Using Large Language Models to Generate Distractors for Language Games

Iztok Kosem[1,2,3], Špela Arhar Holdt[1,2]

[1] Faculty of Arts, University of Ljubljana, Aškerčeva 2, 1000, Ljubljana, Slovenia
[2] Faculty of Computer and Information Science, University of Ljubljana, Večna pot 113, 1000, Ljubljana, Slovenia
[3] Institut "Jožef Stefan", Jamova cesta 39, 1000, Ljubljana, Slovenia
E-mail: Spela.ArharHoldt@ff.uni-lj.si, Iztok.Kosem@fri.uni-lj.si

Abstract

This paper presents two tasks involving large language models (LLMs)—Gemini-2.0-flash and GPT-4o—used to generate distractors (i.e., incorrect options) for synonym and collocation questions in a language game. The lexical data for both tasks was sourced from the *Digital Dictionary Database of Slovene* (DDDS). Prompts were initially tested on a sample dataset with both models, and the better-performing model was selected for each task: Gemini-2.0-flash for synonyms, and GPT-4o for collocations. Evaluation results showed strong performance of the models, with over 80% of the generated distractors rated as appropriate. Common issues included non-existent or rare words and legitimate synonyms in the synonym task, and common collocations or distractors that improperly altered collocational structure in the collocation task. Additional filtering of the data was required to ensure game readiness. Further plans include using LLMs for the production of data for other games, as well as using LLM in the preparation of lexicographic data in the DDDS.

Keywords: language game, LLM, synonym, distractor, collocation, dictionary database

1. Introduction

Language games are a valuable tool for both testing and enhancing linguistic competence. In recent years, dictionary publishers have increasingly integrated games into their platforms, combining lexicographic content with interactive learning. At the Centre for Language Resources and Technologies, University of Ljubljana (CJVT UL), we have been exploring this intersection for over a decade. Our first game for Slovene, Game of Words, a collocation guessing game, was initially launched online and later adapted into a mobile app and expanded to additional languages during the ELEXIS project. It also introduced a synonym module, available for Slovene only (Arhar Holdt et al., 2021).

Building on this foundation, we have recently launched a new portal[1] that serves as a central hub for language games developed at CJVT UL. The games featured on the

[1] https://igre.cjvt.si/

portal primarily draw data from the Digital Dictionary Database of Slovene (DDDS; Gantar et al., 2016; Gantar, 2020; Kosem et al., 2021). A major challenge in game development is the preparation of high-quality data. Even with curated lexical resources, additional annotation and filtering—especially of inappropriate vocabulary —are often required. While semi-automatic methods can help, manual oversight remains essential to ensure the quality of game content.

When developing our newest game, which tests the players in the knowledge of synonyms and collocations, we needed a combination of the data from DDDS (correct answers) and incorrect language data, which we call distractors. We decided to test the potential of Large Language Models (LLMs) in the preparation of this distractor data.

The paper first offers an overview of related work using LLMs for generating or curating lexicographic data. Next, we present our task on generating distractors for synonyms. We also present preliminary findings of a task on generation distractors for collocations. In the conclusion, we summarize the main findings and outline our plans for both the game portal and further use of LLMs in data preparation.

2. LLMs for lexicographic/linguistic purposes

The impact of LLMs in lexicography and lexicographic research has been unprecedented. More and more studies in using LLMs, particularly chatbots such as ChatGPT, are being conducted every year, with LLMs and AI in general also taking over the focus of many (lexicographic) conferences.

In 2023, de Schryver (2023) wrote an overview paper of the LLM-related work in lexicography, summarizing the findings of several papers and presentations (de Schryver & Joffe, 2023; Barrett, 2023; McKean & Fitzgerald, 2023; Jakubiček & Rundell, 2023; Tran et al., 2023; Nichols, 2023; Lew, 2023; de Schryver, 2024), which conducted experiments using LLMs in various lexicographic tasks. The majority of studies focused on the generation of definitions and examples, reporting mixed results, while good results were reported on tasks such as translation, and synonym classification. It is noteworthy that the majority of these studies conducted experiments on English.

More recently, researchers working on lexicographic resources for languages other than English have reported on their studies testing LLMs, e.g., Tiberius et al. (2024) on Dutch, Kosem et al. (2024) and Arhar Holdt et al. (2025) on Slovene, Špica & Perak (2024) on Japanese, Tuulik et al. (2024) on Estonian, and Kosem et al. (2024) on Estonian, Dutch, Portuguese, and Slovene. Promising results were reported for tasks such as the distribution of examples and synonyms under senses and the generation of definitions for neologisms, and mixed results for generating sense division and identifying good examples for sensitive meanings.

While none of the aforementioned studies tested LLMs for generating non-language or atypical language, some elements of these studies are relevant for the purposes of this paper. For example, studies conducted by Kosem et al. (2024) and Arhar Holdt et al. (forthcoming) included identification of non-synonyms among the synonym candidates. Also relevant are experiments that include providing frequency information about words (de Schryver, 2023), with results being reported "convincing". Important in this regard are studies by Davies (2025a), in which predictions by two LLMs (GPT-4o and Gemini 1.5 Pro) are compared with the data from large corpora. Davies reports that a) LLMs are good at generating high frequency lists and ranking words by frequency, and fair or rather poor in generating lists of low frequency words (Davies, 2025b); b) LLMs are not that good at generating phrases that match the corpus data (Davies, 2025c), but are much better at providing collocates, especially for low frequency words (Davies, 2025d).

3. Generating distractors with LLMs

The need for the generation of distractors (i.e., wrong answers) arose in the framework of our language game portal. The portal was launched in early 2025 and is currently featuring three language games, each game drawing on different aspects of lexical knowledge (see Arhar Holdt & Kosem, 2025, for a more detailed presentation of the games). The games utilize data from a lexicographic resource, i.e., Digital Dictionary Database of Slovene (DDDS), which is constantly being improved and enhanced. The experiments of using LLMs for lexicographic tasks such as sense division, definition generation, and entry summarization are currently under way.

There is a considerable body of research on distractors, especially in educational settings such as language testing and language learning, and mostly related to multiple-choice questions (e.g., Thissen et al., 1989; Collins, 2006; Haladyna & Rodriguez, 2013; see Gierl et al., 2017 for an overview). Relatedly, automatic distractor generation is a well-researched area in the fields of NLP and computational linguistics (see Alhazmi et al., 2024 for an overview); especially known is the method by Mitkov et al. (2006). As shown by literature, two strategies of distractor development are in use: in the first one, distractors are plausible but incorrect alternatives, often derived from common mistakes or misconceptions. In the second one, distractors are similar in semantic or structural properties.

For our purposes, the second strategy seemed to be better, given that synonyms represent the first, and supposedly easier part of the game. In addition, there is little information on common mistakes in synonym use in Slovene. We decided not to use any of the existing computational approaches as we were lacking a comprehensive (and open access) semantic resource – DDDS is still under development and Slovenian Wordnet is not of such quality to be used for these purposes. Finally, the motivation was also to test the potential of LLMs for such tasks.

The distractor data is part of our newest game Kombinator (named after on the way words combine or are related in a language), which is currently being developed. In the game, the players are tested on their knowledge of synonyms and collocations[2]. In the synonym part, the player is offered a headword and from two to four potential answers, with one or two being correct ones. In the collocation part, the player is offered the headword in a particular syntactic relation and a missing slot before or after the headword, with again two to four potential answers.

With LLMs showing promising results in lexicographic and linguistic tasks, we decided to test their potential for creating incorrect language data, i.e., non-synonyms and non-collocations. We report on the results of both tasks, although the evaluation for the collocation task is still in progress.[3] We tested two models, GPT-4o by OpenAI and Gemini-flash-2.0 by Google, using API.

The output was produced in the JSON format, and we then wrote a script that produced the combined version of the input data and LLM output in the tsv format. The data in the tsv format was then imported into an Excel file and used for analysis.

3.1 Generating distractors for synonyms

For this task, we used the synonym data from the DDDS, specifically from the Thesaurus of Modern Slovene (Arhar Holdt et al., 2023). The Thesaurus is an automatically generated resource which is constantly being updated with lexicographers' and user contributions. It is based on a large bilingual resource and the reference corpus of Slovene (see Krek et al., 2017; Arhar Holdt et al., 2018; Krek et al., 2020), and linguistic evaluation has shown that the synonym data is fairly reliable. The Thesaurus contains entries in various stages of completion, from fully automatic ones, hybrid ones (sense division available, synonyms not yet or only partly validated and distributed under senses) and completed ones.

3.1.1 Methodology

A selection of 5,000 headwords from the Thesaurus of Modern Slovene was made for the game. The criteria for selecting the headwords (nouns, adjectives, verbs, and adverbs) were that they had to be frequent and had to have several synonyms, preferably more than five. The headwords belonged to all three types of entries in the Thesaurus (manual, hybrid, automatic). The data was extracted from the DDDS, including information such as headword ID, sense ID, relation ID, and relation type (we used only core synonyms). The dataset contained 51,023 headword-synonym

[2] There are also plans to add antonym data at a later point.

[3] We will present the full results at the conference.

pairs.

For testing the prompt and evaluating the performance of the two LLM models, we prepared a sample of 30 headwords and all their synonyms from the dataset. The sample reflected the word class distribution of the entire dataset.

The prompt used:

> *You are given headword and a synonym. Create a distractor — a word that looks similar to the synonym but has a different meaning.*
>
> *The distractor must be the same part of speech as the synonym (e.g., if the synonyms are verbs in their base form, the distractor must also be a verb in its base form).*
>
> *The distractor must not include sensitive vocabulary (e.g., words related to minorities, religion, sexual content, violence, etc.).*
>
> *The distractor must be a frequent word in the Slovene language.*
>
> *The distractor must look similar to the synonym but have a different meaning.*
>
> *Write the distractor in the same line as the headword and synonym, following this format: živahen - vesel - resen. These are the headword and synonym: {word} - {synonym}*
>
> *The distractor cannot be one of these words: {synonym_set}.*

The last line was added during testing as both models sometimes generated legitimate synonyms (other than the one offered) as distractors. To avoid this, we included the list of all the synonyms of the headword in the database (both core and near). There were still other occasional problematic data generated by one or both models, e.g., repeating the headword or part of the instructions, and generating the same distractor for several synonyms of the same headword (not considered as an error). The final evaluation of test data showed a marginally better performance by the Gemini-2.0-flash model, which was then used on the entire dataset.

3.1.2 Results

The evaluation of the Gemini-2.0-flash output was conducted in several steps. Firstly, we automatically identified all the distractors that were already synonyms in the Thesaurus (there were 201 or 0.4%)[4], or were not found in the reference written corpus Gigafida 2.0 (there were 5,795 or 11.4%). The corpus frequency information was not always reliable as we were using a lemma-based frequency list, and some of the distractors were not provided in lemma form (e.g., the adjective was provided in the definite form). Consequently, we decided to leave the decision on their inclusion to the manual evaluation.

[4] This included 25 cases where the distractor was the same word as the synonym.

The manual evaluation of all the distractors (with the exception of the distractors that were existing synonyms) was conducted by two lexicographers. 40,892 (80% of the total dataset) were identified as good, while 9,930 were labelled as bad. The bad distractors belonged to three groups: 4,763 were not found in the corpus, 2,446 were legitimate synonym candidates (but not already in the Thesaurus), and 1,896 were non-synonyms but exhibited other issues (e.g., the distractor was a very rare or obscure word, a derogatory or vulgar word, or contained a combination of Slovene and English text).

Additionally, we performed evaluation of headword-synonym pairs for the purposes of the game, as the manual analysis pointed to some problematic cases in which the synonymy would be difficult to understand without more context (e.g., biti – pomeniti / *to be – to mean*; biti – nahajati se / *to be – to lie*; iti – pridružiti se / *to go – to join*). We also took a stricter look at all multi-word synonyms, especially the ones with final preposition (e.g., kazati – jasno govoriti o / *indicate – clearly speak of*). After excluding the problematic cases, the final dataset for the game included 34,843 headword-synonym-distractor triplets.

In terms of evaluating the LLM model only (not considering the specific requirements of the game), Table 1 shows the breakdown of all the problems in which the Gemini output contradicted the instructions provided in the prompt. However, not all of these cases were automatically excluded from the dataset for the language game. For example, many cases where a distractor was a different part of speech than the synonym were found acceptable, as long as the distractor was a legitimate (and frequent) Slovenian word and was similar to the synonym in form.

The frequency of the distractors deserves a special analysis. We have already mentioned the problem of some distractors not being in the lemma form. To obtain a better idea of the overall frequency of the distractors provided, we focussed on all the distractors (good or bad) that were found in the corpus. Out of 44,026 distractors, only 1,720 (3.9%) had very low frequency (less than 100 occurrences or 0.09 occurrences per million words in the corpus), with 508 of them occurring 10 times or fewer in the corpus, and 3,301 distractors had fairly low frequency (between 0.09 and 0.77 occurrences per million words)[5]. Over half of the distractors (61%) were very frequent in the corpus (9 or more occurrences per million words).

We also compared the frequency of single-word synonyms with the frequency of single-word distractors, where both words were found in the corpus. Out of 36,468 pairs, 23,917 (65.6%) had the distractor with a higher corpus frequency than the synonym. There was only one case where the synonym and the distractor were equally frequent in the corpus. In 21,360 out of 23,917 cases (89%) where the distractor was more frequent than the synonym, the distractor was evaluated as good. Similarly, in 10,867 out of 12,551 cases (86.6%) where the synonym was more frequent than the distractor,

[5] These frequency limits are used in our dictionary resources when determining frequency ranks of headwords.

the distractor was evaluated as good. Therefore, the relation between the distractor and synonym frequency does not seem to be relevant for the distractor quality.

Type of problem	**number of cases**	**examples**
different part of speech	565	zanesljiv – zvest – **vest** *(reliable – loyal – conscience)* žalosten – beden – **preden** *(sad – pathetic – before)* zadovoljiv – dober – **bober** *(satisfactory – good – beaver)*
sensitive vocabulary	13	posrati – to shit (5 cases) posiliti – to rape (3 cases) poscati – to piss (1 case)
not found in the reference corpus	4,763	*reseti, držeti, sveteti, propovedujoč, prostiran, obisten, poslasten, preteza, lepušina, plestenje, neznanež*
already a synonym in the Thesaurus	201	64 nouns, 68 verbs, 56 adjectives, 13 adverbs
new synonym candidates	2,446	strašljiv – **grozen** *(scary – horrifying)* uglajen – **kulturen** *(cultured – civilised)* zdrav – **pozdravljen** *(of good health; well – healed)* akontacija – **nakazilo** *(advance – transfer)* blebetanje – **klepetanje** *(gabbing – natter)*

Table 1: Problems of the Gemini-2.0-flash output consider the instructions in the prompt.

One of the instructions in the prompt was also that the distractor must look similar to the synonym. We used a similarity formula to calculate the similarity between the

distractor and the synonym, as well as between the distractor and the headword. We tested various metrics (Levenshtein, Jaccard, Cosine, Jaro-Winkler, Gestalt pattern matching) and found Gestalt pattern matching (Ratcliff & Metzener, 1988) to be the most suitable for Slovene in this particular case. For example, Gestalt showed much better results when identifying similarity beyond the first couple of letters, e.g., *izdelati - predelati, izmeriti – premetiti, odpadajoč – pripadajoč.* We found the value of 0.75 or above to be the best indicator of similarity in form. As shown in Table 2, slightly less than one fifth of all the distractors displayed similarity to synonyms, with a large percentage of them (73%) being good distractors. There was also a considerable number of distractors that were similar to the headword. There were 472 cases in which the distractor was similar to both the synonym and the headword (e.g., izbrisati – zbrisati – izrisati; izsiliti – izpuliti – izpiliti; opustitev – popustitev – odpustitev).

	distractor-synonym	**% of all**	**distractor-headword**	**% of all**
good distractors	6,885	13.5	2,980	5.8
bad distractors	2,571	5.0	948	1.9
- synonyms	552	1.1	248	0.5
TOTAL	**9,456**	**18.5**	**3,928**	**7.7**

Table 2: Distractors and synonyms/headwords with similarity score of 0.75 or higher.

3.2 Generating distractors for collocations

Collocation data was also taken from the DDDS, specifically from the Collocations Dictionary of Modern Slovene (Kosem et al., 2023). The Collocations dictionary, similarly as the Thesaurus, contains three types of entries (automatic, hybrid and manual).

3.2.1 Methodology

We selected 3,354 headwords, all manually completed by lexicographers, for the task. The headwords were nouns, verbs, adjectives, and adverbs. The metadata extracted included information on collocation ID, syntactic structure, frequency, and salience, as well as lemma and morphosyntactic tags of all the elements in the collocation.

For testing the prompt and evaluating the performance of the two LLM models, we prepared a sample of 25 headwords and all their collocations (566 in total) The sample consisted of headwords from all four word classes selected.

The prompt used:

> *We are preparing a language game where the player will be given a headword, a collocation (combination of the headword and another word) and a distractor (a collocation that has the same headword, but the other word is not a collocate of the main word). For example "huge victory" is a collocation of "victory", but "rotten victory" is not so "rotten is a good distractor. The rules for forming distractors are the following:*
>
> *1. Distractor has to be a single word.*
>
> *2. Distractor has to have the same part of speech as the word being replaced (e.g. if the word next to the headword is a noun, the distractor should also be a noun).*
>
> *3. Distractor should not include sensitive vocabulary, e.g. related to minorities, nationalities, religion, sexual content and similar.*
>
> *4. Distractor has to be a word that is frequent in the Slovene language.*
>
> *5. Distractor has to be a word that is completely unlikely to occur with the headword.*
>
> *Return the distractor in the same format as the examples below:*
>
> *Example: hiter - hitre rešitve (hiter + rešitve) - hitre težave (hiter + težava)*
>
> *Example: obljuba - držati obljubo (držati + obljuba) - najti obljubo (najti + obljuba).*
>
> *This is the headword: {headword}. This is the collocation: {collocation} ({all_collocation_parts}). The distractor has to be a collocation that contains the headword {headword} but is unlikely to occur with it. Only return the distractor in the correct format with the given headword {headword}. No explanations, no other text.*

The prompt consisted similar instructions as were used in the synonym task, with some additional instructions and rules specific to collocations. For example, we wanted the model to include in the output the lemmas of all the elements in the collocation, as this helped with validation and automatic evaluation. Also, rule number 5 was added during testing as both models often produced valid collocations as distractors; the addition of the rule produced much better results.

The evaluation of test data showed a better performance by the GPT-4o model, which was then used on the entire dataset. Based on the evaluation, we have also decided to limit the number of syntactic structures used, as some exhibited high degree of bad distractors, regardless of the model used. Thus, 17 syntactic structures were used for the main task, among them adjective + noun, noun + noun in genitive, and verb + noun in accusative, which also represented the largest share of collocations in the dataset. The final dataset consisted of 59,496 collocations.

3.2.2 Preliminary results

So far, we have only conducted automatic evaluation of distractors, where we compared

the distractors (using the lemma and syntactic structure information) with the collocations stored in our data warehouse. The data warehouse contains all the collocations extracted from the reference corpus Gigafida, including those with frequency of 1. It is used as a repository of collocational data, as not all the data ends up in the DDDS.

The evaluation showed that there were 51,116 (86%) distractors that were not found among the collocations in the data warehouse. Further 4,934 (8.3%) distractors had frequency of 5 or less in the reference corpus. Based on this, we can observe that the model's performance was quite good.

There were 5,531 distractors that occurred more than once in the dataset (see Table 3 for the list of most frequently repeated ones). The repetition of distractors is not necessarily a sign of poor performance of the model. The headword digitalen (*digital*) seem to have been particularly challenging for the model, with two distractors used for 83 out of 96 collocations – digitalni krompir (*digital potato*) or digitalna juha (digital soup). Yet, the majority (4,583 or 82.9%) of these repeated distractors seem to be good, as they were not found in the reference corpus.

collocation distractor	**number of occurrences**
digitalni krompir *(digital potato)*	42
digitalna juha *(digital soup)*	41
čokoladna miza *(chocolate table)*	19
porečje mize *((river) basin of the table)*	17
cerkveni računalnik *(church computer)*	17
alkoholna miza *(alcohol table)*	17
debelina pesmi *(song thickness)*	15
kronična zabava	14

(chronic party)	
izrecno zaspati *(specifically fall asleep)*	14
avtomatično plavati *(automatically swim)*	14

Table 3: Top 10 distractors offered more than once.

Among the problematic aspects of the model's performance are 5,714 cases where the model replaced the wrong part of the collocation, i.e., the headword rather than the collocate. Furthermore, in 169 cases, the distractor generated by the model was exactly the same as the collocation. These problems seem to be easily solvable with some prompt fine-tuning.

As in the synonym task, we are also conducting further manual evaluation of the data for the purposes of the game to exclude problematic content (sensitive language etc.). Furthermore, due to the game specifications, some collocation-distractor pairs will need to be excluded from the final dataset because they will not share the headword in the same for, e.g., will differ in gender or number (e.g. **abstraktno** kiparstvo – **abstrakten** sendvič).

4. Conclusion

The results of two tasks in which LLMs (Gemini-2.0-flash or GPT-4o) were used in creating distractors for synonyms and collocations in the language game respectively show good performance by the specific LLM used for the task. In the synonym task, 80% of the obtained distractors were identified as good. Among the problematic parts of the Gemini-2.0 flash output, considering the instructions in the prompt, were non-words, valid synonyms (either ones already in the Thesaurus or potential new ones), distractors belonging to a different part of speech, and distractors with a sensitive meaning.

In the collocation task, the performance of LLM (GPT-4o) was even better, with 86% of distractors being non-collocations (not appearing in the corpus). There were many cases where the task was incorrectly performed, i.e., the wrong part of the collocate replaced, while the cases with the same distractor being offered repeatedly are not necessarily problematic, but will be less useful for the game.

It is important to note that the good distractor data produced by the LLMs needs further filtering for the purposes of the game. This can be either due to the way in which the data is presented in the game, or due to the problematic aspects of the original data (e.g., demanding nature of the synonym relation).

Our future plans include finishing the manual evaluation of the collocation distractors, and preparing the final datasets for the Kombinator game. We will be repeating the distractor generation in the future with new data, so we intend to further fine-tune the prompt, also by providing several good and bad examples of the expected output. We plan to add the antonym data to the game, and intend to use a similar LLM-based task to produce the antonym distractors. Furthermore, we have several language games in the pipeline and have already identified tasks in which LLM can be used to prepare the data needed. Finally, we have been using and will continue testing LLM in various tasks related to the production of the lexicographic data which can then be used for language games.

5. Acknowledgements

The research program Language Resources and Technologies for Slovene (P6-0411), the infrastructure Network of Research and Infrastructural Centers UL (I0-0022) and the projects Large Language Models for Digital Humanities (GC-0002) are funded by the Slovenian Research and Innovation Agency. Ministry of Culture of the Republic of Slovenia is funding the JR-infrastruktura-SJ-2024-2025 project Data completion and gamification of dictionary resources at CJVT UL (PODVIG). We thank the reviewers for their constructive and valuable suggestions.

6. References

Alhazmi, E., Sheng, Q. Z., Zhang, W. E., Zaib, M. & Alhazmi, A. (2024). Distractor Generation in Multiple-Choice Tasks: A Survey of Methods, Datasets, and Evaluation. *The 2024 Conference on Empirical Methods in Natural Language Processing*. https://doi.org/10.48550/arXiv.2402.01512

Arhar Holdt, Š., Čibej, J., Dobrovoljc, K., Gantar, A., Gorjanc, V., Klemenc, B., Kosem, I., Krek, S., Laskowski, C. & Robnik Šikonja, M. 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

Arhar Holdt, Š., Gantar, P., Kosem, I., Pori, E., Robnik Šikonja, M. & Krek, S. (2023). Thesaurus of Modern Slovene 2.0. In *Proceedings of eLex 2023: Electronic Lexicography in the 21st Century*, Brno, pp. 366–381. https://elex.link/elex2023/wp-content/uploads/82.pdf

Arhar Holdt, Š., Gapsa, M., Gantar, P. & Kosem, I. (forthcoming). Potencial ChatGPT-ja pri razvoju Slovarja sopomenk sodobne slovenščine. *Contributions to Contemporary History*.

Arhar Holdt, Š. & Kosem, I. (2025). CJVT Igre: New Word Games Based on the Digital Dictionary Database of Slovene. Proceedings of eLex 2025.

Arhar Holdt, Š., Logar, N., Pori, E., Kosem, I. (2021). Game of words: play the game, clean the database (2021). In Z. Gavriilidou, L. Mitits & S. Kiosses (eds.) *Lexicography for inclusion: EURALEX XIX: Congress of the European Association for Lexicography*: 7-9 September 2021, virtual: proceedings book. Vol. 2. Komotini: Democritus University of Thrace. pp. 41-49. https://euralex.org/publications/game-of-words-play-the-game-clean-the-database/

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.

Collins, J. (2006). Writing multiple-choice questions for continuing medical education activities and self-assessment modules. *Radiographics*, 26, 543-551. doi:10.1148/rg.262055145

Davies, M. (2025a). Comparing the predictions of Large Language Models to actual corpus data. (White papers). English-Corpora.org. https://www.english-corpora.org/ai-llms/

Davies, M. (2025b). Corpora and LLMs: comparing data on word frequency. (White paper). English-Corpora.org. https://www.english-corpora.org/ai-llms/word-frequency.pdf

Davies, M. (2025c). Corpora and LLMs: comparing data on phrase frequency. (White paper). English-Corpora.org. https://www.english-corpora.org/ai-llms/phrase-frequency.pdf

Davies, M. (2025d). Corpora and LLMs: comparing collocates data. (White paper). English-Corpora.org. https://www.english-corpora.org/ai-llms/collocates.pdf

de Schryver, G-M. (2023). Generative AI and Lexicography: The Current State of the Art Using ChatGPT. *International Journal of Lexicography* 36, pp. 355–387.

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.' 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. (2024). The Future of the Dictionary. In E. Finegan & M. Adams (eds.) *The Cambridge Handbook of the Dictionary*. Cambridge: Cambridge University Press.

Gantar, P. (2020). Dictionary of Modern Slovene: From Slovene Lexical Database to Digital Dictionary Database. Rasprave Instituta Za Hrvatski Jezik i Jezikoslovlje, 46(2), 589–602. https://doi.org/10.31724/rihjj.46.2.7.

Gantar, P., Kosem, I. & Krek, S. (2016). Discovering automated lexicography: the case of Slovene lexical database. *International Journal of Lexicography*, 29(2), pp. 200–225. https://doi.org/10.1093/ijl/ecw014

Gierl, M. J., Bulut, O., Guo, Q. & Zhang, X. (2017). Developing, Analyzing, and Using Distractors for Multiple-Choice Tests in Education: A Comprehensive Review.

Review of Educational Research 87(6), pp. 1082-1116. https://doi.org/10.3102/0034654317726529

Haladyna, T. M., & Rodriguez, M. C. (2013). Developing and validating test items. New York, NY: Routledge.

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.* Brno: Lexical Computing, pp. 508–523. https://www.youtube.com/watch?v=8e52vvDpdfQ

Kosem, I., Arhar Holdt, Š., Gantar, P. & Krek, S. (2023). Collocations Dictionary of Modern Slovene 2.0. In M. Medveď et al. (ed.) *eLex 2023 : electronic lexicography in the 21st century (eLex 2023): proceedings of the eLex 2023 conference*, 27–29 June 2023. Brno: Lexical Computing CZ, pp. 491-507. https://elex.link/elex2023/wp-content/uploads/100.pdf.

Kosem, I., Gantar, P., Arhar Holdt, Š., Gapsa, M., Zgaga, K. & Krek, S. (2024). AI in Lexicography at the University of Ljubljana: case studies. In S. Krek (ed.) Book of abstracts of the workshop Large Language Models and Lexicography. 8. October 2024, Cavtat, Croatia, pp. 29-32.

Kosem, I., Krek, S. & Gantar, P. (2021). Semantic data should no longer exist in isolation: the Digital Dictionary Database of Slovenian. In Z. Gavriilidou, L. Mitits & S. Kiosses (eds.) *EURALEX XIX, Congress of the European Association for Lexicography, Lexicography for inclusion: 7–9 September 2021, virtual: abstracts book*. Komotini: Democritus University of Thrace, pp. 81–83. https://euralex2020.gr/wp-content/uploads/2021/09/EURALEX2020_BookOfAbstracts-Preview-1.pdf

Kosem, I., Zingano Kuhn, T., Arhar Holdt, Š., Koppel, K., Tiberius, C., Zviel-Girshin, R., Waszink, V. & Zgaga, K. (2024). Can AI assist in Selecting Dictionary Examples? A Case Study in Four Languages. In K. Štrkalj Despot, A. Ostroški Anić & I. Brač (eds.) *Lexicography and Semantics. Book of abstracts of the XXI EURALEX International Congress*. Institut za hrvatski jezik, pp. 128-130.

Krek, S., Arhar Holdt, Š., Erjavec, T., Čibej, J., Repar, A., Gantar, P., Ljubešić, N., Kosem, I. & Dobrovoljc, K. (2020). Gigafida 2.0: the reference corpus of written standard Slovene. In N. Calzolari (ed.). *LREC 2020: Twelfth International Conference on Language Resources and Evaluation*: May 11-16, 2020, Marseille, France. Paris: ELRA - European Language Resources Association, pp. 3340-3345. http://www.lrec-conf.org/proceedings/lrec2020/LREC-2020.pdf

Krek, S., Laskowski, C. & Robnik-Šikonja, M. (2017). From translation equivalents to synonyms: creation of a Slovene thesaurus using word co-occurrence network analysis. In I. Kosem et al. (eds.) *Proceedings of eLex 2017: Lexicography from Scratch*, 19-21 September 2017, Leiden, Netherlands. https://elex.link/elex2017/wp-content/uploads/2017/09/paper05.pdf

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

Sciences Communications 10, no. 704. https://doi.org/10.1057/s41599-023-02119-6

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

Mitkov, R., Ha, L. A., & Karamanis, N. (2006). A computer-aided environment forgenerating multiple-choice test items. Natural Language Engineering,12, 177–194. doi:10.1017/S1351324906004177

Nichols, W. (2023). Invisible Lexicographers, AI, and the Future of the Dictionary. Paper presented at the eLex 2023 Conference: Electronic Lexicography in the 21st Century. Brno, Czech Republic, 27-29 June 2023. https://www.youtube.com/watch?v=xYpwftj_QQI

Ratcliff, J. W. & Metzener, D. (1988). Pattern Matching: The Gestalt Approach. *Dr. Dobb's Journal* (46).

Špica, D. & Perak, B. (2024). Enhancing Japanese Lexical Networks Using Large Language Models – Extracting Synonyms and Antonyms with GPT-4o. In K. Štrkalj Despot, A. Ostroški Anić & I. Brač (eds.) *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress*. Institut za hrvatski jezik, pp. 283-303.

Thissen, D., Steinberg, L., & Fitzpatrick, A. R. (1989). Multiple-choice models: Thedistractors are also part of the item. *Journal of Educational Measurement*, 26, 161–176. https://doi.org/10.1111/j.1745-3984.1989.tb00326.x

Tiberius, C., Heylen, K., de Does, J., Vanroy, B., Vandeghinste, V. & van Doeselaar, J. (2024). LLMs and Evidence-Based Lexicography: Pilot studies at INT. In S. Krek (ed.) Book of abstracts of the workshop Large Language Models and Lexicography. 8. October 2024, Cavtat, Croatia, pp. 49-52.

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.* Brno: Lexical Computing, pp. 19–38. https://www.youtube.com/watch?v=rQC3Rz04b20

Tuulik, M., Risberg, L., Koppel, K., Aedmaa, E., Prangel, E., Zupping, S., Vainik, E. & Langemets, M. (2024). Who are Better at Semantics – Experienced Lexicographers or LLMs? In K. Štrkalj Despot, A. Ostroški Anić & I. Brač (eds.) *Lexicography and Semantics. Book of abstracts of the XXI EURALEX International Congress*. Institut za hrvatski jezik, pp. 219-221.

Resources:

OpenAI. (2023). *ChatGPT* (March 2025 version, ChatGPT-4o) [Large language model]. https://chat.openai.com/chat

Google Gemini: https://gemini.google.com/ (version 2.0-flash; API; February 2025)