Collocations Dictionary of Modern Slovene

***Iztok Kosem*[1,2], *Simon Krek*[2], *Polona Gantar*[1], *Špela Arhar Holdt*[1], *Jaka Čibej*[1,2,3], *Cyprian Laskowski*[1]**

[1]*Faculty of Arts, University of Ljubljana,* [2]*Jožef Stefan Institute,* [3]*Faculty of Computer and Information Science, University of Ljubljana*

E-mail: iztok.kosem@ff.uni-lj.si, simon.krek@guest.arnes.si, apolonija.gantar@guest.arnes.si, Spela.ArharHoldt@ff.uni-lj.si, jaka.cibej@ff.uni-lj.si, CyprianAdam.Laskowski@ff.uni-lj.si

Abstract

The paper presents the compilation of the Collocations Dictionary of Modern Slovene, a new resource targeting the language production needs of Slovene speakers. An important aspect of the compilation of the dictionary is the immediate publication of all the entries, from automatic, postprocessed, finalized by lexicographers and so on, and indicating to the users their status, i.e. the stage in the compilation process. Furthermore, we discuss the introduction of crowdsourcing into the lexicographic workflow. The paper also focuses the development and presentation of the interface, which introduces new approaches to collocation presentation. The aim was to develop a collocation-driven interface that would allow different types of users a great deal of flexibility and customizability in exploring collocational information about words. In this way, the interface represents a hybrid between a more corpus-based presentation of collocations (e.g. in tools such as Word Sketch) and a traditional sense-driven presentation of collocations as found in existing collocations dictionaries.

Keywords: collocations, dictionary, database, interface

1 Background

In recent years, collocations have received a great deal of attention in Slovenian lexicography, initially mainly in relation to the conceptualization of a new monolingual dictionary of modern Slovene (Gorjanc et al. 2015, 2017). Relatedly, procedures for the automatic extraction of collocations and their examples have been developed and continuously improved (e.g. Kosem et al. 2013; Gantar et al. 2016). This means that lexicographers can now very quickly obtain large quantities of collocational information about words, which has facilitated the compilation of dictionaries, both general and terminological (e.g. Logar et al. 2013). However, despite these methodological advances, existing Slovene dictionaries, many of which are also outdated, offer users little help with language production tasks such as writing.

Interestingly, despite the advances in methods for collocation identification and the advantages offered by digital media, not many born-digital collocations dictionaries (i.e. dictionaries developed with a digital medium in mind) have been published. The examples the authors are familiar with include the Estonian Collocations Dictionary (Kallas et al. 2015; the dictionary will be published in 2018), the German Collocations Dictionary (Roth 2013; Häcki Buhofer et al. 2014),[1] and the Spanish Collocations Dictionary (DiCE; Vincze et al. 2011; Vincze & Alonso Ramos 2013). Furthermore, projects such as automatic collocation dictionaries (see Kilgarriff et al. 2013) and SkeLL have shown that even automatically extracted data can be useful for language users. All of the aforementioned dictionaries target L2 learners, as collocations specifically tend to pose significant problems for language learners (e.g. Granger & Meunier 2008; Schmitt 2004; Nation 2001); however, even L1 users

1 https://kollokationenwoerterbuch.ch/web/. The dictionary was also published in paper format.

encounter challenges in language production and can benefit from having such resources at their disposal.

The lack of productively-oriented dictionaries for Slovene, regardless of the types of users, prompted the compilation of the Collocations Dictionary of Modern Slovene (CODICT, Gantar et al. 2015). A great deal of attention was paid to the design and customizability of the interface. One of the important decisions was not to publish only completed entries, but also include entries in various stages of completion, mainly in order to avoid causing user frustration due to the lack of collocational data for those headwords without complete entries in CODICT. As we wanted to inform design-related decisions with empirical data as much as possible, we first conducted a small test on a sample database to get feedback on the presentation of collocational information, and to investigate how Slovene users react to automatically created (not manually curated) content.

2 Initial Test with a Sample Collocations Database and Its Interface

In 2016, a sample of 2,500 automatically extracted collocational entries for Slovene (Krek et al. 2016) was extracted and published at http://bkssj.cjvt.si/. Each entry included collocations grouped by grammatical relation and their corpus examples (extracted using GDEX; Kilgarriff et al. 2008). Some additional post-processing was conducted, e.g. putting the collocates in the required case, gender, removing duplicate examples, etc.

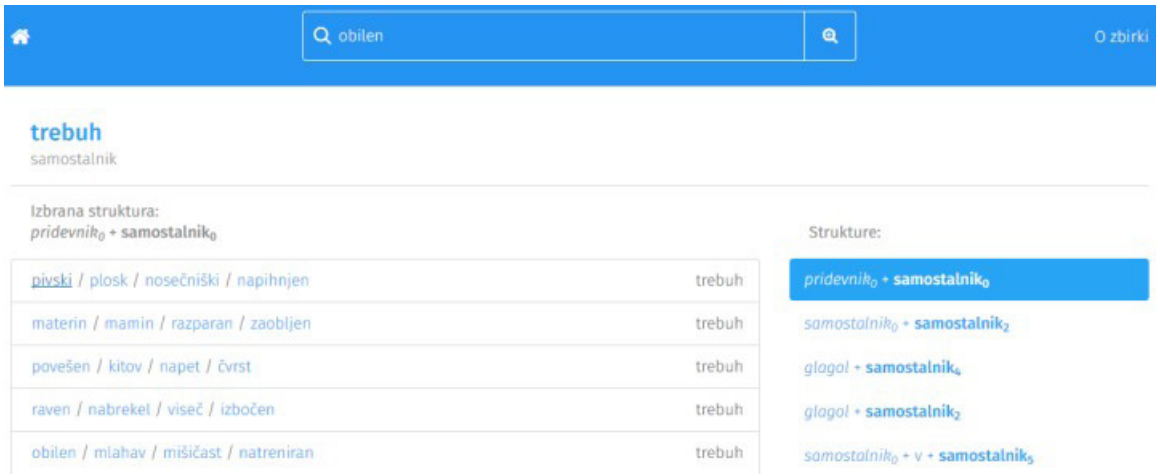

Figure 1: The interface of the sample collocations database.

The interface and the contents of this sample collocations database were then evaluated by a group of linguists and linguistics students. The data was found to be useful despite the fact that it contained a certain degree of noise. The users commented on several issues, which can be grouped into two categories. The first is related to content, such as large quantities of data, lack of data structure (e.g. missing sense information), small number of entries, distracting irrelevant or incorrect information (i.e. noise). The second category involves the presentation of lexical information, for example the lack of (statistical) information on collocations to help assess their relevance, the lack of options for sorting collocations, and so on. In sum, while the automatically extracted information was well- received among the users, they also noted a need for more structure being given to the data, and more options with regard to manipulating it.

3 Collocations Database and Collocations Dictionary of Modern Slovene

After concluding the initial test, we first extracted a much larger dataset, containing 35,989 entries with automatically extracted collocational data (nearly eight million collocations and nearly 37 million corpus examples), using the Sketch Engine API. The data was obtained from Gigafida, the 1.2-billion-word reference corpus of written Slovene (Logar Berginc et al. 2012). Compared to the sample database, several improvements have been introduced, both to the data extraction procedure and the post-processing of the extracted data. For example, good examples, five per collocation, were extracted using an updated GDEX configuration (e.g. penalties for sentences ending with an ellipsis and containing only upper-case letters have been introduced). Secondly, additional filtering of collocational data was used in the post-processing stage; this excluded all collocations containing the verb *biti* ('*to be*') and removed prepositional grammatical relations (preposition + noun in a specific case) that were not in accordance with the rules of the Slovene Orthography (Toporišič 2001).[2] This dataset presented a basis for CODICT.

Nowadays, dictionaries mainly use two approaches in the way they publish entries. One is to wait until the entire dictionary is compiled, and the other one, which has become the norm for online dictionaries, is to publish newly compiled entries at regular intervals (e.g. once a year) – this is what Klosa (2013) calls a dictionary under construction. For our purposes, none of these approaches seemed suitable, so we decided to use the approach proposed by Krek et al. (2013), where all the entries are made available to the users immediately, with a clear indication of their status in the lexicographic process. We introduced five stages: automatically extracted entry, postprocessed entry (semi-automatically cleaned and improved data based on lexicographic decisions), entry with validated collocations, entry with collocations distributed under senses, and final entries. There are several reasons for this, such as improvements in (semi-)automatic methods of processing of language data, especially collocations, the way that users can benefit from all this information immediately, the fact that in many cases clustered collocations already clearly indicate different senses, and so on. In addition, the dictionary will benefit from the results and findings of a research project called KOLOS (Collocations in Slovene), which is focused on collocations and aims to improve methods of collocation detection, collocation clustering, and the use of collocations in comparing synonyms.

The ultimate aim is to have all the entries manually edited, i.e. validated and cleaned data obtained with automatic extraction, with added information such as sense division, labels, collocate groupings (clusters), as well as providing collocations in their typical form (e.g. nouns in plural, adjectives in superlative, verbs in negative). For each collocation, at least two good examples should be provided. Rather than using definitions for senses, we decided to use short indicators, similar to signposts (used first by the Longman Dictionary of Contemporary English) and short definitions in menus (introduced by the Macmillan English Dictionary for Advanced Learners) (see also Kosem et al. 2017). The most important role of the indicators is thus not to explain the meaning, but to help the users clearly distinguish between different senses.

It is important to note that not all collocations that are validated by lexicographers are included in the final entries. This is because of the difference between statistical collocation, i.e. any combination of two or more words that is statistically relevant, and a collocation that is deemed relevant for inclusion in a collocations dictionary. Our inclusion criteria for collocations (and headwords) are less strict than the criteria used to make works such as the *Macmillan Collocations Dictionary*, where the authors excluded headwords such as *house, buy, good* on account of the fact they do not have any strong

2 Some of these excluded prepositional grammatical relations may contain valid information, but we will conduct a detailed analysis before including them in the database.

collocates.[3] In the digital age there is no longer any need to limit the number of headwords, but there remains a need to determine which collocations to include; for example, do we include numerous modifiers of the word *prestolnica* ('capital') related to countries, e.g. *Austrian*, *German*, etc., or do we include only the most salient ones, or none at all? In any case, even when statistically relevant collocations are excluded from CODICT, they are still kept in the database as they will be of use for the compilation of other resources, e.g. general language dictionaries, valency dictionaries, etc.

3.1 Implementing Crowdsourcing into Lexicographic Workflow

Considering the high number of collocations per headword and financial and time constraints related to the compilation of CODICT, we decided to introduce crowdsourcing methods into the lexicographic workflow. Previous tests (Kosem et al. 2013) have shown that tasks such as assigning collocations, via examples, to the relevant senses are not very demanding (even for non-linguists) and provide highly reliable results. We conducted such a task on 6,590 collocations (microtasks) for 88 sample headwords in CODICT, using four annotators (students of linguistics) and requiring three answers per microtask (see Figure 2). In addition to senses, the annotators were given the answer options "None of the above" (if the example indicated a sense of the headword not covered by the ones provided) and "I don't know". The results showed a high degree of annotator agreement: each pair of annotators agreed in 79-86% of the cases (83% on average, with an average Cohen's kappa of 0.83). A total of 4,258 collocation examples (65%) showed perfect agreement.

Figure 2: The crowdsourcing task in Pybossa (an example of a microtask).

In addition to saving lexicographers' time and consequently speeding up the compilation of entries, the crowdsourcing method provides important feedback on sense division even during the process of dictionary compilation. As the experience from our crowdsourcing task has shown, the analysis of annotators' responses can reveal which indicators need to be improved, and potentially identify groups of collocates that might require their own sense, or senses that might have been overlooked. Furthermore, there were instances where the annotators' responses indicated that the sense division was too fine-grained.

3 http://www.macmillandictionaries.com/features/how-dictionaries-are-written/macmillan-collocations-dictionary/

3.2 Designing the Interface

A great deal of attention has been paid to the development of the interface,[4] which has to be easy to use and makes it clear to the users the status of the entry they are consulting. The status is indicated both directly, with the use of a pyramid icon with completed stages colored in red, and indirectly, for example the Sense filter is introduced only in the last two stages of entry compilation. The design has been greatly informed by user feedback on the interface of the aforementioned sample collocations database. An important decision made in the design process was that the interface would be collocation-driven rather than sense-driven; as a result, the user would be initially given a more general overview of the collocations of the word, and would then be able to explore collocational information further using sorting options and filters (by sense, grammatical relation, frequency, etc.). The interface was designed for different devices, i.e. a computer, tablet, and smartphone; some adjustments, e.g. exclusion of certain functions, had to be implemented for smaller screens such, as phones.

The initial view (i.e. general overview) provides a quick summary of most relevant collocations, divided by grammatical relations. We therefore attempt to maintain some grammatical diversity of the collocational overview; normally, there is one line per grammatical relation, although multiple lines can be allocated to a single relation if the number of (salient) collocations it contains is significantly higher than the number of collocations found in other relations. In this initial view, the users can select a specific grammatical relation to get a view of all the collocations in it, or they can click on a particular collocation to see its examples of use.

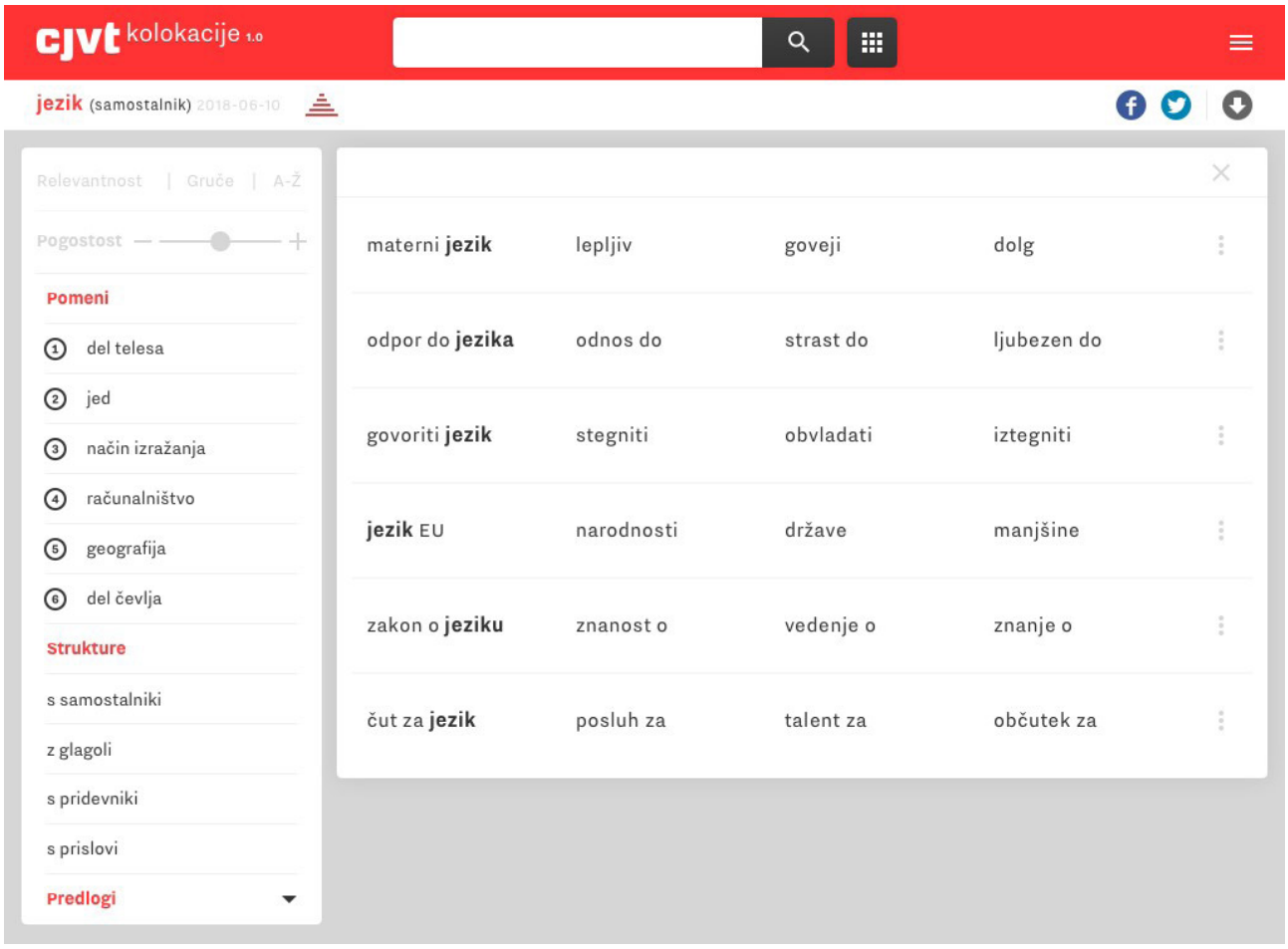

Figure 3: The interface of CODICT (first page of the entry *jezik*).

4 http://viri.cjvt.si/kolokacije/.

The left-hand panel (located at the top in the mobile version) contains sorting and filtering options. Sorting is available only for the single relation view, and enables the user to sort collocates by relevance (default setting), semantic characteristics (clustering), and alphabetical order. There are four types of filters available:

- The frequency filter enables the users to filter collocates according to their frequency in the corpus. The filter can thus help the users to focus on a more frequent or rare collocates.
- *Pomeni* – the sense filter (available for entries in the final two stages) provides an overview of the senses in which the collocations can be found.
- *Strukture* – the grammatical relation filter provides the users with the option to limit the results to relations according to the word class of the collocate (e.g. noun, adjective, verb, adverb), and subcategories such as case or degree (e.g. superlative).
- *Predlogi* – the preposition filter applies to prepositional relations (trinary) in which collocations can be found, and is offered as a separate filter because it transcends different top-level categories in the grammatical relation filter.

As the idea is to give the users flexibility in limiting the amount of data displayed on the screen in order to facilitate finding the relevant information, multiple types of filters, as well as sorting, can be active at the same time. However, only one category within a certain type of filter, e.g. prepositions in the *Predlogi* filter, can be selected by the users at a time. There is a difference in behavior of the sense filter, where all the senses of a word always remain visible, even if some of the sense do not apply to the selection (those are then are greyed out), and the grammatical relation and preposition filters, where only the selected or relevant relation (and its subcategory) is shown.

Even if the users are not using filters and are conducting their activities only in the right-hand of the interface (main panel), the filters remain dynamic and in that way informative, as they provide information on the current selection in the main panel. For example, if the user selects one of the grammatical relations in the main panel, they are taken to the collocations of that relation, and at the same time, the senses in which the collocations belonging to that particular relation are not found are greyed out, and grammatical relation and its subcategory in the grammatical relation filter is selected. This solution has been formed based on users' feedback to the sample collocations database interface, as they found listing all available grammatical relations on the right side overwhelming, and the names of the relations confusing (e.g. verb + $noun_4$ meant verb followed by a noun in the accusative case).

There are additional filters available in the right-hand panel (in the main view), which are based on the characteristics of the headword. For example, the collocations of the adjective headword in the grammatical relation adjective + noun can be filtered by gender. So, as shown in Figure 4, selecting *okusna* (the feminine form of the adjective *okusen*, 'tasty') shows only feminine nouns as collocates (*sladica* 'dessert', *hrana* 'food', etc.).

A very important feature of the interface is its search functionality, which offers searches by both headwords and collocations. It is important to note that if the user searches for a specific collocation, the result differs from the output obtained if opening the same collocation within a particular entry. This is due to the difference in user focus – if a specific collocation is selected within a certain headword, the user's focus comes from the headword, whereas when a specific collocation is searched for, the information on all its elements, and related collocations, is useful.

Finally, the interface also provides links to other resources. A page of each individual collocation, which contains corpus examples, contains a direct link to the corpus concordances for the collocation. Furthermore, the main features of the interface are shared by all the resources of the Centre for Language Resources and Technologies at the University of Ljubljana, so the user can, by clicking

on a specific button in the search row, obtain a quick view of links to all the available resources that contain entries related to his or her search.

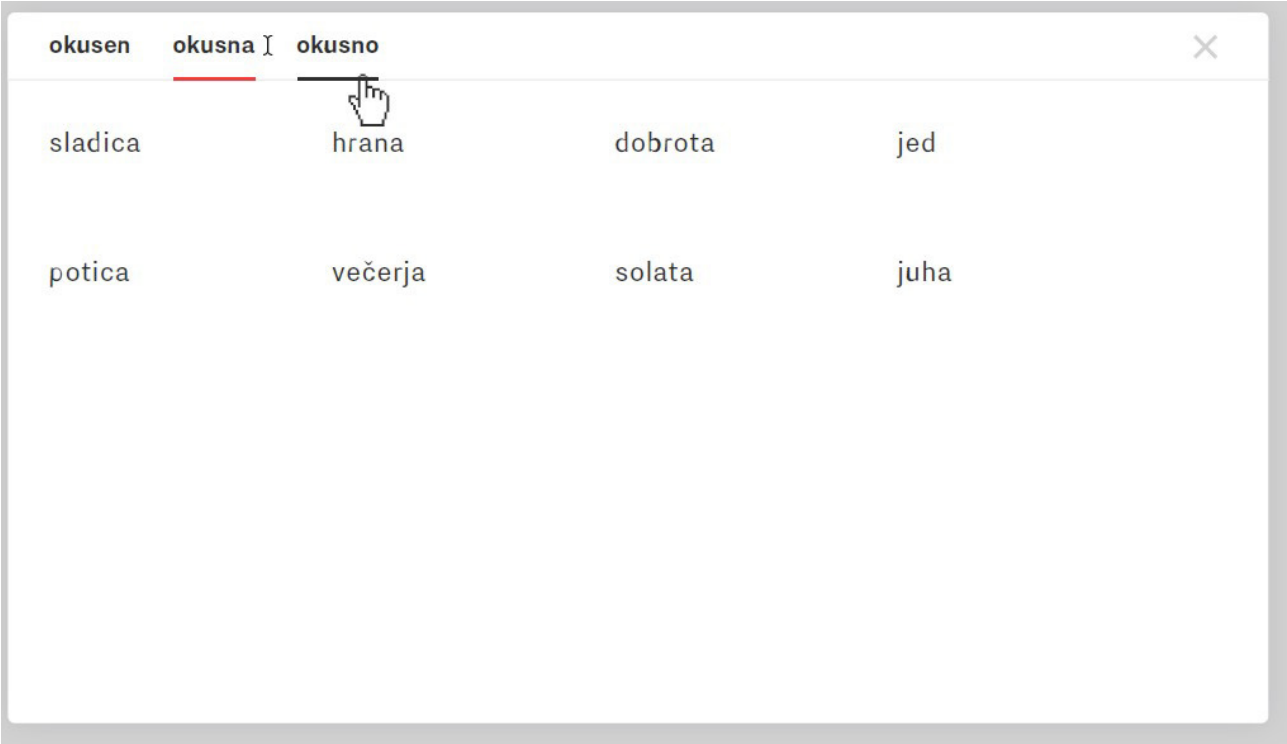

Figure 4: An additional filter in the main collocations view (single relation).

4 Future Plans

Future plans for this project involve making several improvements to the methodology of compilation of collocational data. This involves improving the precision of collocation detection using approaches such as distributional semantics and extraction of collocations from parsed corpora. In addition, we plan to update CODICT with the information from a new version of the Gigafida corpus, which is due to be published at the end of 2018. We will also explore gamification forms of crowdsourcing to identify valid collocations.

Equally important as improving the methodology is testing the interface with different user groups. At the time of writing this paper we are already preparing a survey that will be conducted among the users, and will be complemented with interviews. The study will focus mainly on the content of the initial view, i.e. what different users would expect or want to be offered when opening the entry. Other plans include adding new filters based on the metadata of corpus texts, e.g. text type and year of publication.

References

Häcki Buhofer, A., Dräger, M., Meier, S. & Roth, T. (2014). *Feste Wortverbindungen des Deutschen. Kollokationenwörterbuch für den Alltag*. Tübingen: Francke.

Gantar, P., Gorjanc, V., Kosem, I. & Krek, S. (2015). Going semi-automatic and crowdsourced: collocation dictionary of Slovene. In I. Kosem (ed.) *Electronic lexicography in the 21st century: linking lexical data in the digital*

age. *eLex 2015, book of abstracts.* Ljubljana: Trojina, Institute for Applied Slovene Studies; Brighton: Lexical Computing. 2015, p. 37.

Gantar, P., Kosem, I., & Krek, S. (2016). Discovering Automated Lexicography: The Case of the Slovene Lexical Database. *International Journal of Lexicography*, 29 (2): 200–225.

Gorjanc, V., Gantar, P., Kosem, I. & Krek, S. (eds.) (2015). *Slovar sodobne slovenščine: problemi in rešitve.* Ljubljana: Univerza v Ljubljani, Filozofska fakulteta.

Gorjanc, V., Gantar, P., Kosem, I. & Krek, S. (eds.) (2017). *Dictionary of Modern Slovene: problems and solutions.* Ljubljana: Univerza v Ljubljani, Filozofska fakulteta.

Granger, S., & Meunier, F. (eds.) (2008). *Phraseology. An interdisciplinary perspective*. Amsterdam: Benjamins.

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 I. Kosem, M. Jakubiček, J. Kallas & S. (eds.) *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.

Klosa, A. (2013). The lexicographical process (with special focus on online dictionaries). In R. H. Gouws, U. Heid, W. Schweickard and H. E. Wiegand (eds.) *Dictionaries. An international Encyclopedia of Lexicography.* Supplement Volume: Recent Developments with Focus on Electronic and Computational Lexicography. Berlin in Boston: de Gruyter, pp. 517–524.

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M. & Rychly, P. (2008). GDEX: Automatically Finding Good Dictionary Examples in a Corpus. In E. Bernal & J. DeCesaris. (eds) *Proceedings of the Thirteenth EURALEX International Congress*. Barcelona, Spain: Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra, pp. 425–432.

Kilgarriff, A., Husak, M., Jakubíček, M. (2013). Automatic collocation dictionaries. *Presentation at eLex 2013 conference*, *Tallinn, Estonia*. Available at: https://youtu.be/b3KyhPBeoLU.

Kosem, I., Gantar, P., Krek, S. (2013): Automation of lexicographic work: an opportunity for both lexicographers and crowd-sourcing. In I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets & M. Tuulik (eds.) *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of eLex 2013 Conference, 17-19 October 2013, Tallinn, Estonia*. Ljubljana: Trojina, Institute for Applied Slovene Studies; Tallinn: Eesti Keele Instituut, pp. 32-48.

Kosem, I., Gantar, P., Krek, S. (2017). Sense menus in collocations dictionary of Slovene. *Electronic lexicography in the 21st century: lexicography from scratch*. Leiden: Dutch Language Institut; Brno: Lexical Computing; Ljubljana: Trojina Institute for Applied Slovene Studies, p. 43.

Krek, S., Gantar, P., Kosem, I., Gorjanc, V. & Laskowski, C. (2016). Baza kolokacijskega slovarja slovenskega jezika. In T. Erjavec & D. Fišer (eds.) *Zbornik konference Jezikovne tehnologije in digitalna humanistika, 29. september - 1. oktober 2016, Filozofska fakulteta, Univerza v Ljubljani, Ljubljana, Slovenija = Proceedings of the Conference on Language Technologies & Digital Humanities, September 29th - October 1st, 2016 Faculty of Arts, University of Ljubljana, Ljubljana, Slovenia*. Ljubljana: Znanstvena založba Filozofske fakultete: = Ljubljana University Press, Faculty of Arts, pp. 101-105.

Krek, S., Kosem, I., Gantar, P. (2013). *Predlog za izdelavo Slovarja sodobnega slovenskega jezika*, *version 1.1*. Accessed on 10 May 2018. http://www.sssj.si/datoteke/Predlog_SSSJ_v1.1.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.

Nation, P. (2001*). Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Roth, T. (2013). Going Online with a German Collocations Dictionary. In I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets & M. Tuulik (eds.) *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia*. Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut, pp. 152-163.

Rundell, M. (ed.) (2010). *Macmillan Collocations Dictionary*. Oxford, United Kingdom: Macmillan Education.

Schmitt, N. (ed.) (2004). *Formulaic sequences: Acquisition, processing and use*. Amsterdam: Benjamins.

Toporišič, J. (ed.) (2001). Slovenski pravopis. Ljubljana: Založba ZRC, ZRC SAZU.

Vincze, O., Mosqueira, E., & Alonso Ramos, M. (2011). An online collocation dictionary of Spanish. In I. Boguslavsky & L. Wanner (eds.) *Proceedings of the 5th International Conference on Meaning-Text Theory*. Barcelona, pp. 275– 286.

Vincze, O. & Alonso Ramos, M. (2013). Testing an electronic collocation dictionary interface: Diccionario de Colocaciones del Espanol. In I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets & M. Tuulik (eds.) *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of eLex 2013 Conference, 17-19 October 2013, Tallinn, Estonia*. Ljubljana: Trojina, Institute for Applied Slovene Studies; Tallinn: Eesti Keele Instituut, pp. 328-337.

Acknowledgements

The paper was prepared as part of the two projects, Kolokacije kot temelj jezikovnega opisa: semantični in časovni vidiki (Collocations as a Basis for Language Description: Semantic and Temporal Perspectives, J6-8255) and Nova slovnica sodobne standardne slovenščine: viri in metode (New grammar of contemporary standard Slovene: sources and methods, J6-8256), which were financially supported by the Slovenian Research Agency. The paper was also supported by ELEXIS, a project that has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 731015.

The authors would like to acknowledge the financial support from the Slovenian Research Agency's infrastructure programmes Centre for Applied Linguistics, Trojina Institute (I0-0051), and Centre for Language Resources and Technologies, the University of Ljubljana.

The interface was developed by Studio Kruh in collaboration with Leon Noe Jovan.