Collocations Dictionary of Modern Slovene 2.0

Iztok Kosem[1,2], Špela Arhar Holdt[1], Polona Gantar[1], Simon Krek[1,2]

[1] Faculty of Arts, University of Ljubljana, Slovenia
[2] Jožef Stefan Institute, Slovenia
E-mail: Iztok.Kosem@ff.uni-lj.si, Spela.ArharHoldt@ff.uni-lj.si,
Apolonija.Gantar@ff.uni-lj.si, simon.krek@ijs.si

Abstract

In this paper, we present the Collocations Dictionary of Modern Slovene 2.0, which is a substantial upgrade of the first version, both in terms of content and the interface. The Colocations Dictionary contains 81,445 headwords, nearly 4.5 million collocations, and more than 17 million examples. Relevant findings of user studies and other related research, as well as the development of new methodology for automatic extraction of collocations from corpora, which is based on the syntactically parsed corpus data, have been used to improve the contents of the dictionary. The interface has undergone some important changes such as the immediate view of all the collocations in the entry, and the easy-to-understand three levels of entry completion. In terms of the data storage, a crucial development has been the introduction of the combination of the Digital Dictionary Database, which allows sharing the data among various resources produced at the Centre for Language Resources and Technologies at the University of Ljubljana, and a data warehouse, where all the automatically extracted collocations and additional metadata are stored.

Keywords: collocations dictionary; responsive dictionary; crowdsourcing; examples; post-editing lexicography

1. Introduction

In 2018, the first version of the Collocations Dictionary of Modern Slovene was published (Kosem et al., 2018).[1] The dictionary contained automatically extracted collocations, and their examples, using (at that point) state-of-the-art tools such as Sketch Grammar and GDEX, customised for Slovene (Gantar et al. 2016). A selection of entries was provided in the finalised form, using post-editing methodology.

Over the past four years, a great deal of research related to the Collocations dictionary and the phenomenon of collocations in Slovene has been conducted, from the analysis and improvement of automatic extraction methods, lexicographic workflow, and data modelling, to user experience and participation. A project named Upgrading

[1] Collocations Dictionary of Modern Slovene 1.0 is available as a database at http://hdl.handle.net/11356/1250.

fundamental dictionary resources and databases of CJVT UL funded by the Slovene Ministry of Culture in 2021-22 provided the opportunity to implement the improved methods and new solutions into the next version of the Collocations Dictionary.

In this paper, we first present the developments since the launch of version 1.0 of the Collocations Dictionary of Modern Slovene. These developments include the results of various studies with the users of the dictionary and the improvement of collocation extraction methods, as well as the relevance of the latest trends in data storage and resource linking. Then, we look in detail at the new features of version 2.0 of the Collocations Dictionary, including the data extraction (and selection) method, and the inclusion of collocational data into the Digital Dictionary Database for Slovene. Furthermore, we also take a closer look at the changes in the interface, especially in terms of data visualisation and user participation, i.e., the crowdsourcing module. We conclude the paper with a short outline of future plans, both short-term and long-term.

2. Collocations dictionaries

The importance of collocation has been known since Firth's (1957: 11) famous statement "You shall know the word by the company it keeps", and the phenomenon has been analysed in detail since the arrival of large corpora. However, the compilation of collocation dictionaries for languages other than English, and especially the systematic inclusion of collocational information in general language dictionaries is a more recent trend. There are numerous collocations dictionary projects, either completed or ongoing, and we focus on those that have influenced the further development of the Collocations Dictionary of Modern Slovene. The first one to mention is the Estonian Collocations Dictionary (Kallas et al., 2015) which was compiled using the same methodology as we have been using in the compilation of the Collocations Dictionary for Modern Slovene, namely post-editing lexicography (curation of automatically extracted data). The Estonian Collocations Dictionary does differ in certain characteristics, for example, it was aimed at non-native speakers of Estonian, offers definitions only for polysemous words etc. The Estonian Collocations Dictionary is no longer available as a standalone source, as it has been integrated into the EKI Combined Dictionary.[2]

Similar to the Estonian Collocations Dictionary in terms of target audience is Woordcombinaties (Colman and Tiberius, 2018), a Dutch Collocations Dictionary. This is an ongoing project, which is in the process of switching to post-editing methodology, i.e., the selection of collocations is still done manually from the Sketch Engine corpus tool. Currently, the main focus of the dictionary are verbs. The users can choose from three different views: collocations (divided by syntactic structures), examples of use, and patterns (based on the Corpus Pattern Analysis by Hanks, 2004).

[2] https://sonaveeb.ee/

Targeted at native speakers such as the Collocations Dictionary of Modern Slovene is Croatian Web Dictionary – Mrežnik (Hudeček & Mihaljević, 2020a),[3] currently available in a demo version (letters A-F). Mrežnik is a general language dictionary with a significant section in each entry dedicated to collocations (Hudeček and Mihaljević, 2020b). Collocations are divided into blocks introduced by collocational questions and phrases, modelled after the elexiko project (Haß, 2005; Storjohann, 2005; Klosa, 2015). Methodologically, Mrežnik is more similar to the Woordcombinaties, using a combination of manual insertion of collocations into the dictionary-writing system TLex from the Sketch Engine tool (Hudeček & Mihaljević, 2020b).

The reports by the authors of the abovementioned projects, as well as of other similar projects, point to several common issues of using collocations for dictionary purposes. One of the main ones is the abundance of data, both good and bad. While examining (long) lists of collocation candidates, the lexicographers need to identify the good ones, discard the bad ones, and then also often make a further selection among the good ones. This is far from straightforward; while some bad collocation candidates can be immediately identified, others can be confirmed as bad only after examining corpus examples. Similarly, there are levels of good collocation candidates; cut-off points need to be made not only in terms of how much data the lexicographers need to analyse but also how many collocations one wishes to present to the users. In this respect, it is also crucial to have the criteria for what constitutes a collocation, and what is its relation to other multi-word units, clearly delineated from the onset. The approach we used is described in Kosem et al. (2019) and Gantar et al. (2019).

A related issue is the origin of corpus data and the quality of annotation, which affects the quality of collocation candidates. The origin of bad collocation candidates can often be attributed to the problematic contents of the corpus (e.g., machine-translated texts from the web, Koppel et al., 2019) or errors in lemmatisation, part-of-speech tagging or parsing (Koppel et al., 2019; Pori and Kosem, 2021).

Another challenge is the data model, i.e., where and how is the collocational data stored, which lexicographic decisions are stored (only good candidates or also bad), how are the latest changes in the language monitored and incorporated into the existing data etc. The approach of editing data directly in relational databases where the data can be shared across headwords (i.e., lexical items) is being used by an increasing number of institutions, however editing the dictionary in the XML format still seems to dominate (Tiberius et al., 2022: 9).

Even after addressing all these issues and publishing the dictionary, there is one other aspect to consider, namely the dictionary user. In the next section, we present the findings of the studies conducted among the users of the Collocations Dictionary of

[3] https://rjecnik.hr/mreznik/

Modern Slovene, as well as other relevant research on the use and consultation of collocations.

2.1 User studies

The most influential for the development of the second version of the Collocations Dictionary was the study by Pori et al. (2020; 2021), which investigated the attitudes of four different groups of users (teachers of Slovene as L1, teachers of Slovene as L2, proofreaders and translators, lexicographers) towards the Collocations Dictionary, and the way in which they used the dictionary. Using the evaluation interview based on the guided think-aloud method, the users were asked to conduct random searches of their own choice, conduct pre-determined searches, and comment on the general usefulness of the dictionary and its look. The most important findings can be summarized as follows:

- the attitudes towards the inclusion of automatic collocations were overwhelmingly positive, under the condition that the users are provided with corpus examples for context and a clear warning about the nature of such data (this being particularly stressed by language teachers).

- the pyramid icon indicating the level of entry completeness was considered by many to be not noticeable enough, the information it conveys should have been presented more clearly.

- the dictionary interface was evaluated as very good, all the features were found to be very useful and easy to use. An often-mentioned suggestion was the use of clear headlines or descriptions instead of icons, or at least adding descriptions of icons.

- while initially showing a selection of the top four most salient collocations of each syntactic structure, and having all the collocations in the structure available on a click was considered useful by participants, there were some doubts over whether most of the users ever get to the additional content. This can be considered problematic given that corpus examples are only provided at the stage of seeing all the collocations.

- the links to the corpus were considered very important, crucial even.

- some users wanted additional information on collocations, for example the information on frequency or saliency.

- the crowdsourcing part was considered useful by some participants, especially proof-readers and translators, although they usually lack time to contribute. On the other hand, teachers expressed concerns about the usefulness of the feature if used by less advanced language users.

Another relevant study was conducted by Arhar Holdt (2021) who looked at the preferences (and expectations) of 415 users of the Collocations Dictionary on the ordering of collocations in the dictionary interface. The questionnaire consisted of asking the participants to: list by memory three collocations of a given headword; select the top three syntactic structures they would like to see in the entry; select five collocations among the ones offered for a given headword and order them according to the perceived importance; provide the criteria used for ordering; provide other comments. The findings showed that the user expectations in terms of preferred syntactic structures more or less matched the order of structures provided in the dictionary. On the other hand, the users clearly preferred, and expected, the collocations to be ordered by frequency rather than by saliency; this is in contrast to how the collocations were ordered in the interface of Collocations Dictionary 1.0. Interestingly, other dictionaries are also not unified in this approach: the Estonian Collocations Dictionary orders collocations by frequency, and the Dutch Woordcombinaties, Mrežnik and the Macmillan Collocations Dictionary by alphabetical order.

Relevant to the crowdsourcing aspects of the Collocations Dictionary was the study by Pori and Kosem (2021), which included an experiment with six linguists who voted on the suitability of collocation candidates based on the collocation and its randomly selected example. The possible answers to the question of whether a candidate is a collocation were Yes, No, I don't know. While the main aim was to evaluate the reliability of the automatic extraction method, the study also revealed that one needs to have a clear definition of collocation to be able to decide on its relevance/suitability. Furthermore, in the pilot study, the participants often pointed out that many collocations seem perfectly fine and only a highly skilled person who knows what to look for can spot issues such as collocation not matching the syntactic structure (e.g., "angažirati izvedenca", eng. to hire an expert, found in the syntactic structure verb + noun in genitive whereas it is in fact verb + noun in accusative). One other finding was that often more than one example was needed to be able to validate the collocation.

Valuable experience for crowdsourcing collocations was gained when developing the Game of Words (Arhar Holdt et al., 2021). Testing various game modes showed that for crowdsourcing collocations an implicit, gamification method is much more appropriate than an explicit method. In other words, much better and more reliable results are obtained if the users (players) are not aware they are providing collocational information, for example by listing collocates or distributing them to relevant headwords, as opposed to being asked directly whether something is a collocation or not. Relatedly, we also conducted an experiment where a group of students was asked to assign examples of collocations to relevant senses of selected headwords; the findings proved such a task to be extremely reliable (there was 100 % annotator agreement in over 80 % of cases) for various purposes: determining the understandability of indicators and sense division, indicating whether examples have enough context and indirectly determining their quality/suitability for dictionary purposes, and to some extent confirming the relevance of the collocation (even though this was not the primary goal).

The findings of all these studies provided a point of departure in our planning of the second version of the Collocations Dictionary of Modern Slovene.

3. Collocations Dictionary of Modern Slovene 2.0

The second version of the Collocations Dictionary of Modern Slovene (Kosem et al., 2022)[4] contains 81,445 headwords, nearly 4.5 million collocations, and more than 17 million examples. In comparison with version 1.0, there are more than twice as many headwords (35,989 in version 1.0), but 40% fewer collocations and nearly 50% fewer examples. This is a direct consequence of newly introduced extraction parameters, which is only one of the many changes introduced in version 2.0.

3.1 Data extraction – a new methodology

One of the important methodological differences from the first version of the Collocations Dictionary is the method of automatic extraction, of both collocations and examples. Collocations are entirely new, i.e., they were extracted from syntactically parsed corpus data (Krek et al., 2022; Krek et al., 2021), as opposed to an extraction based on POS-tagged data which was used for the first version. A new formalism defines dependency syntactic relation within a collocation, and also defines "constraints on any level of annotation, from morphology (parts-of-speech and their properties), syntactic dependency relations, concrete lexical items, and any other types of annotation that can be used for other purposes, e. g. semantic roles, semantic types, word senses, etc." (Krek et al., 2022: 241). These constraints can be also used to specify the form of each component found in the corpus to be used in a specific collocation, an option that is very important for storing the collocation in the database as well as its presentation to the users. With a new formalism, we were able to separate verbal structures in terms of negation and reflexiveness, adding more syntactic structures to the list. The total number of syntactic structures is currently 82, and they include collocators belonging to four word classes: nouns, verbs, adjectives and adverbs.

With the new method giving more reliable results, combined with the fact that certain structures excluded from version 1.0 proved to be very important for certain headwords (e.g., the first version did not include 'subject + verb' due to many bad collocation candidates), we decided to include all 82 syntactic structures in the second version. It is important to note that on the one hand, headwords only contain structures which include the headword's part of speech s (e.g., 'noun + preposition + noun in accusative' is found only for nouns), and on the other hand, the number of structures is even higher

[4] The dictionary is available at https://viri.cjvt.si/kolokacije/eng/.

if we take the position into account (e.g., noun headword can be found in the aforementioned structure in the initial or final position). However, this in return meant reducing/limiting the number of collocations per structure to avoid information overload for the users. While the maximum number of collocations per syntactic structure in version 2.0 is 10, more collocations (up to 25) are offered for the structures that proved more collocationally-productive in the research studies (e.g. verb + noun in the accusative, adjective + noun, noun + noun in the genitive).

As far as headwords are concerned, the decision was made to extract collocations for all the nouns (excluding proper nouns), adjectives, adverbs and verbs in the Slovene Digital Dictionary Database (see the next section). The only other parameter used was a minimum frequency of 4 for collocations. Out of 138,032 candidate headwords, 81,445 met this condition; most of the headwords were single words, only 128 were compounds.[5] For the automatic extraction, we imposed the aforementioned limits per syntactic structure, except for the 1,608 headwords that were selected for full manual validation (see the next section).

A new approach was also used in the automatic extraction of corpus examples. For version 1.0, we used different GDEX configurations for different parts of speech, with configurations being optimized for the extraction of good examples for collocations. While this approach produced good results, it took a great deal of processing, plus the GDEX score of a corpus sentence depended on a given headword rather than the sentence as a whole. Consequently, we decided to devise one GDEX configuration for an entire corpus - with the help of the Sketch Engine team, we ran the script on the Gigafida 2.0 corpus and assigned a GDEX score to each sentence in the corpus. Part of the automatic collocation extraction was thus also the extraction of the list of all corpus IDs of the sentences in which each collocation appeared; based on that, we extracted for the Collocations Dictionary up to four examples with the highest GDEX score per each collocation.

4. Storing collocational data: Digital Dictionary Database and a data warehouse

Collocations, along with other types of lexical information, are stored in the Slovene Digital Dictionary Database (Kosem et al., 2021), which aims to become a one-for-all database for the Slovenian language, to be used for both in the compilation of language resources and natural language processing tasks. The plans for the database have been described in detail by Klemenc et al. (2017). This trend of data consolidation can be observed across Europe, with the most noticeable case studies being the attempts for

[5] There are many more compounds in the Digital Dictionary Database, however for now only 128 have collocations.

Estonian (Tavast et al., 2018), German (Geyken, 2019), Polish (Żmigrodzki, 2018), and Dutch (Colman, 2016).

The first version of the Collocations Dictionary was part of the DDDS from the very beginning. However, due to many changes introduced by the data in the second version (method of collocation extraction, new corpus etc.), we had to first completely remove from the database the automatic collocational data from the first version and then import the new data. While we were preparing for the import of new data, other data had been imported, i.e., synonyms from the Thesaurus of Modern Slovene[6] (Arhar Holdt et al., 2018), and bilingual data from the Comprehensive Slovenian-Hungarian Dictionary[7] (Kosem et al., 2021), the latter also containing collocations. It is worth noting that the lexicographic process of the compilation of the Comprehensive Slovenian-Hungarian Dictionary includes a separate step of compiling entries from scratch for various purposes, which means that much more information (especially collocations and examples) is included than is needed, and ends up, in a bilingual dictionary.

Another relevant resource for the import of collocations was a data warehouse, which served as a storage for all the collocation candidates extracted from the corpus (over 63 million collocation candidates in total). The Digital Dictionary Database thus contains a subset of collocations from the data warehouse. In the data warehouse, we keep additional information such as IDs of corpus sentences in which the collocation is found, sense(s) under which the collocation belongs, the relevance of the collocation for the Collocations Dictionary for each of its components etc. Using the data warehouse facilitates the analysis of data, statistics, data extraction, and maintaining the link to corpus metadata. Having a record of not only good but also bad collocation candidates is crucial to preventing the duplication of work in the future.

A significant challenge at the import stage of new automatic collocations from the data warehouse proved to be matching the already identified collocations found in the digital dictionary database with newly automatically extracted ones, which had to be done to prevent duplication. Among other things, this also included analysing compounds, which may have received a status of a compound in a bilingual dictionary, but were considered legitimate collocations in a collocations dictionary. This process resulted in two types of entries - the ones with fully automatic collocations only, and others with a combination of manually inspected and automatically extracted collocations.

For a selection of 1,608 headwords,[8] we compiled fully manually validated entries. For these headwords, we did not use the same limitations in terms of a number of automatic

[6] https://viri.cjvt.si/sopomenke/eng/

[7] https://viri.cjvt.si/slovensko-madzarski/eng/

[8] The initial number was 2,000 but we ended up with fewer entries due to time constraints and work being needed on the matching of automatic collocations with existing manually

collocations per syntactic structure but rather exported all the collocations with the frequency of 4 and above. We, therefore, aimed to inspect all the collocations of a headword, however for frequent headwords with a great number of collocations (over a thousand) we set a minimum value logDice ≥ 4.0 for analysis. This roughly meant that whenever this threshold was applied, we ended up analysing under 300 collocations. We had three types of decisions: is a collocation, is a collocation but not relevant for the collocations dictionary, is not a collocation. The collocations in the first group ended up in the Collocations Dictionary, and the collocations in the second group ended up in the Digital Dictionary Database but not in the Collocations Dictionary.

A thorough analysis of collocations for 1,608 entries also served as an evaluation of the quality of automatic data in each syntactic structure. The results show high relevance of many structures (i.e. many structures contain many good collocation candidates) but also very poor results in certain structures. Table 1 and 2 show the top five syntactic structures with the highest percentage of good collocation candidates, and the top five syntactic structures with the highest percentage of bad collocation candidates, respectively.[9]

structure	**percentage of good collocation candidates**	**number of examined collocations**
adjective + preposition + noun in instrumental	90.91	396
adjective + noun	90.85	33271
verb + noun in accusative	87.72	6783
reflexive verb + noun in accusative	85.67	317
adjective + noun in dative	84.76	105

Table 1: Top five syntactic structures with the highest percentage of good collocation candidates.

validated collocations in the database.

[9] Syntactic structures with fewer than 100 collocations were excluded from these lists.

structure	**percentage of bad collocation candidates**	**number of examined collocations**
adjective + *and/or* + adjective	85.62	1210
noun + negative verb	81.17	154
noun + noun in dative	84.13	252
noun in nominative + verb in 3rd person	84.03	4722
noun + *and/or* + noun	76.56	9789

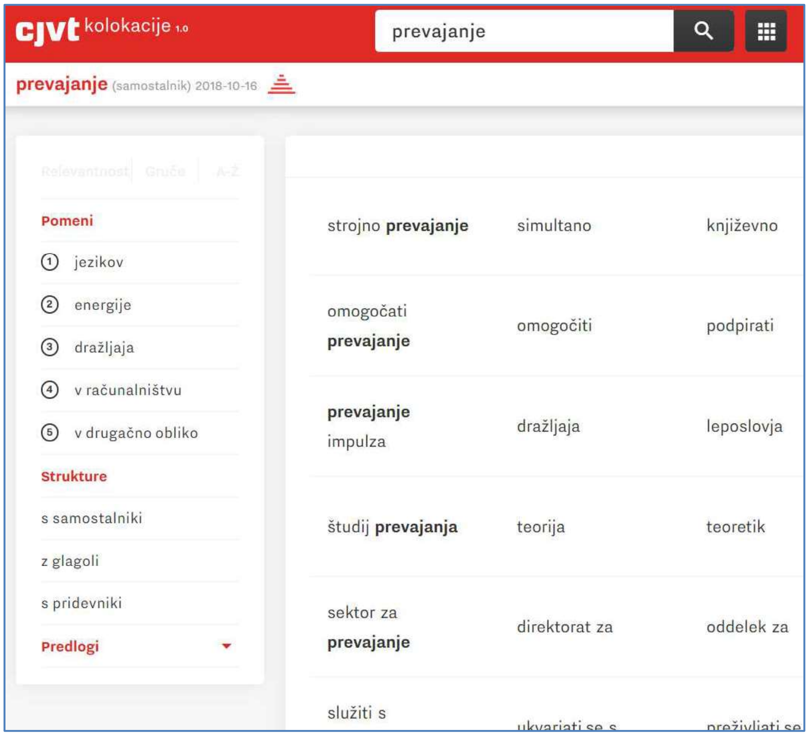
Table 2: Top five syntactic structures with the highest percentage of bad collocation candidates.

4.1 Interface and data presentation

The interface of the Collocations Dictionary has undergone some significant changes, on account of the harmonization with the interface of other language resources of the Centre for Language Resources and Technologies at the University of Ljubljana (CJVT UL), and, more importantly, of the findings of the studies with the users. The former changes were widening the page layout (to reduce scrolling and show more content initially), changing the font (to a more online-friendly one which supports many different characters and languages), and moving the menu box (with sense menu and structure filter) from left-hand column position to the top line above the content (see Figure 1b). The Collocations Dictionary 2.0 has also adopted the entry layout from other CJVT UL dictionaries (and according to the approach observed in foreign collocations dictionaries), abandoning the previous approach where the collocations were never clearly distributed under senses in the main window (the user had to use the sense filter to get the information of which collocations belonged to each sense) - the comparison is provided in Figures 1a and 1b.

The layout change is already quite noticeable, but even more noteworthy and relevant for the users are some other changes, which were informed by user studies. For example, there is now less clicking in general: all the collocations are offered immediately, with various data manipulation options available on the click of a button. These options include: limiting the view to a selection of most frequent collocations (Less/More icon); ordering collocates by frequency (the default option), alphabetical order, reversed alphabetical order, and length; filtering collocates to only 4768 lemmas on the Reference

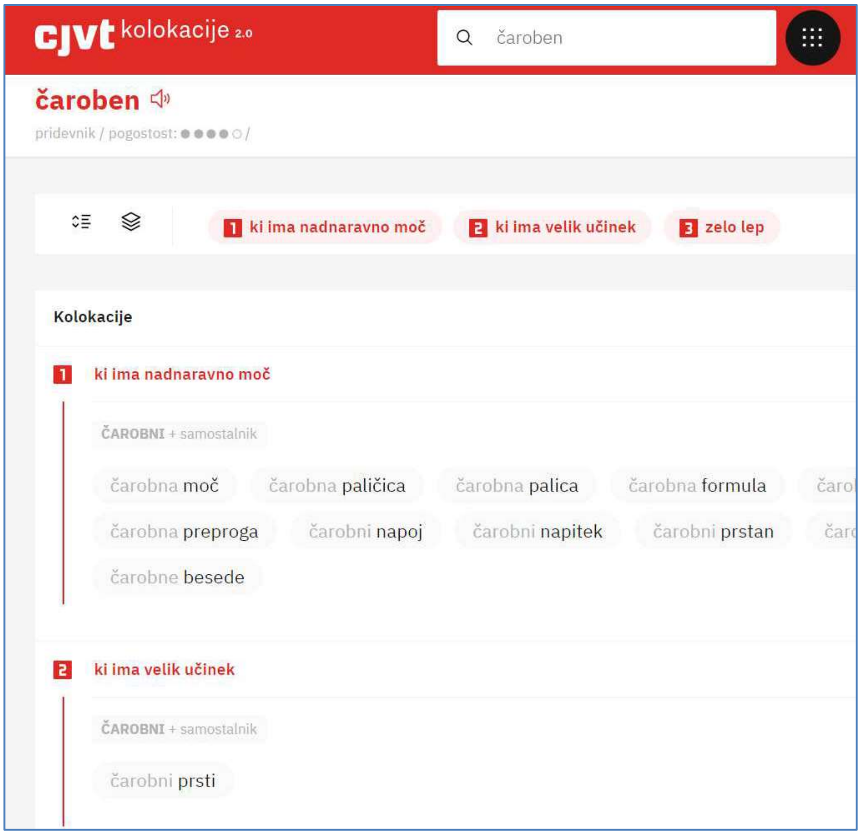
List of Slovene Frequent Common Words (Pollak et al., 2020; Arhar et al. 2020); and showing or hiding the headword in the collocation (the headword is shown by default). With the exception of the Less/More option, all the options are part of the Settings row and are thus used for all subsequent searches once set.



prevajanje (samostalnik) 2018-10-16

strojno **prevajanje**	simultano	književno
omogočati **prevajanje**	omogočiti	podpirati
prevajanje impulza	dražljaja	leposlovja
študij **prevajanja**	teorija	teoretik
sektor za **prevajanje**	direktorat za	oddelek za
služiti s	ukvarjati se s	preživljati se

Figure 1a. Entry layout in version 1.0.



cjvt kolokacije 2.0

čaroben

čaroben

pridevnik / pogostost: ●●●●○ /

1 ki ima nadnaravno moč   2 ki ima velik učinek   3 zelo lep

Kolokacije

1 ki ima nadnaravno moč

ČAROBNI + samostalnik

čarobna moč   čarobna paličica   čarobna palica   čarobna formula   čaro

čarobna preproga   čarobni napoj   čarobni napitek   čarobni prstan   čar

čarobne besede

2 ki ima velik učinek

ČAROBNI + samostalnik

čarobni prsti

Figure 1b. Entry layout in version 2.0.

A lot of thought and effort has been put in improving the clarity of presentation in the interface. The phase pyramid has been abandoned, and instead we added clear headings for two boxes with different types of collocational information. Collocations that have been manually validated and distributed under senses are found in the "Collocations" box, whereas automatic and not yet inspected collocations are found in the box titled "Automatically extracted collocations." In this way, we reduced the previous five-stage entry progress shown by the pyramid icon (which was often missed or considered unclear by the users) to a three-type entry status which is immediately apparent and needs no additional status icon. The three types of entries in the Collocations Dictionary of Modern Slovene are:

- entries with sense division and only manually validated collocations. These entries have only the "Collocations" box.

- entries with sense division and manually validated collocations in these senses, but also with automatically extracted collocations without an assigned sense. These entries contain both the "Collocations" and "Automatically extraction collocations" boxes.

- entries with only automatically extracted collocations. These entries contain only the “Automatically extraction collocations” box.

We also changed the presentation of syntactic structure titles, as now they are clearly presented as titles under which collocations are grouped (in version 1.0, the structure name was only made available on mouseover). The presentation of examples remained the same; they can be viewed by clicking on a collocation. The link to the corpus showing all the examples of a particular collocation is also available at that point.

Another more significant change, which is related to the user experience, is the enhancement of the crowdsourcing aspect of the dictionary. In the first version of the dictionary, the only crowdsourcing feature was the option to mark collocations as good or bad (using upvote and downvote) on the page of each structure. The feature was rarely used, and as shown by research, such a task is far too demanding for an average user. In the second version, we opted to introduce crowdsourcing at an example level; the users can now not only confirm the validity of the collocation in each example provided but also select the relevant sense (if sense division for a particular headword has already been made). This is in line with our findings that examples rather than collocations are much more suitable for direct crowdsourcing.

5. Conclusions and future plans

The Collocation Dictionary of Modern Slovene, version 2.0, has introduced many changes to both the collocational data it contains, and to the way the data is presented to the user. The changes took into account the latest developments in automatic collocation extraction from corpora, and the findings of various user studies. The dictionary has reaped the benefits of storing the data in the Digital Dictionary Database and in a data warehouse, not only because of avoiding the duplication of work but also because we were able to utilize the lexical data produced in other dictionary projects.

Short-term plans include the preparation of the dictionary database in the XML format and its upload to the CLARIN.SI repository. In line with the policy at the CJVT UL, the database will be available under the CC BY-SA 4.0 license (Creative Commons - Attribution-ShareAlike 4.0 International). Moreover, we are currently working on making the user voting information immediately available next to each collocation; the idea is to show the sense number(s), or the tick or cross icon next to the collocation as soon as the user vote is cast.[10]

Long-term, we would like to add other types of grouping of the collocations, for example by questions such as *Mrežnik* and *elexiko*, and/or by semantic properties (e.g., using semantic types). There are also plans to conduct further user studies to identify further

[10] The hold up is mainly technical as we are solving some performance issues.

improvements to the interface. Based on the evaluation of the data of 1,608 manually completed entries, improvements to the automatic extraction method will be made.

An important development expected in the next months will be the introduction of an editor for the Digital Dictionary Database which will facilitate entry compilation and publication, enabling us to make updates to the Collocations Dictionary on a more regular basis.

6. Acknowledgements

The authors acknowledge that the project Empirical foundations for digitally-supported development of writing skills (J7-3159) and the programmes Language Resources and Technologies for Slovene (P6-0411) and Slovene Language – Basic, Contrastive, and Applied Studies (P6-0215) were financially supported by the Slovenian Research Agency. The project Upgrading fundamental dictionary resources and databases of CJVT UL was funded by the Ministry of Culture of the Republic of Slovenia in the period 2021–2022.

7. References

Arhar Holdt, Š., Pollak, S., Robnik Šikonja, M. & Krek, S. (2020). *Referenčni seznam pogostih splošnih besed za slovenščino.* Proceedings of the Conference on Language Technologies and Digital Humanities, pp. 10-15.

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

Arhar Holdt, Š. (2021) Razvrstitev kolokacij v slovarskem vmesniku: uporabniške prioritete. In I. Kosem, Iztok (ed.) *Kolokacije v slovenščini.* Ljubljana: Znanstvena založba Filozofske fakultete, pp. 125-157. Available at: https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/318/465/6974-1.

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

Colman, L. & Tiberius, C. (2018). A good match: a Dutch collocation, idiom and pattern dictionary combined. *Proceedings of the XVIII EURALEX International Congress*, pp. 233-246.

Colman, Lut. (2016). Sustainable lexicography: where to go from here with the ANW

(Algemeen Nederlands Woordenboek, an online general language dictionary of contemporary Dutch)? *International Journal of Lexicography*, 29/2, pp. 139-155.

Firth, John. A synopsis of linguistic theory, 1930-1955. Studies in linguistic analysis, 1957, pp. 10-32.

Gantar, P., Kosem, I. & Krek, S. (2016): Discovering automated lexicography: the case of Slovene lexical database. *International journal of lexicography*, 29 (2), pp. 200–225.

Gantar, P., Krek, S. & Kosem, I. (2021). Opredelitev kolokacij v digitalnih slovarskih virih za slovenščino. In I. Kosem (ed.). *Kolokacije v slovenščini.* Ljubljana: Znanstvena založba Filozofske fakultete, pp. 15-41. Available at: https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/318/465/6969-1.

Geyken, A. (2019). The Centre for Digital Lexicography of the German Language: New Perspectives for Smart Lexicography. I. Kosem & T. Zingano Kuhn (eds.) *Electronic lexicography in the 21st century (eLex 2019): Smart lexicography. Book of abstracts*. Lexical Computing CZ s.r.o., Brno, Czech Republic.

Hanks, P. (2004). Corpus pattern analysis. *Proceedings of the XI EURALEX International Congress*, pp. 87-98.

Haß, U. (ed.). (2005). Grundfragen der elektronischen Lexikographie. elexiko – das Online-Informationssystem zum deutschen Wortschatz. (Schriften des Instituts für Deutsche Sprache). Berlin/New York: de Gruyter.

Hudeček, L. & Mihaljević, M. (2020a). The Croatian Web Dictionary – Mrežnik Project – Goals And Achievements. *Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje* 46/2, pp. 645–667.

Hudeček, L. & Mihaljević, M. (2020b). Collocations in the Croatian Web Dictionary – Mrežnik. *Slovenščina 2.0.* 8/2, pp. 78–111.

Kallas, J., Kilgarriff, A., Koppel, K., Kudritski, E., Langemets, M., Michelfeit, J., Tuulik, M., & Viks, Ü. (2015). Automatic generation of the Estonian Collocations Dictionary database. In *Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference*, 11-13 August 2015, Herstmonceux Castle, United Kingdom Ljubljana/Brighton: Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd., pp. 1—20.

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

Klosa, A. (2015). Wortgruppenartikel in elexiko: Einneuer Artikeltyp im Onlinewörterbuch. *Sprachreport Jg*, 31(4), pp. 34–41.

Koppel, K., Kallas, J., Khokhlova, M., Suchomel, V., Baisa, V. & Michelfeit, J. (2019). SkELL corpora as a part of the language portal Sõnaveeb: problems and perspectives. In I. Kosem, T. Zingano Kuhn, M. Correia, J.P. Ferreira, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek & C. Tiberius (eds.) *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference.* 1-3 October 2019, Sintra, Portugal. Brno: Lexical Computing CZ, s.r.o., pp. 763-782.

Kosem, I., Arhar Holdt, Š., Krek, S., Gantar, P., Pori, E., Čibej, J., Klemenc, B., Laskowski, C., Dobrovoljc, K., Gorjanc, V. & Ljubešić, N. (2022). *Kolokacijski slovar sodobne slovenščine 2.0.* Znanstvena založba Filozofske fakultete Univerze v Ljubljani.

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

Kosem, I., Krek, S. & Gantar, P. (2020). Defining collocation for Slovenian lexical resources. Slovenščina 2.0, 2, pp. 1-27. DOI: 10.4312/slo2.0.2020.2.1-27.

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

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

Krek, S., Gantar, P., Kosem, I. (2022). Extraction of collocations from the Gigafida 2.1 corpus of Slovene. In A. Klosa (ed.). *EURALEX 2022, Proceedings of the XX EURALEX International Congress*, 12-16 July 2022, Mannhein, Germany. [S. l.]: IDS-Verlag, pp. 240-252. https://euralex.org/wp-content/themes/euralex/proceedings/Euralex%202022/EURALEX2022_Pr_p230-239_Kosem.pdf.

Krek, S., Gantar, P., Kosem, I. & Dobrovoljc, K. (2021). Opis modela za pridobivanje in strukturiranje kolokacijskih podatkov iz korpusa. In Š. Arhar Holdt (ed.) *Nova slovnica sodobne standardne slovenščine.* Ljubljana: Znanstvena založba Filozofske fakultete, pp. 160–194.

Pollak, S., Arhar Holdt, Š., Krek, S. & Robnik Šikonja, M. (2020). Reference List of Slovene Frequent Common Words, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1346.

Pori, E. & Kosem, I. (2021). Evalvacija avtomatskega luščenja kolokacijskih podatkov iz besednih skic v orodju Sketch Engine. In I. Kosem (ed.). *Kolokacije v slovenščini.* Ljubljana: Znanstvena založba Filozofske fakultete, pp. 43-77. https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/318/465/6974-1.

Pori, E., Čibej, J., Kosem, I. and Arhar Holdt, Š. (2020): The attitude of dictionary users towards automatically extracted collocation data: a user study. Slovenščina 2.0, 8(2): 168–201. DOI: https://doi.org/10.4312/slo2.0.2020.2.168-201

Pori, E., Kosem, I., Čibej, J. & Arhar Holdt, Š. (2021). Evalvacija uporabniškega vmesnika Kolokacijskega slovarja sodobne slovenščine. In I. Kosem (ed.). *Kolokacije v slovenščini*. Ljubljana: Znanstvena založba Filozofske fakultete, pp. 235-268.

Storjohann, P. (2005). elexiko: A Corpus-Based Monolingual German Dictionary. *Hermes, Journal of Linguistics*, 34, pp. 55–82.

Tavast, A., Langemets, M., Kallas, J., Koppel, K. (2018). Unified Data Modelling for Presenting Lexical Data: The Case of EKILEX. In J. Čibej, V. Gorjanc, I. Kosem, S. Krek (eds.). *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*. Ljubljana: Znanstvena založba Filozofske fakultete, pp. 749-761.

Tiberius, C., Munda, T., Repar, A. and Krek, S. (2022). Lexicographic data in ELEXIS. Deliverable of the ELEXIS project. https://elex.is/wp-content/uploads/ELEXIS_D1_6_Lexicographic_data_in_ELEXIS.pdf

Żmigrodzki, P. (2018). Methodological issues of the compilation of the Polish Academy of Sciences Great Dictionary of Polish. In J. Čibej, V. Gorjanc, I. Kosem, S. Krek (eds.). *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*. Ljubljana: Znanstvena založba Filozofske fakultete, pp. 209-219.