Crowdsourcing for Slovene Language Resources

COST Action CA16105 (enetCollect): SIMON KREK, IZTOK KOSEM, ŠPELA ARHAR HOLDT, JAKA ČIBEJ

Centre for Language Resources and Technologies, University of Ljubljana, Slovenia

Funded by the Horizon 2020 Framework Programme of the European Union

CENTRE FOR LANGUAGE RESOURCES AND TECHNOLOGIES

The Centre for Language Resources and Technologies at the University of Ljubljana (CJVT) is a research unit of the University of Ljubljana aimed at scientific research, and the development and maintenance of key digital language resources and language technology applications for contemporary Slovene. CJVT was founded amidst concerns about the unfavourable state of research on modern Slovene as well as language technologies and resources for Slovene. The University of Ljubljana took on the responsibility to improve the situation by forming an institution that could ensure a systematic long-term development of technologies, resources and tools for Slovene, enabling it to keep up with other languages in the digital world.

CJVT promotes a user-oriented development of language resources and technologies by systematically conducting empirical studies on user habits, preferences, and needs, and by examining the opportunities for user participation and involvement at different levels of the process. In this context, crowdsourcing offers possibilities for workflow optimisation and the formation of user communities that serve not only as recipients but also as active contributors in the development of language resources.

DICTIONARY OF MODERN SLOVENE

The construction of the Dictionary of Modern Slovene (Gorjanc et al. 2017) is an on-going project led by CJVT in an effort to design a modern, corpus-based dictionary database of Slovene, with the aim of providing speakers of Slovene with a reliable source of lexico-grammatical information as well as a fundamental lexical and grammatical resource for the development of new language technologies for Slovene.

COLLOCATION DICTIONARY OF SLOVENE

Currently in the pre-processing phase, the Collocation Dictionary of Slovene aims to be the first collocation dictionary of Slovene, based on the Collocation Dictionary Database of Slovene, which consists of collocation candidates automatically extracted from the Gigafida reference corpus of standard Slovene.

The dictionary aims to implement crowdsourcing in the data pre-processing phase in order to:

- filter tokenisation errors,
- categorise collocations according to headword sense,
- filter out unsuitable n-grams,
- check syntactic structures.

Several crowdsourcing tasks are currently being tested using a local installation of the PyBossa crowdsourcing platform (http://pybossa.com/).

PyBossa Task: Sorting collocations under headword senses.

The screenshot above shows an example of a crowdsourcing task aimed at sorting collocations in the correct headword sense. In this case, the crowdsourcer must determine whether the headword *bazar* of the collocation *obiskati bazar* (to visit a bazaar) in this context signifies an oriental marketplace (*orientalska tržnica*) or an event (*prireditev*). In case none of the senses apply, the crowdsourcer can also click 'none of the above', 'bad example', 'incorrect structure', or 'extended collocation'.

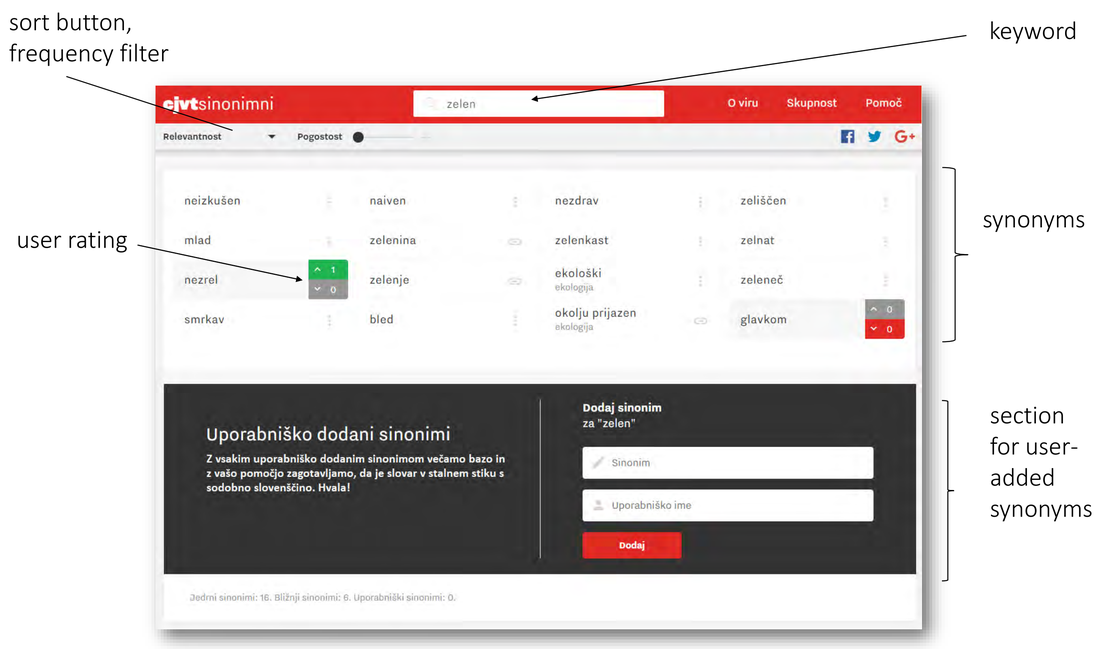
SYNONYM DICTIONARY OF SLOVENE

Currently close to completion, the Synonym Dictionary of Slovene is the largest open-source automatically generated collection of Slovene synonyms (Krek et al. 2017). In the current version, it contains approximately 78.000 entries and 327.000 synonyms. Unlike similar projects, the Synonym Dictionary of Slovene is based on an array of different databases and offers important new features:

- the possibility to compare synonyms in terms of their context and use,
- links to concordances in the Gigafida reference corpus of standard Slovene.

The dictionary will be freely accessible in a specially designed online interface and will feature user involvement in the post-processing phase. Dictionary users will be able to:

- directly add missing synonyms in a separate section for user-added synonyms,
- rate the relevance of synonyms (including user-added synonyms) in the dictionary. Ratings will be taken into account when the dictionary is upgraded.
- provide suggestions for further interface/resource development in a dedicated Facebook group or via a dedicated e-mail address.



Synonym Dictionary of Slovene: Interface.

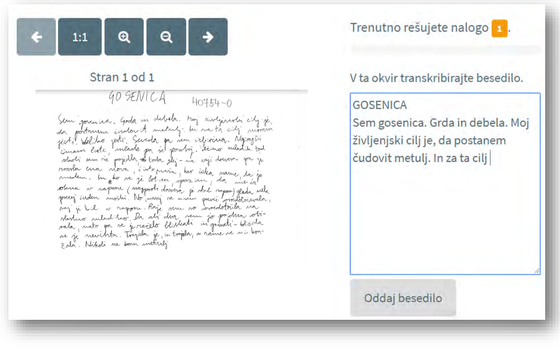
THE ŠOLAR CORPUS AND PEDAGOGICAL GRAMMAR PORTAL

www.korpus-solar.net http://slovnica.slovenscina.eu/

The Šolar corpus and the Pedagogical grammar portal (Arhar Holdt et al. 2017) were created as part of the Communication in Slovene (www.slovenscina.eu) and Šolar 2.0 (http://solar.trojina.si/) projects.

Crowdsourcing tasks that will ease further development:

- transcription and digitalisation of student texts,
- categorisation of teacher corrections,
- selection and further preparation of examples for grammar chapters and language exercises.



PyBossa Task: Transcription into digital form.

REFERENCES

Arhar Holdt, Š., Kosem, I., Gantar, P. 2017. Corpus-Based Resources for L1 Teaching: The Case of Slovene. In A. Marcus-Quinn & T. Hourigan (Eds.), Handbook on Digital Learning for K-12 Schools (pp. 91–113). Springer International Publishing.

Čibej, J., Fišer, D., Kosem, I. 2015. The role of crowdsourcing in lexicography. In Kosem, I., Jakubiček, M., Kallas, J., Krek, 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. 70-83.

Gorjanc, V., Gantar, P., Kosem, I., Krek, S. (eds.). 2017. Dictionary of Modern Slovene: Problems and Solutions. Ljubljana: Ljubljana University Press, Faculty of Arts. Available online: http://www.ff.uni-lj.si/sites/default/files/Dokumenti/Knjige/e-books/dictionary_of_modern_slo.pdf

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. Ljubljana: Trojina, Institute for Applied Slovene Studies; Brighton: Lexical Computing.

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.