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

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

1Faculty of Computer and Information Science, University of Ljubljana,
2Faculty of Arts, University of Ljubljana,
3"Jožef Stefan" Institute
Thesaurus of Modern Slovene

- A freshly published (March 2018) **fully automatically generated** resource.
- 105,473 headwords, 368,117 synonyms, 3,353,061 collocations and 2,505,472 corpus examples = largest **open-access** collection of Slovene synonyms.
- Freely available for use at [http://viri.cjvt.si/sopomenke/eng/](http://viri.cjvt.si/sopomenke/eng/) with both Slovene and English interface: **you are welcome to test it now!**
- The database is openly available at the Clarin.si repository (CC-Attribution-ShareAlike 4.0 International) – data can be used for further development: [https://www.clarin.si/repository/xmlui/handle/11356/1166](https://www.clarin.si/repository/xmlui/handle/11356/1166).
Centre for Language Resources and Technologies

• Interdisciplinary research unit at University of Ljubljana (Faculty of Arts, Faculty of Computer and Information Science, Faculty of Social Sciences, Faculty of Education, Faculty of Electrical Engineering).

• Aimed at scientific research and the development of language infrastructure (resources and technologies) for contemporary Slovene.

• Focus on user research (empirical approaches to problems, opinions, habits, and needs of language users in the context of applied linguistics), and user involvement into the development of digital language resources.

• https://www.cjvt.si/en/
Responsive Dictionary

• Responds to language change and the feedback from the language community.
• Developed specifically for the digital medium.
• The database is constructed using advanced computational methods,
• released to the community as soon as it is linguistically evaluated as relevant (albeit somewhat noisy),
• the community can contribute to improve the resource.
• A clearly defined methodology for including user contributions and information collected from user activities in the dictionary.
• The data evolves, the changes are transparently tracked and archived.
• The database is openly accessible under an appropriate license.
Data reusability and connectivity

- The Thesaurus is based on existing language resources: The Oxford®-DZS Comprehensive English-Slovenian Dictionary (Šorli et. at 2006) and the Gigafida reference corpus of written Slovene (Logar and Krek 2012).
- Methodology for automatic identification, extraction and organisation of synonym candidates: a detailed description in (Krek et al. 2017).
- The Thesaurus provides collocational data to compare the use of different synonyms with the help of their typical collocates (Kosem et al., in print), automatically selected and imported corpus examples (Kilgarriff et al. 2008, Kosem 2017), and links to corpus data in the Gigafida corpus.
User Involvement: synonym evaluation and addition of new synonyms

User activities in the interface are tracked and user suggestions are incorporated into regular updates.
User Activities in the Dictionary

• Data on user activities in the Thesaurus are recorded.
• User activities are defined as events of different types.
  • `static_page_visit`, `link_click`, `headword_select`, `headword_not_found`, `synonym_select`, `collocation_select`, `synonym_go_to_gigafida`, `example_go_to_gigafida`, `synonym_suggest`, `headword_download`, `synonyms_frequency_slider_change`, `synonyms_sort`, `synonym_vote`, `language_change`, `share`
• For every event, the timestamp and the IP of the user are recorded. In this manner, the sequence of events in a specific session can be reconstructed and potentially malicious entries identified.
User Motivation

• Motivational elements in the interface: the Hall of Fame (users with most suggested synonyms) and the Tasks-of-the-Day section (4 random headwords with no user votes).

• Community building: A website, a Facebook profile and a newsletter were created in order to inform the public about the Thesaurus.

• A separate promotional project funded by the Slovene Ministry of Culture is being carried out in 2018-2019 – bookmarks, brochures, posters, promotional events, educational material, a set of podcasts ...
The Role of Dissemination

- As the concept is new compared to existing language resources, dissemination activities address **questions and doubts** of the language community.

- Typical questions:
  - Is automatically extracted data at all reliable?
  - To what extent can automatically created and crowdsourced resources be normative?
  - Are users competent enough to contribute?
  - How do you handle malicious (and politically incorrect) entries?
  - Is the Slovene community large enough for the project to succeed?
  - How do you keep the users motivated to participate?
So far so good

<table>
<thead>
<tr>
<th></th>
<th>19. 4. 2018</th>
<th>19. 6. 2018</th>
<th>15. 7. 2018</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of user synonyms</td>
<td>409</td>
<td>1,930</td>
<td>2,336</td>
</tr>
<tr>
<td>User upvotes</td>
<td>1,797</td>
<td>3,207</td>
<td>3,412</td>
</tr>
<tr>
<td>User downvotes</td>
<td>699</td>
<td>1,213</td>
<td>1,318</td>
</tr>
<tr>
<td>Number of user nicknames</td>
<td>79</td>
<td>153</td>
<td>161</td>
</tr>
<tr>
<td>Number of headwords not found</td>
<td>635</td>
<td>2,179</td>
<td>2,830</td>
</tr>
<tr>
<td>Number of headwords searched</td>
<td>30,195</td>
<td>37,631</td>
<td>39,391</td>
</tr>
<tr>
<td>Number of IP-addresses</td>
<td>7,101</td>
<td>12,879</td>
<td>14,816</td>
</tr>
</tbody>
</table>
So far so good

<table>
<thead>
<tr>
<th>Username</th>
<th>Headword</th>
<th>Suggested synonym</th>
</tr>
</thead>
<tbody>
<tr>
<td>rozaa</td>
<td>skladiščni prostor</td>
<td>shramba</td>
</tr>
<tr>
<td>Red Mosquito</td>
<td>rajati</td>
<td>plesati</td>
</tr>
<tr>
<td>flaša</td>
<td>napiti se</td>
<td>nažgati se ga</td>
</tr>
<tr>
<td>Poulonca</td>
<td>eksplicitno</td>
<td>neposredno</td>
</tr>
<tr>
<td>helena</td>
<td>internet</td>
<td>net</td>
</tr>
<tr>
<td>Vestalka</td>
<td>zgodovinar</td>
<td>historiograf</td>
</tr>
<tr>
<td>Nava</td>
<td>zelen</td>
<td>neuk</td>
</tr>
<tr>
<td>Anja</td>
<td>sigurno</td>
<td>zagotovo</td>
</tr>
<tr>
<td>Deda</td>
<td>odlično</td>
<td>Bolano</td>
</tr>
<tr>
<td>terand58</td>
<td>zdravilo</td>
<td>arcnije (pogovorno)</td>
</tr>
</tbody>
</table>
User suggestions for improvements

- videotutorials on interface functionalities,
- labels (e.g. in the form of a dropdown menu),
- the possibility to add headwords not yet included,
- the possibility to evaluate all user synonyms in one place,
- a more informative Hall of Fame.

- ... in addition to our planned future work on the database.
Before the first major upgrade we want to

1. Improve the data extraction methodology to extend the database.
2. Use the Pybossa crowdsourcing platform to prepare microtasks to remove noise from the database.
3. The KOLOS national project (Collocations as a Basis for Language Description: Semantic and Temporal Perspectives, ARRS J6-8255) will help improve the quality of the collocational data in the Thesaurus.
4. Studies to further address different aspects of user participation and involvement, e.g. the perception of the language community regarding the novelties brought by the responsive dictionary, their attitude towards user involvement in lexicography, etc.

Euralex, Ljubljana, 18. 7. 2018
Thank you!

Špela Arhar Holdt\textsuperscript{12}, Jaka Čibej\textsuperscript{123}, Kaja Dobrovoljc\textsuperscript{13}, Polona Gantar\textsuperscript{2}, Vojko Gorjanc\textsuperscript{2}, Bojan Klemenc\textsuperscript{1}, Iztok Kosem\textsuperscript{23}, Simon Krek\textsuperscript{23}, Cyprian Laskowski\textsuperscript{2}, Marko Robnik-Šikonja\textsuperscript{1}

\textsuperscript{1}Faculty of Computer and Information Science, University of Ljubljana, \textsuperscript{2}Faculty of Arts, University of Ljubljana, \textsuperscript{3}"Jožef Stefan" Institute
References


