algorithm might improve the quality of citation forms generation, by taking advantage of their surface form frequency in order to generate correctly capitalized output in the correct number (e.g. *United States of America*).

```
define(`common_noun', `[tag="NC.*"]')
define(`preposition', `[lc="a|al|con|de|del|en|entre|para|por|sin|sobre"]')
define(`adjective', `[tag="A.*" | tag="VMP.*"]')
define(`agree', `$1.gender=$2.gender & $1.number=$2.number')

*COLLOC "%(1.lemma) %(2.1c) %(3.1c) %(4.1c)"
1:common_noun 2:preposition 3:common_noun 4:adjective & agree(3, 4)
# example: reducción de ojos rojos
```

Figure 2: Simplified example of a rule from the new Spanish term grammar, along with definitions of the used macros. The head noun in position 1 is output as lemma, the noun and adjective in positions 3 and 4 must agree in gender and number. The shown example term means “reduction of red eyes”

When writing a term grammar, we have found it useful to divide the rules into blocks, depending on the number of tokens in the produced term candidates (note that single tokens are not considered terms, but keywords). Within each such block of same-length rules, interactions among the rules are possible, which may lead to overlaps and possibilities to generalize. We try to order the rules within a block by decreasing frequency, although this constraint is sometimes broken in favour of similar rules (such as all starting with a noun) being listed next to each other. For the processing of the term grammar in OneClick Terms, the order of rules in the term grammar, as well as their possible overlaps are irrelevant.

To make orientation in the term grammar and the editing thereof easier, we make use of macros in the rule definitions and show example terms next to each rule. Macros such as `noun` (instead of `[tag="NN"]`) or `modif` (meaning *noun or adjective*) have been used also in the existing term grammars, ever since the adoption of the m4 macro language for term grammars has enabled this, but with the increased complexity of terms recognized by the next-generation term grammars, their usefulness and variety has risen substantially. One and sometimes more examples of terms matched by a rule are included as comments in the term grammar file and provide the possibility of noting that a noun phrase of a certain morphological structure may correspond to two or more syntactic structures, as already explained above.

Many times, incorrect tagging comes into play too, because some rules may partially or fully match terms that have been assigned incorrect part-of-speech or morphological tagging in the corpus. If this is the case, we note this fact in the term grammar by providing an extra example marked as such, but we do not feel obliged to cover all such cases, for the inconvenience of doing so and for the belief that in such cases the respective taggers should be improved instead.

4. Development

The initial design consisting of writing rules corresponding to the most frequently observed patterns in the term corpus is followed by testing the resulting term grammar draft against an actual focus corpus and a reference corpus. We have asked the collaborating speakers to come up with a domain-specific focus corpus of their own preference, expecting that subject knowledge can lead to better results. These focus corpora have varied in size from about 700,000 to 2,000,000 tokens and most were built specifically for this purpose using WebBootCaT (Baroni et al., 2006). To speed up the iterative evaluation process (i.e. each change in the term grammar requires the terms to be recompiled for both the focus corpus and the reference corpus), we did not use the full standard reference corpus (i.e. one of the TenTen corpora), but a downsized sample thereof instead (approximately 200 million tokens) as a sufficient approximation.

Since for each processed language, there had been an existing term grammar before and our aim was to improve it, we did not stop at generating a list of terms in the focus corpus with the new term grammar, but we also ran term extraction from the same focus corpus with the old term grammar. Then we could visualize the differences by putting the two lists side by side and marking for each item in each list whether it is present in the other list or not, and if it is in both, then how much did its ranking (i.e. position in the list) change. See Figure 3 for an example of such comparison. A term’s ranking can easily have changed due to factors such as inflection or incorrect tagging when different tokens (or differently tagged tokens) share the same citation form. For example, the old term grammar could only match the term in the nominative, while the new term grammar matches it in all cases (and outputs it in a lemmatized form, i.e. the nominative, thus increasing the term’s frequency and therefore ranking).

It is natural that some term grammar rules produce more terms than others, and some terms may have been contributed to by multiple rules. In the regular list of extracted terms, it is impossible to make such distinctions. In order to evaluate each rule performance separately, we split the created term grammar into a set of single-rule mini-grammars and run term grammar extraction separately with each of them. This process is time-consuming (tens of minutes for longer term grammars), but it provides useful data not available in a different way. The term lists generated in this way can be combined to form the full grammar term list, with the extra information on which rule(s) produced each term. With such per-rule lists, it is also easy to spot when some rule does not produce any terms at all, which means it should be either fixed or discarded.

Importantly, per-rule lists allow us to quickly review the top-scored terms for each rule with the aim of making sure that no rule produces invalid terms with scores high enough so that they risk spoiling the overall list of terms. The presence of invalid terms is common due to noise in corpora (typos, foreign words, broken language etc.) and inaccuracies in the processing (incorrect tagging, inherently ambiguous rules, incorrectly created citation forms etc.) and we limit our effort to making sure that the top terms produced by each rule are correct. If a rule produces problematic output and all of it is low-scored (compared to the top scores found in the full list), it can be considered for deletion, because its removal is not going to substantially change the overall results of term extraction. All in all, the effort spent at fixing a rule should be proportional to the score of the terms it generates. The full list of all term candidates, produced by all rules as a whole, may contain tens of thousands of items and is never used in practice, because it is the normalized-frequency scoring which

1. pasta sfoglia \uparrow -1	1. pasta al forno \uparrow
2. secondo piatto \uparrow -2	2. pasta sfoglia \downarrow +1
3. primo piatto \uparrow -11	3. ricetta facile \uparrow -1
4. ricetta facile \downarrow +1	4. secondo piatto \downarrow +2
5. pasta fillo \uparrow -1	5. tempo di cottura \uparrow -25
6. forno vegetariana \uparrow -3	6. pasta fillo \downarrow +1
7. tempi di cottura --	7. verdura al forno \uparrow
8. verdure in padella --	8. ricetta vegetariana \uparrow -30
9. prossimo commento \uparrow -2	9. forno vegetariana \downarrow +3
10. cookie salvi --	10. cookie salvo \uparrow
11. ricette antipasti --	11. prossimo commento \downarrow +2
12. torta in padella \uparrow -54	12. antipasto veloce \uparrow -90
13. verdure miste --	13. pasta al forno vegetariana \uparrow
14. cottura in padella \uparrow -17	14. primo piatto \downarrow +11
15. maria bonaccorso --	15. torta salata \uparrow -124
16. cottura in forno \uparrow -2	16. verdura in padella \uparrow -4641
17. forno statico \uparrow -2	17. antipasto sfizioso \uparrow -35
18. padella antiaderente \uparrow -2	18. cottura in forno \downarrow +2
19. email necessario \uparrow -2	19. forno statico \downarrow +2
20. indirizzo email necessario \uparrow -2	20. padella antiaderente \downarrow +2
21. informazioni di profilo --	21. email necessario \downarrow +2
22. informazioni di profilo pubbliche --	22. indirizzo email necessario \downarrow +2
23. profilo pubbliche --	23. informazione di profilo \uparrow
24. ricette di antipasti --	24. informazione di profilo pubbliche \uparrow
25. pasta fredda \uparrow -8	25. informazione di profilo pubbliche fornite \uparrow
26. piatto unico \uparrow -11	26. network scelto in base \uparrow
27. campi obbligatori --	27. profilo pubblica fornita \uparrow
28. social login --	28. network scelto \uparrow
29. ultime ricette --	29. social network scelto \uparrow
30. tempo di cottura \downarrow +25	30. profilo pubblica \uparrow
31. nostra ricetta --	31. cottura in padella \downarrow +17
32. tue impostazioni --	32. verdura mista \uparrow
33. ricette vegetariane --	33. pasta fredda \downarrow +8
34. mio consenso --	34. impostazione sulla privacy \uparrow
35. patate in padella --	35. cottura della pasta \uparrow
36. peperoni ripieni --	36. verdura cotta \uparrow -506
37. carta forno \uparrow -9	37. piatto unico \downarrow +11
38. ricetta vegetariana \downarrow +30	38. filo d'olio \uparrow
39. pollo in padella \uparrow -15	39. ricetta antipasti \uparrow
40. ricetta semplice \uparrow -3	40. patata al forno \uparrow
41. antipasti veloci --	41. dieta vegetariana \uparrow -41
42. domus spa --	42. cottura al forno \uparrow
43. editoriale domus spa --	43. ricetta semplice \downarrow +3
44. g. mazzocchi --	44. patata in padella \uparrow
45. r.e.a. di milano --	45. carta da forno \uparrow
46. cookie completa --	46. carta forno \downarrow +9
47. informativa cookie completa --	47. peperone ripieno \uparrow
48. proprietà di maria bonaccorso --	48. metodo di cottura \uparrow -28
49. x fonte --	49. campo obbligatorio \uparrow
50. proprietà di maria --	50. antipasto vegetariano \uparrow -181

Figure 3: Top of a side-by-side comparison of terms generated from an Italian vegetarian cuisine corpus using the old (left) and the new (right) Italian term grammar: Newly identified terms are marked with a plus sign, discarded terms are marked with a minus sign. Each type is further highlighted in a corresponding color (green and red, respectively). For terms generated by both grammars, the difference in their ranking across the two sets is marked with an up arrow or a down arrow, followed by the change of ranking expressed as a signed integer.

makes term extraction in OneClick Terms so powerful, as it helps to distinguish actual terms from mere term candidates. Because of this, during the development, we only strive to have the first few hundred items in the list as clean as possible, increasingly tolerating noise further down the list.

5. Evaluation

In order to estimate the coverage of terms in IATE by the produced term grammars, we ran each rule’s CQL query on the term corpus (with a restriction that the full paragraph/term must be matched) and calculated the number of unique matches in the output. When compared with the total number of terms in the term corpus, this says what portion of IATE terms is recalled by our term grammar. We ran the same calculation also with the old term grammar to be able to observe if there has been progress. Results for each language are shown in Table 1.

Language	IATE terms	Old grammar		New grammar	
English	635,700	367,693	57.8%	505,431	79.5%
Estonian	37,485	7,624	20.3%	24,884	66.4%
French	585,112	136,783	23.4%	425,133	72.7%
German	227,652	110,418	48.5%	169,558	74.5%
Italian ^a	378,133	176,836	46.8%	277,246	73.3%
Portuguese	302,843	176,836	58.4%	277,246	91.5%
Spanish	365,066	201,990	55.3%	265,435	72.7%

Table 1: Recall of multi-word terms in IATE by old and new term grammars

^a The existing Italian term grammar used the TreeTagger tagset, but because Sketch Engine was switching to FreeLing for Italian at the time, the new term grammar was written for this tagset. The figure for the old grammar in this table was produced by an unpublished rewrite of the old grammar for the new tagger.

Please note that when performing these calculations, we did not consider in any way the selection bias of terms found in IATE, which might over-represent terms from a particular domain or of a particular lexical structure and thus make the results less applicable to general terminology extraction. The calculated numbers are also representative only of the term corpus, i.e. recognized isolated terms. More authentic results would be achieved if we were to search for these terms inside running text, in which they would be used in sentence context and possibly inflected.

The figures in Table 1 demonstrate that we have managed to achieve our goal, namely that we have improved the coverage of actual terms by OneClick Term’s term extraction grammars. The observed differences of recall rank from 17.4% for Spanish to 49.3% for French. Except for Estonian, whose dataset in IATE is smaller by an order of magnitude, all other languages have more than 72% of the multi-word IATE terms covered by the

newly developed term grammars. Importantly, recall has risen from 57.8% to 79.5% of multi-word IATE terms for English, which is the most requested language by OneClick Term users.

Some factors that contribute to the recall not being 100% are:

- Ambiguous or incorrect tagging which hides important information that could be used to identify a term candidate
- Ambiguity in language and lack of information on syntax which makes it impossible to distinguish actual lexical structures from mere token sequences that span across syntactic borders
- Low-frequent patterns in term candidate structure that are ignored to reduce term grammar complexity
- Terms longer than the longest rule in the term grammar (e.g. 8.1% of the English IATE terms are longer than 5 tokens and 1.8% of terms are 10 tokens or longer, e.g. *communal estate of husband and wife comprising only property acquired after their marriage*)
- Terms of type deliberately not supported by the term grammar (verbal terms, e.g. Italian *fare click* – “to click” – constitute approximately 1% of IATE data but their inclusion in term extraction is questionable)

Language	Number of rules	Maximum term length
English	21	5
Estonian	61	5
French	47	8
German	73	6
Italian	40	7
Portuguese	64	9
Spanish	52	8

Table 2: Number of rules and maximum supported length of terms (in tokens) in the new term grammars

The size of each term grammar (expressed in the total number of rules in it), as listed in Table 2, depends on several factors:

- Precision with which rules were written (less strict rules often mean tolerance to small errors in tagging and lead to less complex term grammars while letting in no or very little extra noise)
- Level of detail in the used tagset
- Maximum term length defined in the term grammar (which itself is influenced by the following factor:)
- Variety of the language’s morphology and syntax (e.g. Romance languages typically chain nouns by means of a preposition like *de* and possibly an article, so their

terms tend to be longer than English terms which often expresses the same with adjectives or noun juxtaposition)

In general, we strived to keep the number of rules a two-digit number in order to keep the term grammar friendly to a human editor and the computation of term candidates fast enough (each extra rule means an extra query that has to be executed on the corpus). The number of rules can be somewhat reduced during final optimization of the term grammar, e.g. by creating macros that combine conditions which are often both applicable in a context like having a macro meaning *adjective or past participle*, or by relaxing some rules in order to merge them with other similar rules without causing any actual damage by such generalization.

During finalization, each produced term grammar was tested with several other focus corpora, including different domains and one rather small corpus, to ensure that it performs reasonably well in real-life situations. The final term grammars are made available under the Creative Commons Attribution NonCommercial license. All the new grammars are already installed in OneClick Terms at the time of writing and can be used also in Sketch Engine. Feedback received from both creators and users of these tools suggests that the change has been to their satisfaction and that the quality of term extraction for these languages has noticeably improved in their opinion.

6. Future Work

The fact that we work with isolated terms is a source of inconvenience, both in the design stage and during the evaluation of a term grammar. In authentic use cases, terms are extracted from running text, composed of full sentences. In running text, terms can appear nested within more complex syntactic structures and possibly inflected. The collaborating speaker's linguistic knowledge is likely to mitigate this issue to some extent because of forethought. For instance, rules can be written with all grammatical cases in consideration, even if in the studied list of terms, only the nominative is used. However, if we were able to look up the IATE terms and their possible inflected forms inside full sentences, we could produce a performance estimate that would be more representative of real-life situations. Sentences containing the IATE terms in use could possibly be found and extracted from large corpora, such as those of the TenTen Corpus Family.

More strikingly, the inconvenience of using isolated terms manifests in the term corpus which we use as a gold standard. Although morphological taggers should in theory be able to handle non-sentences such as titles or list items and process them correctly, this is not always the case. For instance, the FreeLing tagger for Spanish had the tendency to sometimes mark nouns at the start of a term as verbs: e.g. in *aduanas de primera entrada* (“customs office of first entry”), the first word is asserted to start with the third person singular of the verb *aduanar*, i.e. “(he) pays the customs”, rather than the correct noun meaning “customs” or “customs office”. Similarly, capital letters in proper names at the start of terms would get confused for sentence-start capitals, possibly influencing the tag assigned to the word (the FreeLing tagset distinguishes common and proper nouns).

In an effort to prevent these problems, we experimented with enclosing the terms into *sentence frames* before the tagging and removing these frames afterwards. For instance, English terms could be prefixed with the words *I know the* or Spanish terms with the

word *Hay* (“There is/are”), creating a full sentence in which the term constituents get tagged more accurately. It is, however, not always possible to come up with such universal sentence frame in a language which would work with all or almost all terms; many times, such a frame would need to be differentiated in form by the grammatical number or gender of the term that follows it, which is information unknown to us at the time and not easy to derive. Our research so far has been inconclusive in whether the creation and usage of such sentence frames is desirable and worth the effort.

There are also some intended deficiencies in the produced term grammars, due to situations we could not handle without letting in too much noise. Many terms that include a conjunction, mainly “and” or “or”, are not covered by the new term grammars because these conjunctions are frequently used to join lexical structures and even sentences and therefore most of the output generated by rules that feature a conjunction would in fact be spanning across these syntactic borders and not represent an actual lexical structure. In rare cases, we could allow conjunctions in rules with confidence due to it clearly being situated inside, rather than possibly at the edge of a lexical structure. An example is the French *système de séparation et de tri* (“separation and sorting system”), in which the conjunction *et* (“and”) is followed by the preposition *de* (“of”), indicating that it is joining the two attributes of the preceding headword (*système*).

The IATE term base is a unique, large and freely accessible source of terms in multiple languages, but an alternative needs to be identified when writing term grammars for languages not present in IATE. Our ongoing effort at developing a term grammar for the Ukrainian language has shown that resources similar to IATE are scarce and it might be necessary to adopt a different approach and start identifying terms where they are highlighted in running texts rather than collected in ready-made term bases.

7. Conclusion

We have designed a procedure for the creation of a new generation of term extraction grammars, which are inspired primarily not by someone’s linguistic judgement, but by an existing term base such as IATE, which serves as a model of what lexical structures are likely to be considered terms by end users. The existing term base, which serves as a gold standard, also provides a way of evaluating the quality of the new term grammars. The development of each new term grammar happens in a standardized process, in the cooperation of a computer linguist with a speaker of the respective language. In the article, we have described possible challenges during term grammar design presented by specific languages or linked to cases of inaccuracies or ambiguities, along with recommendations of how they should be handled.

By the time of writing, we had produced such next-generation term grammars for seven European languages (English, Estonian, French, German, Italian, Portuguese, Spanish). Evaluation showed that recall indeed increased after the new grammars had been designed with IATE in mind, as on average three fourths of the (cleaned, multi-word) IATE terms can now be detected during term extraction. Most of these new *evidence-based* term grammars have been already installed in OneClick Terms and Sketch Engine and positive feedback from users confirms that they are actually getting higher-quality results than with the old term grammars. Lack of negative comments suggests that, while the number and versatility of term extraction rules increased, we managed to avoid polluting the term

extraction results with incorrect terms, or, more specifically, with sequences of words which are matched by some of the new rules *and* would be lifted high enough by OneClick Term scoring algorithm, but which would not be considered proper terms by the user.

We plan to produce term grammars for more languages using the described method in the future, including languages not represented in IATE. For other languages than the 24 covered by IATE, another similar term base or another approach at gold standard compilation will need to be identified.

8. Acknowledgements

This work has been partly supported by the Ministry of Education of CR within the LINDAT-CLARIAH-CZ project LM2023062. This publication was written with the support of the Specific University Research provided by the Ministry of Education, Youth and Sports of the Czech Republic.

We would like to thank Eleri Aedmaa, Hanna Barabakh, Kristína Koláčková, Ondřej Matuška, Merily Plado and Emma Romani for assistance with particular languages.

9. References

- Aker, A., Paramita, M.L. & 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)*. pp. 402–411.
- Baisa, V., Michelfeit, J. & Matuška, O. (2017). Simplifying terminology extraction: OneClick Terms. *The 9th International Corpus Linguistics Conference*.
- Baisa, V., Ulipová, B. & Cukr, M. (2015). Bilingual Terminology Extraction in Sketch Engine. In *RASLAN*. pp. 61–67.
- Baroni, M., Kilgarriff, A., Pomikálek, J., Rychlý, P. et al. (2006). WebBootCaT: a web tool for instant corpora. In *Proceeding of the EuraLex Conference*, volume 1. pp. 123–132.
- Gojun, A., Heid, U., Weissbach, B., Loth, C. & Mingers, I. (2012). Adapting and evaluating a generic term extraction tool. In *LREC*. pp. 651–656.
- 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*. pp. 53–56.
- 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/>.
- Jakubíček, M., Kilgarriff, A., McCarthy, D. & Rychlý, P. (2010). Fast Syntactic Searching in Very Large Corpora for Many Languages. *PACLIC*, pp. 741–47.
- 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, pp. 7–36.
- Kovář, V., Baisa, V. & Jakubíček, M. (2016). Sketch engine for bilingual lexicography. *International Journal of Lexicography*, 29(3), pp. 339–352.
- Rychlý, P. (2008). A Lexicographer-Friendly Association Score. In *RASLAN*. pp. 6–9.
- Zorrilla-Agut, P. & Fontenelle, T. (2019). IATE 2: Modernising the EU’s IATE terminological database to respond to the challenges of today’s translation world and beyond. *Terminology*, 25(2), pp. 146–174.

**Workshop on Lexicography and CEFR: Linking
Lexicographic Resources and Language Proficiency
Levels**

Building a CEFR-Labeled Core Vocabulary and Developing a Lexical Resource for Slovenian as a Second and Foreign Language

Matej Klemen¹, Špela Arhar Holdt^{1,2}, Senja Pollak³, Iztok Kosem¹, Eva Pori¹, Polona Gantar¹, Mihaela Knez¹

¹ Faculty of Arts, University of Ljubljana, Aškerčeva ulica 2, 1000 Ljubljana, Slovenia

² Faculty of Computer and Information Science, University of Ljubljana, Večna pot 113, 1000 Ljubljana, Slovenia

³ Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia

E-mail: matej.klemen@ff.uni-lj.si, spela.arharholdt@ff.uni-lj.si, senja.pollak@ijs.si, iztok.kosem@ff.uni-lj.si, eva.pori@ff.uni-lj.si, apolonija.gantar@ff.uni-lj.si, mihaela.knez@ff.uni-lj.si

Abstract

This article introduces two newly available datasets: the KUUS 1.0 corpus and the list Core Vocabulary for Slovenian as L2 1.0. The KUUS 1.0 corpus consists of seventeen textbooks published by the Center for Slovene as a Second and Foreign Language at the University of Ljubljana, and it contains a total of 520,796 words accompanied by various linguistic tags and metadata. Using the KUUS 1.0 corpus, we compiled the list Core Vocabulary for Slovenian as L2 1.0. The list includes 350 words labeled as A1-core, 864 words as A1-larger, 1,451 words as A2, and 2,608 words as B1. The A1 vocabulary was used as pilot data for a project focused on developing a lexical description for learning Slovenian as a second and foreign language. Our methodology involved combining the data from the new datasets with existing, openly available lexical information on modern Slovenian, with the aim of achieving didactic adaptation and maximal reusability of the results.

Keywords: Lexicography and CEFR; Slovenian; second and foreign language; textbook corpus; core vocabulary

1. Introduction

Most existing CEFR-based language documents and curricula for Slovenian as a second and foreign language—for example, *Preživetvena raven v slovenščini* (Breakthrough Level in Slovenian, 2004, revised version 2016); *Sporazumevalni prag za slovenščino* (Threshold Level for Slovenian, 2004), and *Slovenščina kot drugi in tuji jezik: Izobraževalni program za odrasle* (Slovenian as a Second and Foreign Language: Adult Education Program, 2020)—are based on consensual expert group knowledge. The documents present a general description of language skills and contain a list of illustrative vocabulary. Over the past twenty years, these documents have been the basis for developing learning materials aimed at different target groups (e.g., adolescents, students, adult speakers, etc.) learning Slovenian as a second and foreign language in Slovenia and in other countries. The different communicative needs of these learners are reflected in the different choice of vocabulary in the learning materials.

Our aim was to create a corpus-based list of core vocabulary¹ covering different CEFR levels. This article presents KUUS 1.0, a corpus of textbooks for learning Slovenian as a second and foreign language, and how it was used to create the corpus-based list Core Vocabulary for Slovenian as L2 1.0,² which contains single-word vocabulary labeled as A1, A2, and B1. We then present how the newly available datasets have been included in developing the CEFR-labeled lexical resource for Slovenian as a second and foreign language.

2. The KUUS 1.0 corpus

The work presented in this article is based on the KUUS 1.0 corpus, which is a collection of seventeen textbooks specifically created for teaching Slovenian as a foreign and second language. These textbooks, published between 2002 and 2022 by the Center for Slovene as a Second and Foreign Language (Sln. *Center za slovenščino kot drugi in tuji jezik*, CSDTJ) at the University of Ljubljana, are widely used in both Slovenia and other countries to teach Slovenian to learners of different ages at various CEFR levels (Gril, 2022: 123; Knez et al., 2021: 261–262, 342–343). The KUUS corpus was developed as a companion project to the CSDTJ’s publishing of graded readers series and aims to provide a standardized, linguistically annotated, and openly accessible dataset of this nature for Slovenian.

KUUS 1.0 includes metadata for each textbook, including the title, subtitle, authors, year of publication, publisher, CEFR level, target audience, and estimated number of lessons. The corpus was linguistically annotated with the CLASSLA v1.1.1 pipeline (Ljubešić & Dobrovoljc, 2019) at the levels of tokenization, sentence segmentation, lemmatization, MULTEXT-East v6 MSD-tags, JOS dependency syntax, and named entities. The current version of the corpus comprises 520,796 words and is available as a database at the CLARIN.SI repository (Klemen et al., 2022a).

The selection of textbooks was made to cover different CEFR levels, contain the bulk of the textbook production of the CSDTJ, and comprise a significant part of the current textbooks for learning Slovenian as a foreign and second language. The texts were converted from PDF or DOC format into TXT format. Parts of the textbooks that are not intended for the student or for direct use in teaching were manually removed. These typically included the introduction, table of contents, colophon, and sources of pictures and texts. In addition, any recurring text in the header or footer of pages was deleted, except for page numbers. Foreign-language instructions were marked with special codes so that they are easily separable from the Slovenian part of the text. We furthermore

¹ Similar to Volodina et al. (2022), we understand core vocabulary as vocabulary known to most learners at a certain level of language proficiency. In terms of building a vocabulary list, we understand the “core” as a consensually agreed-upon and stable but expandable starting point for learning.

² In the article, we refer to the official names of institutions and published resources, which can result in some discrepancies, such as the contrast between “Slovene as a Second and Foreign Language” and “Slovenian as L2.”

corrected any errors that occurred during the conversion process, including problems with characters such as č, š, ž, upper- and lower-case letters, punctuation, and hyphenated words. In some cases, we had to add text that was erroneously omitted during the conversion due to specific fonts or layouts. The preparation of the KUUS corpus is presented in greater detail by Klemen et al. (2022).

Some of the textbooks included in the KUUS corpus have a part of the book that is structurally similar to workbooks and includes grammar exercises. These parts of textbooks have been included in the corpus because they are part of a single publication. However, in the current version, the corpus only includes textbooks and not the corresponding workbooks.

3. Core Vocabulary for Slovenian as L2 1.0

Using the KUUS 1.0 corpus, we prepared the list Core Vocabulary for Slovenian as L2 1.0 (Klemen et al. 2022b). The list comprises 5,273 lemmas, classified into the first three CEFR levels: 350 lemmas with the assigned label A1-core, 864 words with the label A1-larger, 1,451 words with the label A2, and 2,608 words labeled B1.³ The current version of the list is available at CLARIN.SI in a tab-separated format containing the lemma, part-of-speech (following the MULTEXT-East tagset for Slovenian), information on whether the lemma appears in the Reference List of Slovene Frequent Common Words (Pollak et al., 2020), and the relative average frequency. An example of the data is presented in Table 1.

CERF level	Lemma	POS	Lemma in Reference List of Slovene Frequent Common Words	Sum of relative frequencies across textbooks
A1-core	<i>biti</i> ‘to be’	g	Yes	124.87740
A1-core	<i>v</i> ‘in’	d	Yes	38.03003
A1-core	<i>se</i> ‘oneself’	z	Yes	34.44841
A1-core	<i>in</i> ‘and’	v	Yes	34.28150

³ In the article, we intentionally make a distinction between “level” and “label.” Here, “level” refers to the CEFR level, while “label” pertains to the corpus-based annotation of a specific lemma in the core vocabulary list. In our methodology, the current labels serve as a baseline and are subject to potential modifications in subsequent stages of our work.

When creating the list of core vocabulary, we did not distinguish between criterion levels and plus levels (e.g. A2 and A2+) as conceived in the CEFR Companion Volume, as the labels on the textbooks do not differentiate between them. Therefore, we have used the labels B1 and A2, and for A1 we have introduced two labels: A1-core and A1-larger. The former was assigned to words that appear in all five A1 textbooks included in the KUUS 1.0 corpus, the latter to words that appear in four or fewer A1 textbooks.

A1-core	<i>na</i> ‘on, at’	d	Yes	26.39539
A1-larger	<i>ki</i> ‘which’	v	Yes	6.74070
A1-larger	<i>svoj</i> ‘one’s one’	z	Yes	3.67359
A1-larger	<i>če</i> ‘if’	v	Yes	3.17442
A1-larger	<i>človek</i> ‘human’	s	Yes	3.16109
A1-larger	<i>res</i> ‘really’	r	Yes	3.14526
A2	<i>treba</i> ‘necessary’	r	Yes	0.98788
A2	<i>saj</i> ‘because’	v	Yes	0.96636
A2	<i>pomemben</i> ‘important’	p	Yes	0.95674
A2	<i>zaradi</i> ‘because of’	d	Yes	0.92400
A2	<i>svet</i> ‘world’	s	Yes	0.92355
B1	<i>nekdo</i> ‘someone’	z	Yes	0.32863
B1	<i>glede</i> ‘regarding’	r	Yes	0.28649
B1	<i>sodoben</i> ‘contemporary’	p	Yes	0.27790
B1	<i>lastnost</i> ‘characteristic’	s	Yes	0.26160
B1	<i>dejanje</i> ‘action’	s	Yes	0.24973

Table 1: First five lemmas for each CEFR level in the Core Vocabulary for Slovenian as L2 1.0 with associated data.

In summary, our approach involved importing the corpus into the Sketch Engine tool (Kilgarriff et al., 2014), exporting the frequency lists for each separate textbook, and calculating the relative frequency of each word (lempos) across the seventeen textbooks. We compiled these data (23,068 words of different types) into a single table that included information on word frequency and occurrence across textbooks at each CEFR level. Next, we compared the data to the Reference List of Common Frequent Words (Pollak et al., 2020). This reference list consists of 4,768 common general lemmas compiled by comparing the most frequent 10,000 lemmas by word type from four Slovenian text corpora: Kres 1.0, GOS 1.0, Janes 1.0, and Šolar 2.0.⁴ We found an

⁴ The Kres corpus (Logar et al., 2012) is a balanced sub-corpus of the Gigafida reference corpus, with almost 100 million words from various written sources. The Janes corpus (Fišer et al., 2020) consists of online user-generated content, and the Šolar corpus (Kosem et al., 2016) consists of written texts created independently by primary- and secondary-school

overlap of 4,603 words between the two lists, with only 166 words appearing solely in the list of common general vocabulary but not in the KUUS corpus, and 18,465 words appearing only in the KUUS corpus (Klemen et al., 2022a: 170).

After conducting a comprehensive first review of the data, we established robust numerical criteria with the aim of obtaining core (i.e., relevant or typical) vocabulary for each level from the textbook material. The criteria were used to assign a baseline CEFR-level label to the words. The criteria were considered sequentially, starting with the A1-core criteria, followed by the A1-larger criteria check, and so on. When preparing the criteria, we considered that there are fewer textbooks available for B1 than for A1 and A2, and that a textbook covering two levels (A2–B1, see Klemen et al. 2022) also appears in the material.

- For the A1-core label, the word must appear in all five level-A textbooks (e.g., *nov* ‘new’, *dober* ‘good’, *slovenski* ‘Slovenian’, *star* ‘old’, *velik* ‘big’, *lep* ‘beautiful’, *majhen* ‘small’, *mlad* ‘young’, *sam* ‘alone’, *zanimiv* ‘interesting’).⁵
- For the A1-larger label, the word must appear in four, three, or two level-A textbooks (e.g., *ustrezen* ‘relevant’, *srednji* ‘middle’, *prijazen* ‘friendly’, *prost* ‘free’, *visok* ‘high’, *beseden* ‘word’, *ženski* ‘feminine’, *naslednji* ‘next’, *deloven* ‘working’, *oseben* ‘personal’).
- For the A2 label, the word appears in no more than one A1 textbook, but it appears in five, four, three, or two A2 textbooks (e.g., *pomemben* ‘important’, *znan* ‘known’, *različen* ‘different’, *svetoven* ‘global’, *evropski* ‘European’, *kulturen* ‘cultural’, *šolski* ‘school’, *osnoven* ‘basic’, *zadovoljen* ‘satisfied’, *posloven* ‘business’). (If a word appears in two textbooks at level A2, and one of them is the A2–B1 textbook, then it is considered a B1 word.)
- For the B1 label, the word does not appear in A1 textbooks and can appear in at most one A2 textbook. It must appear in one or two B1 textbooks, and it must have a frequency of at least 2 in the entire corpus (e.g., *sodoben* ‘contemporary’, *državen* ‘state-owned’, *družben* ‘social’, *socialen* ‘social’, *skupen* ‘common’, *lasten* ‘own’, *današnji* ‘today’s’, *prepričan* ‘convinced’, *vprašan* ‘asked’, *posamezen* ‘individual’).

We manually reviewed the labeled words and eliminated any irrelevant instances that we considered to be noise, such as erroneously lemmatized or POS-tagged data, proper nouns, and numerals that would require separate addition because they are not represented systematically in the corpus. However, we decided to retain linguistic terminology and metalanguage commonly found in textbooks, symbols, and

students. The GOS corpus (Verdonik & Zwitter Vitez, 2011) is a spoken Slovenian reference corpus with 120 hours of recordings, spanning a wide range of contexts.

⁵ As an example, the first ten adjectives of each tag are given. The English glosses do not necessarily cover all the meanings and are for general information only. Because of the identical form of adjective and noun in English, certain adjectives may appear as nouns in translation, e.g. *šolski* ‘school’ as in *šolske počitnice* ‘school holidays’.

abbreviations. During our examination of the words in a wider textual context, we encountered some cases that belonged to a higher level than B1 due to mislabeling, homonymy, or polysemy. Nonetheless, in the vast majority of cases, the automatically assigned CEFR labels were found adequate. It is worth noting that our methodology in the first step is purposefully permissive because we prefer to include a word too many rather than too few.

4. Developing a Lexical Resource for Slovenian as a Second and Foreign Language

This section presents how the corpus and the list described in sections 2 and 3 are being utilized to develop a new lexical resource for Slovenian as a second and foreign language. Because the resource is still a work in progress, we explain the methodological considerations and present the work on sample entries.

4.1 Project framework

An opportunity to utilise the newly prepared datasets was offered as part of the project *Nadgradnja učnega gradiva Čas za slovenščino 1 v digitalnem okolju in prilagoditev gradiva za pouk slepih in slabovidnih mladostnikov* (Expanding the Teaching Material *Čas za slovenščino 1* in the Digital Environment and Adapting the Material for Teaching Blind and Partially Sighted Adolescents), led by the CSDTJ. As part of the project, funded by the Slovenian Ministry of Culture,⁶ we committed ourselves to enriching the vocabulary previously labeled as A1 (see section 3) with user-adapted grammatical, semantic, and multimedia information (e.g., pronunciation recordings) in Slovenian, and to including translations of the headwords into three foreign languages (i.e., Albanian, English, and Hungarian),⁷ thus combining monolingual and multilingual dictionary approaches. For this purpose, the project envisages using all relevant information in the lexicographical and other resources produced by the Center for Language Resources and Technologies (Sln. *Center za jezikovne vire in tehnologije*, CJVT) at the University of Ljubljana, revisiting them through the approaches developed at the CSDTJ on the basis of experience in teaching Slovenian as a second and foreign language.

As part of the project, we aim to prepare a lexical resource that could be used by A1 users of Slovenian because no such dictionary for Slovenian has been developed yet. We

⁶ The aims of the project are threefold: (a) preparation of a digital platform with interactive activities for learning, (b) development of a lexical resource (as described in this article), and (c) adaptation of the textbook for teaching blind and partially sighted learners (cf. <https://centerslo.si/za-otroke/projekti/nadgradnja-ucnega-gradiva-cas-za-slovenscino-1-v-digitalnem-okolju-in-prilagoditev-gradiva-za-pouk-slepih-in-slabovidnih-mladostnikov/>).

⁷ The three languages were chosen for the following reasons: Albanian is a non-Slavic language spoken by migrants that have moved to Slovenia from Kosovo (cf. Knez et al., 2021); English is a lingua franca; and Hungarian is used in the Slovenian cross-border area and, as a non-Indo-European language, is the least similar to Slovenian among the four neighboring languages (German, Italian, Croatian, and Hungarian).

are thus targeting users able to understand the explanation (which we perceive as both the basic description and the illustrative material) of the headword, provided that it uses common everyday expressions and very basic phrases in simple grammatical and sentence structures referring to particular concrete situations (e.g., the most basic personal and family information, everyday routine activities and tasks, schooling, or employment) in which the users communicate in Slovenian in their everyday life and which they need to meet their concrete needs and perform linguistic tasks relevant to them (cf. Companion Volume, 2020: 54, 56, 60, 131–132, 175).

Furthermore, the idea is that the resource could be systematically expanded for users of Slovenian as a second and foreign language at higher levels of language proficiency (A2–C1) in the future.

4.2 Methodological background

The list Core Vocabulary for Slovenian as L2 1.0 provided the candidates for the lexical description, consisting of 350 words labeled as A1-core and 864 words as A1-larger. To ensure connectivity between the new lexical resource and the digital platform with interactive exercises (also being developed as part of this project) we supplemented the list with 247 additional words found in the textbook *Čas za slovenščino 1*, which did not meet the criteria for inclusion in the core vocabulary list.⁸ The final wordlist included 1,461 words of various types (e.g., content and function words) requiring distinct lexicographic treatment.

In the first step, we selected fifteen headwords with different part-of-speech categories (common nouns, adjectives, verbs, and adverbs) and created diverse sample entries with the grammatical and semantically structured features (semantic indicators, collocations, and usage examples) to develop a model for a lexical description of Slovenian as a second and foreign language that is suitable for users at level A1 and can be expanded in the future with more complex semantic information relevant for higher-level users.

⁸ The textbook *Čas za slovenščino 1* is aimed at teenagers, especially migrant children who are joining the Slovenian school system. Thus, it also includes specific vocabulary that is relevant for them at the beginner level (e.g., *radirka* ‘eraser’, *ravnilo* ‘ruler’), but is less relevant for other users of Slovenian as a second and foreign language at this level and is therefore not included in other textbooks and consequently not part of the core vocabulary.

The additional 247 words have already been included in the baseline data but have remained unlabeled and will thus remain without a label in the new lexical resource for the time being. We plan to review them and assign them appropriate level in the subsequent stages of our work (see section 5).

The new lexical resource will be linked to the interactive exercises accompanying the textbook *Čas za slovenščino 1* (see footnote 6). This will allow learners to use it simultaneously while solving the exercises.

The test entries were created using a localized and customized version of the dictionary tool Lexonomy.⁹ The grammatical and semantic information for the enrichment of the vocabulary list was taken from the following sources: the Slovene Morphological Lexicon Sloleks 2.0 (Čibej et al. 2022),¹⁰ which contains essential information on Slovenian words (e.g., their part-of-speech category and their grammatical features) as well as recordings of word pronunciations and automatically generated recordings; the Collocation Dictionary of Modern Slovene 1.0 (Kosem et al. 2019)¹¹ with information on the most common and statistically typical collocations and collocations for the selected vocabulary; the Thesaurus of Modern Slovene 1.0 (Krek et al. 2018),¹² which offers synonyms as well as certain antonyms; and the Comprehensive Slovenian–Hungarian Dictionary 1.0 (Kosem et al. 2021)¹³ with information on semantic indicators, dictionary labels, manually reviewed corpus examples, and Hungarian translations of words. For words not covered by these sources, the data were updated in accordance with the methodology used. Currently automatically prepared data were also manually reviewed and corrected.

The concept of the lexical resource for Slovenian as a second and foreign language includes the following elements: (a) a **semantic indicator**; (b) a **set of collocations**; (c) **usage examples**; (d) **translations of the headword** into Albanian, English, and Hungarian, and, where possible, (e) **multimedia elements** (images and recordings) that effectively illustrate the sense of the headword (Figure 1).

⁹ <https://lexonomy.cjvt.si/>

¹⁰ <https://viri.cjvt.si/sloleks/slv/>

¹¹ <https://viri.cjvt.si/kolokacije/slv/headword/69883#>

¹² <https://viri.cjvt.si/sopomenke/slv/>

¹³ <https://viri.cjvt.si/slovensko-madzarski/slv/>

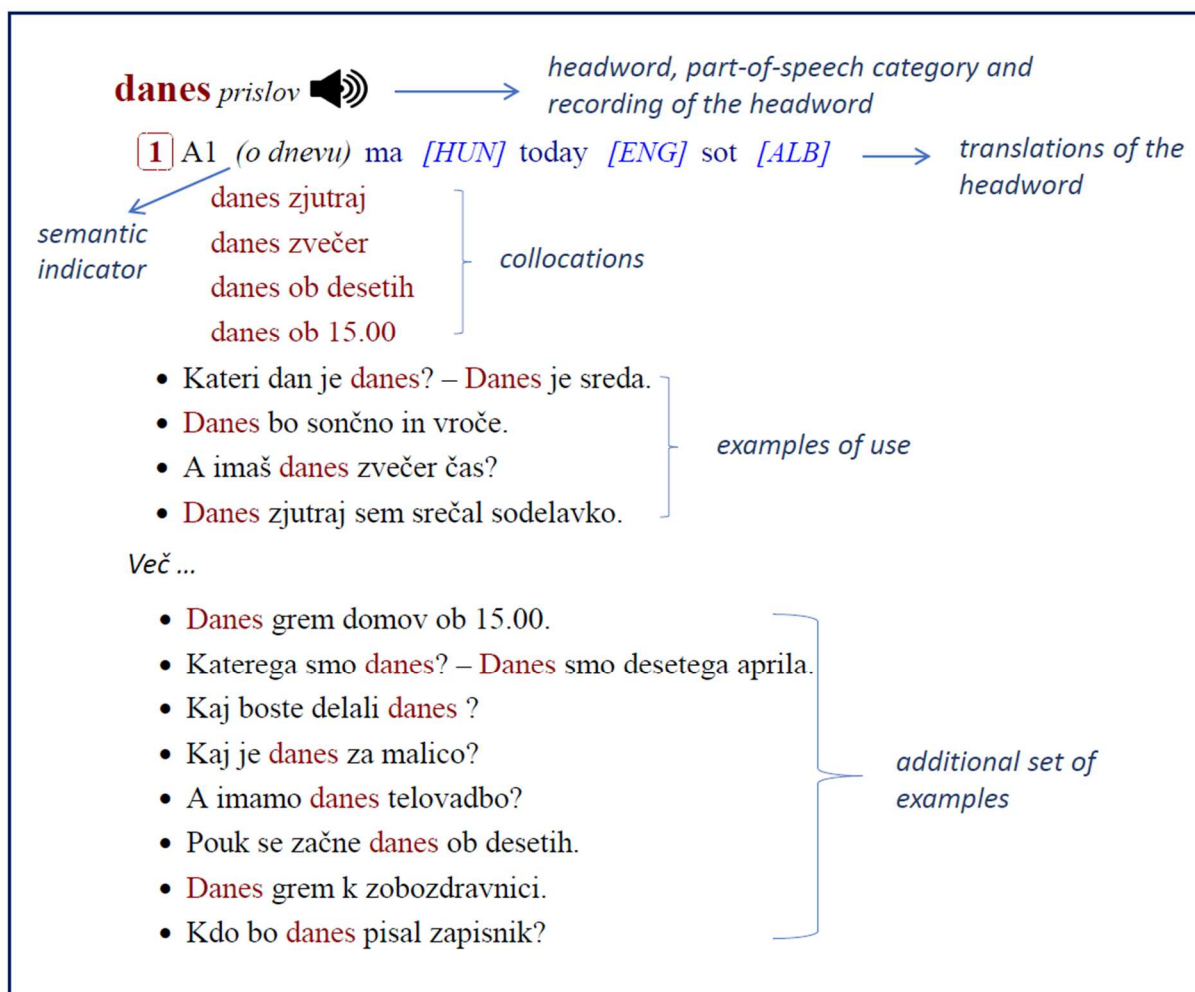


Figure 1: Sample entry for the adverb *danes* ‘today’.

4.3 Elements of the semantic description in the model presented

Considering the target users and the tool being user friendly, certain guidelines have been found relevant when creating entries and semantic descriptions for level A1. These guidelines are explained below.

The **semantic indicator** is one of the three segments of semantic information in the CJVT dictionary resources.¹⁴ It defines the meaning of a word concisely and distinctively in relation to its other meanings. For this purpose, semantic indicators are primarily used to create a “sense menu,” which is familiar from foreign language dictionaries and introduced in CJVT dictionary resources (cf. Collocation Dictionary of Modern Slovene 1.0, Comprehensive Slovenian–Hungarian Dictionary 1.0).

The semantic indicator should be informative for the target users. It should be short and clear. The semantic indicator is either a thematic category (as in similar language learning resources—e.g., English Profile—these are in line with language documents

¹⁴ In addition to indicators, the semantic description includes a label and an explanation.

such as *Threshold Level for Slovenian*)¹⁵ or hypernym (e.g., for the headword *bel* ‘white’, the indicator is *lastnost – barva* ‘characteristic – color’). Where this proves to be relevant and helpful, the semantic indicator should be supplemented by a synonym (e.g., in describing the meaning of some nouns: *punca:dekle* ‘girl’) and/or antonym (e.g., in describing the meaning of qualitative adjectives: *lep:grd* ‘beautiful:ugly’).

The **set of collocations** provides information about the most typical textual environment of the headword. When creating the set of collocations, we took into account collocations from the KUUS corpus that also showed semantic and statistical relevance in the Gigafida 2.0 reference corpus of written Slovenian (Krek et al. 2020). For example, in the case of the entry *bel* ‘white’, the collocation *bela barva* ‘white color’ was accepted because it has been verified as a collocation in both corpora. In addition to the aforementioned criterion of typicality, the criterion of a variety of syntactic relations or structures was taken into account; for example, the use of a noun in different cases, with different prepositions, verb valency, and so on (e.g., for the entry *babica* ‘grandmother’: *draga babica* ‘dear grandmother’, *obiskati babico* ‘to visit one’s grandmother’, *počitnice pri babici* ‘vacation at grandmother’s’, *dedek in babica* ‘grandfather and grandmother’). However, the linguistic competence of the target user has been taken into account.

The **usage examples** are taken from the level-A1 textbooks included in the KUUS corpus. They are typically one-sentence utterances. Regarding the form, declarative sentences (*Rad bi naročil pico.* ‘I would like to order a pizza.’), negative sentences (*Naša učilnica ni velika.* ‘Our classroom is not big.’), and interrogative sentences (*A greš zvečer na pijačo?* ‘Are you going for a drink tonight?’) are included as usage examples for each headword, if relevant. In principle, one-clause sentences have been included. In some cases, simple coordinated and subordinated sentence structures (e.g., *Kaj delaš, ko prideš domov?* ‘What do you do when you get home?’) with the conjunctions *in* ‘and’, *ali* ‘or’, *ampak* ‘but’, *ker* ‘because’, *ko* ‘when’ (limited to expressing time), *če* ‘if’ (limited to use with present and future forms), or *ki* ‘which’ (limited to the nominative case) have been used if the usage has been documented in the corpus. Especially when expressing time or location, dialogue forms were also included in the usage examples so the question word could provide contextual clues and/or illustrate the grammatical limitations of use (e.g., *Kje si? – Doma.* ‘Where are you? – At home.’; *Kdaj greš na dopust? – Avgusta.* ‘When are you going on vacation? – In August.’).

Vocabulary and morphosyntactic patterns in the usage examples correspond to the expected lexical and morphosyntactic ability of the target user. Inflectional word types are shown in their various forms (e.g., *Danes je sreda.* ‘Today is Wednesday.’; *V sredo imamo angleščino.* ‘We have English on Wednesday.’; *Tečaj imamo ob sredah.* ‘We have classes on Wednesdays.’). We have taken into account the fact that users usually have

¹⁵ Cf. the documents *Preživetvena raven za slovenščino* (Pirih Svetina, 2004, 2016) and *Sporazumevalni prag za slovenščino* (Ferbežar et al., 2004).

a slightly higher receptive ability than productive ability, and that they are able to use some reception strategies, especially if the examples are supported by pictures, if they can use their general knowledge or first language to help them understand the meaning, if the examples show a predictable communicative situation, or if the circumstances are familiar to the user (cf. CEFR Companion Volume, 2020: 54, 59–60, 175).

Due to transparency and the pedagogical maxim of progressivity, the usage examples are presented in two categories. The first set of examples, the three to five “core examples”, appear on the screen automatically, whereas an additional set of examples appears only on demand. Within the core examples, the headword is presented in various general domains or contexts (e.g., for the headword *danes* ‘today’: *Danes je sreda*. ‘Today is Wednesday.’; *Kateri dan je danes?* ‘What day is today?’; *Danes bo sončno in vroče*. ‘Today will be sunny and hot.’; *A imaš danes zvečer čas?* ‘Do you have time this evening?’). In the additional set of examples, the use of the headword in specific domains is illustrated; for example, in the context of school or work (e.g., *Kaj je danes za malico?* ‘What’s for (school) lunch today?’; *Danes imamo geografijo*. ‘Today we have geography.’; *Danes ne morem priti na sestanek*. ‘Today I can’t come to the meeting.’), and some usage examples beyond level A1 are shown (e.g., *Danes ponoči sem sanjala o tebi*. ‘Last night I dreamed about you.’).

The newly created lexical resource includes usage examples that function as self-sufficient even in isolation from a wider textual context, and that are comprehensible, accessible, and useful for the user (e.g., *Komaj čakam počitnice*. ‘I can’t wait for vacation to begin.’). Examples that were not semantically coherent without a context were not included (e.g., *Lepo, komaj čakam*. ‘Nice, I can’t wait’). The usage examples are selected from the level-A1 textbooks included in the KUUS corpus. In some entries the examples from the workbooks that complement the textbooks have also been manually included in the resource because the plan is to expand the corpus with workbooks (see section 5).

Where possible, **multimedia elements** (i.e., photographs and/or recordings of the headword) are included. As mentioned in the previous paragraph, these have an important explanatory function for users with limited linguistic ability.

4.4 Import of data into the Digital Dictionary Database for Slovene (DDDS)

Because it is essential for languages with fewer speakers to facilitate optimal connectivity and reusability of language resources and data, special care has been taken to ensure that all newly produced data are available for further use. The presented lexical information will be included in the Digital Dictionary Database for Slovene (DDDS) (Kosem et al., 2021a), which is being developed at the CJVT at the University of Ljubljana. The main aim of the DDDS is to offer a uniform set of concepts (i.e., senses) for various monolingual and bilingual dictionaries (with Slovenian as the source

language) and similar resources. Naturally, the integration of a resource targeted at nonnative speakers requires a few special features in the database; features that have been predicted since the beginning. For example, each sense in the DDDS can have more than one definition, depending on the type of resource. Similarly, examples can be attributed to one or many (or all) resources drawing on the data in the DDDS.

In the case of dictionary entries for nonnative speakers with CEFR-labeled senses, we will use the sense indicators already found in the DDDS. We expect to find most of the collocations from our entries in the DDDS already; as we already observed during the entry compilation, the collocations that are “new” are often those that are less typical in the reference (written) corpus and more typical of spoken language. Examples selected for the entries will initially be linked to this particular resource only. The information on which CEFR label should be attributed to which sense(s) is based on the KUUS corpus. Currently, the focus is on A1, and the senses already present in A1 textbooks are thus labeled as such. Sometimes a sense that is suitable for A2 or higher levels can potentially occur in A1, however we found such cases to be rare. Overall, the majority of headwords have only one A1 sense, and few two A1 senses. Expectedly, A1 senses are almost always the first senses of the headword. It is worth noting that the lexical resource includes both single-word and multi-word headwords with CEFR level labels. At the moment, multi-word headwords consist of compounds only (e.g. *bela kava*, literally ‘white coffee’ meaning ‘caffè latte’), but there are plans to add phraseology in the future.

5. Future work and Conclusion

This article presented the KUUS 1.0 corpus of textbooks for learning Slovenian as a second and foreign language, the core vocabulary list that was created on the basis of this corpus, and the construction of a lexical resource for Slovenian as a second and foreign language, currently with the vocabulary labeled as A1.

For the next version of the corpus, we aim to also include the workbooks because they contain examples of language use that are very valuable for the work we describe in section 4. The inclusion of workbooks will take place under the umbrella of a project called *Nadgradnja korpusov za slovenščino kot drugi in tuji jezik KOST in KUUS* (Expanding the KOST and KUUS Corpora of Slovenian as a Second and Foreign Language), with the improved version of the corpus available at the end of 2023.

For our next version of the list, we plan to manually review and label the words that were not included in the current list. These mainly consist of candidates for the levels B2 and C1, as well as some for lower levels that did not meet the inclusion criteria. In addition, we aim to confirm the CEFR labels assigned to each word by obtaining a wider consensus from experts that teach Slovenian as a second and foreign language.

The KUUS corpus has proven to be an invaluable resource in the development of a lexical resource for Slovenian as a second and foreign language. It is the first of its kind

for Slovenian. Moving forward, our aim is to expand this corpus-based lexical resource by incorporating vocabulary entries (both single-word and multi-word) for higher proficiency levels. Additionally, we plan to enhance the existing level-A1 explanations by including senses that are relevant to higher proficiency levels. The process of constructing the lexical descriptions, as described in this article, involves manual review and editing of the automatically extracted data. However, the subsequent major step, which involves creating pedagogically tailored sense definitions, will require more input from the authors. A specialized interface for the DDDS (Digital Dictionary of Slovene) is currently under development, which will streamline and enhance the efficiency of all stages of the lexicographic work. Once the lexical resource is published, we intend to evaluate its usability and gather feedback on user experience. This assessment will help identify priorities for further development, ensuring that future enhancements align with user needs and expectations.

6. Acknowledgements

This work was supported by the Slovenian Research Agency (ARRS) via the core programs Language Resources and Technologies for Slovene (P6-0411), Knowledge Technologies (P2-0103), and Slovene Language – Basic, Contrastive, and Applied Studies (P6-0215), and via the projects Empirical Foundations for Digitally Supported Development of Writing Skills (J7-3159), Quantitative and Qualitative Analysis of Unregulated Corporate Financial Reporting (J5-2554), and Computer-Assisted Multilingual News Discourse Analysis with Contextual Embeddings (J6-2581).

The KUUS corpus of textbooks for learning Slovenian as a second and foreign language and the vocabulary lists for levels A1, A2, and B1 were supported by CLARIN.SI. The project Expanding the Teaching Material *Čas za slovenščino 1* in the Digital Environment and Adapting the Material for Teaching Blind and Partially Sighted Adolescents is financially supported by the Slovenian Ministry of Culture.

7. References

- Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume* (2020). Strasbourg: Council of Europe Publishing. Available at: <https://rm.coe.int/common-european-framework-of-reference-for-languages-learning-teaching/16809ea0d4>
- Čibej, J., Gantar, K., Dobrovoljc, K., Krek, S., Holozan, P., Erjavec, T., Romih, M., Arhar Holdt, Š., Krsnik, L. & Robnik-Šikonja, M. (2022). *Morphological lexicon Sloleks 3.0*, Slovenian language resource repository CLARIN.SI, <http://hdl.handle.net/11356/1745>.
- Gril, P. (2022). Na tečaj: učenje in poučevanje slovenščine kot drugega in tujega jezika v Sloveniji. In N. Pirih Svetina & I. Ferbežar (eds.) *Na stičišču svetov: slovenščina kot drugi in tuji jezik, Obdobja 41*. Ljubljana: Založba Univerze v Ljubljani, pp. 117–127.
- Ferbežar, I., Knez, M., Markovič, A., Pirih Svetina, N., Schlamberger Brezar, M.,

- Stabej, M., Tivadar, H. & Zemljarič Miklavčič, J. (2004). *Sporazumevalni prag za slovenščino*. Ljubljana: Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovenistiko Filozofske fakultete, Ministrstvo RS za šolstvo, znanost in šport.
- Fišer, D., Ljubešič, N. & Erjavec, T. (2020). The Janes project: language resources and tools for Slovene user generated content. *Lang Resources and Evaluation* 54, pp. 223–246. Available at: <https://doi.org/10.1007/s10579-018-9425-z>
- 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(1), pp. 7–36.
- Knez, M., Ferbežar I., Kern Andoljšek, D. & Stabej M. (2021). *Evalvacija modelov učenja in poučevanja slovenščine kot drugega jezika za učence in dijake, ki jim slovenščina ni materni jezik. Zaključno poročilo*. Ljubljana: Center za slovenščino kot drugi in tuji jezik.
- Klemen, M., Arhar Holdt, Š., Pollak, S., Kosem, I., Huber, D. & Lutar, M. (2022). Korpus učbenikov za učenje slovenščine kot drugega in tujega jezika. In N. Pirih Svetina & I. Ferbežar (eds.) *Na stičišču svetov: slovenščina kot drugi in tuji jezik, Obdobja 41*. Ljubljana: Založba Univerze v Ljubljani, pp. 165–174.
- Klemen, M., Kosem, I., Arhar Holdt, Š., Pollak, S., Huber, D. & Lutar, M. (2022a). *Corpus of textbooks for learning Slovenian as L2 KUUS 1.0*, Slovenian language resource repository CLARIN.SI, <http://hdl.handle.net/11356/1696>.
- Klemen, M., Arhar Holdt, Š. & Pollak, S. (2022b). *Core vocabulary for Slovenian as L2 1.0*, Slovenian language resource repository CLARIN.SI, <http://hdl.handle.net/11356/1697>.
- Krek, S., Arhar Holdt, Š., Erjavec, T., Čibej, J., Repar, A., Gantar, P., Ljubešič, N., Kosem, I. & Dobrovoljc, K. (2020). In N. Calzolari (ed.) *Gigafida 2.0: the reference corpus of written standard Slovene. LREC 2020: Twelfth International Conference on Language Resources and Evaluation, Palais du Pharo, Marseille, France*. Paris: ELRA – European Language Resources Association, pp. 3340–3345. Available at: <http://www.lrec-conf.org/proceedings/lrec2020/LREC-2020.pdf>
- Krek, S., Laskowski, C., Robnik Šikonja, M., Kosem, I., Arhar Holdt, Š., Gantar, P., Čibej, J., Gorjanc, V., Klemenc, B. & Dobrovoljc, K. (2018). *Thesaurus of Modern Slovene 1.0*, Slovenian language resource repository CLARIN.SI, <http://hdl.handle.net/11356/1166>.
- Kosem, I., Čibej, J., Dobrovoljc, K., Erjavec, T., Ljubešič, N., Ponikvar, P., Šinček, M. & Krek, S. (2022). *Monitor corpus of Slovene Trendi 2022-10*, Slovenian language resource repository CLARIN.SI, <http://hdl.handle.net/11356/1681>.
- Kosem, I., Bálint Čeh, J., Ponikvar, P., Zaranšek, P., Kamensšek, U., Koša, P., Gróf, A., Böröcz, N., Harmat Császár, J., Szijártó, I., Šantak, B., Gantar, P., Krek, S., Roblek, R., Zgaga, K., Logar, U., Pori, E., Arhar Holdt, Š. & Gorjanc, V. (2021). *Comprehensive Slovenian-Hungarian Dictionary 1.0*, Slovenian language resource repository CLARIN.SI, <http://hdl.handle.net/11356/1453>.
- Kosem, I., Krek, S. & Gantar, P. (2021a). Semantic data should no longer exist in isolation: the Digital Dictionary Database of Slovenian. In Z. Gavriilidou, L.Mitits

- & S. Kiosses (eds.), EURALEX XIX, Congress of the European Association for Lexicography, Lexicography for inclusion. Komotini: Democritus University of Thrace, pp. 81–83. Available at: https://euralex2020.gr/wp-content/uploads/2021/09/EURALEX2020_BookOfAbstracts-Preview-1.pdf
- Kosem, I., Gantar, P., Krek, S., Arhar Holdt, Š., Čibej, J., Laskowski, C., Pori, E., Klemenc, B., Dobrovoljc, K., Gorjanc, V. & Ljubešić, N. (2019). *Collocations Dictionary of Modern Slovene KSSS 1.0*, Slovenian language resource repository CLARIN.SI, <http://hdl.handle.net/11356/1250>.
- Kosem, I., Rozman, T., Arhar Holdt, Š., Kocjančič, P., Laskowski, C. A. (2016). Šolar 2.0: nadgradnja korpusa šolskih pisnih izdelkov. In T. Erjavec & D. Fišer (eds.) *Zbornik konference Jezikovne tehnologije in digitalna humanistika*. Ljubljana: Znanstvena založba Filozofske fakultete, pp. 95–100. Available at: http://www.sdjt.si/wp/wp-content/uploads/2016/09/JTDH2016_Kosem-et-al_Solar-2-0-nadgradnja-korpusa-solskih-pisnih-izdelkov.pdf
- 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, Florence, Italy*. Association for Computational Linguistics, pp. 29–34.
- Logar, N., Grčar, M., Brakus, M., Erjavec, T., Arhar Holdt, Š. & Krek, S. (2012). *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida inccKRES: gradnja, vsebina, uporaba*. Ljubljana: Trojina; Fakulteta za družbene vede.
- Pirih Svetina, N., Rigler Šilc, K., Lavrič, M., Ferbežar, I. & Jerman, T. (2004). *Preživetvena raven v slovenščini*. Krakov: TAIWPN Universitas.
- Pirih Svetina, N. (2016). *Preživetvena raven za slovenščino: za potrebe programa Opismenjevanje v slovenščini za odrasle govorce drugih jezikov*. Ljubljana: Center za slovenščino kot drugi in tuji jezik. Available at: https://centerslo.si/wp-content/uploads/2016/07/IC_Preživetvena_2016.pdf
- Pollak, S., Arhar Holdt, Š., Krek, S., Robnik-Šikonja, M. (2020). *Reference List of Slovene Frequent Common Words*, Slovenian language resource repository CLARIN.SI, <http://hdl.handle.net/11356/1346>.
- Verdonik, D. & Zwitter Vitez, A. (2011). *Slovenski govorni korpus Gos*. Ljubljana: Trojina, zavod za uporabno slovenistiko.
- Volodina, E., Alfter, D. & Lindström Tiedemann, T. (2022). Crowdsourcing ratings for single lexical items: a core vocabulary perspective. *Slovenščina 2.0*, 10(2), pp. 5–61.

Author Index

- Abdelzaher Esra, 545
Appel Kirsten Lundholm, 392
Arhar Holdt Špela, 178, 376, 501
- Bajčetić Lenka, 638
Benko Vladimír, 75
Blahuš Marek, 613, 650
- Challis Kate, 141
Chambó Santiago, 587
Cukr Michal, 613, 650
- Declerck Thierry, 638
Denisová Michaela, 1
Drusa Tom, 141
Ducasse Mireille, 293
- Eckart Thomas, 280
Elizbarashvili Archil, 293
Ene Vainik, 476
Erjavec Tomaž, 449
Eva Pori, 376
- Gantar Polona, 376, 501, 663
Gapsa Magdalena, 178
Girnat Boris, 466
Grasmanis Mikus, 410
Grūzītis Normunds, 410
Günther Luke, 246
- Hassert Naïma, 216
Heid Ulrich, 466
Herold Axel, 280
Holdt Špela Arha, 663
Héja Enikó, 160
- Isaacs Loryn, 258
- Jakubíček Miloš, 518, 613, 650
Jürviste Madis, 104
- Kern Boris, 429, 449
Khachidze Manana, 293
Klemen Matej, 663
Knez Mihaela, 663
- Kocek Jan, 75
Koppel Kristina, 104
Kosem Iztok, 376, 501, 663
Kovář Vojtěch, 613, 650
Krashtan Tamila, 308
Krek Simon, 376, 501
Kruse Theresa, 466
Körner Erik, 280
- Langemets Margit, 104
Lareau François, 216
Lazić Daria, 322
León-Araúz Pilar, 587
Ligeti-Nagy Noémi, 160
Lipp Veronika, 160
Lohk Ahti, 476
Lugli Ligeia, 201
- Malčovský Peter, 75
Marjanović Saša, 567
Martinc Matej, 201
Matijević Maja, 322
Matuška Ondřej, 650
Mondaca Francisco, 230, 246
- Neuefeind Claes, 230
Nimb Sanni, 392
- Ohlídalová Vlasta, 613
Ostroški Anić Ana, 322
- Paikens Pēteris, 410
Paulsen Geda, 476
Pavić Martina, 322
Pelicon Andraž, 201, 449
Phoodai Chayanon, 345
Podpečan Vid, 19
Pollak Senja, 19, 201, 429, 449, 663
Pori Eva, 663
Pranjić Marko, 449
Pretkalniņa Lauma, 410
- Rau Felix, 246
Rezania Kianoosh, 230
Rikk Richárd, 345

Rituma Laura, 410
Robnik-Šikonja Marko, 376
Rundell Michael, 518
Rychlý Pavel, 1

Sass Bálint, 534
Schildkamp Philip, 246
Simon László, 160
Stramljič Breznik Irena, 449
Strankale Laine, 410
Synchak Olena, 118
Sérasset Gilles, 638
Sørensen Nathalie Hau, 392
Sørensen Nicolai Hartvig, 392

Tavast Arvi, 104
Tiberius Carole, 53
Tittel Sabine, 39
Tomazin Mateja Jemec, 19

Tran Thi Hong Hanh, 19
Tsintsadze Magda, 293
Tuulik Maria, 476
Tóth Ágoston, 545

Ulčar Matej, 429

Van Huyssteen Gerhard, 53
Vapper Silver, 104
Voršič Ines, 429

Widmann Thomas, 91
Wiegand Frank, 280

Znotiņš Artūrs, 410

Čéplö Slavomír, 230
Škrabal Michal, 75

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

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



