Simon Krek (Ed.)

Book of Abstracts

of the Workshop

Large Language Models and Lexicography

8 October 2024

Cavtat, Croatia

PRE-PUBLICATION

Title:	Book of Abstracts of the Workshop Large Language Models and Lexicography
Edited by:	Simon Krek
Technical editor:	Tina Munda

FOREWORD

The overall topic of the workshop is the use of Large Language Models (LLMs) in lexicography. The workshop aims to explore how these models aid in linguistic analysis and generation of dictionary data, enhancing dictionary development through automation of processes. The topics also include identifying new word usages and trends, and how LLMs facilitate multilingual lexicography, as well as the ethical implications of AI in lexicography, including concerns about bias and cultural sensitivity, or any other topics related to the use of LLMs in lexicography. The workshop is of interest to lexicographers and language technology experts, offering insights into the trends of AI-assisted lexicography and preparing them for digital transformation.

Dr Simon Krek
Chair of the workshop

PROGRAMME COMMITEE

Simon Krek, Jožef Stefan Institute, Chair

Iztok Kosem, Faculty of Arts, University of Ljubljana

Ilan Kernerman, K Dictionaries

Polona Gantar, Faculty of Arts, University of Ljubljana

Bolette Sandford-Pedersen, University of Copenhagen

Sussi Olsen, Centre for Language Technology, University of Copenhagen

Carole Tiberius, Instituut voor Nederlandse Lexicologie

Sanni Nimb, Det Danske Sprog- og Litteraturselskab (DSL)

Ranka Stanković, University of Belgrade

John P. McCrae, University of Galway, Ireland

Rusudan Makhachashvili, Borys Grinchenko Kyiv University

Annette Klosa-Kückelhaus, Leibniz-Institut für Deutsche Sprache

ORGANISING COMMITEE

Simon Krek, Jožef Stefan Institute

Tina Munda, Centre for Language Resources and Technologies, University of Ljubljana

ORGANISERS

Centre for Language Resources and Technologies, University of Ljubljana

ELEXIS Association

TABLE OF CONTENTS

Nataliia Cheilytko, Ruprecht von Waldenfels

Semantic Change and Lexical Variation in Ukrainian with Vector Representations and LLM

Keywords Semantic change; lexical variation; word sense disambiguation; Ukrainian; vector representation; vector space; large language model

Ukrainian is a sizeable European language with a high degree of variability and versatile semantic developments. Before WWII, Ukrainian was a multi-standard language (using the (Auer, 2021) terminology), having active language contacts with neighboring Polish and Russian. This caused significant variation at the time and continues to the present despite the convergence of standard Ukrainian during the second part of the XX century. Moreover, today, due to intense and active war-driven processes within the country, Ukrainian exposes a new wave of innovative semantic phenomena. To get a bird's-eye view of those trends in the XX — XXI century, we follow the methodological principles formulated in (Geeraerts et al., 2023). The project aims to explore the interplay of semasiological and onomasiological changes in Ukrainian based on the bottom-up large-corpus-based distributional approach.

To automate the identification of variation and semantic changes of lexemes, we developed an R&D pipeline with token-level vector space models, accounting for second-order word co-occurrence, similar to (Hilpert & Correia Saavedra, 2020). As a result, for each lexeme under examination, we extracted its occurrences (taking into account six contextual tokens on the left and the right) from a significantly large corpus of Ukrainian - the General Regionally Annotated Corpus of Ukrainian, aka GRAC, created by (Shvedova et al., - 2024). Each occurrence was represented as an embedding (an aggregated vector) and visualized in a semantic vector space. Fig. 1-2 provides an example of a vector space built for the lexemes *bavovna* 'cotton', *lion* 'flax', and *vybuh* 'explosion'. According to the distributional hypothesis, utterances with similar meaning should cluster together, and embeddings of synonyms and thematically close words would overlap in a semantic space. In contrast, utterances with different meanings would take different locations.

The proposed pipeline was successful in identifying various cases of semantic changes. For example, it was sensitive to the rapid development of a fundamentally new sense for the noun *bavovna*. Traditionally, this word has two meanings: 'cotton as a flower' and 'cotton as a material'. With the ongoing war in Ukraine since 2022, this lexeme has started denoting a completely different idea of an explosion in media texts. The ground for the new sense stemmed from a creative rethinking of the fact that the Russian word *hlopok* (a Russian equivalent for cotton, but also having a homograph with the sense ‘a clap’) was widely used in the Russian media to veil and diminish the events of explosions by calling them non-significant clap-line sounds.

A B

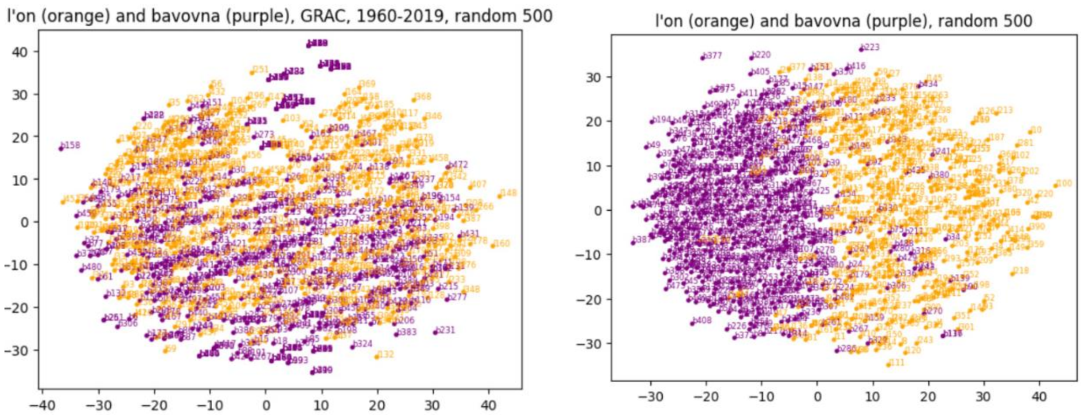

Fig. 1: Embeddings of lion ‘flax’ and bavovna ‘cotton’ before 2020 (A) and after (B)

Fig. 1 shows a 2D view obtained with the Multidimensional Scaling technique for the contextualized vectors of 500 random occurrences for the words *bavovna* (‘cotton’) and another closely related noun l’on (‘linen’ as a material and ‘flax flower’). During 1960-2019 (Fig. 1 A) *bavovna* was used in quite similar contexts as *l’on.* An overlap of dots corresponding to occurrences of each word confirms that. Thus, it indicates the similarity of their meanings. However, starting the war in 2022 (Fig. 1 B), their occurrences cluster apart, which shows discrimination of the meanings.

A B

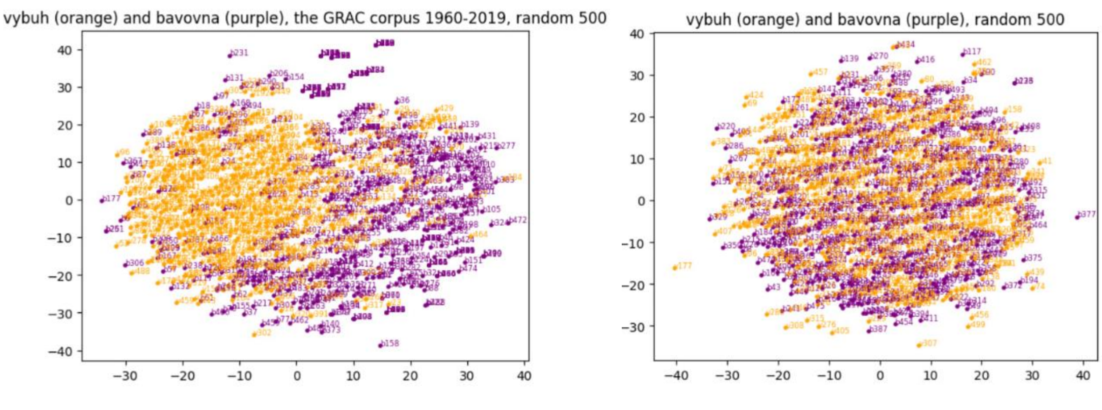

Fig. 2: Embeddings of vybuh ‘explosion’ and bavovna ‘cotton’ before 2020 (A) and after (B)

Fig 2 compares occurrences of *bavovna* (‘cotton’) and *vybuh* (‘an explosion’), which shows that before the war, their occurrences were separate areas in the semantic space (Fig. 2 A, whereas they have a significant overlap from 2022 (Fig. 2 B). A more detailed analysis of the occurrences confirmed that *bavovna* had acquired a new meaning — ‘explosion’ and is being used more frequently in this sense in media texts than in other senses — ‘cotton flower’, ‘cotton material’. Moreover, the opposite picture for *bavovna* and *vybuh* (‘an explosion’) —

from 1960 to 2019, those lexemes were separate clusters (the plot on the left), but from 2022, they have similar usage (the plot on the right).

Having such a lexical change monitoring pipeline for Ukrainian makes it possible to explore trends in language development — region- and register-wise — and reveal how semantic change characterizes the standardization process of Ukrainian.

Among other typical semantic change cases is the reactivation of a word sense previously used in relatively narrow discourses. For example, since 2022, *tryvoga* has been used more often to denote an air alarm than anxiety**,** compared to texts before 2022. Back then, the primary usage of *tryvoga* was close to such words as *zanepokojennja ‘*concern*’, hvyluvannja* ‘anxiety’*,* and *tuga* ‘longing’. However, since the invasion in 2022, *tryvoga* is used to denote a different sense uncommon for decades. Thus, its nearest synonyms are *nebezpeka ‘*danger*’, vidkljuchennja ‘*shutdown*’*. Another example is the synonyms ***motor*** and ***dvygun***, which were used similarly to denote ‘engine’ in the first half of the XX century, however starting 1950s the concept structure changed: the *motor* prototypical usage is to demote an engine of everyday common-life devices (a car, a small boat, a motorbike, a household device). In contrast, *dvyhun* is used to name heavy industrial engines (Cheilytko & Waldenfels, 2023).

Other types of semantic change revealed with the proposed workflow are the disappearance of a regional difference in using a word in a particular sense (the adjective *povazhnyi* lost the sense ‘severe, dangerous’*,* typical for the Western variety of Ukrainian before WWII), and the predominance of a particular synonymous variant for a concept (***kvytok*** over ***bilet*** both meaning ‘ticket, pass’ , *vitrylo* over *parus* meaning ‘sail’ by the last quarter of the XX century).

In addition, we performed experiments with LLMs (GPT-4o and GPT-4o-mini) to verify whether those models can capture a sense of a word in a context. The scope of the was 107 concepts represented by 280 lexemes given six sub-corpora formed from the GRAC by two dimensions - region and time: West and East regions, 1920 - 1939 (before WWII), 1940 - 1969 (after WWII), and 1970 - 1990 (before Ukraine became an independent state). Table 1 reports on the size of the sub-corpora in tokens and the number of concordance lines (equal to a sentence) processed by the models.

Considering the cost of the GPT-4o model, those prompt requests that assumed a large text response were limited to 20 most promising lexemes (requests to define word senses of a word, to list its synonyms, to explain a difference and commonalities between synonyms). A larger experiment formulated as a word sense classification task, thus resulting in a short response with a sense label, was performed for approx. 450,000 occurrences of 280 words. For the large dataset, we used GPT-4o-mini since accuracy assessment for random 500 occurrences from 1920-1939 (as the most problematic period) showed that GPT-4o-mini performs even better than GPT-4o: 0,9 for GPT-4o-mini vs. 0,84 for GPT-4o.

Table 1: Data for the Word Sense Disambiguation Task

	West 1920-39	**East 1920-39**	**West 1940-69**	**East 1940-69**	**West 1970-90**	**East 1970-90**	**Total**
Tokens in a sub-corpus	1,717,292	2,262,798	4,271,384	12,481,935	3,369,667	34,372,568	**58,475,644**
Sentences with chosen lemmas	31,844	42,438	65,709	102,277	65,709	143,465	**451,442**

To organize the experiment, we made the following decisions:

- Use an unbalanced dataset to get more utterances for rare words.
- Limit frequent words (more than 1k per corpus) to 1k random occurrences.
- The input sentences comprise a sequence of raw tokens.
- Predefine sense labels by analyzing Ukrainian dictionaries and senses proposed for a word by GPT. Add an option 'other' to allow potential new senses.
- Define sense labels and formulate the prompt in English acting as a metalanguage.

A prompt example:

{"Classify the meaning of the Ukrainian word ***батько*** in the following sentence:

Батько мій інженер !.. *into one of the following senses: ['Father', 'Founder or 'Patriarch', 'Older Man', 'Benefactor, patron', 'other']*

Output format - a sense."}

Output: 'Father'

(A comment: *'Батько мій інженер !..'* translates to English as *'My father is an engineer'*).

Both models are quite successful in detecting senses of ambiguous lexemes for the most recent state of Ukrainian. However, they lack the knowledge to deal with regionally specific senses and older occurrences, especially before WWII. For example, it was hard for GPT-4o to recognize an older sense of **baba** ('woman', 'elder woman', 'grandmother') – ***'woman'*** rather than ***'elder woman'.*** Meanwhile, GPT-4o-mini handled this difference more successfully (accuracy rate: 0,69 for GPT-4o vs. 0,75 for GPT-4o-mini). Therefore, for such cases, more traditional count-based vectors bring more value. Alternatively, GPT has to be additionally fine-tuned.

To conclude, the initial set of experiments applied to known test cases, including synonymous pairs representing certain concepts, as well as individual polysemous lexemes, proved that the proposed distributional techniques, including the semantic vector models together with LLMs, are able to capture both similarities and distinctions in the semantics of the word occurrences, trace semantic change and variation, which can serve as a solid ground for the subsequent Ukrainian semantic change and lexical variation research.

Combining two models (count-based vectors and LLMs) makes it possible to deal with weak parts of each of them: the former encounters issues with consistent solving of the word sense disambiguation task, which is compensated by utilizing LLMs. Meanwhile, the LLMs turned out to be insensitive to regional and diachronic variations of Ukrainian, causing model hallucinations. Applying vector modelling and visualization to specialized regional datasets overcomes the issue.

For the next R & D phases of the project, the ambition is to automate the creation of onomasiological profiles to model the structure of the Ukrainian regional varieties, explore their peculiarities, and measure their from each other.

References

Cheilytko, N. & von Waldenfels, R. (2023). Exploring Word Sense Distribution in Ukrainian with a Semantic Vector Space Model. In Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP), pages 73–78, Dubrovnik, Croatia. Association for Computational Linguistics. https://aclanthology.org/2023.unlp-1.9/

Hilpert, M., & Correia Saavedra, D. (2020). Using token-based semantic vector spaces for corpus-linguistic analyses: From practical applications to tests of theoretical claims. Corpus Linguistics and Linguistic Theory, 16(2), 393-424. https://doc.rero.ch/record/323542/files/Hilpert_Martin_-_Using_token-based_semantic_vector_spaces_for_corpus-linguistic_20181126.pdf

Auer, P. (2021). Reflections on linguistic pluricentricity. Sociolinguistica, vol. 35, no. 1, 2021, pp. 29-47. https://doi.org/10.1515/soci-2021-0003

Geeraerts, D., Speelman, D., Heylen, K., Montes, M., De Pascale, S., Franco, K., and Lang, M. (2024). Lexical Variation and Change. A Distributional Semantic Approach. (Oxford, 2023; online edn, Oxford Academic, 23 Nov. 2023). https://doi.org/10.1093/oso/9780198890676.001.0001.

Shvedova, M., von Waldenfels, R., Yarygin, S., Rysin, A., Starko, V., Nikolajenko, T. (2017-2024). GRAC: General Regionally Annotated Corpus of Ukrainian, from https://uacorpus.org/

Contact information

Nataliia Cheilytko

Friedrich Schiller University Jena

natalia.cheilytko@gmail.com

Ruprecht von Waldenfels

Friedrich Schiller University Jena

ruprecht.waldenfels@uni-jena.de

Gilles-Maurice de Schryver

The Road towards Fine-Tuned LLMs for Lexicography

Keywords LLMs; ChatGPT; out-of-the-box GPT; custom GPT; fine-tuned GPT; English; Portuguese; Bantu; Xhosa; Zulu; Lusoga; Luganda

Ever since the first LLM has been put into the hands of the general public, by means of the release of a nifty interface to 'chat' with it, the world has been in a frenzy. We are referring, of course, to the release of ChatGPT in November 2022. Already in February 2023, during a live demo from Tokyo, the first uses in the field of lexicography were extolled. Convinced that the future had arrived, AI functionality was added to the dictionary writing system TLex using OpenAI's GPT-3 API with which a public online dictionary was updated in real time with new entries generated by the AI a few minutes prior (de Schryver & Joffe, 2023). In the months that followed, colleagues on all continents joined the fray, at the conferences of the DSNA (Barrett, 2023), ASIALEX (McKean & Fitzgerald, 2023; Rundell, 2023), and eLex (de Schryver et al., 2023; Jakubíček & Rundell, 2023; Nichols, 2023; Tran et al., 2023), followed by the first publications in professional journals (de Schryver, 2023b; Lew, 2023).

By and large, the most successful experiments were reported for applications using English,[1] with critical notes regarding the usefulness for other languages mostly missing from the conversation. A notable exception is Jakubíček and Rundell (2023, pp. 522-523) who briefly looked at Czech (11 million speakers) and derided the use of ChatGPT for it. The results of the first in-depth and detailed attempt to study the lexicographic potential of using ChatGPT for a language other than English, were presented at the inaugural Americalex-S conference in São Paulo in October 2023 (de Schryver, 2023a). Although the assumption was that ChatGPT would be able to do an excellent job for Portuguese (260 million speakers), it was shown that it got things very wrong for the compilation of both monolingual Portuguese as well as bilingual Portuguese-English dictionaries. By way of illustration, Figure 1 displays one of the slides from that talk. While this Portuguese material may at first appear rather convincing (it fooled two professors!), it is riddled with errors. The reason? One is actually looking at English lexical analysis through the eyes of Portuguese. Reformulated, everything one sees there was created internally *in* English, *for* and *about* English, and was merely translated back into Portuguese (the prompt was in Portuguese). If a further analogy is allowed: Given the overwhelming amount of English in an LLM, the model is simply incapable to 'think' in any other language, let alone analyse it linguistically.

[1] See for instance the supplementary material for a COBUILD-style monolingual English dictionary compiled by ChatGPT here https://academic.oup.com/ijl/article/36/4/355/7288213#supplementary-data.

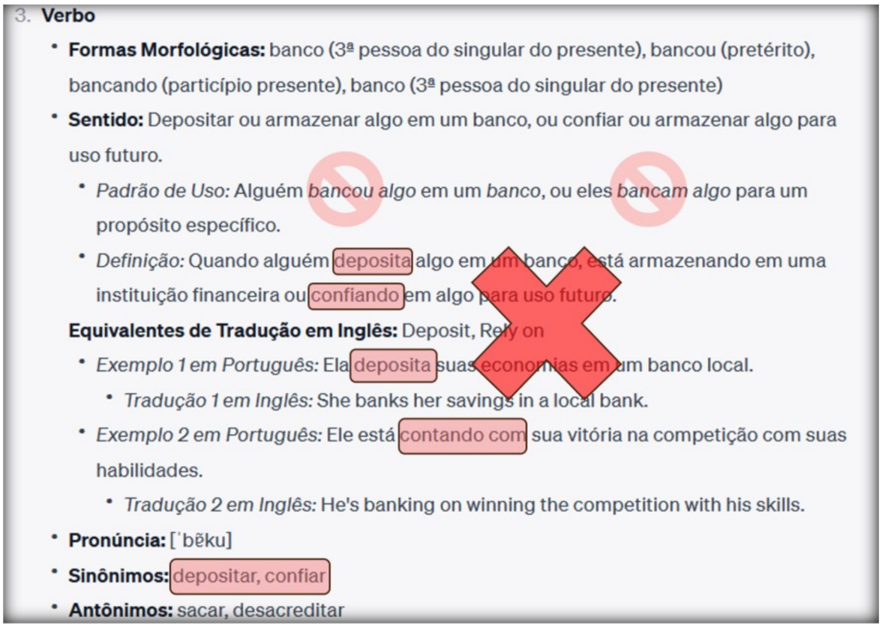

Fig. 1: The 3rd sense for the entry ***banco*** in a COBUILD-style bilingual Portuguese-English dictionary as proffered by ChatGPT-3.5. This sense, 'to bank money', simply does not exist in Portuguese, but ChatGPT not only 'created' this sense, it went as far as inventing conjugations for it to then use them as well (**bancou*, **bancam*), proceeds with a convincing COBUILD-style full-sentence definition but forgets to use the lemma and instead uses the (correct!) verb '*depositar*', and finally does the same when providing the example. Likewise for the phrasal verb 'bank on' which also does not use ***banco*** in Portuguese, but rather '*contar com*' as seen in the second example. ChatGPT's cover is blown when it offers its list of synonyms, repeating what it had used all along in the microstructure already.

However, during the remainder of 2023, and the start of 2024, colleagues in lexicography (and terminography) continued to sing the praises of ChatGPT, and fabulous results indeed kept coming in, see for instance Rees and Lew (2023, December 13), Franceschi and Pinnavaia (2023, December 20), San Martín (2024, February 25), Ptasznik and Lew (2024, March 25) and Cai et al. (2024, April 9) — needless to say, all focused on English.

With the belief that an LLM like ChatGPT is 'fantastic' for English, and the knowledge that it is often 'fake' for bilingual lexicography, the next step was to show that it simply 'fails' for exotic languages. This is exactly what was done in April 2024 during a talk at the University of Missouri (de Schryver, 2024b), where various attempts were made at having ChatGPT compile proper dictionary articles for a variety of Bantu languages, amongst others Swahili (100 million speakers) and Xhosa (19 million speakers). Figure 2 reproduces one of the slides from that talk, from which it is clear that ChatGPT cannot even pick 20 frequent words from Xhosa, as it includes a word from the neighbouring and bigger language Zulu (spoken by 27 million people). In current off-the-shelf LLMs English not only crushes other big languages, but small languages crush the even smaller ones, making them useless for lexicography.

More recently, colleagues have also begun to give attention to lexicographic applications in languages other than English, such as for Spanish (600 million speakers) (Fuertes-Olivera, 2024; Tarp & Nomdedeu-Rull, 2024), as well as exotic varieties of English such as Singlish (4 million speakers) (Chow et al., 2024). Concurrently, the pendulum is swinging back, and the need for the human touch in lexicography is being stressed again (Lew, 2024, pp. 6-7).

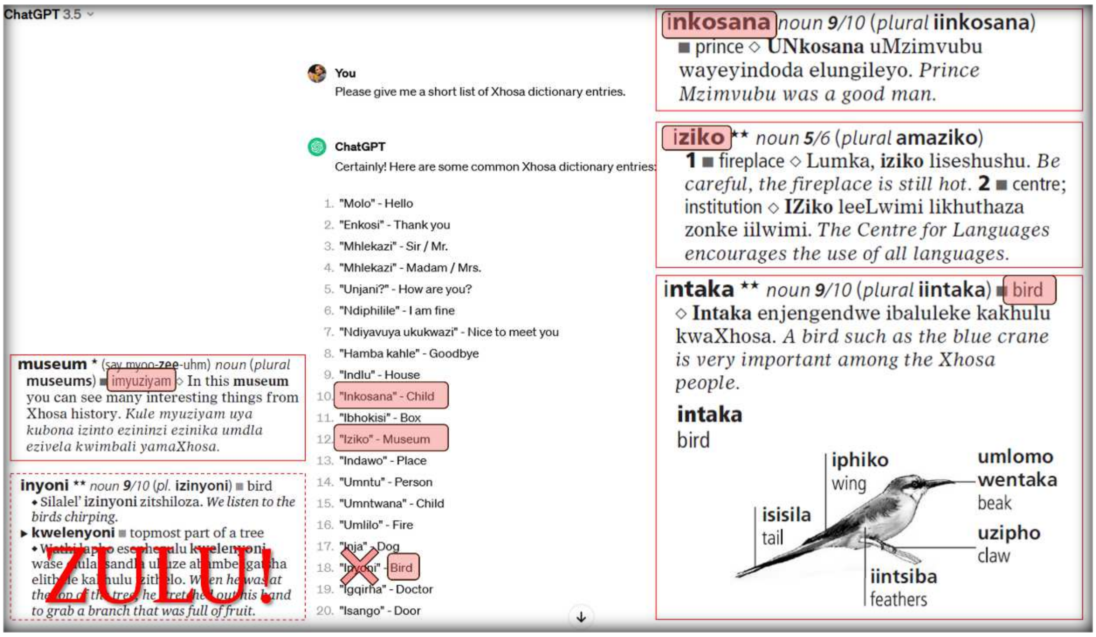

Fig. 2: Asking ChatGPT for a short list of Xhosa dictionary entries is underwhelming. Of the 20 'common' words two are wrongly translated: The noun ***inkosana*** is a 'prince', not a '*child', and ***iziko*** is a 'fireplace; centre, institution', not a '*museum' for which the loanword '*imyuziyam'* is used — as may be seen from the top three dictionary articles taken from OUP's Xhosa-English School Dictionary (de Schryver & Reynolds, 2014). Additionally and worryingly, ChatGPT even includes the Zulu word ***inyoni*** for 'bird' where it is actually ***intaka*** in Xhosa — as may be seen from the bottom two dictionary articles, respectively taken from OUP's Zulu-English School Dictionary (de Schryver, 2010) and the said Xhosa dictionary.

Even though Chatbot arenas now exist (e.g. https://lmsys.org/) where one can currently pair any two of up to a 100 LLMs — and thus choose from models like GPT (OpenAI), Claude (Anthropic), Gemini & Bard (Google), Llama (Meta), Command R+ (Cohere), Mistral, etc. — OpenAI's GPT models are still 'winning' overall, so it is not surprising that lexicographers have stuck by ChatGPT. Also, all the studies reviewed so far took LLMs straight 'out-of-the-box', but since the release of GPT-4o in mid-May 2024, easy customisation has been put into the hands of users. That is why, one week after the release of GPT-4o a serious attempt was made to customise this model by trying the impossible: Could it compile dictionary articles for neologisms in the extremely under-documented Bantu language Lusoga (3 million speakers, with no written tradition), by feeding GPT-4o everything that exists on Lusoga, viz. a corpus of 3.7 million tokens (de Schryver & Nabirye, 2022), and a full comprehensive monolingual dictionary in XML, the *Eiwanika* (Nabirye, 2012)? The outcome, presented at LREC-COLING in Turin (de Schryver, 2024a), was tantalising: whereas GPT-4o generated only nonsense out-of-the-box, it started to make sense with this customisation, see Figure 3. While there is still a long way to go to get to proper lexicography, and while the neighbouring language Luganda (spoken by 20 million people) still creeps in, seeing that GPT-4o can actually learn a new language and generate material in it is promising. Moreover, when asked to explain where the data for the concocted examples come from, GPT-4o is able to point to the various corpus sections from where it sourced and summarised information, meaning that the LLM truly stops being the black box feared by lexicographers. At the same time, customisation for an unknown exotic language is not good enough, so we need more.

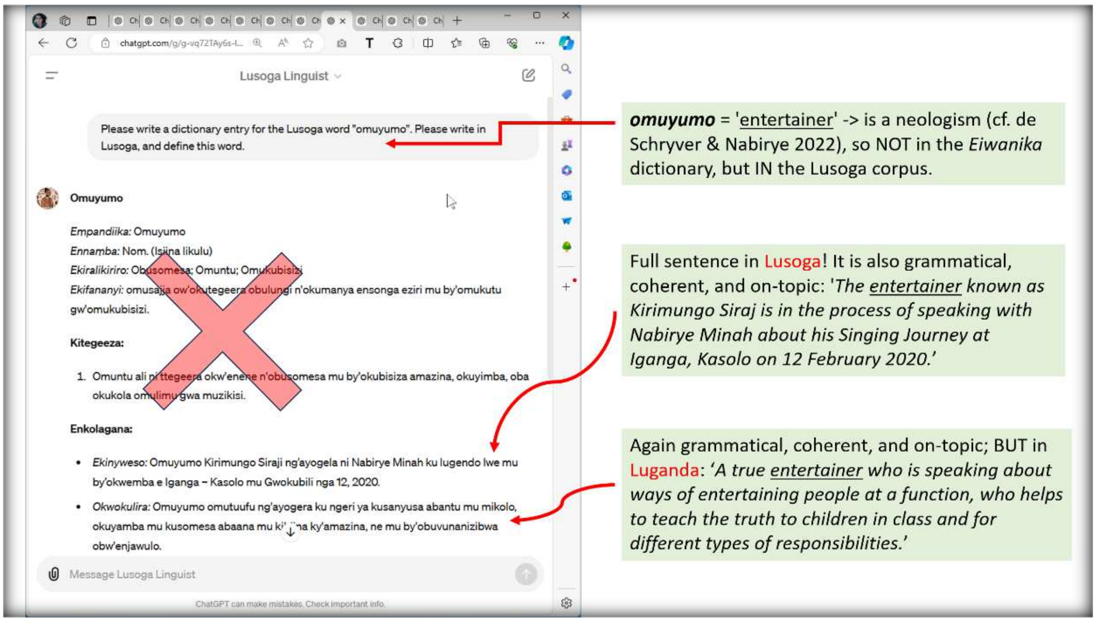

Fig. 3: Customising GPT-4o for Lusoga lexicography, a language for which it generates only rubbish out-of-the-box. Here ChatGPT is asked to compile an article for ***omuyumo*** 'entertainer', a word absent from the dictionary but present in the corpus – both of which were added to its 'knowledge base'. While the opening section makes no lexicographic sense, the first example is grammatical, coherent and on-topic Lusoga!

Armed with all the facts presented so far, and having proceeded from an out-of-the-box GPT (in February 2023), to a customised GPT for lexicographic purposes (in May 2024), we are now ready and willing to take the next step, namely the fine-tuning of an LLM to compile or even act as a dictionary in its own right (in October 2024). The differences between the three types are summarised in Table 1.

Table 1: Comparing out-of-the-box, customisation and the fine-tuning of GPT models for lexicography

Feature	Out-of-the-Box GPT	Custom GPT	Fine-Tuning
Modifies Core Model	No	No	Yes
Training Required	No	No, leverages pre-trained model	Yes, on a specific dataset
User Input	Text prompt	Instructions and/or relevant documents	New data for the model to learn from
Complexity	Least complex, easy to use	Less complex, potentially no-code/low-code	More complex, requires expertise
Cost	Least expensive	More expensive	A lot more expensive
Output Control	Limited control over the generated text	More control over the direction of the output	High degree of control over the generated text
Suitability for 'exotic lexicography'	Limited, may not be familiar with 'exotic language'	Can be adapted to the 'exotic language' through prompts and instructions	More directly addresses the 'exotic language' through fine-tuning

The most important difference is listed in the first line: we will need to modify the core model. In keeping with seeking a true challenge, as hinted at in the last line: we will proceed with an exotic language about which the model has initially no clue. At this stage, we are still running our experiments. But whatever the outcomes, we intend to report on them during the workshop, warts and all.

References

Barrett, G. (2023). *Defin-o-bots: Challenging A.I. to create usable dictionary content*. Paper presented at the 24th Biennial Conference of the Dictionary Society of North America, Boulder, CO, USA, 31 May – 3 June 2023.

Cai, B., Ng, C., Tan, D., & Hotama, S. (2024, April 9). *Low-cost generation and evaluation of dictionary example sentences*. Paper presented at the 2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Mexico City, Mexico, 16-21 June 2024. https://doi.org/10.48550/arXiv.2404.06224

Chow, S. Y., Shin, C.-U., & Bond, F. (2024). This word mean what: Constructing a Singlish dictionary with ChatGPT. Proceedings of the 2nd Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia (EURALI-II), co-located with the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), Turin, Italy, 25 May 2024.

de Schryver, G.-M. (2010). *Oxford Bilingual School Dictionary: Zulu and English / Isichazamazwi Sesikole Esinezilimi Ezimbili: IsiZulu NesiNgisi, Esishicilelwe abakwa-Oxford*. Oxford University Press Southern Africa.

de Schryver, G.-M. (2023a). *Contemporary lexicography: A case study in AI [Keynote Lecture]*. Paper presented at the Inaugural Conference of the Association of Lexicography for the Americas: South, Central, Caribbean, and Mexico (Americalex-S), São Paulo, Brazil, 20-25 October 2023. https://youtu.be/watch?v=yAxzTx7A2LU

de Schryver, G.-M. (2023b). Generative AI and lexicography: The current state of the art using ChatGPT. *International Journal of Lexicography*, *36*(4), 355–387 + 10 pp. of supplementary material online.

de Schryver, G.-M. (2024a). *Customising LLMs for lexicography [Keynote Lecture]*. Paper presented at the 8th International Workshop on Cognitive Aspects of the Lexicon (CogALex-VIII), co-located with the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), Turin, Italy, 20 May 2024.

de Schryver, G.-M. (2024b). *Generative Artificial Intelligence & Three Lexicographic Fs: FANTASTIC for monolingual English dictionaries, FAKE for translation dictionaries, FAIL for exotic dictionaries [Invited Talk]*. Paper presented at the Department of English, University of Missouri, Columbia, MO, USA, 5 April 2024.

de Schryver, G.-M., & Joffe, D. (2023). *The end of lexicography, welcome to the machine: On how ChatGPT can already take over all of the dictionary maker's tasks [Invited Talk]*. Paper presented at the 20th CODH Seminar, Center for Open Data in the Humanities, Research Organization of Information and Systems, National Institute of Informatics, Tokyo, Japan, 27 February 2023. https://youtu.be/watch?v=mEorw0yefAs

de Schryver, G.-M., & Nabirye, M. (2022). Towards a monitor corpus for a Bantu language. A case study of neology detection in Lusoga. In A. Klosa-Kückelhaus, S. Engelberg, C. Möhrs, & P. Storjohann (Eds.), *Dictionaries and Society. Book of Abstracts of the XX EURALEX International Congress, 12-16 July 2022, Mannheim, Germany* (pp. 233). IDS-Verlag.

de Schryver, G.-M., & Reynolds, M. (2014). *Oxford Bilingual School Dictionary: IsiXhosa and English / Oxford isiXhosa-isiNgesi English-isiXhosa Isichazi-magama Sesikolo*. Oxford University Press Southern Africa.

de Schryver, G.-M., Rundell, M., Tavast, A., Rychlý, P., Kokol, M., & Krek, S. (2023). *Round table on "Large language models and AI in lexicography"*. Paper presented at the 8th Electronic Lexicography in the 21st Century Conference, Brno, Czech Republic, 27-29 June 2023. https://youtu.be/watch?v=dNkksTDYa_s

Franceschi, D., & Pinnavaia, L. (2023, December 20). New developments in English for Special Purposes (ESP) lexicology and lexicography. *International Journal of English Linguistics*, *13*(7), 1–4.

Fuertes-Olivera, P. A. (2024). Making lexicography sustainable: Using ChatGPT and reusing data for lexicographic purposes. *Lexikos*, *34*, 123–140.

Jakubíček, M., & Rundell, M. (2023). The end of lexicography? Can ChatGPT outperform current tools for post-editing lexicography? In M. Medveď, M. Měchura, C. Tiberius, I. Kosem, J. Kallas, & M. Jakubíček (Eds.), *Proceedings of the eLex 2023 conference: Electronic lexicography in the 21st century* (pp. 508–523). Lexical Computing. https://youtu.be/watch?v=8e52vvDpdfQ

Lew, R. (2023). ChatGPT as a COBUILD lexicographer. *Humanities and Social Sciences Communications*, *10*(704).

Lew, R. (2024). Dictionaries and lexicography in the AI era. *Humanities and Social Sciences Communications*, *11*(426).

McKean, E., & Fitzgerald, W. (2023). The ROI of AI in Lexicography [Keynote Lecture]. In *Proceedings of the 16th International Conference of the Asian Association for Lexicography: "Lexicography, Artificial Intelligence, and Dictionary Users"* (pp. 10–20). Yonsei University.

Nabirye, M. (2012). *e-Eiwanika ly'Olusoga [Digital Lusoga Dictionary]*. Menha Publishers / TshwaneDJe HLT. http://www.menhapublishers.com/dictionary/

Nichols, W. (2023). *Invisible lexicographers, AI, and the future of the dictionary [Keynote Lecture]*. Paper presented at the eLex 2023 conference: Electronic lexicography in the 21st century, Brno, Czech Republic, 27-29 June 2023. https://youtu.be/watch?v=xYpwftj_QQI

Ptasznik, B., & Lew, R. (2024, March 25). A learners' dictionary versus ChatGPT in receptive and productive lexical tasks. https://doi.org/10.31219/osf.io/9n685

Rees, G. P., & Lew, R. (2023, December 13). The effectiveness of OpenAI GPT-generated definitions versus definitions from an English learners' dictionary in a lexically orientated reading task. *International Journal of Lexicography*, *37*(1), 50–74.

Rundell, M. (2023). Automating the creation of dictionaries: Are we nearly there? [Keynote Lecture]. In *Proceedings of the 16th International Conference of the Asian Association for Lexicography: "Lexicography, Artificial Intelligence, and Dictionary Users"* (pp. 1–9). Yonsei University.

San Martín, A. (2024, February 25). *What Generative Artificial Intelligence means for terminological definitions*. Paper presented at the 3rd International Conference on Multilingual Digital Terminology Today. Design, representation formats and management systems, Granada, Spain, 27-28 June 2024. https://doi.org/10.48550/arXiv.2402.16139

Tarp, S., & Nomdedeu-Rull, A. (2024). Who has the last word? Lessons from using ChatGPT to develop an AI-based Spanish writing assistant. *Círculo de Lingüística Aplicada a la Comunicación*, *97*, 309–321.

Tran, H. T. H., Podpečan, V., Jemec Tomazin, M., & Pollak, S. (2023). Definition extraction for Slovene: Patterns, transformer classifiers and ChatGPT. In M. Medveď, M. Měchura, C. Tiberius, I. Kosem, J. Kallas, & M. Jakubíček (Eds.), *Proceedings of the eLex 2023 conference: Electronic lexicography in the 21st century* (pp. 19–38). Lexical Computing. https://youtu.be/watch?v=rQC3Rz04b20

Contact information

Gilles-Maurice de Schryver

BantUGent, UGent Centre for Bantu Studies

Language and Translation Technology Team, Ghent University

GillesMaurice.DeSchryver@UGent.be

Ivana Filipović Petrović, Slobodan Beliga

Lexicographic Treatment of Idioms and Large Language Models: What Will Rise to the Surface?

Keywords large language models; artificial intelligence; idioms; lexicographical treatment of idioms; post-editing lexicography; idiomatic meaning

Just as lexicographers had become accustomed to tools and technologies that allowed them to automate many steps in the creation of dictionaries (e.g. corpora and concordances in conjunction with dictionary writing systems) and focus on post-editing lexicographic work, large language models (LLMs) and AI-based tools such as chatbots emerged for general use. As other professions tried them out for various purposes, lexicographers had no choice but to explore their possibilities (for an overview of the initial contributions in this regard, see de Schryver, 2023).

Lexicographers and computational linguists have tested the performance of some lexicographic tasks by AI, such as the formulation of definitions and usage examples (e.g. Rundell, 2023; Tran et al., 2023; Lew, 2023; Gantar, 2024). They mostly concluded that AI performs very well in formulating definitions, while its performance in generating usage examples is somewhat weaker.

With 15 years of experience in compiling both a historical dictionary of literary quotations and a contemporary specialized electronic dictionary, the co-author of this abstract has gone through various technologies as a lexicographer: from manual transcription from dusty card indexes and printed books to the use of automatic methods for searching and identification in corpora and semi-automatic methods in the creation of entries. In a constant race against time, out of all proportion to expectations and purpose, this lexicographer is more than interested in a full-fledged lexicographer's assistant (as de Schryver called ChatGPT in 2023) available at the click of a button.

In this research, we focus on the metaphorical aspect of language and its description for dictionary creation and explore the possibilities for AI tools to take over some lexicographic tasks. Specifically, we have experience creating an online phraseological dictionary that still needs some improvements. Our main focus is on linking the meanings of idioms, which is important for language learning, translation, and communication. Both printed dictionaries and digital resources like machine translation struggle to effectively link similar meanings within the same language and across different languages. The traditional organization of entries, limited search options, and literal translation all contribute to this issue. This difficulty is not surprising, as dealing with multi-word expressions with figurative meanings has always been a challenge in natural language processing.

Significant progress has been made in linking phraseological equivalents across different languages through the LIdioms dataset (Moussallem, 2018), which encompasses idioms in English, German, Italian, Portuguese, and Russian. Recently, this dataset has been expanded to include a Croatian dataset (Filipović Petrović, López Otal, and Beliga, 2024). This is an area that requires further development, particularly taking into account the use of extensive language models. Concerning linking phraseological synonyms in Croatian, the Online Dictionary of Croatian Idioms (Filipović Petrović and Parizoska v.2, 2023) is currently working on integrating a thematic index containing concepts or semantic fields, allowing users to search for idioms based on their meanings. To develop this resource, an experiment was conducted to test the potential of large language models in automatically categorizing idioms into semantic fields, based on a sample created by human lexicographers (Beliga and Filipović Petrović, forthcoming). Further compilation of phraseological entries and expanding the list of idioms in the dictionary is necessary for this purpose, and corpus research of idioms is currently underway. Additionally, during the exploration of automatic identification of verbal idioms in the Croatian corpus (Filipović Petrović and Kocijan, forthcoming), a task was identified that could be entrusted to artificial intelligence. Namely, despite automating the process of finding sentences in the corpus that contain idiom components, the manual evaluation involving linguistic and lexicographic analysis was time-consuming as it required separating the results of literal and idiomatic usage.

Therefore, for this research, we selected a construction that was automatically identified 1073 times in the latest Croatian web corpus, the CLASSLA corpus (Ljubešić and Kuzman, 2023): *isplivati na površinu* ('to come to light, to surface'). Out of 1073 instances, in 703 instances, it appeared as an idiom, meaning 'to appear, suddenly become visible, noticeable,' as in example (1), while in 370 instances, it had the literal meaning 'to emerge from a liquid' (as in example 2).

(1) *Nije trebalo dugo da na površinu isplivaju dobre i loše strane takvog načina rada.* 'It didn't take long for the good and bad sides of such a working method to come to light.'

(2) *Njoke pažljivo stavite u posudu i kuhajte par minuta dok ne isplivaju na površinu.* 'Carefully place the gnocchi in a pot and cook for a few minutes until they float to the surface.'

We aimed to automate the process using LLMs that can distinguish literal and idiomatic meanings of multiword expressions. Unfortunately, we haven't found any models specifically designed for the Croatian language for this task, or any that are available for free use. Potential models that could be adapted for this task include current models such as Llama2 (Touvron et al., 2023), Mistral-7b-Instruct (Jiang et al., 2024), or XLM-RoBERTa (Conneau et al., 2020). However, adapting these models would necessitate a significant experimental effort. This process would encompass the preparation of extensive data sets, annotation, fine-tuning of the model, evaluation, and potentially the application of

Retrieval-Augmented Generation (RAG) techniques, among other demanding tasks. In a recent study, De Luca Fornaciari (2024) presented a similar idiom detection initiative.

Taking into account the limitations of our small research team and the constraints mentioned above, we aimed to examine how the model performs when dealing with idiomatic structures that are unique to Croatian. For this reason, we decided to use the widely recognized GPT-4 model for our preliminary experiment, as it would allow us to gain an initial understanding of its potential and identify any potential challenges.

We employed the GPT-4 model (OpenAI, 2023) in conjunction with prompt engineering to devise an automated procedure for distinguishing idioms based on the presence or absence of idiomatic meaning. We used the prompt with the following settings. The system has been assigned the role of a language expert (for the Croatian language). It has been tasked with identifying whether a specific expression is an idiom when it appears in a defined context. The system must classify the expression as either having an idiomatic or literal meaning, as well as indicate the reliability of the classification: completely safe, partially safe, or uncertain.

Although we have 1073 instances of the idiom *isplivati na površinu* (‘to come to light, to surface’), due to the commercial nature of the GPT-4 model and its availability, we examined 380 samples in the experiment. Out of 380 examples, the system correctly identified 248 (refer to Figure 1) and gave incorrect answers for 132 examples. When asked about its confidence in its answers, it stated that it was completely confident for 362 examples and unsure for 18. Among the 362 examples where it claimed to be confident, it was correct for 240 examples and incorrect for 122. Out of the 122 incorrect examples where it expressed confidence, it marked 2 literal uses as idiomatic, and in 120 instances, it marked idioms as literal meanings.

To be truly helpful for lexicographers, complete accuracy is crucial for this task. The sample size that was tested is small and only relates to a single idiom, while there are much larger lists of idioms and examples available. Therefore, for the research to continue, it will be necessary to refine the prompt further and possibly provide a broader context for each usage to improve the results potentially.

This leads to the question of whether we should first focus on improving the performance of the original tool that identifies constructions in the corpus so that it only detects idiomatic examples. Even if large language models could accomplish this, from a lexicographic standpoint, it's still important to prioritize the method of identifying constructions in the corpus for gathering dependable linguistic data on which dictionaries rely.

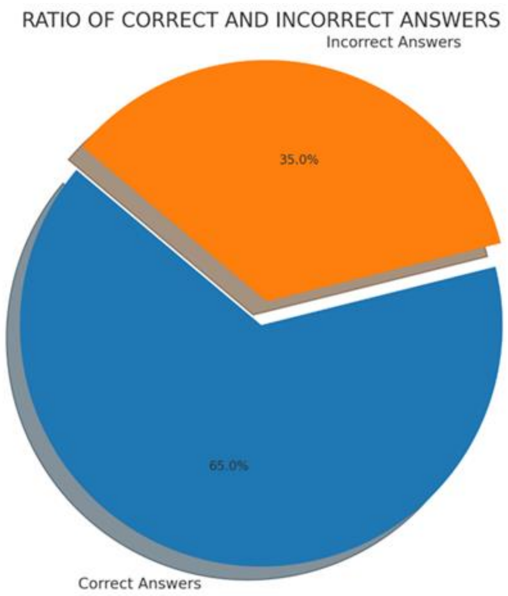

Fig. 1: The ratio of correctly and incorrectly identified instances of literal and idiomatic usage of the construction *isplivati na površinu* 'to surface'.

References

Beliga, Slobodan & Filipović Petrović, Ivana. *Large language models supporting lexicography: conceptual organization of Croatian idioms.* Accepted as a full paper for the conference Language Technologies and Digital Humanities 2024.

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 D. Jurafsky, J. Chai, N. Schluter & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 8440–8451). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.747.

De Luca Fornaciari, F., Altuna, B., Gonzalez-Dios, I. & Melero, M. (2024). A Hard Nut to Crack: Idiom Detection with Conversational Large Language Models. In D. Ghosh, S. Muresan, A. Feldman, T. Chakrabarty & E. Liu (Eds.), *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)* (pp. 35–44). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.figlang-1.5.

Filipović Petrović, Ivana & Kocijan, Kristina. *Creating the dataset of Croatian verbal idioms: automatic identification in a corpus and lexicographic implementation*. Accepted as a full paper for Euralex conference 2024.

Filipović Petrović, Ivana & Parizoska, Jelena. (2022). *Frazeološki rječnik hrvatskoga jezika.* Hrvatska akademija znanosti i umjetnosti. https://lexonomy.elex.is/#/frazeoloskirjecnikhr.

Filipović Petrović, Ivana; López Otal, Miguel & Beliga, Slobodan. (2024). Croatian idioms integration: Enhancing the LIdioms multilingual linked idioms dataset. In: N. Calzolari, M.–Y. Kan, V. Hoste, A. Lenci, S. Sakti & N. Xue (eds.), *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (lrec-coling 2024)* (pp. 4106–4112). Torino, Italia: ELRA and ICCL. https://aclanthology.org/2024.lrec-main.366.

Gantar, Polona. *Izdvajanje semantičkih informacija korišćenjem veštačke inteligencije na primeru slovenačkih frazeoloških jedinica.* Presented at the conference Leksikografski susreti in Belgrade 2024.

Hanh Thi Hong, Tran; Podpečan, Vid; Tomazin; Mateja Jemec & Pollak, Senja. (2023). Definition Extraction for Slovene: Patterns, Transformer Classifiers and ChatGPT'. In Medveď, Marek, Michal Měchura, Carole Tiberius, Iztok Kosem, Jelena Kallas & Miloš Jakubíček (Eds.), *Proceedings of the eLex 2023 Conference: Electronic Lexicography in the 21st Century* (pp. 19–38). Brno: Lexical Computing.

Jiang, Albert Q.; Sablayrolles, Alexandre; Roux, Antoine; Mensch, Arthur; Savary, Blanche; Bamford, Chris; Singh Chaplot, Devendra; De Las Casas, Diego et al. (2024). *Mixtral of Experts.* ArXiv, abs/2401.04088.

Lew, R. (2023). ChatGPT as a COBUILD lexicographer. *Humanities and Social Sciences Communications,* 10(704). doi: 10.1057/s41599-023-02119-6.

Ljubešić, Nikola; Rupnik, Peter & Kuzman, Taja. (2024). *Croatian web corpus CLASSLA-web.hr 1.0.* Slovenian language resource repository CLARIN.SI.

Moussallem, Diego; Sherif, Mohamed Ahmed; Esteves, Diego; Zampieri, Marcos & Ngonga Ngomo, Axel-Cyrille. (2018). LIdioms: A multilingual linked idioms data set. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (pp. 2468–2474). Miyazaki, Japan. European Language Resources Association.

OpenAI; Achiam, Josh; Adler, Steven; Agarwal, Sandhini; Ahmad, Lama; Akkaya, Ilge; Ale man, Florencia Leoni; Almeida, Diogo; Altenschmidt, Janko et al. (2023). *Gpt-4 technical report.* Computer Science 2303.08774, arXiv.

Rundell, Michael. (2023). Automating the Creation of Dictionaries: Are We Nearly There? In *Proceedings of the 16th International Conference of the Asian Association for Lexicography: Lexicography (Asialex 2023 Proceedings),* 22–24 June 2023, Seoul, Korea: Artificial Intelligence, and Dictionary Users: 1–9. Seoul: Yonsei University.

Schryver, Gilles-Maurice de. (2023). Generative AI and Lexicography: The Current State of the Art Using ChatGPT. *International Journal of Lexicography*, 36. 10.1093/ijl/ecad021.

Touvron, Hugo; Martin, Louis; Stone, Kevin R.; Albert, Peter; Almahairi, Amjad; Babaei, Yasmine; Bashlykov, Nikolay; Batra, Soumya; Bhargava, Prajjwal et al. (2023). *Llama 2: Open Foundation and Fine-Tuned Chat Models*. ArXiv, abs/2307.09288.

Contact information

Ivana Filipović Petrović
Croatian Academy of Sciences and Arts
Linguistic Research Institute
ifilipovic@hazu.hr

Slobodan Beliga
University of Rijeka
Faculty of Informatics and Digital Technologies
Center for Artificial Intelligence and Cybersecurity
sbeliga@uniri.hr

Radovan Garabík, Vladimír Benko

An Experiment with LLM for Lexicography

Keywords Slovak; GPT; lexicography; dictionary; LLM

Since the availability of generative large language models (LLMs) approximately two years ago (Wei at al., 2022), we can observe their rapid improvement with the current versions exhibiting reasonably good language knowledge. Their use in lexicography can be therefore expected to be beneficial (de Schryver, 2023), though currently it cannot replace professional lexicographic work and inherits the usual objections – closed, commercially available LLMs lack training data transparency and hinder the ability to reproduce the results in the future, if the model becomes unavailable. Since the dominant language of many multilingual LLMs is English, the results in a less widespread language (in the sense of being part of the training data) are often influenced by English or other close languages.

In this article, we describe using LLMs as a helper tool within our Project of creating a multi-volume monolingual Slovak dictionary (SSSJ I–IV, 2006–2021).

The Slovak language is notably absent from almost all "Open Source" LLMs, with only Llama-3[1] and Gemma[2] (at the time of writing) offering acceptable output, with the only available instruction model being Llama3-Instruct. We also performed some experiments using this model, as well as with gemini-pro[3] and Claude Opus[4].

We are using a simple web interface utilising the API access to the latest OpenAI GPT (OpenAI, 2024) models, which generates lexicographic entries (definition and entry structure) in zero-shot or few-shot scenarios. Currently, we are using gpt-4o, gpt-4o-mini, gpt-4-turbo-preview, or gpt-3.5-turbo. The interface is named "Lexicographic Hallucinations" to remind the user of the nature of the generated text.

[1] https://llama.meta.com/

[2] https://ai.google.dev/gemma

[3] https://deepmind.google/technologies/gemini/

[4] https://www.anthropic.com/claude

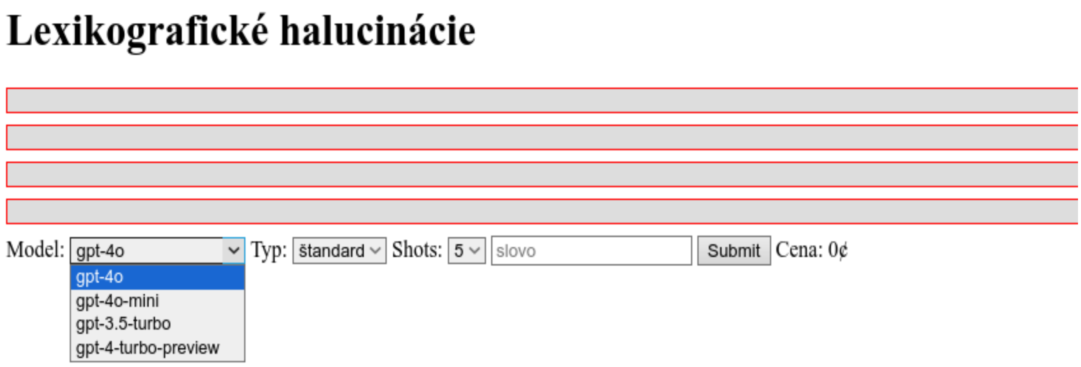

Fig. 1: Screenshot of the interface

The LLM-generated definitions are reasonably usable for casual lexicographic referencing; nevertheless, they sometimes exhibit surprising subtle interference from other languages. However, our evaluation consisted only of manual (and subjective) assessment of the quality of the output definitions of several (up to ten) random nouns.

We generate four chat completions and display them simultaneously, to provide the users with a direct way to assess the variability and quality of the output, without the need to play with the temperature or other model settings.

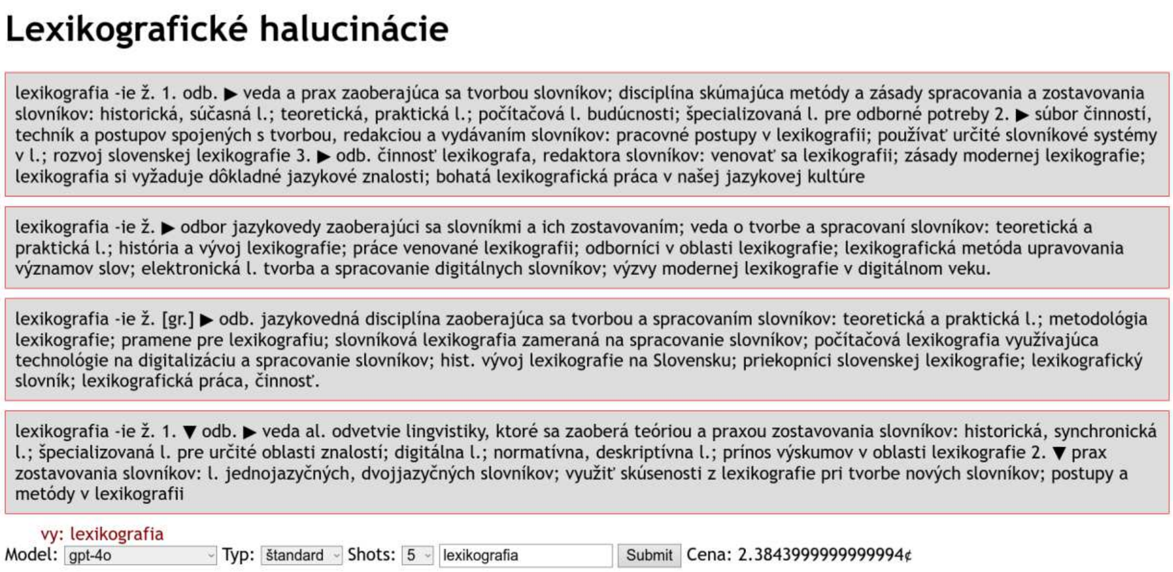

Fig. 2: Example of the output (headword *lexikografia*)

The parameters tuneable in the interface are deliberately kept to the minimum. The list of available parameters and inputs:

- The language model name. The “best”, currently the most recent model is the default one.
- The type of the output.
- Whether to use 5-shot (the default) or 0-shot strategy.
- The input word.
- The *Submit* button.

There are three different types of the output. They differ by the system prompt and definitions used in a 5-shot scenario. The default one (*štandard*) uses standard, unmodified definitions from the SSSJ as the 5-shot examples. The *stručný* uses abridged definitions for the examples, shortened to approximately half. The *vtipný* type uses definitions modified to be witty, sarcastic and funny (as perceived by us; your mileage may vary) and is meant for demonstrations and PR purposes. The interface also includes a cumulative price to remind the users of the paid character of the OpenAI API. The lexicographers are instructed to use predominantly the default settings, i. e. the most recent model, 5-shot strategy and the *štandard* type.

The prompt we use is in English, with an instruction to keep the conversation in Slovak. We observed that the quality of the output does not change compared to the prompt being in Slovak, keeping the communication in English is more versatile (e.g. when testing output in other languages), and the use of a 5-shot examples influences the output much more than the exact wording of the prompt.

Earlier, we noticed an interesting behaviour: if the model started to hallucinate, i.e. to generate incorrect or fictitious definitions, the language of the definition deteriorated from correct Slovak, containing an admixture of hallucinated words, sometimes with a faux Slovenian/Croatian orthography. This way, it was possible to immediately see if the definition is hallucinated. Unfortunately, this changed from the version gpt-4-0125-preview onwards, and the language is more or less in correct Slovak all the time.

Slovník súčasného slovenského jazyka A – G, H – L, M – N, O – Pn z r. 2006, 2011, 2015, 2021.

lexikografia *-ie* *ž.* ⟨gr.⟩ ▶ vedecká disciplína jazykovedy zaoberajúca sa teóriou a tvorbou slovníkov, spracovaním slovnej zásoby v podobe slovníka; tvorba slovníkov: *dvojjazyčná, viacjazyčná l.; terminologická l.* terminografia; *súčasná lexikológia a l.; dejiny slovenskej lexikografie; počítačová l.* spracovanie slovnej zásoby jazyka pomocou počítačových nástrojov

Fig. 3: The actual dictionary entry (headword *lexikografia*) as is displayed at the Institute's Dictionary portal (SP)

While the entry compiled by a human lexicographer contains just one sense – the scientific field and a specific sub-sense of the process of dictionary creation (following the semicolon), the AI-generated ones contain two or three senses, and this splitting is, strictly speaking, not incorrect. All entry variants contain the correct PoS label, and most of the definitions are acceptable. An amusing example appeared in earlier models: "work at the lexicographic department of the Institute of Czech language".

References

SSSJ I: Buzássyová, K. & Jarošová, A. (Eds.) (2006). *Slovník súčasného slovenského jazyka. A – G.* . Veda, vydavateľstvo SAV.

SSSJ II: Buzássyová, K. & Jarošová, A. (Eds.) (2011). *Slovník súčasného slovenského jazyka. H – L*. Veda, vydavateľstvo SAV.

SSSJ III: Jarošová, A. (Ed.) (2016). *Slovník súčasného slovenského jazyka. M – N*. Veda, vydavateľstvo SAV.

SSSJ IV: Jarošová, A. (Ed.) (2021). *Slovník súčasného slovenského jazyka. O – Pn*. Veda, vydavateľstvo SAV.

OpenAI (2024). GPT-4 Technical Report, arXiv:2303.08774v6 [cs.CL] 4 Mar 2024. Available at: https://arxiv.org/pdf/2303.08774. Retrieved June 3, 2024.

de Schryver, G.-M. (2023). Generative AI and lexicography : the current state of the art using ChatGPT. *International Journal Of Lexicography*, 36(4), 355–387.

Wei et al. (2022). Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research,* (08). Available at: https://openreview.net/pdf?id=yzkSU5zdwD. Retrieved June 3 2024.

SP: *Slovníkový portál*. Retrieved September 13, 2024, from https://slovnik.juls.savba.sk/?w=lexikografia&d=sssj

Contact information

Radovan Garabík

Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences

radovan.garabik@kassiopeia.juls.savba.sk

Vladimír Benko

Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences

vladimir.benko@juls.savba.sk

Timotej Knez, Tim Prezelj, Slavko Žitnik

SemSex: Automated Assessment of Sex Education Representation in Slovene Curricula

Keywords Knowledge graph; Natural language processing; Ontology development

Introduction

The automatic assessment of topic representation within documents is crucial for accurately depicting subject matter across a wide array of fields. Automating this process is essential to manage and analyze large volumes of documents, ensuring a comprehensive and nuanced understanding of the covered topics. Our project specifically addresses the coverage of sex education topics within Slovene curriculum documents by designing an ontology covering all topics of interest and then designing a model for recognizing the coverage of identified topics in curriculum documents. By automating the recognition and analysis of these topics, we aim to ensure balanced and comprehensive representation across all educational subjects. Such a tool is invaluable for curriculum developers, allowing them to identify gaps, ensure thorough coverage, and align educational content with modern standards and requirements.

Related work

Using machine learning to recognize the topic of documents has been an active area of research in recent years. The recent approaches for recognizing topics tend to rely on transformer models to recognize the topics (Terragni et al., 2021; Zhao et al., 2021; Grootendorst, 2022). For example, Grootendorst (Grootendorst, 2022) proposes to solve this issue by designinga model called BERTopic, which uses Sentence-BERT (Reimers and Gurevych, 2019) to generate document embeddings and uses them to cluster documents into several topic groups. In our research, we use a similar approach for recognizing sexual education concepts in curriculum documents. We expand the approach further by connecting the recognized concepts to an ontology and presenting the documents as knowledge graphs. This allows for more accurate analysis of extracted information. A similar idea was presented by Liu and El-Gohary (Liu and El-Gohary, 2017) as they extracted information from building reports using an ontology.

Methodology

Our goal when designing the methodology for the project was to make it as general as possible. This way we could use a similar approach for other domains and applications. We outlined the methodology in Figure 1. Our methodology begins with the creation of a domain-specific ontology for sex education. The initial step involves manually defining a hierarchical structure of concepts, creating a foundational framework. This structure is enriched by adding relationships between concepts, thereby developing a basic knowledge graph. To enhance the ontology's utility and interconnectedness, we link the concepts to their corresponding entries in Wikidata. This connection allows us to leverage the extensive relational data within the Wikidata knowledge graph, providing a robust and enriched dataset. The ontology is presented in RDF format and is publicly accessible through the clarin.si repository (http://hdl.handle.net/11356/1895) and a GitHub repository (https://github.com/clarinsi/SemSex).

To ensure clarity and usability, we documented the ontology using the Widoco tool, generating comprehensive HTML documentation for each concept. This documentation includes detailed descriptions and attributes of each concept. Additionally, we created a network diagram illustrating the connections between concepts, providing a clear and intuitive visual representation of the ontology's structure. This documentation facilitates easy navigation and understanding of the ontology for users and developers.

The ontology was further enriched automatically through its connection to the Wikidata knowledge graph. This automatic enrichment process involved gathering concept descriptions and pre-existing relationships from Wikidata, significantly enhancing the depth and detail of the ontology. By incorporating data from Wikidata, we ensured that our ontology was not only comprehensive but also aligned with a widely recognized and used knowledge base.

An important objective of our research was to enable the automatic recognition of sex education concepts within curriculum documents. To achieve this, we developed two sentence-level classifiers. The first classifier detects whether a sentence pertains to any sex education concept, while the second identifies the specific concept being discussed. The training dataset was constructed by manually annotating Slovenian curriculum plans, resulting in 814 sentences, 166 of which contained mentions of sex education concepts. Despite our ontology encompassing 196 unique concepts, only 20 were present in the curriculum plans. In addition to the manually annotated dataset, we also automatically constructed a second dataset, by linking all of the concepts in the knowledge graph to their respective Wikipedia sites. As the project was meant for the Slovene language, we used Slovene Wikipedia pages where possible, while automatically translating the English pages when Slovene was not available. Using solely the dataset constructed from Wikipedia pages proved ineffective due to structural differences between curriculum documents and Wikipedia articles. On the other hand, we have found that augmenting the manually labeled data with automatically generated examples improved the results of the models. This suggests, that when using a similar approach on a new domain, we only need to manually annotate a relatively small set of data if we combine it with an autaomatically gathered dataset.

For the effective analysis of documents, it is crucial to represent the extracted information in a structured format. We designed a pipeline that processes PDF curriculum documents, identifies sentences related to sex education concepts, determines the specific concepts discussed, and constructs a knowledge graph aligned with the SemSex ontology. This structured representation allows for systematic analysis and visualization of the curriculum content, facilitating better insights and decision-making.

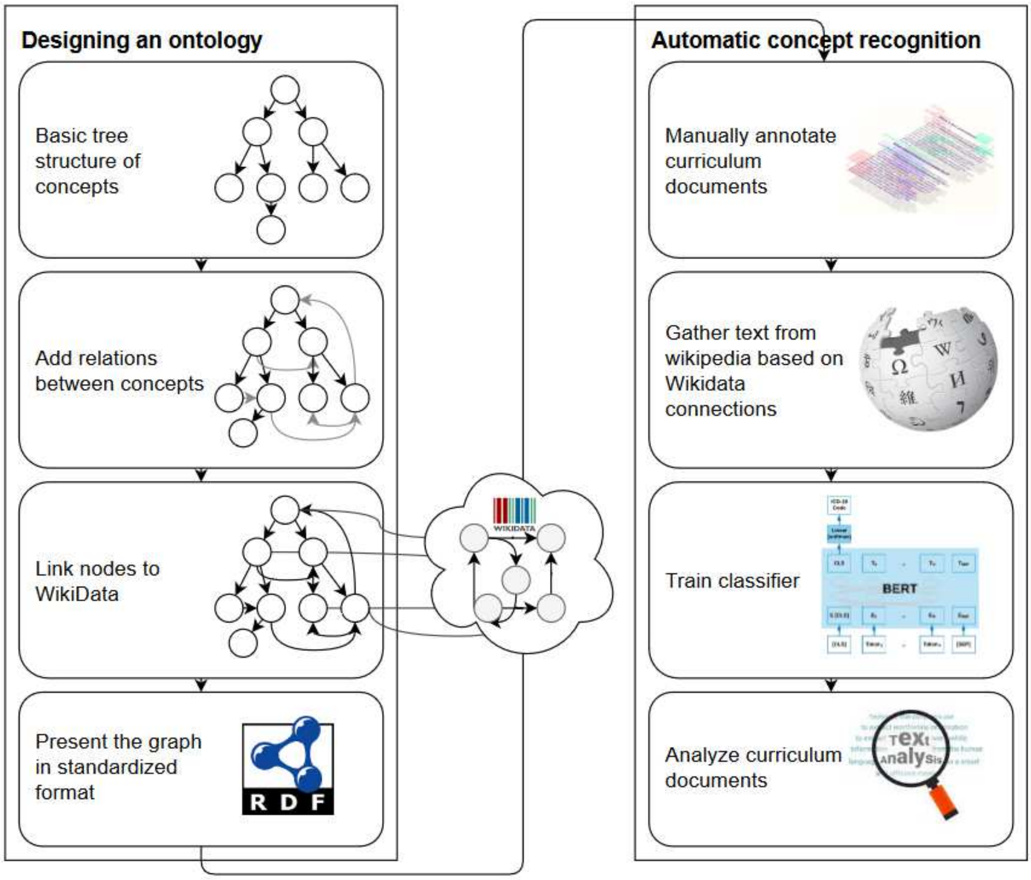

Fig. 1: An illustration of the proposed methodology

Results

We evaluated two parts of the project. The first is the constructed ontology, and the second is the two models for recognizing the presence of sexual education concepts in documents.

We successfully constructed a basic tree structure of sex education concepts, resulting in an ontology that includes 196 concepts and enriched it with semantic links, adding nine object property types and seven data property types. Each concept is linked to the Wikidata knowledge graph, ensuring comprehensive and interconnected data. This enriched ontology serves as a robust foundation for analyzing sex education coverage in curricula. The structure of the ontology is visualized in Figure 2.

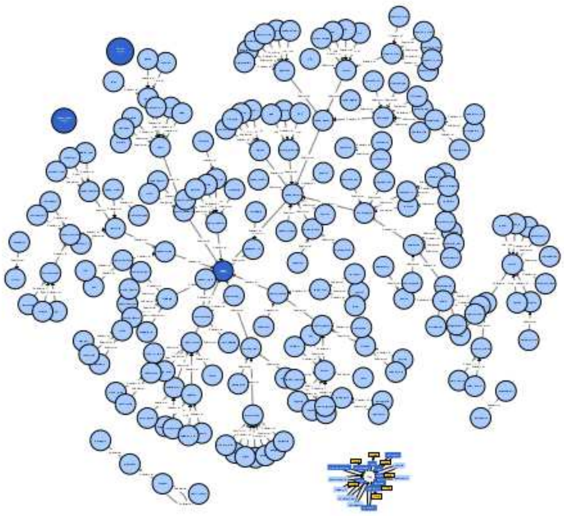

Fig. 2: An illustration of the SemSex ontology structure

We evaluated two transformer models for concept recognition on the manually annotated dataset: crosloengual-bert, a multilingual model, and Sloberta, a RoBERTa model trained on Slovene text. In identifying sentences describing a concept, crosloengual-bert achieved 91% accuracy, significantly outperforming Sloberta, which achieved 69% accuracy. Both models were trained and tested on a balanced dataset, ensuring robust evaluation. For recognizing specific concepts, Sloberta achieved 52% accuracy, while crosloengual-bert achieved 46% accuracy, with the most common class occurring in 18% of examples. These results indicate that while both models are effective, crosloengual-bert performs better in identifying relevant sentences, whereas Sloberta shows slightly better accuracy in pinpointing specific concepts. All of the results are presented in Table 1.

Table 1: Classification results on our dataset. The models were trained and tested using text from curriculum documents

Model	Test Accuracy	Majority classifier
Concept Detection		
crosloengual-bert	91.2%	50.0%
sloberta	69.3%	50.0%
Concept Classification		
crosloengual-bert	46.3%	18.1%
sloberta	52.9%	18.1%

In addition to using only the manually annotated training examples for recognising which concepts are described in a sentence, we also experimented on using the automatically constructed examples, we gathered from Wikipedia pages about the concepts. When constructing these examples, we assume that each sentence on a Wikipedia page about a concept talks about that concept. While this assumption might not always hold, we empirically found that the data quality was sufficient for training the models.

When training and testing the models exclusively on the examples from Wikipedia, the models achieved results presented in Table 2. In this test, the Sloberta model performed the best in both tasks. When training the models on Wikipedia data and testing them on recognizing concepts in curriculum documents, we found that the performance was not significantly better than that of a majority classifier. However; by training the models on the examples from both the manually annotated documents and the Wikipedia pages, we managed to improve the performance of the models in recognizing concepts in curriculum documents. The results in this scenario are also presented in Table 2.

Table 2: Results of the models when using the data gathered from Wikipedia documents. In the first test, we used Wikipedia data to train and test the models. In the second test, we combined the manually annotated curriculum text with automatically generated Wikipedia examples for training. For testing, we used only the curriculum documents

Model	**Accuracy**	**Majority classifier**
Training and testing on Wikipedia data		
CroSloEngual-BERT	66.0%	14.2%
Sloberta	**74.0%**	14.2%
Training on curriculum and Wikipedia, testing on curriculum		
CroSloEngual-BERT	56.1%	18.1%
Sloberta	**58.5%**	18.1%

Discussion

A broader goal of our project was to design a generalizable process for constructing similar systems across various domains. While the ontology and models we developed are specific to sex education in school curricula, the underlying methodology can be applied to other fields. The critical aspect of designing the ontology is its connection to the Wikidata knowledge graph, which facilitates the utilization of existing resources and enables automatic enrichment. This approach ensures that the ontology remains dynamic, up-to-date, and interconnected with a broader knowledge base. We also benefited from connecting each concept to a Wikipedia page describing it as it allowed us to automatically construct a simple dataset for training models for recognizing the sex education concepts described in documents.

By establishing a robust, automated system for topic representation and recognition, our research contributes to more effective curriculum development. This system provides curriculum developers with a powerful tool for analyzing and enhancing educational content, ensuring comprehensive and balanced coverage of essential topics. Furthermore, the methodology outlined in this research offers a template for similar endeavors in other subject areas, demonstrating the potential for wide-ranging applications and impact.

Conclusion

This paper presents an innovative approach for automatically assessing the representation of sex education topics in Slovene curricula. By constructing and enriching a detailed ontology linked to the Wikidata knowledge graph, we created a robust framework for analyzing curriculum content. Our development of sentence-level classifiers allowed for precise identification and categorization of sex education concepts, despite initial challenges with training data.

The designed pipeline effectively transforms curriculum documents into structured knowledge graphs, providing valuable insights into topic coverage. Our methodology, while focused on sex education, is generalizable and can be adapted to other educational domains, offering a scalable solution for automated curriculum analysis.

In conclusion, our work significantly advances the automated evaluation of educational content, ensuring comprehensive coverage of essential topics like sex education. This approach not only improves curriculum development but also provides a versatile framework for similar applications in other fields.

References

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. ArXiv preprint arXiv:2203.05794.

Liu, K. and El-Gohary, N. (2017). Ontology-based semi-supervised conditional random fields for automated information extraction from bridge inspection reports. Automation in construction, 81:313–327.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084.

Terragni, S., Fersini, E., Galuzzi, B. G., Tropeano, P., and Candelieri, A. (2021). Octis: Comparing and optimizing topic models is simple! In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, pages 263–270.

Zhao, H., Phung, D., Huynh, V., Jin, Y., Du, L., and Buntine, W. (2021). Topic modelling meets deep neural networks: A survey. arXiv preprint arXiv:2103.00498.

Contact information

Timotej Knez

University of Ljubljana, Faculty of Computer and Information Science

Timotej.knez@fri.uni-lj.si

Tim Prezelj

University of Ljubljana, Faculty of Education

Tim.prezelj@pef.uni-lj.si

Slavko Žitnik

University of Ljubljana, Faculty of Computer and Information Science

Slavko.zitnik@fri.uni-lj.si

Iztok Kosem, Polona Gantar, Špela Arhar Holdt, Magdalena Gapsa, Karolina Zgaga, Simon Krek

AI in Lexicography at the University of Ljubljana

Case Studies

Keywords lexicography; ChatGPT; synonyms; sense division; phraseological units; Slovenian

The Centre for Language Resources and Technologies at the University of Ljubljana (CJVT UL) has started developing a Digital Dictionary Database (DDD) with the aim to create an open-access comprehensive repository of information on modern Slovenian language. This database is intended for use in both the compilation of language resources and natural language processing tasks. Detailed plans for the database were outlined by Klemenc et al. (2017). One of the key decisions in the development of DDD is that the same concepts, which translate into dictionary senses, are used for all the resources coming out of the database. This has already been implemented for the Collocations Dictionary of Modern Slovene (Kosem et al., 2018), the Thesaurus of Modern Slovene (Arhar Holdt et al., 2018), and the Comprehensive Slovenian-Hungarian Dictionary (Kosem et al., 2024), with additional resources, particularly bilingual ones, in development.

With most information in DDD compiled from scratch, the lexicographers at CJVT UL have consistently utilised and tested the most advanced tools and methods available, such as GDEX, Tickbox lexicography, post-editing approaches, and sense induction. The arrival of Large Language Models, particularly ChatGPT, has provided another powerful method that can potentially facilitate dictionary compilation. We have conducted, or are in the process of conducting, several studies to determine ChatGPT's usefulness in various parts of the lexicographic workflow.

In the first study, we tested how well ChatGPT-4 cleans the list of automatically retrieved synonym candidates and distributes the synonyms under appropriate lexical senses. As a gold standard, we considered the lexicographic decisions made when updating the Thesaurus of Modern Slovene to version 2.0. We compared the results for 246 dictionary entries. For 41.9% of entries, ChatGPT processed the data in the same way as lexicographers, while for 58.1%, it made a different decision: 43.5% of entries contained differences in the removal of noisy data, and 28.9% in the mapping of synonyms to lexical senses. When assessing the relevance of synonym candidates, ChatGPT was more permissive than the gold standard (recall 0.33), while precision was higher (0.75), but the errors were more difficult to explain. Differences in synonym placement (incorrect placement in 14.6% of entries, missing placement in 19.9%) can partly be attributed to features of the input data, such as task complexity and brevity of semantic indicators. Future work will focus on further improvements and validation of the method for speeding up lexicographic work.

Our second study focuses on devising senses (or concepts) for headwords. Many headwords in the DDD contain automatically obtained corpus information, such as collocations and their examples, but lack sense division. Our initial attempts with ChatGPT (version 3.5 at the time), where the prompt provided the headword and asked ChatGPT to devise dictionary entries, returned mixed results, with problems ranging from far too fine-grained sense division and hallucinations (non-existent senses) to unnatural examples and non-Slovenian grammar and syntax in both definitions and examples. Subsequent testing and evaluation showed that ChatGPT performs much better if context is provided. Consequently, we decided to include more information in both the prompt and the system instructions, providing a selection of automatically extracted collocations and their examples. When selecting collocations, we pick more collocations per syntactic structure from those syntactic structures that are more semantically relevant, e.g. adjective + noun, verb + noun in accusative, noun + noun in genitive.

In addition, we are including a combination of definitions obtained from the Open Slovene WordNet and the English-Slovenian Bridge dictionary. The nature of these two resources, and the way the definitions were obtained, has resulted in the fact that many candidate headwords have several similar definitions or definitions explaining the same sense. Thus, the instruction to take definitions only as a point of departure, and to consider merging them or ignoring the near-duplicates, was included in the prompt.

One other thing that was noticed when testing this new approach was that ChatGPT (in this case, version 4 and later 4o) predominantly produced the pattern "X means …" for the definitions, regardless of the part-of-speech category. The usefulness of this pattern is limited to certain verbs. To address this, we made two adjustments: a) we included sample definition patterns in system instructions, along with an example of expected input and output, and b) given that different part-of-speech categories use different definition patterns, we devised slightly different system instructions for each part-of-speech category.

For the study, we selected 115 headwords with completed entries from the DDD, representing different part-of-speech categories, with most headwords being polysemous and a few monosemous. For most headwords there are definitions available in the Open Slovene WordNet and/or the English-Slovenian Bridge dictionary, for some, there are not; this is intentional as we also want to test ChatGPT performance with and without the definitions provided. At the workshop, we will report on the results of this study.

In the next experiment, we wanted to test the possibilities of using AI in the lexicographic treatment of phraseological units (PU), which in DDD are understood as multi-word lexical units that show a certain degree of structural stability and semantic opacity and have a demonstrated expressive or pragmatic role in the language (Kosem et al., 2020). To (a) obtain the most useful, i.e. reliable, information on PU semantics and (b) determine how LLMs behave in identifying and understanding complex multi-word units such as PUs, we used a list of 35 PUs as a starting point, which we constructed based on freely available dictionary sources. The list includes PUs of different part-of-speech categories and structural and semantic complexity. We also considered PUs with culturally specific lexical constituents and PUs with a distinctive pragmatic role. All PUs had to have some semantic interpretation in existing dictionary sources and appear more than ten times in the Gigafida corpus. For the preliminary study, we created a zero-shot linguistic prompt that required ChatGPT to produce a lexicographic definition for each of the PUs in the list while also considering potential multi-sense. We also needed it to provide two relatively short dictionary examples for each PU's meaning that will explain the typical usage of the PU in context.

In the analysis of the results, we checked through the generated definitions whether ChatGPT had adequately identified at least one phraseological meaning of the PU, identified the multiple meanings of the PUs, and whether the examples offered met the criteria of a good dictionary example. We additionally observed the adequacy of understanding the meaning of the PUs concerning the degree of idiomaticity of the PUs and the presence of language-specific lexical items. Preliminary results showed that in more than half of the cases, the artificially generated definitions are adequate or even better than the lexicographic ones. The degree of semantic (in)transparency and the presence or absence of language-specific elements did not play a significant role here: among the PUs for which ChatGPT did not provide an adequate definition of the meaning, there were both extremely idiomatic PUs, e.g. *iti rakom žvižgat*, and relatively transparent ones, e.g. *na vse ali nič*. In all cases where ChatGPT did not adequately identify the meaning, the proposed definition defined the literal rather than the phraseological meaning. ChatGPT performed much worse in identifying multiple meanings (in fact, in only one case, and even then, it identified one meaning incorrectly) and providing dictionary examples that are virtually useless for lexicographic work.

The pilot study has shown that using AI to generate dictionary definitions is worthwhile even for complex PUs. There is a lot of potential in improving the linguistic prompt and adding a good example of what the definition and examples should look like. On the other hand, the study also reflected on including positive and negative examples in the data sets on which large language models are trained.

References

Arhar Holdt, Š., Čibej, J., Dobrovoljc, K., Gantar, P., Gorjanc, V., Klemenc, B., Kosem, I., Krek, S., Laskowski, C., Robnik Šikonja, M. (2018): Thesaurus of Modern Slovene: By the Community for the Community. In: J. Čibej, V. Gorjanc, I. Kosem, S. Krek (eds.): *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*. Ljubljana: Znanstvena založba Filozofske fakultete, pp. 401-410. https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/118/211/3000-1

Klemenc, B., Robnik-Šikonja, M., Fürst, L., Bohak, C., Krek, S. (2017): Technological Design of a State-of-the-art Digital Dictionary. In: V. Gorjanc, P. Gantar, I. Kosem, S. Krek (eds): *Dictionary of modern Slovene: problems and solutions*. 1st ed. Ljubljana: Ljubljana University Press, Faculty of Arts, pp. 10-22.

Kosem, I., Krek, S., Gantar, P. Defining collocation for Slovenian lexical resources. In: Kosem, I. (ur.), Gantar, P. (ur.). *Kolokacije v leksikografiji : obstoječe rešitve in izzivi za prihodnost = Collocations in lexicography : existing solutions and future challenges*. Ljubljana: Znanstvena založba Filozofske fakultete, 2020. Letn. 8, št. 2, str. 1-27, ilustr. Slovenščina 2.0, 2, 2020. https://revije.ff.uni-lj.si/slovenscina2/article/view/9338/9069

Kosem, I., Krek, S., Gantar, P., Arhar Holdt, Š., Čibej, J., Laskowski, C. (2018): Collocations dictionary of modern Slovene. In J. Čibej, V. Gorjanc, I. Kosem, S. Krek (eds.): *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*. Ljubljana: Znanstvena založba Filozofske fakultete, pp. 989-997. https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/118/211/3000-1

Kosem, I. Bálint Čeh, J., Ponikvar, P., Zaranšek. P., Kamenšek, U., Koša, P., Gróf, A., Gantar, P., Krek, S., Arhar Holdt, Š., Gorjanc, V. (2024*). Veliki slovensko–madžarski slovar: zvezek 2024*. Založba Univerze v Ljubljani. https://ebooks.uni-lj.si/ZalozbaUL/catalog/book/573

Contact information

Iztok Kosem

Centre for Language Resources and Technologies, University of Ljubljana

Iztok.Kosem@fri.uni-lj.si

Polona Gantar

Centre for Language Resources and Technologies, University of Ljubljana

Apolonija.Gantar@ff.uni-lj.si

Špela Arhar Holdt

Centre for Language Resources and Technologies, University of Ljubljana

Spela.ArharHoldt@fri.uni-lj.si

Magdalena Gapsa

Centre for Language Resources and Technologies, University of Ljubljana

magdalena.gapsa@cjvt.si

Karolina Zgaga

Centre for Language Resources and Technologies, University of Ljubljana

karolina.zgaga@cjvt.si

Simon Krek

Centre for Language Resources and Technologies, University of Ljubljana

Simon.Krek@fri.uni-lj.si

Pauline Sander, Simon Hengchen, Wei Zhao, Xiaocheng Ma, Emma Sköldberg, Shafqat Virk, Dominik Schlechtweg

The DURel Annotation Tool

Using Fine-Tuned LLMs to Discover Non-Recorded Senses in Multiple Languages

Keywords DURel tool; annotation; lexicography; semantic proximity; large language model; word-in-context; word sense induction; clustering

The concept of semantic proximity has long been present in Cognitive Semantics (Blank, 1997). It quantifies how much the meanings of two word uses “have in common” (Schlechtweg, 2023, cf. p. 25). Semantic proximity is also recognized in Lexicography (Kilgarriff, 1997), where it has been used as a criterion in the lexicographic clustering process (Kilgarriff, 2007). Semantic proximity is essential for identifying word senses and creating dictionary entries, as well as research building on senses such as lexical semantic change or semantic variation (Schlechtweg, 2023).

After advances in modeling the meaning of word uses with contextualized embeddings from language models trained on large amounts of textual data (Peters et al., 2018; Devlin et al., 2019), it has become possible to estimate the semantic proximity between word uses using so-called Word-in-Context (WiC) models (Pilehvar and Camacho-Collados, 2019; Armendariz et al., 2020), which are specifically optimized on human-annotated semantic proximity training data. These models achieve high performance (He et al., 2020; Raffel et al., 2020) and serve as an excellent starting point for any practical task that relies on semantic proximity, such as finding novel/unrecorded senses or identifying words that change their meaning.

To make these new techniques accessible to researchers outside of Computational Linguistics, we have developed the DURel tool (Schlechtweg et al., 2024). The basic annotation data gathered in the system are judgments of semantic proximity between word uses (Blank, 1997; Erk et al., 2013), created using the DURel relatedness scale (Schlechtweg et al., 2018; Schlechtweg 2023, p. 33).

DURel’s computational annotators enable us to generate word sense clusters for large sets of words and word uses, and to systematically search unlabelled data for new senses. The most important annotator is XL-Lexeme, a bi-encoder that vectorizes the input sequences using a

XLMR-based Siamese Network (Cassotti et al., 2023), which has been trained to minimize the contrastive loss with cosine distance on several WiC datasets and predicts either cosine similarities or relatedness scores derived from these.

The DURel tool then creates Word Usage Graphs (WUGs) out of these proximity judgments, where nodes represent word uses, and weights on edges represent the semantic relatedness of two nodes. Various graph clustering techniques can subsequently be used to identify word senses.

To showcase the potential of our computational methods, we explore how DURel can be used to identify potentially outdated dictionary entries. Our experiments are inspired by the work of Sköldberg et al. (2024), who selected a set of random 281 headwords from a Swedish dictionary with only one sense registered. For each headword, they sampled 25 occurrences from a Swedish corpus, predicted their semantic proximity and clustered them into sense clusters using Correlation Clustering (Bansal et al., 2004; Schlechtweg et al., 2020). It appeared that 66 out of 281 words (23%) were predicted to have more than one cluster (and hence more than one sense). This can be taken as a signal that the dictionary is not up-to-date. We conduct parallel experiments for three different online dictionaries: WordNet[1] (Fellbaum 1998) and the Oxford English Dictionary (both English) as well as DWDS (German). In all examples we use the clustering parameters determined by Sköldberg et al. (2024).

Table 1: Table with the results from all three experiments

	Sampled	>1 cluster		New sense if >1 cluster	
OED	103	44	42,7 %	21	47%
WordNet	100	40	40 %	4	10 %
DWDS	108	46	42,6 %	17	37 %

The first modern English experiment was accomplished with the help of Wordnet (Fellbaum 1998) from which we randomly sample headwords and keep those assigned to only one synset. The process is carried out on a part-of-speech (POS) basis and not at the headword level, i.e. if a headword exists as different POS we discard those POS that are assigned to more than one synset and from the remaining possible POS (if more than one) randomly selected one. In total, 100 of those monosemous headwords are kept. For each monosemous headword selected we sample 25 usages from the 2023 version of the 1M Leipzig corpus (Leipzig Corpora Collection 2023a). As the sampling occurs for usages of lemmas and not

[1] We downloaded the database (v 3.1) and used it offline.
https://wordnet.princeton.edu/download

surface forms, the corpus had to be lemmatised using SpaCy,[2] and the search was carried out on lemmas. Original sentences are then retrieved from the corpus while keeping lemma position information, and are formatted for the DURel tool. Using DURel, usages for any given word are then judged by an automated annotator for semantic similarity, and the resulting annotations are clustered. Ideally, the clustering of all usages of any of the 100 words should result in one cluster, as those words are monosemous. For WordNet, this is not the case for 40 of them. Fifteen words that have more than one cluster are due to named entities ('euphoria' the state of mind vs 'Euphoria' the TV show), with even the special case of 'Lisbon' where the model made the difference between the capital of Portugal and Lisbon, Ohio (USA). Five words show erroneous clusters: one main cluster and a (or several) non-clustered usage(s) – typically, a very short usage of only a few words. Interesting new senses are usually found in literal vs metaphorical usages ('water seeping' vs 'emotion seeping') or jargon use ('forest logging' vs 'computer logging'). Figure 1 shows the striking example of the verb 'to seep': the left orange cluster is the metaphorical sense, the right blue cluster cluster refers to the literal sense, and the middle two-item cluster is a PoS-error ('seep' as a noun ('hydrothermal seep'), not as a verb).

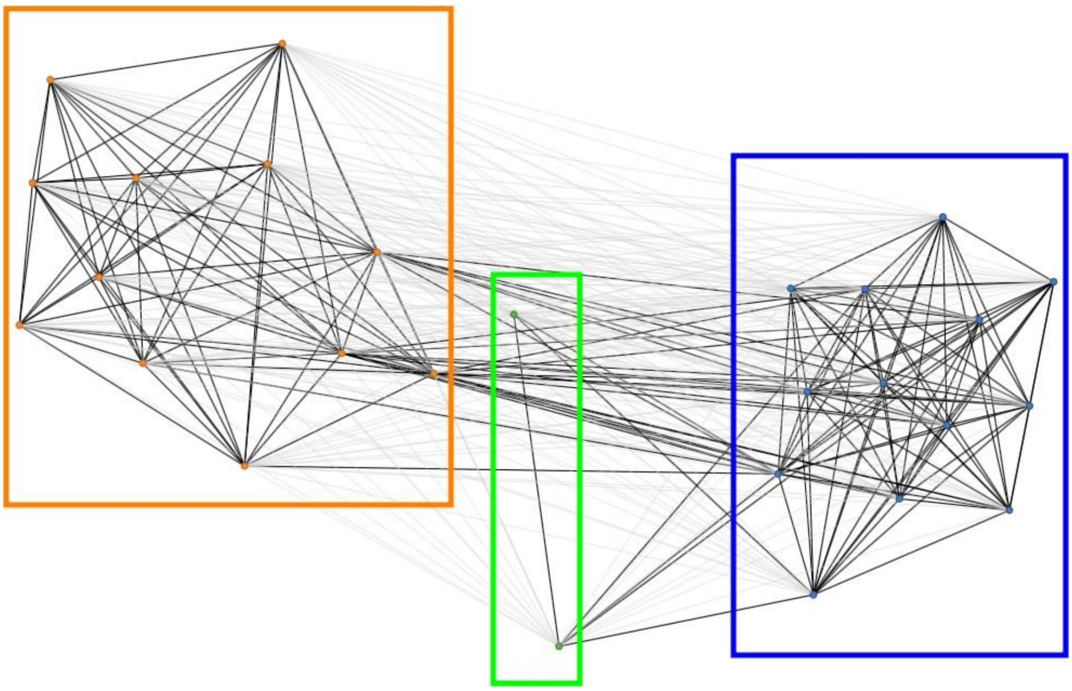

Fig 1: Usage clustering of 'to seep' in three clusters. From left to right: metaphorical use, PoS-errors, literal use

[2] https://github.com/explosion/spaCy, version 3.7.6. The models used were: en_core_web_lg-3.7.1, available at https://github.com/explosion/spacy-models/releases/tag/en_core_web_lg-3.7.1 and de_core_news_sm-3.7.0, available at https://github.com/explosion/spacy-models/releases/tag/de_core_news_sm-3.7.0

For the second English experiment, we randomly sample headwords from the Oxford English Dictionary (OED) updates from 2015 to 2024. Headwords with more than one sense are excluded. This results in 3039 monosemous headwords; however, many of them are neologisms with few usages. To increase data coverage, we resort to a large collection of the 1M Leipzig corpora from 2015 to 2024. In this dataset, only 103 headwords with more than 25 usages are kept, and for each we randomly sample 25 usages. We note that OED headwords are not lemmatized, i.e., a word with varying morphological endings is treated as separate headwords. For this reason, we do not lemmatize the Leipzig corpus. Retrieved sentences from the corpus are formatted for the DURel tool. In our analysis, we find 44 of the 103 headwords selected above with more than one sense cluster, and then we manually compare their sense clusters against OED entries. We find that 17 of 44 headwords (e.g., 'clickbait' and 'JavaScript') are overclustered, i.e., some clusters without meaningful sense distinction could be merged. 2 headwords are underclustered, e.g., 'hass' has a cluster containing uses both as a person name and as the avocado skin. 4 headwords contain faulty clusters due to (a) spelling errors - the corpus use of 'matcha' is a spelling mistake; based on the context, it should be 'march' and (b) part-of-speech tag mismatch - 'cis' as an adjective means 'cissexual', while as a noun it means 'Commonwealth of Independent States'. For 21 headwords, one of the predicted clusters represents a new sense.[3] For instance, 'AGI' refers to both 'Air Quality Index' and 'al-Qaeda in Iraq', while 'Marco Polo' can mean a person name, a hotel/venue and more. An interesting case is 'broken heart' (medical use for heart attack vs. metaphorical use).

For German, we sample 1000 headwords from Digitales Wörterbuch der Deutschen Sprache (DWDS). Again, we exclude entries with more than one sense. Like for the WordNet experiment, we sample 25 uses for each headword from the lemmatized 2023 1M news corpus in the (German) Leipzig Corpora Collection (Leipzig Corpora Collection 2023b), that match POS and lemma. We discard headwords with less than 25 uses in the corpus. The remaining 108 headwords are uploaded to DURel and sense clusters are inferred as described above. In our analysis, we find 46 headwords with more than one cluster. We manually analyze the clusters and assign our own sense definitions, which we check against the corresponding DWDS entries. We find that 26 of the 46 headwords with more than one cluster do not have meaningful sense distinctions, including 'fotografisch' ('photographic') and 'Hemmschwelle' ('inhibition threshold'). 3 graphs are clustered correctly, but contain faulty uses ('Soli' includes uses for both 'Solo' and 'Soli', which are distinct headwords). This leaves 17 headwords for closer examination. The most interesting cases are 'Einlassung' ('statement' vs 'mounting'), 'Einmarsch' ('march-in' vs 'invasion'), 'Segnung' ('the act of blessing' vs 'the blessing itself'), 'Eck' ('street corner' vs 'corner of a goal') and 'Lebenswerk' ('life's work' vs 'charitable organization'). The graphs of 'Segnung' and 'Eck' also contain noise clusters, i.e. additional clusters that contain uses of both senses. For 9 headwords, the

[3] Following OED policy, we regard proper nouns as representing a proper sense. 20 out of 21 new senses are proper nouns.

new sense is an entity or a person, including ‘Alphabet’, which can refer both to the alphabet itself as well as to the company. In three cases, metaphorical/metonymic and literal uses are present in the sampled data (‘Ottawa’, ‘aufwerten’ and ‘KZ’).

Together, these experiments show that the DURel pipeline is a viable tool to identify headwords that need to be updated. Given a large enough corpus, the method can also easily be scaled to several thousand target headwords.

References

Armendariz, C., Purver, M., Ulcar, M., Pollak, S., Ljubešić, N. & Granroth-Wilding, M. (2020). CoSimLex: A resource for evaluating graded word similarity in context. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*, (pp. 5878–5886). European Language Resources Association.

Bansal, N., Blum, A., and Chawla, S. (2004). Correlation clustering. In: *Machine Learning*, 56(1-3), 89–113.

Berlin-Brandenburgischen Akademie der Wissenschaften (Ed.). DWDS – Digitales Wörterbuch der deutschen Sprache. Das Wortauskunftssystem zur deutschen Sprache in Geschichte und Gegenwart. https://www.dwds.de

Blank, A. (1997). Prinzipien des lexikalischen Bedeutungswandels am Beispiel der romanischen Sprachen. Niemeyer.

Cassotti, P., Siciliani, L., de Gemmis, M., Semeraro, G., & Basile, P. (2023). Xl-lexeme: Wic pretrained model for cross-lingual lexical semantic change. In: *Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics*, (pp. 1577–1585). Association for Computational Linguistics

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). Association for Computational Linguistics.

Erk, K., McCarthy, D. & Gaylord, N.. (2013). Measuring word meaning in context. In: *Computational Linguistics*, 39(3), 511–554.

Fellbaum C. (Ed.) (1998). *WordNet: An Electronic Lexical Database*. MIT Press.He, P., Liu, X., Gao, J., & Chen, W. (2020). *Deberta: Decoding-enhanced bert with disentangled attention*. arXiv preprint arXiv:2006.03654.

Kilgarriff, A. (1997). I don’t believe in word senses. *Computers and the Humanities*, 31(2), 91-113.

Kilgarriff, A. (2007). Word Senses. In: Agirre, E. & Edmons, P. (Eds.). *Word Sense Disambiguation - Algorithms and Applications* (pp. 29–46). Springer.

Leipzig Corpora Collection (2023a): *English News corpus based on material from 2023*. Leipzig Corpora Collection. https://corpora.uni-leipzig.de?corpusId=eng_news_2023_1M

Leipzig Corpora Collection (2023b): *German News corpus based on material from 2023*. Leipzig Corpora Collection. https://corpora.uni-leipzig.de?corpusId=deu_news_2023_1M

OED (2009). *Oxford English Dictionary*. Oxford University Press.

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*, (pp. 2227–2237). Association for Computational Linguistics.

Pilehvar, M. & 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)*, (pp. 1267–1273). Association for Computational Linguistics.

Raffel, R., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P., et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. In: *Journal of Machine Learning Research*, 21(140), 1–67.

Schlechtweg, D., Schulte im Walde, S., & Eckmann, S. (2018). Diachronic Usage Relatedness (DURel): A framework for the annotation of lexical semantic change. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (pp. 169–174). Association for Computational Linguistics.

Schlechtweg, D., McGillivray, B., Hengchen, S., Dubossarsky, H. & Tahmasebi, N. (2020). SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection. In: *Proceedings of the 14th International Workshop on Semantic Evaluation*, (pp. 1–23). Association for Computational Linguistics.

Schlechtweg, D. (2023). *Human and Computational Measurement of Lexical Semantic Change.* [Doctoral dissertation, University of Stuttgart].

Schlechtweg, D., Virk, S., Sander, P., Sköldberg, E., Theuer Linke, L., Zhang, T., Tahmasebi, N., Kuhn, J. & Schulte Im Walde, S. (2024). The DURel Annotation Tool: Human and Computational Measurement of Semantic Proximity, Sense Clusters and Semantic Change. In: *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, (pp. 137-149). Association for Computational Linguistics.

Sköldberg, E., Virk, S., Sander, P., Hengchen, S. & Schlechtweg, D. (2024). *Revealing semantic variation in Swedish using computational models of semantic proximity - Results from lexicographical experiments*. To appear.

Acknowledgements

This work has jointly been funded by the research program Change is Key! supported by Riksbankens Jubileumsfond (under reference number M21-0021) and the Swedish Academy.

Contact information

Pauline Sander

University of Stuttgart, Institute for Natural Language Processing

pauline.sander@ims.uni-stuttgart.de

Simon Hengchen

iguanodon.ai and Université de Genève

simon@iguanodon.ai

Wei Zhao

University of Aberdeen, Department of Computing Science

wei.zhao@abdn.ac.uk

Xiaocheng Ma

University of Aberdeen, Department of Computing Science

x.ma1.23@abdn.ac.uk

Emma Sköldberg

University of Gothenburg, Department of Swedish, multilingualism, language technology

emma.skoldberg@svenska.gu.se

Shafqat Mumtaz Virk

University of Gothenburg, Department of Swedish, multilingualism, language technology

shafqat.virk@svenska.gu.se

Dominik Schlechtweg

University of Stuttgart, Institute for Natural Language Processing

dominik.schlechtweg@ims.uni-stuttgart.de

Marko Tadić

Can LLMs Really Generate New Words?

Keywords morpheme-based morphology; neologisms; large language models

Introduction: theoretical background

Instead of using the word-based morphology (e.g. Aronoff (1976)), we believe that the problem we are trying to present here fits better in the morpheme-based morphology approach (e.g. Marantz (1992), Manova et al. (2020)), so this will be our theoretical background. In this approach the words are seen as "morphological objects" i.e. they are results of application of rules that combine lexical morpheme(s) with derivational morpheme(s) taken from two different lists: 1) list of lexical morphemes (traditionally: roots) which is open for expansion by native speakers of a language (by borrowing from other languages or by inventing new lexical morphemes); 2) list of derivational morphemes, which is in principle a closed set.[1] The rules that define how items from these two lists can be combined in a language are called Word Formation Rules (WFRs).

Halle in (Halle, 1973) introduces the difference between potential words and real words: "In other words, I am proposing that the list of morphemes together with the rules of word formation define the set of potential words of the language. It is the filter and the information that is contained therein which turn this larger set into the smaller subset of actual words. This set of actually occurring words will be called the dictionary of the language." Additionally, Tadić in (Tadić, 1994) mentions a "derivational capacity of a language" and sees it as an exhaustive list of all generated combinations of lexical and derivational morphemes in a language following the WFRs applied to two lists of morphemes. Alternatively, that derivational capacity can also be understood in a processual manner, i.e. as an automaton (FSA or FST) that is capable of generating such extensive list of combinations.

Are potential words from Halle (1973) also words? Or more generally: is language just what has occurred or has been attested (already performed) or language is also a potential, a capability (a competence) to generate so far unseen combinations of units? This question follows from Chomsky (1970) where the notion of word has been reinstated in the Generative Grammar and it opened the floor for specialised generative approaches in morphology.

[1] The third list of inflectional morphemes is of no interest here since traditionally new combination of lexical and derivational morphemes (traditionally: stems) are considered to be new words while combinations with inflectional endings are used for morphosyntactic reasons only and do not introduce new lexical item.

Should lexicographers include potential words in dictionaries or not? Up until recently, the usual answer to this question was "no", since traditionally, only the occurred words were submitted to the lexicographic descriptions.

Who actually defines how the potential words get filtered into a dictionary of a language? When native speakers apply WFRs and come up with the potential word and then issue it in a given co-text and con-text, we usually consider this a creative usage of language. This combination can be surprising, but if it's used in an appropriate co-text and con-text transferring the clearly delimited meaning, it may become a real word of a language.

Is it the same when such combination has been generated by LLM?

Example: investigation

Here we'll briefly use as an example the pilot study from Tadić (2023). There it was investigated how a NMT system, while translating from English to Croatian, can generate new, unseen combinations of lexical and derivational morphemes.

As a language resource the Croatian-English Parallel Corpus (Tadić, 2000) was used, a human-translated unidirectional newspaper corpus. Random 10,000 aligned (hr->en) sentence pairs were extracted and they encompassed ca 0.19 Mw in Croatian (*hr*) and 0.25 Mw in English (*en*). The English parts of aligned 10,000 sentence pairs were translated with the NMT system Hrvojka[2] (Vasiļevskis et al., 2023) back in Croatian (*hr-t*).

After matching the tokens from *hr-t* with the Croatian Morphological Lexicon (HML), an inflectional lexicon (Tadić, 2005), more than 4 thousand tokens were marked as unknown. Manual inspection detected that most of them are typos or named entities unknown to the HML.

However, 321 token were composed following completely all derivational and compositional rules in Croatian and yet these tokens couldn't be found in any written or online dictionary, they do not occur in any Croatian corpus, and couldn't be detected online by any search engine.

In Tadić (2023) the more detailed description of this research is provided, but here we'd like to list just few examples to illustrate how bilingual LLM tailored for neural machine translation generates new morpheme combinations when it needs to convey a particular meaning:

- expectable compound: compounds that could be expected having in mind possible combination of compounding parts, e.g. en: *self-denying* / hr-t: *samoopovrgavajući*, en: *late antique* / hr-t: *kasnoantika* instead or hr: *kasna antika*;

[2] https://hrvojka.gov.hr

- unexpectable compound: compounds that are partial errors in translation but convey the general meaning, e.g. en: *five-movement* / hr-t: *petokretni* instead of hr: *petostavačni*, en: *Euro game* / hr-t: *euroigre* instead of hr: *europske igre*;
- alternative derivation: derivation that uses different, but possible, derivation affix, e.g. en: *lace-makers* / hr-t: *čipkaši* instead of hr: *čipkarice*, en: *broker* / hr-t: *burzer*;
- unexpectable derivation: derivations that are partial errors in translation, but convey the general or alternative meaning, e.g. en: *swallow* (bird) / hr-t: *gutljica*, en: *(voucher) holders* / hr-t: *imatelji (vaučera)*.

Conclusion: questions and future directions

After this evidence, could we reiterate the question: Are these new combinations of lexical and derivational morphemes produced by LLM new words or just potential words? Should they be included in dictionaries or not? Particularly in cases that are today, unfortunately, becoming more and more present online. Namely, today vast quantities of texts are generated automatically and are being published online to be later crawled and included in different corpora. Sometimes it's already hard to tell whether a text has been generated by a human or by a machine? There is no firm obligation for watermarking or accompany with meta-data the texts generated by machine.

Could the valid combinations of morphemes, appearing to be composed of morphemes from two mentioned lists by application of WFRs and generated by LLMs, also be considered a creative usage of a language? How will this phenomenon influence lexicological and lexicographical theory and practice, it remains for a thorough discussion and decision. It seems that at this moment of development of LLMs, lexicographers (traditional, or modern ones) could select one of two possible lanes of progress:

1) discard any generated combination of lexical and derivational morphemes that has not been produced by a human speaker;
2) accept all generated combinations of lexical and derivational morphemes produced either by humans or machines if they have been following the WFRs.

Both lanes entail a number of questions and here are just a few: e.g. which combinations should be included in future dictionaries, should we have different dictionaries for humanly-generated and machine-generated words, what if a human sees a novel combination generated by machine and starts to use it with other humans, can we track down that this really happened, etc.?

Additionally, how exactly LLMs in their training procedure induce and adopt segmentation of words at the subword level (LLM-tokenisation), is another question that we yet need to tackle. For that we should develop a battery of intrinsic and extrinsic evaluation tools for LLMs in order to understand their inner structure (if possible and if not deliberately left non-transparent in the form of a black box) and measure their performance and all that for different NLP-tasks.

References

Aronoff, M. (1976). *Word Formation in Generative Grammar*, MIT Press.

Chomsky, N. (1970). Remarks on nominalization. In: R. Jacobs, P. Rosenbaum (Eds.), *Reading in English Transformational Grammar* (pp. 184-221). Ginn.

Halle, M. (1973). Prolegomena to a Theory of Word Formation. *Linguistic Inquiry*, 4(1), 3-16. http://web.mit.edu/morrishalle/pubworks/papers/1973_Halle_Prolegomena_to_a_Theory_of_Word_Formation.pdf

Manova, S., Hammarström, H., Kastner, I., Nie, Y. (2020). What is in a morpheme? Theoretical, experimental and computational approaches to the relation of meaning and form in morphology. *Word Structure* 13(1), 1–21. https://doi.org/10.3366/word.2020.0157

Marantz, A. (1992). Morphology: Word structure in generative grammar – Review. *Language*, 68(2), 413-416.

Tadić, M. (1994). *Računalna obrada morfologije hrvatskoga književnog jezika* [Doctoraal dissertation, University of Zagreb]. http://darhiv.ffzg.unizg.hr/id/eprint/1884/

Tadić, M. (2000). Building the Croatian-English Parallel Corpus. In M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperidis, G. Stainhauer (Eds.), *Proceedings of the Second International Conference on Language Resources and Evaluation* (pp. 523-530). European Language Resources Association.

Tadić, M. (2005). The Croatian Lemmatization Server. *Southern Journal of Linguistics*, 29(1/2), 206-217.

Tadić, M. (2023). Can Large Language Models Tell Us Something about Derivation Processes?. In M. Filko, K. Šojat (Eds.), *Proceedings of the Fourth International Workshop on Resources and Tools for Derivational Morphology (DeriMo 2023)* (pp. 33-38). Croatian Language Technologies Society.

Vasiļevskis, A., Ziediņš, J., Tadić, M., Motika, Ž., Fishel, M., Barkarson, B., Borg, C., Aquilina, K., Spiteri, D. (2023). National Language Technology Platform (NLTP): The Final Stage. In C. Oraşan, G. Corpas Pastor, R. Mitkov, J. Monti (Eds.), *Proceedings of the International Conference HiT-IT2023* (pp. 203-208). INCOMA Ltd.

Contact information

Marko Tadić

University of Zagreb, Faculty of Humanities and Social Sciences, Department of Linguistics

marko.tadic@ffzg.unizg.hr

Carole Tiberius, Kris Heylen, Jesse de Does, Bram Vanroy, Vincent Vandeghinste, Job van Doeselaar

LLMs and Evidence-Based Lexicography

Pilot Studies at INT

Keywords LLM; evidence-based lexicography; lexicographic workflow

The Dutch Language Institute (INT) has corpus-based workflows to compile historic and contemporary dictionaries and other types of lexicographic databases, mainly for Dutch but also for some other languages with a relation to Dutch. The INT is currently exploring how LLMs can be used for optimising different parts of these workflows without compromising data quality and reliability. These experiments are conducted against the background of a gradual integration of the INT's different lexicographic databases into one central lexicographic knowledge base. First, we briefly describe this move towards an integrated workflow and lexical knowledge base and then we describe where LLMs are being tested to facilitate this process. Rather than focussing on the use of LLMs for one specific task in one project, the paper explores their potential for operational improvements at an institutional level on the basis of a few pilot studies:

Can we use an LLM to classify example sentences under the correct sense using our own sense inventory?

- Case study using the sense inventory and example sentences from the project *Woordcombinaties* (Word Combinations). We have carried out a small pilot study with 86 polysemous lemmas (85 nouns and 1 verb) with a total of 250 senses using GPT-4 to classify 25 example sentences for each of those lemmas. In the input, we provide the senses for the lemma (as defined in *Woordcombinaties*), give a few examples and specify the expected output together with a concrete example of the output. In this setting about 91% of example sentences are correctly classified. In a next step, we intend to explore whether these results can be improved including RAG elements.
- Related to the first case study, we are exploring whether we can classify concordance lines from the project *Woordcombinaties* under the correct pattern. See poster on ChatGPT and Corpus Pattern Analysis accepted for presentation at the EURALEX conference.
- Case study using data from the Danish-Dutch dictionary. In this case study we will apply the same technique to restructure the content of this dictionary which was

created in Word. Example sentences are included in a block underneath the senses. To align this resource with the other bilingual resources in the *Vertaalwoordenschat* (Translation Vocabulary), example sentences need to be classified per sense. This task is slightly more complex as some of the example sentences are idiomatic expressions which do not clearly fit under one of the senses.

Can we use LLMs to generate definitions for neologisms in the format of our in-house dictionaries?

- Unknown forms and meanings have traditionally posed challenges for Natural Language Processing (NLP), and hence linguistic innovations serve as an interesting testbed for assessing the inferential capabilities of LLMs over previously unseen data. In collaboration with prof. Tim Van de Cruys and the Master of AI programme at KU Leuven (University of Leuven, Belgium) and in the framework of a master thesis (Kohkle 2024), we investigated the use of 2 different types of transformer models for the automatic generation of dictionary definitions in the style of the *Algemeen Nederlands Woordenboek* (ANW - Dictionary of Contemporary Dutch) and the *Woordenboek van Nieuwe Woorden* (WNW - Dictionary of New Words). The dataset for the experiments is extracted from both dictionaries and contains approx. 35 K triples consisting of (1) a headword, (2) a corpus example illustrating a specific sense and (3) the ANW or WNW's full definition for that sense. We split up the dataset in a training and test set. In the first series of experiments, we used the test set to finetune the multilingual encoder-decoder model mT5-large (Xue et al., 2021) to generate definitions for a headword based on a corpus example. In the second series of experiments, we used the same set-up to finetune the newer open-source Aya model (Üstün et al., 2024) which has been derived from mT5 but additionally pretrained to improve performance for 101 lesser resourced languages, including Dutch. The performance of both models is evaluated automatically, by calculating different similarity scores (BLEU, ROUGE-L, and BERTscore) between the generated and original definition from the dataset, and manually, by scoring a sample for fluency and accuracy. The finetuned Aya model clearly outperforms mT5 and is generally able to generate concise definitions that follow the AMW/WNW-template.

Can we use LLMs to create a core sense inventory for our contemporary and historical dictionaries and link them on the sense-level?

- INT develops and/or hosts lexicographic databases that describe word meaning at quite different levels of granularity, going from very fine-grained in the *Woordenboek der Nederlandsche Taal* (WNT - Dictionary of the Dutch Language), over relatively fine grained in the *Algemeen Nederlands Woordenboek* (ANW - Dictionary of Contemporary Dutch) to quite coarse grained in the *Referentiebestand Nederlands* (RBN - Dutch Reference Database) and the *Vertaalwoordenschat* (Translation Vocabulary). Within our central knowledge base for Dutch, these lexicographic resources have already been linked on the lemma level in the GiGaNT lexicon. To also link the databases on the sense-level, the institute has a 4-year project to construct a core sense inventory with a level of granularity that maximises linkability across resources, somewhat comparable to the semantic part of the COR lexicon for Danish (Pedersen et al., 2022). We are exploring the use of LLMs both for linking between different lexicographic databases and for establishing an appropriate level of sense granularity. In a first phase, a dataset containing different definitions for the same sense is currently being extracted from the lexicographic databases to test the performance of GPT-4 for monolingual dictionary linking as compared to existing tools like NAISC (McCrae et al., 2021). In a second phase, we plan to further enrich the dataset with a ground-truth for sense granularity reduction across multiple monolingual dictionaries.

Can we use an LLM for content simplification?

- The content in our lexicographic databases for modern Dutch (ANW and RBN) is generally aimed at educated native speakers and/or proficient L2 speakers of Dutch. However, the INT is regularly contacted to provide lexical resources that can support language training for lower proficiency speakers. These quite diverse groups include speakers with a linguistic disability, early stage second language learners, or students with an immigration background entering the higher education system. We plan to investigate the use of LLMs for the simplification of our lexicographic content (definitions and examples) by building on our ongoing case study *Duidelijke Taal* 'Clear Language' that looks at text simplification for Dutch more generally. This case study is currently creating a large-scale reference set for Dutch text simplification through crowdsourcing. The dataset consists of relatively complex sentences from the SONAR corpus (with a Leesindex coefficient higher than 60) that were automatically simplified using GPT-4 with the same prompt as used in the UWV/Leesplank project (available on HuggingFace). Through the in-house crowdsourcing application https://duidelijketaal.ivdnt.org/, users are then asked in different tasks to evaluate the automatic simplifications on dimensions like fluency, simplicity, accuracy. The resulting dataset can be used as a benchmark for LLMs and will made available in the CLARIN infrastructure at INT.

References

Brouwer, R.H.M. (1963), Onderzoek naar de leesmoeilijkheden van Nederlands proza [Research into reading difficulties of Dutch proza], Pedagogische Studiën 40, 454–464.

McCrae, J. P., Ahmadi, S., Yim, S. B., & Bajčetić, L. (2021). The ELEXIS system for monolingual sense linking in dictionaries. In *Electronic lexicography in the 21st century. Proceedings of the eLex 2021 conference*. (pp. 542-559). Brno: Lexical Computing CZ, s.r.o.

Kohkle, T. (2024). 'Multilingual Definition Modeling for Neologisms using Transformers'. M.Sc. Thesis, KU Leuven (University of Leuven)

Pedersen, B. Sandford, Hau Sørensen, N., Nimb, S., Flörke, I., Olsen, S., & Troelsgård, T. (2022). Compiling a suitable level of sense granularity in a lexicon for AI purposes: The open source COR lexicon. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, (pp.51-60).

Seidl, T. & Vandeghinste, V. (2024). Controllable Sentence Simplification in Dutch. Computational Linguistics in the Netherlands Journal. Vol 13. 31–61.

Üstün, A.,Aryabumi, V., Yong, Z., Ko, W., D'souza, D., Onilude, G., Bhandari, N. Singh, S., Ooi, H., Kayid, A., Vargus, A., Blunsom, P., Longpre, S., Muennighoff, N., Fadaee, M., Kreutzer, J., Hooker S. (2024). Aya Model: An Instruction Finetuned Open-Access Multilingual Language Model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics,* vol. 1. (pp. 15894–15939).

Xue,et al. (2021). mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (pp. 483–498).

Vandeghinste, V. & Bulté, B. (2019). Linguistic Proxies of Readability: Comparing Easy-to-Read and regular newspaper Dutch. Computational Linguistics in the Netherlands Journal, vol. 9. 81-100.

Dictionaries:

Algemeen Nederlands Woordenboek - ANW [Online service]. Available at the Dutch Language Institute: http://hdl.handle.net/10032/tm-a2-k8

Gigant-Molex (Version 2.0) (2022) [Data set]. Available at the Dutch Language Institute: http://hdl.handle.net/10032/tm-a2-u7

Referentiebestand Nederlands - RBN (Version 2.0.1) (2014) [Data set]. Available at the Dutch Language Institute: http://hdl.handle.net/10032/tm-a2-n2

SoNaR-corpus (Version 1.2.1) (2015) [Data set]. Available at the Dutch Language Institute: http://hdl.handle.net/10032/tm-a2-h5

Vertaalwoordenschat (2017) [Online service]. Available at the Dutch Language Institute: http://hdl.handle.net/10032/tm-a2-d2

Woordcombinaties [Online Service]. Available at the Dutch Language Institute: http://hdl.handle.net/10032/tm-a2-t4

Woordenboek der Nederlandsche Taal - WNT (Version 2.0) (2018) [Online service]. Available at the Dutch Language Institute: http://hdl.handle.net/10032/tm-a2-g3

Woordenboek van Nieuwe Woorden - WNW [Online service]. Available at the Dutch Language Institute: http://hdl.handle.net/10032/tm-a2-y4

Contact information

Carole Tiberius

Instituut voor de Nederlandse Taal

carole.tiberius@ivdnt.org

Kris Heylen

Instituut voor de Nederlandse Taal

kris.heylen@ivdnt.org

Jesse de Does

Instituut voor de Nederlandse Taal

jesse.dedoes@ivdnt.org

Bram Vanroy

Instituut voor de Nederlandse Taal

bram.vanroy@ivdnt.org

Vincent Vandeghinste

Instituut voor de Nederlandse Taal

vincent.vandeghinste@ivdnt.org

Job van Doeselaar

Instituut voor de Nederlandse Taal

job.vandoeselaar@ivdnt.org

Carole Tiberius, Elisabetta Ježek, Annachiara Clementelli,
Lut Colman

LLMs Automating Corpus Pattern Analysis

A Cross-Lingual Pilot Study for Dutch and Italian

Keywords LLM; corpus pattern analysis; lexicography

According to Hanks (2013), words in isolation do not have meaning but have meaning potential. Meaning is activated when words are combined with other words, in sentences. In the Dutch sentence *De bakker bakt brood* ('the baker bakes bread'), a very different meaning of the verb *bakken* is activated than in the expression *De bakker bakt er niets van* ('to make a mess of it' 'lit. the baker bakes nothing of it'). Hanks calls these semantically motivated recurring structures of words, patterns. In his Theory of Norms and Exploitations (TNE), he distinguishes two types of patterns: the normal, prototypical patterns of words, the norms, and the creative use of those normal patterns, the exploitations. An example of a sentence with a normal pattern for the verb *to talk* is a sentence in which someone speaks to someone else about something: "I want to talk to you about that." An example of an exploitation of the verb *to talk* can be found in a statement by the golfer Lee Trevino: "You can talk to a fade but a hook just won't listen." (Hanks, 2013: 213). The words *fade* and *hook* are not of the semantic type [[Human]] normally used with talk, but specific golf terms.

Norms and exploitations are two extremes of the spectrum and there is no sharp dividing line between them. Some norms are more normal than others; some exploitations are more extreme than others (as in the example with *talk* above). To gain insight into the norms and exploitations of a language, TNE studies actual language behaviour as recorded in corpora. The theory offers practical guidelines in the form of the Corpus Pattern Analysis (CPA) methodology to sort and classify the language data. Typically, lexicographers annotate 250 concordance lines identifying recurrent patterns of word usage and determining the semantic types (e.g. [[Human]], [[Animal]], [[Location]]) of the pattern slots of the verb. Semantic types are selected from a corpus-driven hierarchically structured ontology (CPA ontology, cf. Ježek & Hanks, 2010).

TNE is applied in practice in various languages, e.g. English (the Pattern Dictionary of English Verbs, Hanks & Pustejovsky 2005), Spanish (Renau & Nazar, 2016), Italian (Ježek et al., 2014; Giacomini & Rebosio, 2024), Croatian (Marini & Ježek, 2019) and Dutch (Colman & Tiberius, 2018). Creating such pattern dictionaries is still mainly a computer-assisted manual process and therefore very time-consuming.

In a series of small-scale experiments we are exploring whether ChatGPT (GPT-4o) could support the lexicographic pattern editing process. We are specifically considering ChatGPT's performance with regard to pattern generation, detection and annotation of argument structures and semantic type annotation, as well as classification of concordances according to implicature/sense and pattern (for this task, see also the poster on ChatGPT and Corpus Pattern Analysis for Dutch accepted for presentation at the EURALEX conference). We use the manually annotated data from the Italian (https://tpas.unipv.it/) and Dutch (https://woordcombinaties.ivdnt.org/) projects to evaluate the output from ChatGPT. So far, we have explored different prompting techniques (e.g. zero-shot, few-shot (Liu et al., 2023)) for the above tasks, but we intend to run more experiments by the time of the workshop.

Initial results suggest that pattern generation seems too ambitious using straightforward zero or few-shot prompting in both Italian and Dutch. Particularly, experiments with different types of information provided in the prompts (i. only patterns, ii. patterns and senses; iii. patterns, senses and examples) for similar (i.e. It. *vendere* 'sell', *acquistare* 'purchase' > *comprare* 'buy'), and dissimilar (It. *abbaiare* 'bark', *pulire* 'clean' > *comprare* 'buy') verbs show that for similar verbs:

- By providing only patterns, ChatGPT4 replicates the patterns of the verbs given as input but does not independently produce new patterns. The model always adds an explanation to the generated patterns. It claims that it can follow similar patterns and replicates them.
- By providing patterns and senses, although ChatGPT4 can add something new and original and produce more accurate patterns, it does not identify all the patterns the verb has.
- The results are better when we provide only examples and ask Chat GPT to generate patterns. We notice that ChatGPT4 is good at generating new examples that match the semantic types in the patterns.

On the other hand, for dissimilar verbs, the greatest difficulty remains that ChatGPT is unable to generate new patterns, regardless of the stimuli provided: it is tightly bound to the examples given. Furthermore, the patterns that ChatGPT4 produces are not complete.

However, initial results for the other tasks are much more promising, particularly for semantic types annotation of single arguments. ChatGPT can classify nouns into the correct semantic types, having as a reference the CPA ontology. This could be useful for automatic classification of nouns into semantic types and for checking the manual annotation.

In the workshop we intend to provide examples of the different types of experiments we are running and discuss the results cross-linguistically.

References

Colman, L. & Tiberius, C. (2018). A Good Match: a Dutch Collocation, Idiom and Pattern Dictionary Combined. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (Eds.), *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts* (pp. 233-246). Ljubljana, Slovenia.

Giacomini, L. & Rebosio, L. (2024). Introducing PhraseBase: A Linguistic Information System for language learners and for NLP. Presentation at the *PhrasaLex III Workshop on Phraseological Approaches to Lexicography*. Innsbrück. 10-12 July 2024.

Hanks, P. (2013). *Lexical Analysis: Norms and Exploitations*. Cambridge MA: The MIT Press.

Hanks, P., & Pustejovsky, J. (2005). A pattern dictionary for natural language processing. In *Revue Française de Linguistique Appliquée*, 10:2.

Ježek, E. & Hanks, P. (2010). What Lexical Sets Tell Us about Conceptual Categories. *Lexis. Journal in English Lexicology*, 4, 7-22. https://doi.org/10.4000/lexis.555

Ježek, E., Magnini, B., Feltracco, A., Bianchini, A. & Popescu, O. (2014). T-PAS; A resource of Typed Predicate Argument Structures for linguistic analysis and semantic processing. In N. Calzolari, K. Choukri, Th. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis (Eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* (pp. 890-895). Reykjavik, Iceland: European Language Resources Association (ELRA).

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., Neubig, G. (2023) Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing, *ACM Computing Surveys 55*, 9, 1–35, https://doi.org/10.1145/3560815

Marini, C. & Ježek, E. (2019). CROATPAS: A Resource of Corpus-derived Typed Predicate Argument Structures for Croatian. In *Proceedings of the Sixth Italian Conference on Computational Linguistics (CliC-it)*.

Renau, I. & Nazar, R. (2016). Automatic extraction of lexico-semantic patterns from corpora. In T. Margalitadze & G. Meladze (Eds.), *Lexicography and Linguistic Diversity: Proceedings of the XVII EURALEX International Congress* (pp. 823-830). Bari, Italy: Ivane Javakhishvili Tbilisi State University.

Contact information

Carole Tiberius

Instituut voor de Nederlandse Taal

carole.tiberius@ivdnt.org

Elisabetta Ježek

University of Pavia

elisabetta.jezek@unipv.it

Annachiara Clementelli

University of Pavia

annachiara.clementelli01@universitadipavia.it

Lut Colman

Instituut voor de Nederlandse Taal

lut.colman@ivdnt.org