Thesaurus of Modern Slovene 2.0

Špela Arhar Holdt[1,2], Polona Gantar[1], Iztok Kosem[1,3], Eva Pori[1], Marko Robnik-Šikonja[2], Simon Krek[1,3]

[1] Faculty of Arts, University of Ljubljana, Aškerčeva ulica 2, 1000 Ljubljana, Slovenia

[2] Faculty of Computer and Information Science, University of Ljubljana, Večna pot 113, 1000 Ljubljana, Slovenia

[3] Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia

E-mail: Spela.ArharHoldt@ff.uni-lj.si, Apolonija.Gantar@ff.uni-lj.si, Iztok.Kosem@ff.uni-lj.si, Eva.Pori@ff.uni-lj.si, Marko.RobnikSikonja@fri.uni-lj.si, Simon.Krek@ijs.si

Abstract

This paper describes the improvement of the Thesaurus of Modern Slovene from version 1.0 to 2.0. The Thesaurus is a digitally-born, automatically created resource that provides fast access to open data on modern language use and is gradually improved through editing and user participation. The initial version 1.0 lacked metadata, dictionary labels, and semantic information, but was well-received by users. However, a user study identified priorities for improvement, which were addressed in the upgrade funded by the Slovenian Ministry of Culture in 2021-2022. The project aimed to upgrade the dictionary interface design, establish protocols for labeling negative vocabulary, pilot the automatic extraction of antonyms, and supplement the dictionary with semantic indicators for 2,000 entries. This paper presents the upgraded Thesaurus, the methodology for each enhancement, and the challenges and solutions of lexicographic work. The Thesaurus serves as an example of lexical data reuse, interconnectivity, and user involvement, with insights useful for other language communities pursuing similar initiatives.

Keywords: Thesaurus of Modern Slovene; responsive dictionary; automated lexicography; user involvement; post-editing lexicography

1. Introduction

Thesaurus of Modern Slovene, first published in 2018,[1] introduced the concept of a responsive dictionary: a digitally-born, automatically created language resource that provides fast access to open data on modern language use and is gradually improved through editing, which involves both lexicographic work and user participation (Arhar Holdt et al., 2018: 404). The most defining characteristic of the responsive model is its ability to quickly and flexibly respond to both language change and the feedback provided by the community: in the case of the Thesaurus, users can contribute by

---

[1] Thesaurus of Modern Slovene 1.0 is available in the interface at https://viri.cjvt.si/kolokacije/eng/ and as a database at http://hdl.handle.net/11356/1166. Thesaurus 2.0 is available in beta version: https://viri.cjvt.si/sopomenke-beta/slv/.

adding suggestions of missing synonyms and by up- or downvoting existing synonym candidates.

Thesaurus of Modern Slovene 1.0 consists of 105,473 keywords and 368,117 synonyms. It was automatically generated using pre-existing resources: the Oxford®-DZS Comprehensive English-Slovenian Dictionary and the Gigafida reference corpus of written Slovene (Logar et al., 2012). The extraction of data relied on the co-occurrence of words in translation strings of the Oxford-DZS Dictionary. The next step utilized a method that combined balanced co-occurrence graphs and the Personal PageRank algorithm to divide synonyms into subgroups and rank them based on their degree of semantic relatedness (Krek et al., 2017).

The data published in Thesaurus 1.0 was not lexicographically post-processed. The entries and synonym candidates were presented in a form of lemmata (without part-of-speech or other metadata that would help disambiguate between forms), semantic descriptions were replaced by automatically obtained semantic clusters, and the data also lacked dictionary labels, apart from domain ones. Despite these limitations, the community found the new resource and the concept of a responsive dictionary useful (Arhar Holdt 2020: 470), and statistics show the consistent widespread use of the Thesaurus ever since it was published.

However, continuous development is an integral part of the responsive model, and the aforementioned user study also identified priorities for the first upgrade. The upgrade was funded by the Slovenian Ministry of Culture in 2021–2022 and included upgrading the dictionary interface design; ensuring transparent editorial protocols for evaluating user suggestions; piloting the automatic extraction of antonyms and facilitating crowdsourcing of antonyms through the dictionary interface; adding dictionary labels for extremely offensive (hateful) and vulgar vocabulary and allowing users to also provide dictionary labels when contributing synonyms and antonyms; and finally, supplementing the dictionary database with the description of sense distribution including short definitions of senses known as semantic indicators for 2,000 entries. In the following sections, we describe the database and interface improvements and conclude with plans for future dictionary development.

2. Database Improvements

2.1 Data Cleaning and Import into the Digital Dictionary Database

The first step of the project was to import the data from the Thesaurus of Modern Slovene database into the Slovene Digital Dictionary Database (Kosem et al., 2021a), which would allow for the interlinking, easier editing, and optimized reusability of lexical information. To achieve this goal, we had to undertake a series of technical and editing procedures on the Thesaurus data.

Firstly, we extracted synonym pairs containing domain dictionary labels, which were then reviewed, corrected, or upgraded to correspond to the labeling system used in the Digital Dictionary Database. Secondly, we used the results of previously conducted crowdsourcing campaigns aimed at removing noise from the database (Čibej & Arhar Holdt, 2019). We removed 8,878 problematic entries, such as noisy and redundant multi-word units (e.g. *zeleni pas – zeleni pas med vozišči,* 'green belt – green belt between the lanes'; *akrobat na vrvi – plesalka na vrvi,* 'a male tightrope acrobat – a female tightrope dancer').

Lastly, we addressed the issue of headwords and synonyms not containing part-of-speech information in Thesaurus 1.0. This led to homonymous headwords with synonyms placed together. We disambiguated such cases and semantically separated the synonyms (4,560 units in total) accordingly. For instance, the adverb *blago – zmerno, nežno, rahlo* ('mildly – moderately, gently, slightly') vs. the noun *blago – tekstil, material* ('fabric – textile, material').

2.2 Semantic Analysis and Sense Division

We selected 2,000 headwords by merging the headword lists from the Thesaurus of Modern Slovene 1.0 (Krek et al., 2018), the Collocation Dictionary of Modern Slovene 1.0 (Kosem et al., 2019), and the Comprehensive Slovenian-Hungarian Dictionary 1.0 (Kosem et al., 2021b), considering relevant parameters such as part-of-speech categories, single or multiple senses, dictionary labels, and potentially offensive vocabulary. We equipped the headwords with semantic indicators, [2] primarily sourced from the Comprehensive Slovenian-Hungarian Dictionary and supplemented with newly prepared indicators. The synonym candidates for these headwords were then attributed to the corresponding senses. Our original plan was for lexicographers to use the localized and adapted version of Lexonomy,[3] but the testing phase revealed slow data classification and challenging workflow management. As a result, we revised the process by exporting data from the dictionary database to a tabular form (Google Sheets), editing and validating the data, and finally importing it back into the database.

In the process of our work, we developed guidelines for the classification of synonym candidates under the appropriate senses. The guidelines provided relevant information

---

[2] Semantic indicator is one of the three segments of semantic information included in the CJVT dictionary resources. Along with labels and explanations, the semantic indicator aims to define the meaning of a word concisely and clearly in relation to its other meanings. The primary purpose of the indicators is to create a sense menu, a feature introduced in CJVT dictionary resources such as Collocations 1.0 (https://viri.cjvt.si/kolokacije/eng/) and Comprehensive Slovenian-Hungarian Dictionary 1.0 (https://viri.cjvt.si/slovensko-madzarski/eng/).

[3] https://lexonomy.cjvt.si/

for the classification process, including the presentation of the data and the main steps for classifying synonymous material and verifying synonymy.

To ensure the accuracy of the classification, we checked the usage of words in various corpora such as the reference corpus of written Slovene Gigafida 2.0 (Krek et al., 2020), the Monitor corpus of Slovene Trendi (Kosem et al., 2022), the JANES corpus of Slovene user-generated content version 1.0 (Fišer, 2020), and KAS corpus of academic Slovene 2.0 (Žagar et al., 2021). In some difficult cases, we also investigated collocate overlap among the synonyms to help determine which senses they should be attributed to. For this task, we used the Collocation Dictionary of Modern Slovene 1.0, Thesaurus of Modern Slovene 1.0, Gigafida 2.0, and the Sketch Engine tool's Sketch Diff function (Kilgarriff et al., 2014) (see Section 2.5). An example of a headword with distributed synonyms is presented in Table 1.

Headword	Senses[4]	Synonyms
hiteti (glagol)	[1: pri dejavnosti]	pohiteti, brzeti, drveti, dirjati
	[2: o premikanju]	brzeti, drveti, leteti, dirjati, teči, divjati, hitro hoditi, rezati jo, drobencljati, drobencati, planiti, vrveti, dreti, ubirati jo, poditi se, sukati se, švigniti
	[3: minevati]	brzeti, drveti, leteti, hitro minevati, teči

Table 1: Classification of synonym candidates according to the senses of the headword *hiteti* ('to hurry').

If a word's usage could not be confirmed in our resources, we did not consider it a synonym. False candidates appeared in our data due to the methodology we used for the creation of the Thesaurus, where we exported synonym candidates from the Oxford®-DZS Comprehensive English-Slovenian Dictionary (see Section 1). For example, for the headword *brat* ('a brother'), we included synonyms, such as [1: sorodnik, 'a relative']: *bratec* ('a little brother'); [2: pripadnik skupine, 'a group member']: *prijatelj* ('a friend'); but not *kolega, sodelavec* ('colleague, coworker'), as the word *brat* is not used in this sense in Slovene.

---

[4] The presented semantic indicators and synonyms for the verb 'to hurry' are roughly equivalent to the English concepts of [1: activity]: to hasten, to accelerate, to race, to rush [2: movement]: to accelerate, to rush, to fly, to dash, to run, to rampage, to walk quickly, to stride, to scurry, to scamper, to leap, to throng, to swarn, to run around, to twist, to dart, and [3: time]: to accelerate, to rush, to fly, to pass quickly, to run.

Anticipatedly, we encountered numerous borderline candidates, and in such cases, we chose to prioritize inclusion over exclusion. Our decision was based on the expectation that future projects would involve further data cleaning and editing. Additionally, our responsive dictionary concept aims to provide users with as much data as possible, facilitating a vast choice of options. Hence, we also included certain candidates with semantic proximity or similarity, for example, *juha – enoločnica, obara* ('soup – stew, casserole') or *kavč – divan, zofa, postelja* ('couch – loveseat, sofa, bed').

During the classification process, we identified opportunities to propose changes to the existing semantic classification. This included suggesting an additional sense, dividing an existing sense, combining two senses, or changing the semantic indicator. In some cases, the semantic indicators themselves contained one of the synonyms, resulting in repetition within the string, e.g. *besneti* = 'to be furious' [1: jeziti se, 'to get angry']: *peniti se, divjati, jeziti se, kipeti, vreti* ('to foam, to rage, to get angry, to boil, to seethe'). We marked these cases, which enabled us to review them after the project and enhance the creation of semantic indicators in the future.

2.3 Dictionary Labels

In Thesaurus 1.0, headwords and their synonyms lacked any explicit information on usage, stylistic and pragmatic value, except for a limited set of domain labels. A survey conducted with 671 respondents revealed that more than one-third of users (37%) found the absence of dictionary labels problematic (Arhar Holdt 2020: 472). There were two main issues with the absence of labels. Firstly, automatically generated headwords and synonym candidates appeared without labels or usage warnings even in highly problematic cases, such as the word *buzi* with synonyms *peder, buzerant, toplovodar, homič,* all of which are derogatory expressions for 'a gay man'. Secondly, users added marked vocabulary as legitimate synonyms, such as for the marked word *južnjak* ('southerner'), where users suggested similarly marked words like *jugovič, južni brat, jugič, trenirkar*. In some of these cases, users even added a note on usage next to the proposed synonym candidate. Therefore, incorporating a labeling system into the dictionary became a top priority, and it also made sense to upgrade the interface and provide users with the option to label their suggestions in a more systematic way.

We based our labeling system on the dictionary style guide used for developing lexicographic resources in the Digital Dictionary Database at the CJVT, for instance, the Comprehensive Slovenian-Hungarian Dictionary (Kosem et al., 2021b). Negative vocabulary is labeled with three distinctive labels (Table 2) used for hate speech elements (labeled as *hateful*), elements of rudeness and offensiveness (*coarse*), and elements of negative evaluation or connotation (*expresses a negative attitude*). Each label is depicted in the interface with an icon and accompanied by an explanation of the potential impact the use of the labeled word can have (see Section 3).

The labels were assigned manually to the headwords and their synonyms during the lexicographic process, taking into account the distribution of senses and their descriptions, as described in Section 2.2. Arhar Holdt et al. (2023) provide further examples and a detailed explanation of the labeling choices.

Label	Icon	Explanation
hateful	⚠	This word can be used to express a hostile or intolerant attitude towards an individual or social group.
coarse	⚠	This word can seem coarse or inappropriate to many language users due to social and moral norms. Using the word can make people feel uncomfortable, upset, or offended.
expresses a negative attitude	⊙	This word may not be neutral. The word can be used to ridicule, express disapproval, or criticize certain characteristics of individuals, objects, or actions.

Table 2: Labels for negative vocabulary, icons in Thesaurus 2.0, and their explanations.

2.4 Automatic Extraction and Selection of Antonyms

We prepared a prototype sense-antonym extraction methodology based on machine learning using word embeddings and large pre-trained language models. We decided to test using word-sense information in order to prepare antonyms separately for each sense. We started with candidate antonyms without sense information obtained from lexical sources. Our methodology is composed of three approaches. In the first approach, we construct a sense-antonym dataset and cluster a set of contextual embeddings (for words in contexts) to produce sense-clusters; we then assign candidate sense-antonyms to their nearest clusters. In the second approach, meant to obtain antonyms without prespecified antonym-candidate pairs, we first fine-tune a large language model using a dataset of antonyms in context to predict if two words in a context (i.e. a sentence), are antonyms. The third approach is a traditional lexical approach based on a dictionary and WordNet.

For the clustering-based approach, we first constructed a dataset of sense-antonyms used in sentences. For that purpose, we used 2,852 antonym candidates without sense information and extracted examples of their use from the Comprehensive Slovenian-Hungarian Dictionary (Kosem et al., 2021b). We formed the contextual embeddings for each candidate word by concatenating the last four layers of the CroSloEngual BERT model (Ulčar & Robnik-Šikonja, 2020) as recommended by Devlin et al. (2019). To produce the pairs of sense-antonyms on the resulting 3072-dimensional vectors, we used

the k-nearest neighbor method which, for a given word, found the nearest sense-cluster. A portion of candidate pairs was manually checked.

For the classification approach, we first constructed a dataset of antonyms in context, similarly to the process described above. The dataset was split into a training, validation, and test set. We fine-tuned the CroSloEngual BERT language model on the training set. The model achieved 90% precision and 61% recall on the test set. We published the datasets, containing a total of 79,229 records with information about word pairs in specific senses, as an open-access collection on the CLARIN.SI repository (Pegan et al., 2022). Only a fraction of this dataset was manually checked.

To complement the machine learning approaches, we tried a lexical approach producing 4,734 antonym candidates using a bilingual dictionary (English-Slovene) and antonym information available in English Wordnet. Each candidate pair was accompanied by its part-of-speech tags.

All three approaches showed promising results and will undergo further evaluation. In Thesaurus 2.0, we have included only the 2,544 antonym pairs that were manually checked.

2.5 Comparing Collocates of Synonyms

One of the more innovative features of the first version of the Thesaurus of Modern Slovene was the comparison of collocations of the headword and its synonyms. The comparison was made by using the Sketch Diff function in the Sketch Engine tool (Kilgarriff et al., 2014), with specific parameters set for extraction, such as minimum frequency and grammatical relations used for different word classes. The comparison was available only for single-word headwords and single-word synonyms, and only when both compared items were frequent enough to produce the Sketch Diff output. The collocation comparison contained three different sections: joint (collocates typical for both the headword and the synonym), and two individual (collocates used with the headword or the synonym only). Each section included up to 20 collocates; four columns with up to five collocates each. In general, one column per grammatical relation was used, and only the most productive relations were presented in two columns; for example, for the comparison of adjectives, the relation *adjective + noun* was allocated two columns (more examples in Arhar Holdt et al. 2018: 403, 408).

In Thesaurus 2.0, both the data and its presentation received an upgrade. This is a direct result of two factors: a) the change in the methodology used for automatic collocation extraction, and b) the findings of the evaluation study of the informative nature of the collocates provided by the selected grammatical relations. The main change in the methodology of collocation extraction was moving it from POS-tagged to parsed corpus data, and consequently considerably improving the reliability of the results (Krek et al., 2021). This also meant that additional syntactic structures could

be considered for the inclusion into collocational comparison. Relatedly, a study (Arhar Holdt, 2021) was conducted on the informative nature and reliability of the collocations in grammatical relations used in version 1.0. The study did not merely provide an evaluation of existing grammatical relations but also included recommendations on which ones should be removed from the comparison feature, and which new ones should replace them. The two most important changes occurred at the level of (a) nouns, where we replaced the pattern *noun + preposition + noun* (e.g. *program za prihodnost*, 'program for the future') with a more productive *noun + noun in genitive* (e. g. *izvedba programa*, 'the implementation of the program'), and (b) adjectives, where we replaced *adjective + preposition + noun* (e.g. *pozitiven za gospodarstvo,* 'positive for the economy') with *adjective + AND + adjective* (e.g. *pozitiven in optimističen,* 'positive and optimistic').

We have also made a change in the way we obtain collocations for the three sections of the comparison. As we wanted to focus more on the typicality of the collocations, the sections became "joint", "more typically used with the headword", and "more typically used with the synonym". In this way, we excluded from the comparison the less typical (and often more infrequent) collocates, which were causing some problems in version 1.0. For instance, a collocation that was highly characteristic of one item, but also appeared with the other item, could only be included in the joint list. However, in most cases, it would be excluded from the final list due to numerous other collocates that showed a similar level of typicality of usage with both items. The visualization of collocates in the Thesaurus 2.0 is presented in Section 3.

2.6 Editorial Protocols for Evaluating User-suggested Synonyms

As previously noted, users can suggest synonyms and, as of version 2.0, antonyms to the Thesaurus of Modern Slovene. This feature enables dictionary users to actively participate in the development of openly available language infrastructure in a democratic manner. They can contribute to the expansion and refinement of the Thesaurus, helping to make it more comprehensive.

To encourage participation, user suggestions are displayed in the interface immediately after they are submitted, along with the user's chosen username. At this stage, the suggestions are not subject to editorial evaluation but are visually distinguished from the rest of the synonyms and not automatically added to the openly accessible database. Instead, the evaluation process occurs during dictionary upgrades, where lexicographers carefully consider the user suggestions according to editorial guidelines. This approach ensures that user input is taken into account while also maintaining the consistency and quality of the database.

To develop the guidelines, 972 user-suggested synonym pairs were evaluated by a team of six lexicographers and classified as suitable, unsuitable, or conditionally suitable for inclusion in the database. The evaluation was complemented by a larger study that

involved selected user groups (e.g. teachers, translators, and proofreaders) performing the same task (Gapsa, 2023). This study provided valuable insight into the preferences of users compared to lexicographers and the differences between the conditions for synonym selection reported by different groups (Gapsa & Arhar Holdt, 2023).

Based on our analysis, we have developed a protocol for evaluating user suggestions. Firstly, we check whether the proposed word or phrase appears in authentic language use by consulting the corpora already mentioned in Section 2.2. Secondly, we assess whether the suggestion is appropriately categorized under the relevant headword, which we determine by studying the use of the word in resources. Thirdly, we consider any proposals for dictionary labels made by users and add the appropriate label if needed. Finally, we consider feedback on the suggestion from other users, based on their upvotes and downvotes. If there is any uncertainty about the suitability of a user-suggested synonym for inclusion in the database, we err on the side of caution and do not include it. It is worth noting that even if a proposal is not included in the database, it remains accessible via the dictionary interface. We would only remove entries from the interface in rare cases when they are deemed malicious (see also Section 2.3).

Currently, there are 60,976 user-suggested synonyms awaiting evaluation for potential inclusion in the Thesaurus. The implementation of the editorial guidelines will be part of the next edition, and the guidelines will be continuously improved in the process. In the meantime, we have upgraded the interface in version 2.0 to enable users to contribute dictionary labels to proposed synonyms and antonyms. Our evaluation process has shown that many user suggestions include regionally specific, slang, or jargon terms that could benefit from a label, both to facilitate the editorial process and to assist other users in understanding the terms.

3. Interface Improvements

In addition to enhancing the content of the Thesaurus, we also placed a significant emphasis on developing the user interface. The designer has created a library of redesigned interface elements, which has enabled all elements within CJVT dictionary resources to share a consistent visual appearance and logical structure, including colors, icons, typography, and element formatting such as search, toggle, share, and user engagement. Additionally, the implementation of a responsive font size provides improved accessibility for users who may need to adjust zoom levels for different reasons. Working closely with the team, the designer has also introduced a new design for the dictionary headers, footer, and "About" section, as well as interface features not present in version 1.0, including antonyms, sense-separated entries, icons for negative vocabulary, and user mechanisms for adding dictionary labels and including suggested synonyms and antonyms under the relevant sense.

Thesaurus of Modern Slovene 2.0 offers two different entry layouts. Figure 1[5] depicts the layout of the automatically generated entry, while Figure 2 shows an entry with sense division and manually arranged synonyms (Section 2.2). In both layouts, new metadata has been included: the part-of-speech label, an indicator of the headword's frequency in the reference corpus Gigafida 2.0, and an indicator of the entry layout type - either automatically generated or sense-divided. Both layouts also feature a tab for antonyms, although only a limited number of headwords currently have antonyms (Section 2.4).

The layout of the automatically generated entry for the word *jedrnat* ('concise') is depicted in Figure 1. The extracted synonym candidates are presented in two sections: core synonyms that are semantically closer are on a white background, while less related near synonyms are on a grey background. The entry also includes a section for user-suggested synonyms, with the option to add more suggestions. The image shows one such user suggestion: *kratek in jedrnat* ('short and concise'), which was added in version 1.0. The Antonym section currently lists one antonym, *dolgovezen* ('verbose'). Finally, the section for user-added antonyms is currently empty.
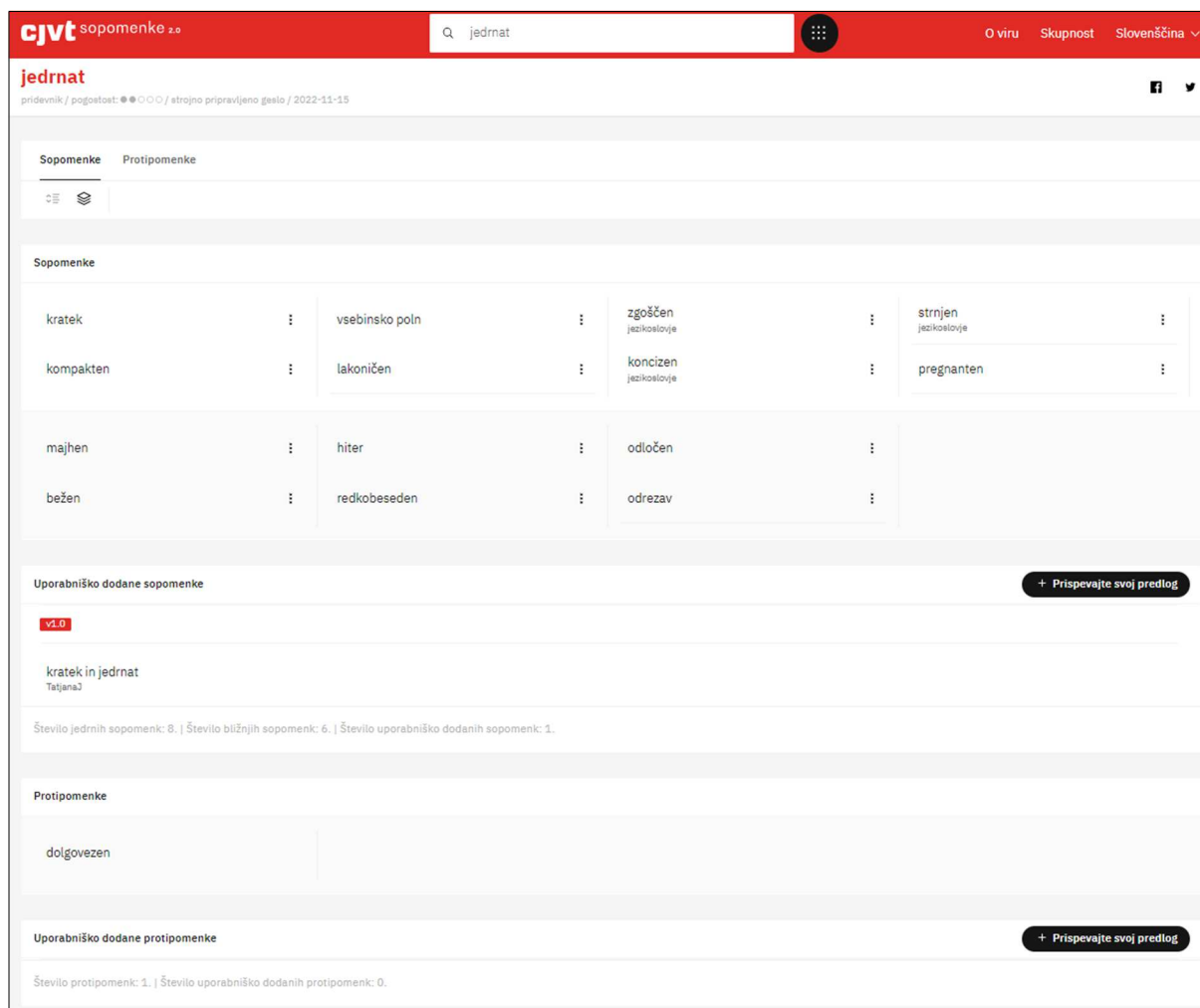
Figure 1: The layout of the automatically generated entry *jedrnat* ('concise').

Figure 2 presents the layout of the sense-divided entry for the word *baba* ('a broad'). This word can be used to refer to a woman and express either a negative or a positive attitude, or to refer to a man and express a negative attitude in the sense of 'a coward'. We chose this example as it includes both the icon for a coarse word (*lajdra,* 'a slut') and a hateful word (*peder,* 'a faggot'). Clicking on the icon opens a longer explanation of the potential impact that the use of the labeled word can have (see Table 2).

As mentioned in Sections 2.3 and 2.6, we have upgraded the user-suggestion protocol to allow users to add dictionary labels to their proposed words or phrases. In the sense-divided entries, they can also attribute the suggestion to a corresponding sense (Figure 3). The default option is for the user's suggestion to be *without label*, while other options are available in the drop-down menu. In Thesaurus 2.0, labels for *hateful, coarse,* and *expresses a negative attitude* are available upon clicking, along with a box where users can type in any other possible label. The meaning and usage of these labels are explained and illustrated with examples, which will help achieve a certain level of consistency in user labeling (information is available upon clicking the icon (i) in Figure

3). It is expected that users will interpret and use these labels differently than lexicographers would in some cases. User suggestions will thus be valuable not only for supplementing the open-access dictionary database but also for the analyses of the perception of the labeling system.

Figure 2: The layout of the sense-divided entry *baba* ('a broad').

Figure 3: Adding potentially labeled synonyms for the noun *brat* ('a brother').

From the initial layout, users can click selected synonyms to open the collocate comparison view. Figure 4 presents this view for the pair *program* and *plan* ('a program - a plan'). As mentioned in Section 2.5, firstly the collocates that appear with both words are presented, followed by collocates that typically occur with only one of the words. For example, *razvojni načrt* and *razvojni program* ('development plan, development program') are both typical collocations, while *kulturni, nacionalni, študijski* tend to collocate with *program* and *prostorski, poslovni, lokacijski* with *načrt* ('cultural, national, study program' and 'spatial, business, location plan').

Figure 4: Collocate comparison for *program* and *plan* ('a program - a plan').

4. Conclusion and Future Work

In this paper, we presented an upgraded version of the Thesaurus of Modern Slovene, which was developed to address the lack of openly available synonym data for modern Slovene. Our work serves as a benchmark for other languages that face similar issues, demonstrating the importance of data reusability and user involvement in language infrastructure development. To address contemporary needs, such as the desire of dictionary users to participate in the development of the language infrastructure, we integrated machine processes and user suggestions into our workflows. This integration

allows for the incorporation of user feedback, ensuring that openly accessible language data is readily available.

The upgraded version of the Thesaurus addressed some of the most significant shortcomings of the previous version. While the project had limitations in scope and not all of the database could be manually edited, Thesaurus 2.0 sets a clear direction for future development. The newly established guidelines for preparing sense divisions, labeling negative vocabulary, and evaluating user suggestions provide a solid foundation for the expansion of the database. In version 3.0, our attention will remain focused on the automatic extraction of antonyms, which has displayed promising results and requires further evaluation and more extensive implementation. Additionally, more detailed work is planned for the selection and visualization of collocations. With the development of the Collocation Dictionary of Modern Slovene (Kosem et al., 2018), the data is improving, and new possibilities for utilization will soon be available.

5. Acknowledgements

6. References

Arhar Holdt. Š. (2023, in press). Negativno zaznamovano besedišče v Slovarju sopomenk sodobne slovenščine 2.0. *Slovenščina 2.0.*

Arhar Holdt, Š. (2021). Kolokacije v Slovarju sopomenk sodobne slovenščine: evalvacija podatkov in predlog za izboljšavo. In I. Kosem (ed.) *Kolokacije v slovenščini.* Ljubljana: Znanstvena založba Filozofske fakultete, pp. 269–296. Available at: https://ebooks.uni-lj.si/zalozbaul//catalog/view/318/465/6967-1

Arhar Holdt, Š. (2020). How users responded to a responsive dictionary: the case of the Thesaurus of Modern Slovene. Rasprave Instituta za hrvatski jezik i jezikoslovlje, 46(2), pp. 465–482. doi: 10.31724/rihjj.46.2.1

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

Devlin, J., Chang, M.W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep

Bidirectional Transformers for Language Understanding. In J. Burstein, C. Doran & T. Solorio (eds.) *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* Minneapolis, Minnesota. Association for Computational Linguistics, pp. 4171–4186. Available at: https://aclanthology.org/N19-1.pdf

Čibej, J. & Arhar Holdt, Š. (2019). Repel the syntruders! A crowdsourcing cleanup of the thesaurus of modern Slovene. A crowdsourcing cleanup of the thesaurus of modern Slovene. In I. Kosem et al. (eds.) *Proceedings of the eLex 2019 conference, Electronic lexicography in the 21st century: Smart lexicography.* Sintra, Portugal. Brno: Lexical Computing, pp. 338–356. Available at: https://elex.link/elex2019/wp-content/uploads/2019/09/eLex_2019_19.pdf

Fišer, D., Ljubešić, N. & Erjavec, T. (2020). The Janes project: language resources and tools for Slovene user generated content. *Language resources and evaluation,* 54(1), pp. 223–246. doi: 10.1007/s10579-018-9425-z

Gapsa, M. (2023, preprint). "But why??" Evaluation of user-suggested synonyms in the Thesaurus of Modern Slovene. Research Square. doi: 10.21203/rs.3.rs-2775161/v1

Gapsa, M. & Arhar Holdt, Š. (2023). How Lexicographers Evaluate User Contributions in The Thesaurus of Modern Slovene in Comparison to Dictionary Users. In *Electronic lexicography in the 21st century. Invisible Lexicography. Proceedings of the eLex 2023 conference.* Brno, Czech Republic.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography,* 1(1), pp. 7–36. doi: 10.1007/s40607-014-0009-9

Kosem, I., Čibej, J., Dobrovoljc, K., Erjavec, T., Ljubešić, N., Ponikvar, P., Šinkec, M. & Krek, S. (2022). *Monitor corpus of Slovene Trendi 2022-10,* Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1681.

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

Kosem, I., Bálint Čeh, J., Ponikvar, P., Zaranšek, P., Kamenšek, U., Koša, P., Gróf, A., Böröcz, N., Harmat Császár, J., Szíjártó, I., Šantak, B., Gantar, P., Krek, S., Roblek, R., Zgaga, K., Logar, U., Pori, E., Arhar Holdt, Š. & Gorjanc, V. (2021b). *Comprehensive Slovenian-Hungarian Dictionary 1.0,* Slovenian language resource repository, CLARIN.SI, http://hdl.handle.net/11356/1453.

Kosem, I., Gantar, P., Krek, S., Arhar Holdt, Š., Čibej, J., Laskowski, C., Pori, E., Klemenc, B., Dobrovoljc, K., Gorjanc, V. & Ljubešić, N. (2019). *Collocations Dictionary of Modern Slovene KSSS 1.0,* Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1250.

Kosem, I., Krek, S., Gantar, P., Arhar Holdt, Š., Čibej, J., Laskowski, C. A. et al. (2018). Collocations dictionary of modern Slovene. In J. Čibej et al. (eds.)

Proceedings of the 18th EURALEX International Congress: lexicography in global contexts. Ljubljana: Ljubljana University Press, Faculty of Arts, pp. 989–997. Available at: https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/118/211/3000-1

Krek, S., Gantar, P., Kosem, I. & Dobrovoljc, K. (2021). Opis modela za pridobivanje in strukturiranje kolokacijskih podatkov iz korpusa. In Š. Arhar Holdt (ed.) *Nova slovnica sodobne standardne slovenščine.* Ljubljana: Znanstvena založba Filozofske fakultete, pp. 160–194. Available at: https://ebooks.uni-lj.si/zalozbaul//catalog/download/325/477/7320-1?inline=1

Krek, S., Arhar Holdt, Š., Erjavec, T., Čibej, J., Repar, A., Gantar, P., 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, Palais du Pharo, Marseille, France.* Paris, ELRA - European Language Resources Association, pp. 3340-3345. Available at: http://www.lrec-conf.org/proceedings/lrec2020/LREC-2020.pdf

Krek, S., Laskowski, C., Robnik Šikonja, M., Kosem, I., Arhar Holdt, Š., Gantar, P., Čibej, J., Gorjanc, V., Klemenc, B. & Dobrovoljc, K. (2018). *Thesaurus of Modern Slovene 1.0.*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1166.

Krek, S., Laskowski, C. & Robnik Šikonja, M. (2017). From translation equivalents to synonyms: creation of a Slovene thesaurus using word co-occurrence network analysis. In I. Kosem et al. (eds.) *Proceedings of eLex 2017: Lexicography from Scratch.* Leiden, Netherlands, pp. 93–109. Available at: https://elex.link/elex2017/wp-content/uploads/2017/09/paper05.pdf.

Logar Berginc, N., Grčar, M., Brakus, M., Erjavec, T., Arhar Holdt, Š. & Krek, S. (2012). *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba.* Ljubljana: Trojina, zavod za uporabno slovenistiko; Fakulteta za družbene vede. doi: 10.4312/9789610603542

Pegan, J., Robnik-Šikonja, M., Kosem, I., Gantar, P., Ponikvar, P. & Laskowski, C. (2022). *Slovenian datasets for contextual synonym and antonym detection*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1694.

Ulčar, M. & Robnik-Šikonja, M. (2020). FinEst BERT and CroSloEngual BERT: less is more in multilingual models. In P. Sojka (ed.) *Proceedings of Text, Speech, and Dialogue: 23rd International Conference.* Brno, Czech Republic. Cham: Springer, pp. 104–111. doi: 10.1007/978-3-030-58323-1_11

Žagar, A., Kavaš, M., Robnik Šikonja, M. (2021). Corpus KAS 2.0: Cleaner and with New Datasets. In *Proceedings of the 24th International Multiconference – IS2021 (Slovenian Conference on Artificial Intelligence).* Available at: https://doi.org/10.5281/zenodo.5562228