# Thesaurus of Modern Slovene: By the Community for the Community

***Špela Arhar Holdt[1,2], Jaka Čibej[1,2,3], Kaja Dobrovoljc[1,3], Polona Gantar[2], Vojko Gorjanc[2], Bojan Klemenc[1], Iztok Kosem[2,3], Simon Krek[2,3], Cyprian Laskowski[2], Marko Robnik-Šikonja[1]***

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

*E-mail: spela.arharholdt@ff.uni-lj.si, jaka.cibej@ff.uni-lj.si, kaja.dobrovoljc@ijs.si, apolonija.gantar@guest.arnes.si, vojko.gorjanc@ff.uni-lj.si, bojan.klemenc@ fri.uni-lj.si, iztok.kosem@ff.uni-lj.si, simon.krek@ijs.si, cyprianadam.laskowski@ff.uni-lj.si, marko.robnik@fri.uni-lj.si*

## Abstract

By presenting the Thesaurus of Modern Slovene, the largest open-access collection of Slovene synonyms, this paper describes the concept of a responsive dictionary, a dictionary that allows its data to continuously respond to the changes in language and the feedback from the language community. We begin by briefly summarizing the method of its construction and its technical aspects. A great deal of deliberation and work has been put into interface design, with the aim to make the Thesaurus as user-friendly as possible for all digital media. This is followed by a more detailed description of the types of user input (e.g. synonym suggestions, synonym votes) and feedback (interface improvement suggestions) collected as part of development, as well as the methodology for their implementation. We also touch upon a series of dissemination activities aimed specifically at community building and user involvement. In conclusion, we describe our plans for the future, such as updates to be implemented in version 1.1 of the Thesaurus.

**Keywords**: responsive dictionary, digital lexicography, community, crowdsourcing, thesaurus, Slovene

## 1 Thesaurus of Modern Slovene

The recently published Thesaurus of Modern Slovene (http://viri.cjvt.si/sopomenke/) was created at the Centre for Language Resources and Technologies of the University of Ljubljana[1] as part of an effort to establish an infrastructure for Slovene that is comparable to the infrastructures of larger languages. In its current version, the Thesaurus contains 105,473 headwords and 368,117 synonyms, making it the largest automatically generated open-access collection of Slovene synonyms. The database was developed using innovative computational approaches for optimizing data reusability and connectivity. The Thesaurus is based on a range of different language resources (Krek et al. 2017) and enables users to compare synonyms in their collocational context, as well as check their use in the Gigafida reference corpus of modern Slovene.[2] In terms of financial resources, the computer-assisted data preparation was significantly less demanding and more economical than manual processing, as well as significantly less time-consuming. This enables regular updates and upgrades of the resource, making the dictionary a dynamically evolving source of language information. Furthermore, the Thesaurus facilitates a number of ways to involve the user community into the database construction process. The newly developed platform allows users to evaluate entries through voting and add

1 https://www.cjvt.si/

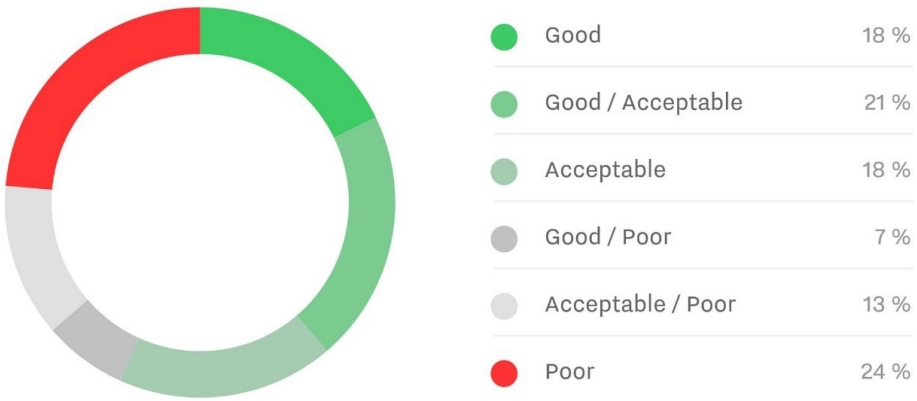
2 http://www.gigafida.net/

suggestions for new synonyms. User activities in the interface are tracked and user suggestions are incorporated into regular updates. The database of the Thesaurus is openly available at the Clarin.si repository under the Creative Commons Attribution-ShareAlike 4.0 International licence.[3]

### 1.1 Creation of the Thesaurus Database

A detailed description of the methodology used in the preparation of the Thesaurus database is presented in (Krek et al. 2017). Here, we summarise the basic idea of the method. The Thesaurus combines language data from two existing reference resources: The Oxford®-DZS Comprehensive English-Slovenian Dictionary (Šorli et. at 2006) and the Gigafida reference corpus of written Slovene (Logar and Krek 2012). Both resources comprise authentic language material published after 1991 and offer a solid basis for a description of modern Slovene. The extraction of synonym candidates was based on the manner in which words co-occurred in translation strings of the Oxford®-DZS Dictionary (e.g. *abandon > zapustiti; opustiti; odpovedati se, odstopiti od*). This information, together with the frequency of word co-occurrences, was the basis for discrimination between 'core' and 'near' synonyms, with 'core' synonyms exhibiting a greater degree of connection to the keyword. The identified links between synonyms were additionally confirmed using the older Dictionary of Standard Slovenian Language (SSKJ). The method formed balanced co-occurrence graphs and used the Personal PageRank algorithm (Page et al 1999) to automatically divide the synonyms into subgroups and rank them according to the degree of semantic relatedness to the keyword. Co-occurrence graphs were used to organize synonyms in the dictionary. The information on the frequency of keywords and synonym candidates in language use was also included in the database.

Expert evaluations (see Figure 1, based on the data from Krek et al. 2017) showed that, although not perfectly accurate, the presented method produced results of sufficient quality. The evaluation involved three linguists who independently rated automatically extracted synonyms from a set of 50 randomly chosen headwords from different part-of-speech categories (with 550 candidates in total). Possible ratings were good, acceptable, and poor. It should be noted that the boundaries of synonymy are difficult to define and heavily depend on the context and the circumstances of language use. Thus, even for expert evaluators, rating the data was not trivial, as seen from data on agreement in Figure 1. However, the percentage of synonyms with at least one positive score was 76.4%, which indicates that it is likely that some dictionary users would be interested in seeing them in the dictionary.




The evaluation has shown that the presence of the majority of synonyms that were rated as poor can be

3 https://www.clarin.si/repository/xmlui/handle/11356/1166

attributed to the methodology used and the content of the initial Oxford®-DZS database. In particular, these included literal repetitions within the phrase (*analiza* 'analysis' – *podrobna analiza* 'detailed analysis', *matematična analiza* 'mathematical analysis'), multi-word paraphrases similar to the headword (*povratno* 'reflexively' – *v refleksivni obliki* 'in the reflexive form'), masculine-feminine pairs (*nagajivka* 'mischievous girl' – *nagajivec* 'mischievous boy'), sense groups corresponding to the headword but not mutually overlapping in terms of meaning (*krutost* 'cruelty' – *nesramnost* 'rudeness', *strogost*, 'strictness', *neprijaznost* 'unfriendliness'), and words present in the Oxford®-DZS database but rarely used in Slovene (*grotesken* 'grotesque' – *hogarthovski* 'Hogarthian').

The current database is organized in a self-contained MySQL database, centered on the headword and synonym data, and enriched with corpus-derived collocations and examples (see Section 1.2). It currently contains 105,473 headwords, 368,117 synonyms, 3,353,061 collocations and 2,505,472 examples. It also stores and updates user contributions (synonym suggestions and votes, see Section 3). The Thesaurus interface (see Section 1.3) interacts with the database through a REST API, written in Python and Flask.

### 1.2 Inclusion of Collocations and Corpus Examples

The Thesaurus provides the possibility to explore synonymy in context with the use of corpus data. An important novelty for Slovene language resources in this regard is the option to compare the use of different synonyms with the help of their typical collocates (see Figure 4, which represents the collocations page of the Thesaurus, where e.g. the adjectives *pozitiven* 'positive' and *brezskrben* 'carefree' can be compared in context through their typical collocates: *pozitivna energija* 'positive energy' and *brezskrbne počitnice* 'carefree vacation'). The information for the comparison was obtained by using the Sketch Difference function in the Sketch Engine tool (Kilgarriff et al. 2004).

For each part-of-speech category, a selection of grammatical patterns was made that can be used when comparing synonyms through collocations. The selection was made in line with the preparation of the Collocations Dictionary of Modern Slovene (Kosem et al., in print). For nouns, the patterns are *adjective* + *noun* (*študijski program* 'study program', *študijski načrt* 'study plan'), *noun* + *preposition* + *noun* (*program za prihodnost* 'program for the future', *načrt za prihodnost* 'plan for the future'), *verb* + *preposition* + *noun* (*vključiti v program* 'include in the program', *vključiti v načrt* 'include in the plan'), and *verb* + *noun* (*predstaviti program* 'present a program', *predstaviti načrt* 'present a plan'). For verbs, the patterns are *verb* + *preposition* + *noun* (*presenetiti pri dejanju* 'surprise in the act', *zasačiti pri dejanju* 'catch in the act'), *verb* + *noun* (*presenetiti vlomilca* 'surprise an intruder', *zasačiti vlomilca* 'catch an intruder in the act'), *adverb* + *verb* (*ponovno presenetiti* 'surprise again', *ponovno zasačiti* 'catch in the act again'). For adjectives, the patterns are *adjective* + *noun* (*pozitivna ocena* 'positive evaluation', *spodbudna ocena* 'encouraging evaluation'), *adjective* + *preposition* + *noun* (*pozitiven za gospodarstvo* 'positive for the economy', *spodbuden za gospodarstvo* 'stimulating for the economy'), and *adverb* + *adjective* (*zelo pozitiven* 'very positive', *zelo spodbuden* 'very encouraging'). For adverbs, the patterns are *adverb* + *verb* (*odločno zanikati* 'firmly deny', *ostro zanikati* 'strongly deny'), *adverb* + *adjective* (*odločno zavrnjen* 'firmly rejected', *ostro zavrnjen* 'strongly rejected'), and *adverb* + *preposition* + *noun* (*odločno proti predlogu* 'firmly against the suggestion', *ostro proti predlogu* 'strongly against the suggestion').

In addition, examples of use were imported into the dictionary using computational methods for the automatic recognition of good (dictionary) examples (Kilgarriff et al. 2008, Kosem 2017). Collocations and examples of use are included in most entries, and all the entries also contain links to the concordances of a particular collocation in the Gigafida corpus, which provide information on the use of the collocation in different contexts. Additional contextual information is provided by domain labels

(e.g. *biologija* 'biology'), which were added from the Oxford-DZS Comprehensive English-Slovenian Dictionary and help explain the context of use for individual synonyms.

### 1.3 Dictionary Interface Design

In recent years, the Slovene digital lexicography has focused on the systematic collection of empirical data on the habits, wishes and needs of Slovene language users, as well as the possibilities for user involvement in the creation of language resources, primarily through crowdsourcing (Arhar Holdt et al. 2017, Arhar Holdt et al. 2016, Čibej et al. 2016, Gorjanc et al. (ed.) 2017), Čibej et al. 2015). The results of these studies were the basis for the design of the Thesaurus interface. In addition, feedback from future evaluations (possibly collected and analysed through a semi-automatic approach) will be used for further improvement. The interface was developed over a period of approximately one year, along with the concept of the responsive dictionary and the visual identity of the resource. When designing the interface, we took into account the following principles: speed and ease of use (e.g. search bar auto-complete, user-friendly data display, minimal number of clicks), clarity (a light, uncrowded display without unnecessary elements), step-by-step navigation, and motivation for user participation (e.g. Hall of Fame, see Section 2).

### 1.4 Responsive Dictionary as a Concept

With the Thesaurus of Modern Slovene, we are introducing a new type of dictionary called the *responsive dictionary,* so named because it enables the data to continuously respond both to changes in language and the feedback from the language community. Furthermore, it is defined by the following features:

- From the first, the responsive dictionary is developed specifically for the digital medium.
- The initial dictionary database is constructed using advanced computational methods, instantly providing the language community with a large quantity of relevant, albeit still somewhat noisy language information.
- The dictionary interface provides a number of ways for the community to contribute to the expansion of the database and clean up noisy elements.
- The development of a responsive dictionary is never concluded as its data constantly evolves in response to the developments in modern language. The changes are tracked using timestamps in individual entries, while the different versions of the database are stored in a dedicated archive.
- Dictionary development follows a clearly defined methodology for including user contributions and implementing improvements based on information systematically collected from user activities in the dictionary.
- The dictionary and its database are openly accessible under an appropriate license (e.g. CC-BY-SA 4.0).

## 2 Community Building and Inclusion

The Thesaurus of Modern Slovene provides a number of different ways for user involvement. While user involvement is not a new concept in lexicography, our approach differs from collaborative lexicographic practices (Abel and Meyer 2013) that anticipate editorial control of user-added content (e.g. Macmillan Open Dictionary, Leo, Openthesaurus.de) in order to prevent the addition of noisy data (Cristea et al. 2014), or editorial interventions that implement user feedback after the dictionary is already completed (De Schryver and Prinsloo 2000). The responsive dictionary integrates user contributions directly into the process of creating and maintaining the dictionary, while potential

editorial decisions can be taken in consensus/comparison with the decisions of language users. In the dictionary, users can evaluate synonyms and directly add their own suggestions. Technical support for automatically importing existing synonym collections into the Thesaurus is also provided. In addition, users can participate in a public debate on dictionary features, and also have the opportunity to join a group of dedicated user evaluators, i.e. community members that are invited to participate in user evaluations of new versions of the Thesaurus. All these activities are described in more detail in subsections 2.1-2.3.

However, user participation cannot be expected without sufficient user motivation. In the current version of the interface, this challenge is tackled by incorporating motivational elements used in crowdsourcing projects, e.g. the Hall of Fame (a list of users who added the most synonyms) and the Tasks-of-the-Day section (a daily random list of headwords with no user votes; see Figure 2). As for community building, a website, a Facebook profile and a newsletter were created in order to inform the public about the Thesaurus. In addition, a separate project[4] aimed at organizing events for the promotion of the Thesaurus is being carried out in 2018-2019 and will hopefully boost the number of participants in the initial stage of the process as well as provide answers to several open issues: what type of user motivation is the most effective (and what group does it influence the most), how do promotional events affect the number of dictionary users, etc.

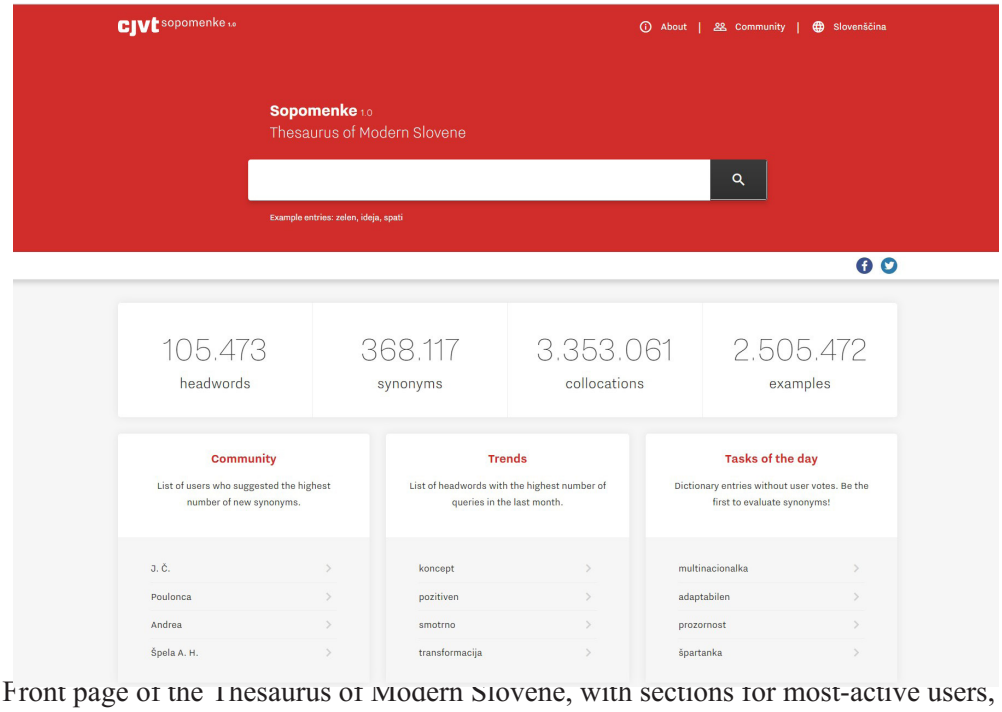



Figure 2: Front page of the Thesaurus of Modern Slovene, with sections for most-active users, trends and tasks of the day.

## 2.1 Synonym Evaluation

As can be seen from Figure 3 showing the synonyms of *pozitiven* ('positive'), the synonym candidates listed under individual thesaurus headwords can be rated as good (appropriate, useful) or poor (noisy, incorrect). In Figure 3, the synonym *pohvalen* ('commendable') has been rated as appropriate, as indicated by the green voting button (displaying a value of 1). Users also have the option to filter the candidates by user rating. With dictionary updates, the ratings will be taken into account in order

4 https://www.cjvt.si/promocija-sopomenk/

to re-arrange or remove data accordingly (e.g. if the users predominantly identify a synonym as inadequate, it can be removed from the entry). User evaluation is not just limited to user-added synonyms but extends to synonyms included in the initial database as well. This allows the community not only to evaluate whether user-added suggestions are valid and constructive, but also to identify potential noisy candidates that are present in the database because of automatic database generation.

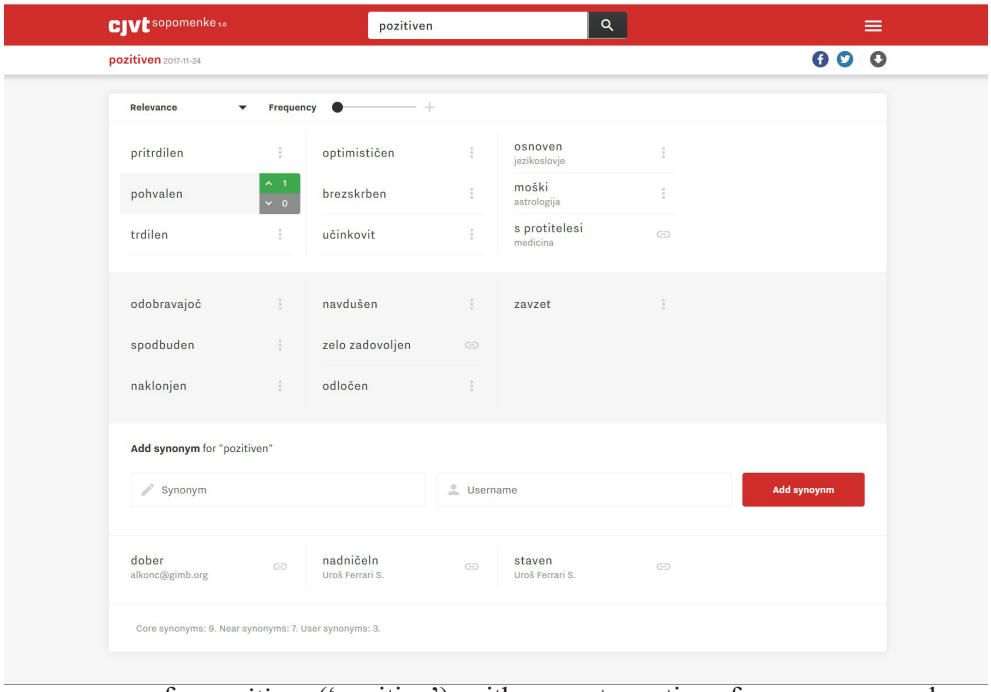



Figure 3: Synonym page for *pozitiven* ('positive'), with separate sections for core, near and user synonyms.

## 2.2 Addition of New Synonyms

Each headword contains a list of core and/or near synonyms (core synonyms are more closely linked to the headword compared to near synonyms; for more detail, see Krek et al. 2017). At the bottom of each entry, a third section is provided for user-added synonyms (Figure 3). This section consists of a form in which users enter their synonym suggestion and user name (no registration is required.). Suggestions are displayed in the entry immediately after submission. As already mentioned, technical support is also provided to automatically import existing synonym collections into the Thesaurus. However, in order to avoid malicious additions and bot entries, the process of adding synonyms is not without restrictions. Once a user has entered a synonym suggestion, the form is substituted by a link that needs to be clicked in order to reopen it. In addition, a number of user activities in the dictionary are logged (more on this in Section 3), which provides enough metadata (e.g. time of submission, IP-address) to remove all contributions from a user that would be identified as malicious. The addition of new synonyms is motivated with the Hall of Fame (a list of users who added the most synonyms), while constructive suggestions are encouraged with the list of best-rated user suggestions.

## 2.3 Discussion of Existing Solutions

Users that want to provide direct feedback on the interface can either send an e-mail to the development team or participate in a discussion that takes place in a dedicated public Facebook group, where users can share links to individual thesaurus entries, ask questions on dictionary data or related language problems, suggest improvements, etc. For instance, since the official publication of the

Thesaurus in March 2018, users have expressed a number of specific needs that should be addressed. First, the users would like to add labels to provide additional contextual information on the suggested synonyms (as shown e.g. by the user suggestion *arcnije (pogovorno)*, '*arcnije* (colloquial)' added to the headword *zdravilo* 'cure'). Second, in addition to adding synonyms to existing entries, the users would like to add new (or missing) headwords to the dictionary. Third, the users would like to have the option to evaluate user suggestions in one place. In the current version of the interface, user suggestions can only be evaluated within their respective headwords and there is no concise overview of user suggestions (either an overview by individual users or in general). User opinions will be taken into account when preparing thesaurus updates.

## 3 User Data Collection and Dictionary Upgrades

A vast array of data on user activities in the Thesaurus are recorded for further use. User activities are defined as events of different types, as presented in Table 1. For every event, a timestamp and the IP of the user are recorded. In this manner, a sequence of events in a specific session can be reconstructed and potentially malicious entries identified.

Table 1: Data on User Activities.

| Event Type | Description | Source Pages |
|---|---|---|
| static_page_visit | When a user visits a specific static page of the interface (e.g. the pages with information on the Thesaurus), the log records the visit. | main_page, about_page, community_page, versions_page, impressum_page, 404_page |
| link_click | The log records when a user clicks on a link that leads to an external page (e.g. the Thesaurus database in the Clarin.si repository). | URL (string) |
| headword_select | When a user enters a headword, the log records: (I) the entered headword; (II) where in the interface the action took place (from a search window on a specific subpage or by clicking on the example entry links, Tasks of the Day; Trends, or links from Facebook and Twitter). | synonyms_page, collocations_page, main_page, main_page_example, main_page_random, main_page_popular, facebook, twitter |
| headword_not_found | When a user searches for a headword missing in the database, the log records the provided search string. | headword (string) |
| synonym_select | When a user selects one of the synonyms on the synonyms page, the log records: (I) the selected synonym (II) the headword under which the synonym is listed; (III) where in the interface the selection took place (e.g. on the synonyms page; from the side menu on the collocations page; from the side menu by using the buttons for previous/next synonym). | synonyms_page, collocations_page, collocations_page_btn_next, collocations_page_btn_prev |
| collocation_select | When a user clicks a collocate to access the corresponding corpus examples, the log records the set of collocational data the collocate is from. | collocations_page |
| synonym_go_to_gigafida | When a user clicks the button with the link to the Gigafida corpus from the synonyms page, the log records: (I) the selected synonym, (II) the synonym's headword; (III) the URL of the link. | synonyms_page |

| Event Type | Description | Source Pages |
|---|---|---|
| example_go_to_gigafida | When a user clicks the button with the link to the Gigafida corpus from the section with corpus examples, the log records: (I) the set of collocational data the collocate is from, and (II) the URL of the link. | collocations_page |
| synonym_suggest | When a user adds a synonym, the log records: (I) the added synonym and (II) the synonym's headword; (III) the provided username; (IV) whether the entry was successful or not (e.g. if the proposed synonym already existed in the database, which prompts a warning). | synonyms_page |
| headword_download | When a user downloads entry data, the log records: (I) the current headword; (II) the subpage from which the download button was clicked. | synonyms_page, collocations_page |
| synonyms_frequency_slider_change | When a user filters synonyms by frequency, the log records: (I) the selected headword, (II) the subpage where the filtering took place, and (III) the type of slider change (left or right). | synonyms_page, collocations_page |
| synonyms_sort | When a user sorts synonyms according to their features, the log records: (I) the selected headword, (II) the subpage where the sorting took place, and (III) the selected type of sorting (alphabetical, by synonym length, etc.) | synonyms_page, collocations_page |
| synonym_vote | When a user votes on a synonym, the log records: (I) the synonym the user is voting on and (II) the synonym's headword, (III) whether the vote was added or removed, (IV) the value of the vote (+1 or –1). | synonyms_page, collocations_page |
| language_change | The log records when a user changes the language of the interface (Slovene to English or vice versa). | language (slv, eng) |
| share | When a user shares the link to the dictionary content on social media, the log records: (I) the shared URL (II) the source medium (Facebook or Twitter) | medium (Facebook, Twitter), URL |

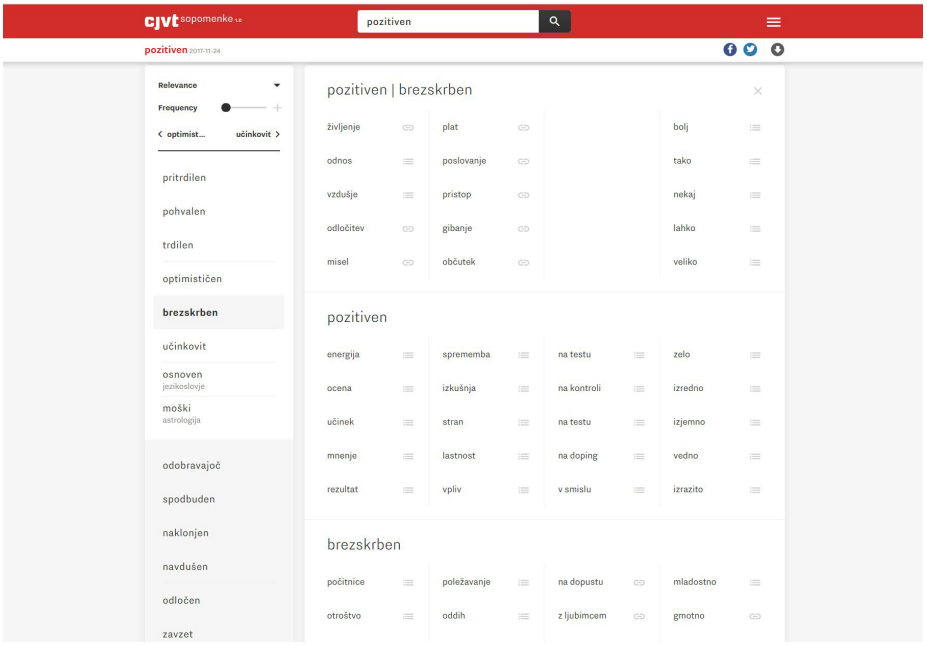



Figure 4: Collocations page with collocates for *pozitiven* ('positive') and *brezskrben* ('carefree').

## 4 Conclusion and Future Work

In the paper, we have presented the Thesaurus of Modern Slovene and the main ideas underlying its construction. The first upgrade of the Thesaurus is planned for the end of 2018 and will provide an opportunity to evaluate and refine the methodology for further steps in dictionary development. Apart from the plans already mentioned in the previous sections of the paper, we point out here the most essential steps of our future work. Firstly, we plan to extend the core database with new synonyms using dense neural vector embeddings (Mikolov et al. 2013), an unexploited source of synonymy detection which projects similar words into a latent vector space where distances preserve the similarity among words. Clustering in this space may provide new candidates for synonyms, especially with variants of embeddings that construct more than one vector per polysemous word (Huang et al. 2012, Peters et al. 2018). Secondly, we intend to use the Pybossa crowdsourcing platform (https://pybossa.com/) to prepare a number of specific crowdsourcing tasks to improve the existing database (e.g. removing archaic or idiosyncratic synonyms from literary works). Thirdly, the KOLOS national project (*Collocations as a Basis for Language Description: Semantic and Temporal Perspectives,* ARRS J6-8255) will provide a framework for improving the quality of the collocational data in the Thesaurus. Finally, studies are planned to further address different aspects of user participation and involvement. As mentioned in Section 1.1, the results of the linguistic evaluation regarding the appropriateness of the synonyms included in the core database were not uniform. This can be explained by the fact that synonymy is a contextual linguistic phenomenon, where the true value of synonyms can be evaluated exclusively in their context and within the exact purpose of the user. The collective authorship of the responsive dictionary can thus provide valuable insight regarding the extent to which user knowledge corresponds to expert linguistic assessments. On the other hand, the perception of the language community regarding the novelties brought by the responsive dictionary concept will also be addressed: a study is underway combining think-aloud protocols and semi-structured interviews with representatives of selected user groups in order to obtain feedback on dictionary functionality (e.g. the usefulness and the clarity of the interface), as well as opinions on the democratization of the dictionary creation process, different issues regarding user involvement (e.g. the question of authorship, ethical issues, and quality control), the dynamic nature of a responsive dictionary, etc. The results will help pinpoint the potential weaknesses of the concept that will have to be given priority in the future.

## References

Arhar Holdt, Š., Kosem, I. & Gantar, P. (2016). Dictionary user typology: the Slovenian case. In T. Margalitadze & G. Meladze (eds) *Lexicography and linguistic diversity: proceedings of the XVII EURALEX International Congress*. Tbilisi: Ivane Javakhishvili Tbilisi State University, pp. 179-187.

Arhar Holdt, Š., Čibej, J. & Zwitter Vitez, A. (2017). Value of language-related questions and comments in digital media for lexicographical user research. *International journal of lexicography*, 30 (3), pp. 285-308.

Cristea, D. Forăscu, C., Răschip, M. & Zock, M. (2014). How to Evaluate and Raise the Quality in a Collaborative Lexicographic Approach. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008*, Marrakech, Morocco.

Čibej, J., Fišer, D. & Kosem, I. (2015). The role of crowdsourcing in lexicography. In I. Kosem et al. (eds.) *Electronic lexicography in the 21st century: linking lexical data in the digital age*. Ljubljana: Trojina, Institute for Applied Slovene Studies; Brighton: Lexical Computing, pp. 70-83.

Čibej, J., Gorjanc, V. & Popič, D. (2016). XVII EURALEX International Congress, 6-10 September, 2016, Tbilisi. Analysing translators' language problems (and solutions) through user-generated content. In T. Margalitadze & G. Meladze (eds) *Lexicography and linguistic diversity: proceedings of the XVII EURALEX International Congress*. Tbilisi: Ivane Javakhishvili Tbilisi State University, pp. 158-167.

De Schryver, G. M. & Prinsloo, D. J. (2000). Dictionary-Making Process with 'Simultaneous Feedback' from the Target Users to the Compilers. In U. Heid, S. Evert, E. Lehmann & C. Rohrer (eds): *Proceedings of the 9th EURALEX International Congress.* Institut für Maschinelle Sprachverarbeitung: Stuttgart, Germany, pp. 197-209.

Gorjanc, V., Gantar, P., Kosem, I. & Krek, S. (eds.) (2017). *Dictionary of modern Slovene: problems and solutions*. Ljubljana: Ljubljana University Press, Faculty of Arts.

Huang, E. H., Socher, R., Manning, C. D. & Ng, A. Y. (2012). Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, pp. 873-882.

Kilgarriff, A., Rychly, P., Smrz, P., Tugwell, D. (2004). The Sketch Engine. In G. Williams & S. Vessier (eds.) *Proceedings of the Eleventh EURALEX International Congress*, EURALEX 2004 Lorient, France July 6–10, 2004. Lorient: Universite de Bretagne-sud, pp. 105–116.

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.

Kosem, I. (2017). Dictionary examples. In V. Gorjanc, P. Gantar, I. Kosem and S. Krek (eds.) *Dictionary of Modern Slovene: problems and solutions.* Ljubljana: Univerza v Ljubljani, Filozofska fakulteta, pp. 174-194.

Kosem, I., Krek, S., Gantar, P., Arhar Holdt, Š., Čibej, J. & Laskowski, C. (in print). Collocations Database and 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 University Press, Faculty of Arts: Ljubljana.

Krek, S., Laskowski, C. & Robnik-Šikonja, M. (2017). From translation equivalents to synonyms: creation of a Slovene thesaurus using word co-occurrence network analysis. In I. Kosem et al. (eds.) *Proceedings of eLex 2017: Lexicography from Scratch, 19-21 September 2017.* Leiden, Netherlands, pp. 93-109.

Logar, N. & Krek, S. (2012): New Slovene corpora within the "Communication in Slovene" project. In *Prace Filologiczne,* 63, pp. 197-207.

Meyer, C. M. & Abel, A. (2017): User Participation in the Era of the Internet. In P. A. Fuertes Olivera (ed.): *The Routledge Handbook of Lexicography*. London: Routledge, pp. p. 735-753.

Mikolov, T., Yih, W.T. & Zweig, G. (2013). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* pp. 746-751.

Page, L., Brin, S., Motwani, R. & Winograd, T. (1999). *The PageRank Citation Ranking: Bringing Order to the Web*. Technical Report. Stanford InfoLab.

Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. and Zettlemoyer, L. (2018). Deep contextualized word representations. *Proceedings of NAACL*.

SSKJ, Slovar slovenskega knjižnega jezika [Electronic version] (2014). A. Bajec et al. (eds.). Ljubljana: Založba ZRC, Znanstvenoraziskovalni center SAZU. Available at: http://www.fran.si.

Šorli, M., Grabnar, K., Krek, S. & Košir, T. (2006). Oxford-DZS comprehensive English-Slovenian dictionary. In *Proceedings XII EURALEX international congress.* Edizioni dell'Orso: Universita di Torino: Academia della Crusca, pp. 631-637.

## Acknowledgements

The paper was prepared as part of two national projects, *Nova slovnica sodobne standardne slovenščine: viri in metode* (*New grammar of contemporary standard Slovene: sources and methods,* ARRS J6-8256) and *Kolokacije kot temelj jezikovnega opisa: semantični in časovni vidiki* (*Collocations as a Basis for Language Description: Semantic and Temporal Perspectives*, ARRS J6-8255). The development of the Thesaurus was funded by the ARRS P6-0215 research program (*Slovene language – basic, contrastive, and applied studies*) and the infrastructural programme of the Centre for Language Resources and Technologies at the University of Ljubljana. The interface was developed by Studio Kruh in collaboration with Leon Noe Jovan. The crowdsourcing development benefited from the COST Action CA16105: *European Network for Combining Language Learning with Crowdsourcing Techniques.*