

Repel the Syntruders! A Crowdsourcing Cleanup of the Thesaurus of Modern Slovene

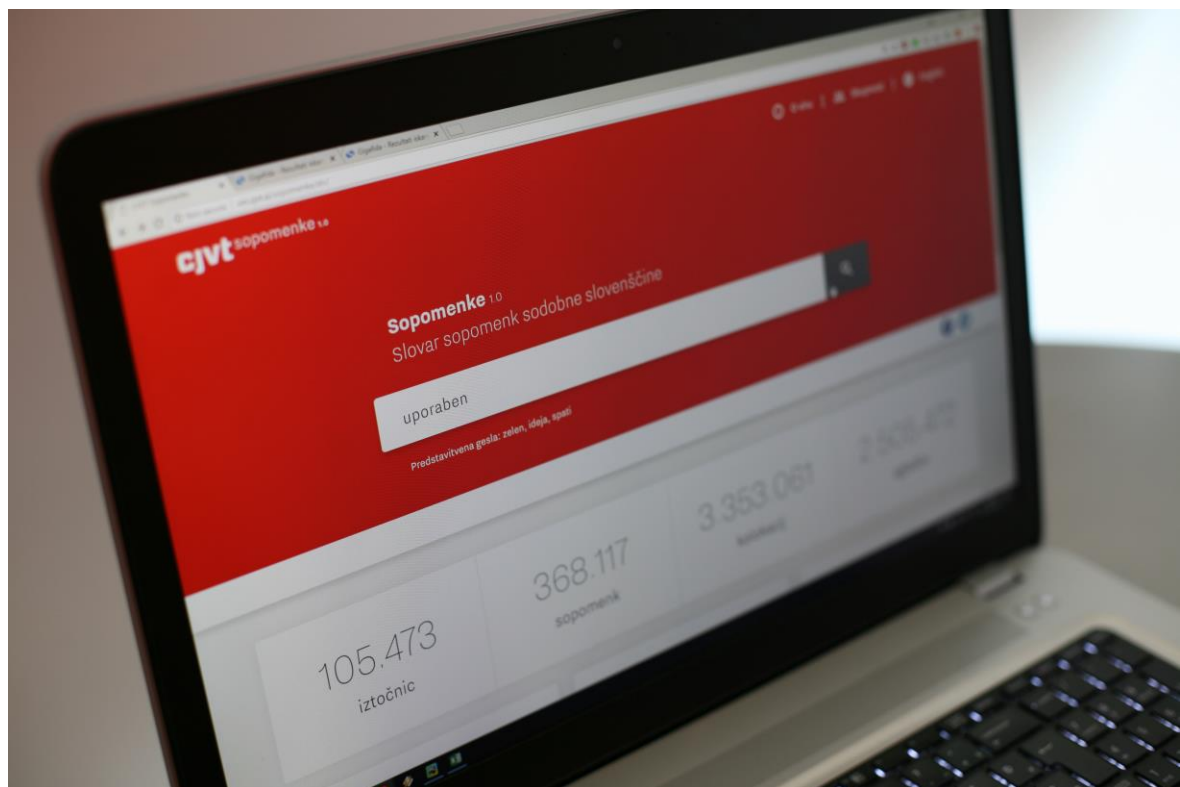
Jaka Čibej, Špela Arhar Holdt

Centre for Language Resources and Technologies, University of Ljubljana
(Faculty of Arts, Faculty of Computer and Information Science)

eLex 2019 – Sintra, Portugal, 1. 10. 2019

The Thesaurus of Modern Slovene

- the **largest open-source digital collection of Slovene synonyms**
- published in **March 2018** by the Centre of Language Resources and Technologies of the University of Ljubljana



- CJVT Sopomenke
- <https://viri.cjvt.si/sopomenke>

A new type of language resource

- **responsive dictionary** (Arhar Holdt et al. 2018)
- born-digital and digital-only dictionary
- initial database compiled automatically (Krek et al. 2017)
- immediately openly accessible
- frequently updated; all changes are tracked
- *user involvement*




- **responds to user feedback and language change**



User involvement in the Thesaurus

- adding suggestions for missing synonyms

Add synonym for "kralj"

 visokost

 eLex2019 Participant

Add synonym

suveren
K10



maziljenec
RD



božji izbranec
RD



Core synonyms: 3. Near synonyms: 1. User synonyms: 3.

- no registration needed

User involvement in the Thesaurus

- up- or downvoting synonym candidates in the dictionary

kralj 2017-11-24

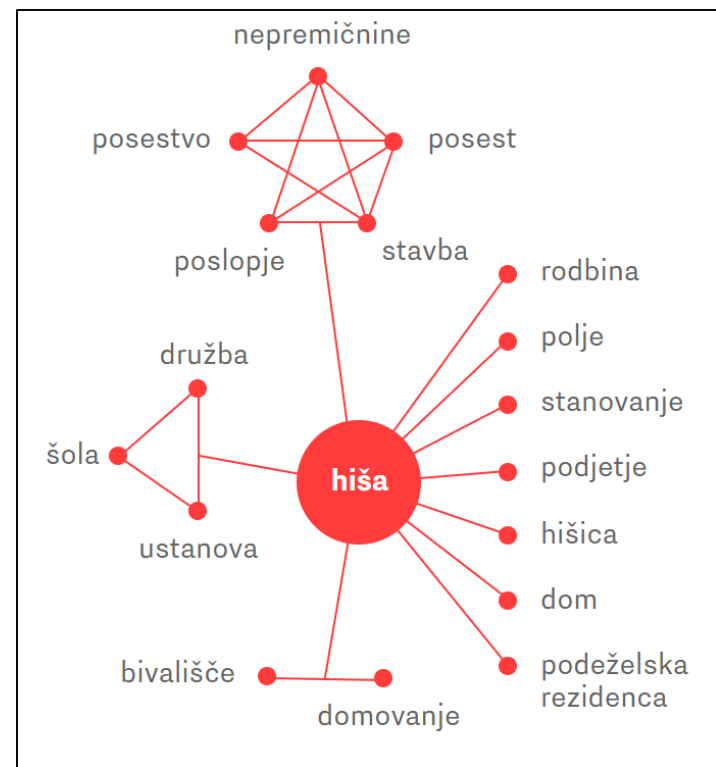
The screenshot shows a user interface for a thesaurus. At the top, there are two controls: 'Relevance' with a dropdown arrow and 'Frequency' with a slider and a plus sign. Below these, a list of synonym candidates is displayed. The first candidate, 'magnat', has a green button with an up arrow and the number '1', and a grey button with a down arrow and the number '0'. The second candidate, 'mogotec', has a vertical ellipsis menu icon. The third candidate, 'šef', also has a vertical ellipsis menu icon.

Relevance	Frequency
magnat	^ 1 v 0
mogotec	⋮
šef	⋮

Why even involve dictionary users?

- automatic extraction → noise
- **multi-word synonym candidates**
- The Oxford®-DZS Comprehensive English-Slovenian Dictionary
 - **extracted through co-occurrence graphs**

dom; bivališče, domovanje; stanovanje; hiša



Problematic multi-word synonym candidates

- **descriptive approximates of concepts that are lexicalized in English but not in Slovene**
 - *rooming-in* (as an English loanword)
 - *24-urno sobivanje novorojenčka in matere* ('a 24-hour cohabitation of a newborn and their mother')
 - **feminine-masculine word pairs**
 - *učitelj* 'teacher [masculine]' – *učiteljica* 'teacher [feminine]'
 - **paraphrases, definitions, partial repetitions, or descriptions**
 - *pozorno opazovati koga skozi monokel* 'to observe someone attentively through a monocle'
- the most obviously problematic category (as well as manageable in size)
- **Task 1: Targeted campaign**
- **Task 2: User votes in the dictionary**

Task 1: Data preparation

- ~368,000 headword-synonym pairs in the Thesaurus
 - each synonym is also a headword!
- ~84,000 unique pairs containing a multi-word string
- **Additional pairs filtered out:**
 - two two-word synonym candidates with the **reflexive pronoun se**
 - *prelomiti se, zlomiti se* 'to break'
 - pairs with problematic words often used in **descriptive synonym candidates** (*biti* 'to be', *začeti* 'to begin', *končati* 'to finish')
 - pairs that **overlapped** to a great extent
 - *hoditi z dolgimi koraki* 'to walk with long steps'
 - *začeti hoditi z dolgimi koraki* 'to begin walking with long steps'
 - pairs with a **terminological label** (*zoologija* 'zoology')
 - **masculine and feminine** synonym candidates
 - *industrijski psiholog – industrijska psihologinja*

Task 1: Targeted crowdsourcing/annotation campaign

- final set of pairs after the automatic preprocessing
 - **18,635** headword-synonym pairs
 - **6 student annotators** (+7th annotator for ambiguous examples)
 - introductory session
 - guidelines not overly specific
 - subjective judgment on whether the given headword-synonym pair would be useful in the Thesaurus (Yes – No – I don't know)
 - 3 responses per pair

PyBossa

- open-source crowdsourcing platform
- custom-made interface (error minimization)
- 1 task = 10 pairs

agitator || dekllica za vse

Sopomenki: Da Ne Ne vem

agitator || deček za vse

Sopomenki: Da Ne Ne vem

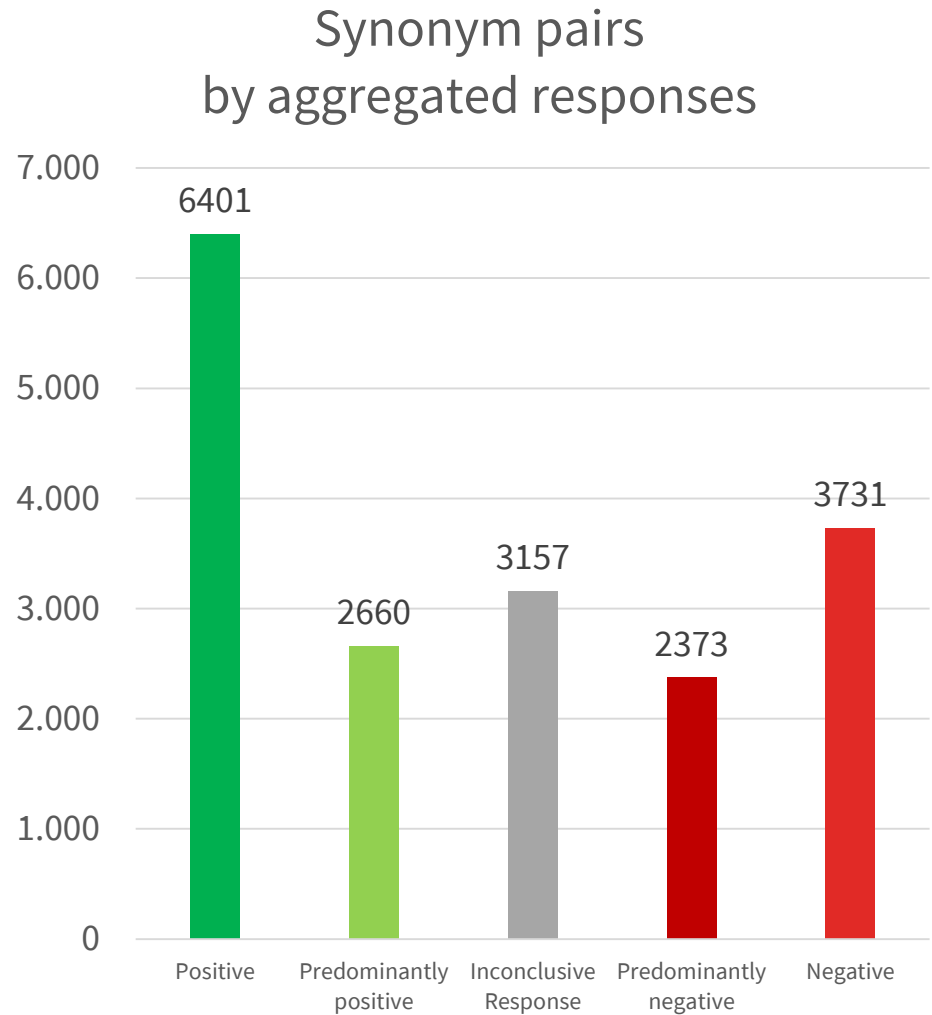
Shrani

Task 1: Results

- **56,745** responses
- approximately **8.4 seconds per task**
- total time spent on tasks: approx. **92 hours** (~15 hours per participant)
- The average percentage of same answers between annotators
 - **71%** (ranging from 63% to 79%)
- inter-annotator agreement (Cohen's kappa):
 - **0.42** (0.33 to 0.55)
 - fair to moderate agreement

Task 1: Results

- Positive – 35%
- Predominantly positive – 15%
- Inconclusive – 17%
- Predominantly negative – 13%
- Negative – 20%

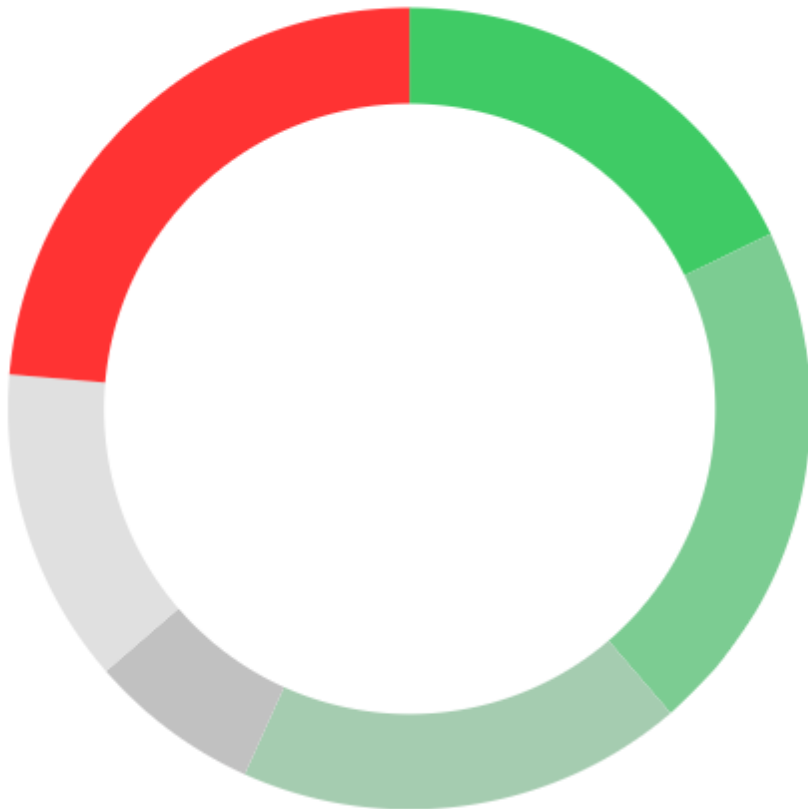


Task 1 vs. Expert Evaluation

- Expert evaluation made before the publication
- [Good – Acceptable – Poor] synonym

Linguistic evaluation of synonyms for a given headword

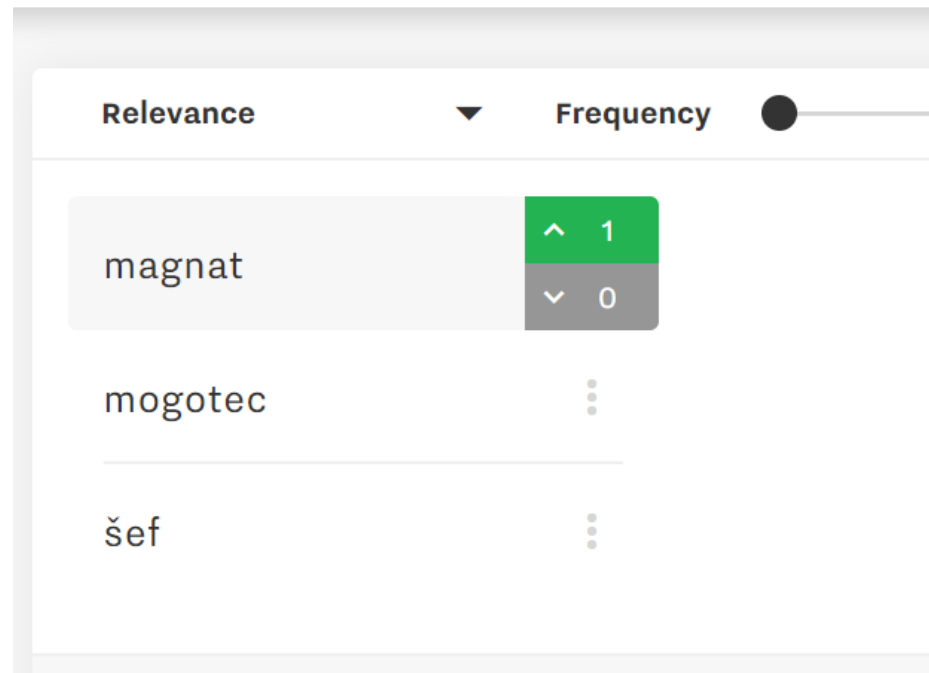
● Good	18 %
● Good / Acceptable	21 %
● Acceptable	18 %
● Good / Poor	7 %
● Acceptable / Poor	13 %
● Poor	24 %



Task 2: Collecting user votes

- User votes were also collected directly from the interface of the Thesaurus (for all synonym candidates, not just multi-word ones)
- March 2018 – 3 June 2019
- 26,253 user votes
 - only **5%** of the Thesaurus so far!
 - **24,214 (92%)** upvotes
 - **2,039 (8%)** downvotes
 - 83% votes for database synonyms
 - 17% votes for user-added synonyms
 - 868 examples with 3+ words have received user votes in the interface.
 - Only 60 include negative votes!

kralj 2017-11-24



Clean-up through user votes – A utopia?

- ~184,000 headword-synonym pairs to be evaluated in the entire dictionary (not counting inverted pairs and user-added synonyms)
- **4 votes** per pair
 - we need a total of **736,000** votes.
- **38,000** different IP-addresses
 - if **25%** users vote on synonyms;
 - if a vote takes **8 seconds** on average;
 - then each individual would have to spend approximately **13 minutes** (not necessarily all at once) voting on synonyms
- Feasible, but ...

User motivation

- users need to be motivated to vote!

Tasks of the day

Dictionary entries without user votes. Be the first to evaluate synonyms!

visokost >

tirnica >

nadzornica >

kdaj >

Community

List of users who suggested the highest number of new synonyms.

Saša Jenko Pahor >

RD >

Ferdo >

TatjanaJ >

More user motivation?

- How do we motivate them more?
 - improve the interface
 - emphasize voting features
 - we need to make voting easier!
 - *pozorno opazovati koga skozi monokel* 'to observe someone attentively through a monocle'
 - more targeted crowdsourcing campaigns?
 - **gamification!**

Igra besed (A Game of Words)

- Collocations
- Synonyms incoming!

× 11

+ uporniški

nič 

+ resen

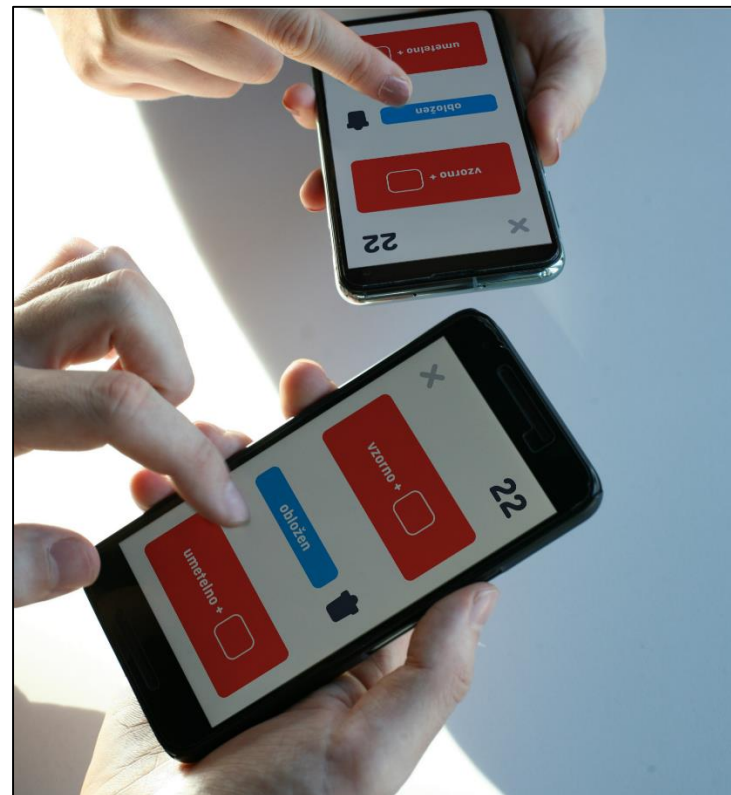
× 24

+ žafran

mlet

spomladanski

divji



Conclusion

- **Annotated set of ~18,000 headword-synonym pairs**
 - for Thesaurus 2.0, the **negative** ones will be hidden from the interface (but kept in the database!)
 - the rest → stay put for now, but will be used to develop better guidelines
- **Future work**
 - user-added synonyms (~20,000)
 - suspicious headwords
 - linguistic analysis (guidelines)
 - test gamification

Thank you!

Jaka Čibej
jaka.cibej@cjvt.si

Špela Arhar Holdt
spela.arhar@cjvt.si

