

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

Jaka Čibej, Špela Arhar Holdt

Centre for Language Resources and Technologies (Faculty of Arts, Faculty of Computer and Information Science), University of Ljubljana, Večna pot 113, 1000 Ljubljana, Slovenia
E-mail: jaka.cibej@cjvt.si, spela.arhar@cjvt.si

Abstract

The Thesaurus of Modern Slovene is 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. The Thesaurus was initially compiled entirely automatically and allows users to contribute toward improving the resource by adding suggestions for missing synonyms and/or by evaluating both the synonym candidates from the initial database as well as the suggestions added by other users. As an automatically generated language resource, however, the initial database of the Thesaurus includes a certain degree of noise. In the paper, we present two crowdsourcing activities aimed at cleaning up the database. The first is a targeted annotation campaign aimed at evaluating multi-word synonym candidates in the Thesaurus, and the second is an analysis of user votes provided directly in the Thesaurus interface. Both scenarios are examples of an effective postprocessing method for an automatically generated language resource and demonstrate that crowdsourcing can play an important role in smart lexicography, especially in the case of less-resourced languages.

Keywords: crowdsourcing; synonyms; Slovene; thesaurus; digital lexicography

1. Introduction

Crowdsourcing has demonstrated its value in numerous scientific endeavours, as demonstrated by a number of successful initiatives that have channelled the power of the crowd to great effect: in the field of linguistics, natural language processing has embraced crowdsourcing as a method to clean noisy datasets (Fišer et al., 2014), annotate language data (Fort et al., 2014), or collect user estimations and judgments (Snow et al., 2008). In the field of lexicography, a number of important steps towards the implementation of crowdsourcing in lexicographic workflows have also been made – see, for example, Čibej et al. (2015) for a proposed modular crowdsourcing workflow model for lexicography, or Abel and Meyer (2013) for an overview of different types of user contributions to online dictionaries – and although user involvement in lexicographic projects for digital dictionaries is not entirely new (as shown by Lew (2014) numerous collaborative lexicographic projects are available online, the most

noted among them being the Urban Dictionary¹ and Wiktionary²), crowdsourcing differs from collaborative lexicography in the fact that the former is usually more restricted in how users can contribute to the compilation of a digital dictionary (i.e. they solve a relatively narrow, predefined task and require a platform on which to solve it, as opposed to the free-for-all approach often employed by collaborative dictionaries). In addition, crowdsourcing can take place at any stage of dictionary compilation, both pre- and post-publication. Although only a handful of good practice examples showcase the implementation of crowdsourcing in lexicographic workflows (Kosem et al., 2018), the rise and proliferation of digital-born dictionaries is paving the way to a more crowd-oriented form of dictionary compilation. The goal of this paper is to present one such dictionary project, the Thesaurus of Modern Slovene (Krek et al., 2018a), and the way crowdsourcing is being used to clean it. We present the results of two instances of crowdsourcing activities aimed at cleaning up the noise in the Thesaurus: (a) the votes provided by the dictionary users and collected directly through the dictionary interface in the first year since its publication, and (b) a more targeted crowdsourcing campaign for students of linguistics in order to evaluate a set of multi-word synonym candidates.

The paper is structured as follows: in Section 2, we present the Thesaurus of Modern Slovene, its compilation and overall design. In Section 3, we describe the results of the targeted crowdsourcing campaign focusing on multi-word synonym candidates. In Section 4, we present an analysis of the upvotes and downvotes on synonym candidates collected directly through the Thesaurus interface. We conclude with a discussion and some directions for future work in Section 5.

2. The Thesaurus of Modern Slovene

The Thesaurus of Modern Slovene is the largest open-source digital collection of Slovene synonyms. It was published in March 2018 by the Centre of Language Resources and Technologies of the University of Ljubljana as the first example of a *responsive dictionary* (Arhar Holdt et al., 2018), a new type of language resource that is defined by the following characteristics: first, it is a born-digital and digital-only dictionary, designed with the needs, requirements, and advantages of the digital medium in mind. Second, its database was initially compiled entirely through automatic methods that were tested and evaluated beforehand. Third, both the database and the language resource were made openly accessible to the language community immediately after the automatic compilation to provide a large amount of automatically extracted language data which is relevant, but this contains a certain degree of noise. Fourth, because of its digital nature, the dictionary is frequently updated and all changes are tracked through versions and with timestamps at the level of entries. Finally, the responsive dictionary features one or more ways to allow users to contribute to its development.

¹ <https://www.urbandictionary.com/>

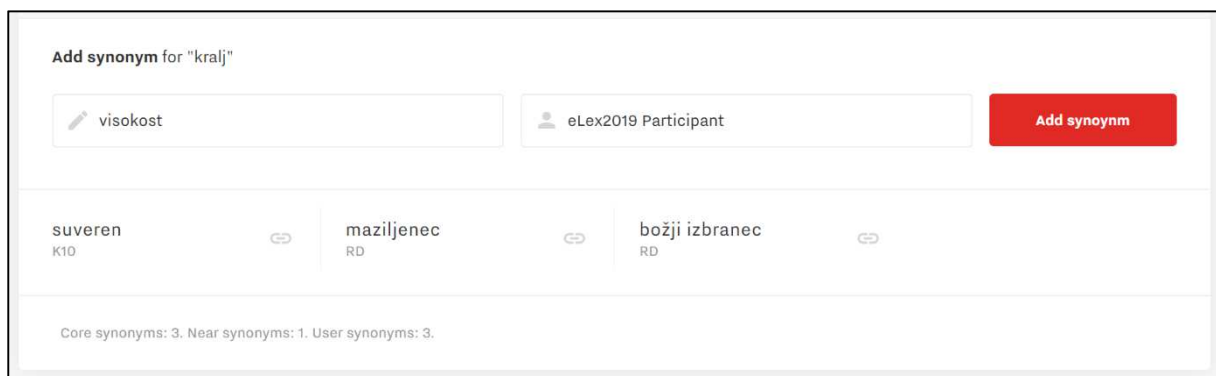
² <https://en.wiktionary.org/>

Through this, it responds to changes in language on the one hand and the knowledge/consensus of its users on the other.

As the first example of a responsive dictionary, the Thesaurus of Modern Slovene was initially compiled automatically with co-occurrence graphs (for a more detailed description of the methodology, see Krek et al., 2017) using existing language resources, namely The Oxford®-DZS Comprehensive English-Slovenian Dictionary and the Gigafida Reference Corpus of Written Slovene. The Thesaurus database was made available in the CLARIN.SI repository (Krek et al., 2018b) under the Creative Commons Attribution-ShareAlike 4.0 International licence (CC BY-SA 4.0).

A custom interface was developed to enable the language community to contribute toward improving and further developing the resource in two ways: (1) by adding their own suggestions of missing synonym candidates to a particular entry; and/or (2) by evaluating both the synonym candidates from the initial database as well as the suggestions added by other users by upvoting or downvoting them.

Users can add synonyms through a special form integrated in the interface (Figure 1). No registration is required – the user can enter a username and the suggested synonym in the designated fields and then click the Add Synonym button. The suggestion is instantly displayed in the user synonym section.



The screenshot shows a web interface for adding a synonym for the word "kralj". At the top, it says "Add synonym for 'kralj'". Below this, there is a text input field containing the word "visokost". To the right of the input field is a dropdown menu showing the user "eLex2019 Participant". A red button labeled "Add synonym" is positioned to the right of the dropdown. Below the input field and dropdown, there are three synonym candidates displayed in a row: "suveren" (K10), "maziljenec" (RD), and "božji izbranec" (RD). Each candidate has a small link icon to its right. At the bottom of the interface, there is a summary: "Core synonyms: 3. Near synonyms: 1. User synonyms: 3."

Figure 1: Adding user synonyms to the Thesaurus.

Users can upvote or downvote existing synonym candidates by hovering over the candidate and clicking the upvote (green) or downvote (red) button (Figure 2). They can also cancel the vote if they misclicked. During dictionary updates, votes are taken into account so that downvoted synonyms can be excluded from the dictionary, while upvoted synonyms can be ranked higher.

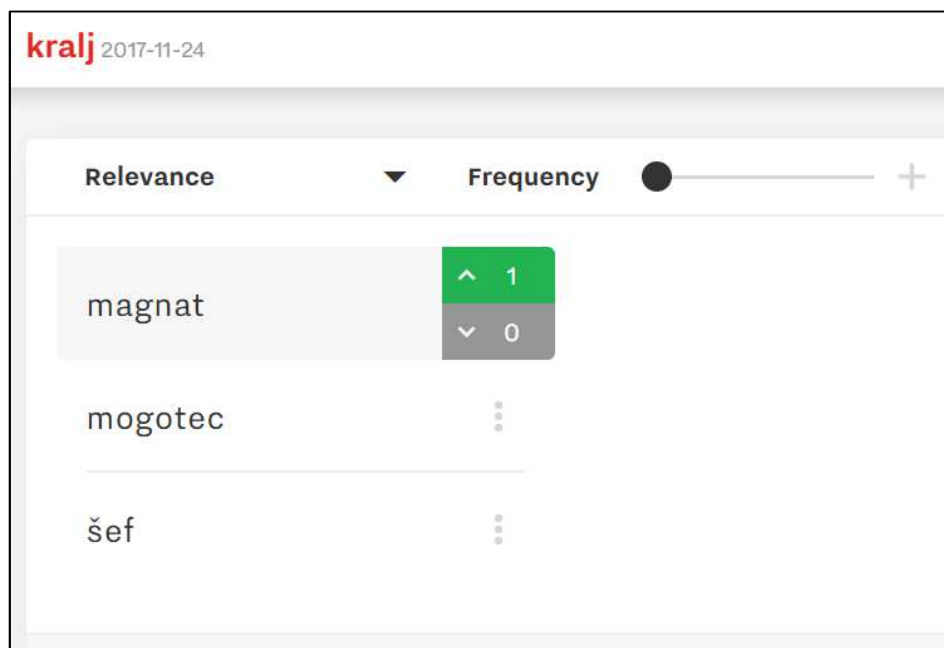


Figure 2: Voting for synonym candidates in the Thesaurus.

The users' reactions to the Thesaurus were predominantly positive. The number of user added synonyms and votes (more on this in Section 4) indicates that users are positively inclined toward user involvement in the Thesaurus. This demonstrates that the automatic compilation of language resources can be efficient both in terms of time and financial investment, particularly when development is continued in the post-publication phase and complemented by user involvement.

2.1 Noise from automatic synonym extraction

As an automatically generated language resource, the initial database of the Thesaurus of Modern Slovene includes a certain degree of noise. These methodology-related problems were clear in the beginning, so we decided to focus on them. Because the synonyms in the Thesaurus were extracted from Slovene translation equivalents of English headwords, multi-word synonym candidates are sometimes only descriptive approximates of concepts that are lexicalized in English but not in Slovene, e.g. *rooming-in* as an English loanword and *24-urno sobivanje novorojenčka in matere* 'a 24-hour cohabitation of a newborn and their mother'. Problematic categories of automatically extracted synonyms include feminine-masculine word pairs that ended up under the same entry (e.g. the word *učitelj* 'teacher [masculine]' is listed as a synonym under the headword *učiteljica* 'teacher [feminine]'), inadequate candidates arising from structural irregularities or inconsistencies in The Oxford®-DZS Comprehensive English-Slovenian Dictionary, and multi-word synonym candidates that border on paraphrases, definitions, partial repetitions, or descriptions.

Because multi-word synonym candidates were easy to identify and were the most obviously problematic category (as well as manageable in size), we decided to organize a targeted crowdsourcing campaign to identify the synonym candidates to be removed in the next Thesaurus update. We describe the campaign in Section 3.

3. Crowdsourcing multi-word synonym candidates

The goal of the crowdsourcing experiment was to exclude inadequate multi-word synonym candidates from the dataset. In this section, we describe the infrastructure utilized in the data and the platform used, the task design process, and the results of the experiment.

3.1 Data preprocessing

Version 1.0 of the Thesaurus of Modern Slovene contains 368,117 headword-synonym pairs (not counting the synonym candidates added by users); 162,719 of these pairs (44%) contain a multi-word synonym candidate or headword. The Thesaurus is structured in such a way that each synonym can also appear as a headword, so the number of unique pairs (in terms of their components) containing a multi-word string is 84,128.

Not all pairs were included in the crowdsourcing task. The data was preprocessed to make sure the workload was manageable and to remove inadequate pairs that were easy to identify automatically (through a set of rules).

The excluded categories were the following:³ (a) pairs containing two two-word synonym candidates, both containing the reflexive pronoun *se* (e.g. *prelomiti se* and *zlomiti se* ‘to break’; 4,510 pairs); (b) pairs containing a number of problematic words often used in descriptive synonym candidates, e.g. the verbs *biti* ‘to be’, *začeti* ‘to begin’, *končati* ‘to finish’, the preposition *brez* ‘without’, and the nouns *prebivalec* ‘inhabitant [male]’ and *prebivalka* ‘inhabitant [female]’ (6,141 pairs); (c) pairs that overlapped to a great extent (e.g. *hoditi z dolgimi koraki* ‘to walk with long steps’ and *začeti hoditi z dolgimi koraki* ‘to begin walking with long steps’, 5,517 pairs); (d) pairs that contained a synonym candidate with a terminological/field label (e.g. *zoologija* ‘zoology’, 14,581 pairs); and (e) pairs that contained masculine and feminine synonym candidates (e.g. *industrijski psiholog* ‘industrial psychologist [male]’ and *industrijska psihologinja* ‘industrial psychologist [female]’, 5,334 pairs). The final set of pairs after the automatic preprocessing contained 18,635 headword-synonym pairs. The pairs not included in the experiment will be further analyzed and most likely removed from the Thesaurus.

³ Some synonym pairs were assigned to multiple categories.

3.2 Crowdsourcing platform

The platform we used in the experiment was PyBossa⁴, an open-access Python-based crowdsourcing platform that features a great deal of flexibility, especially with regard to task design and interface optimization with the aim of greater user-friendliness. It also features an API, and allows data upload/download in .JSON format. The crowdsourced data can be downloaded at any stage of the crowdsourcing process. However, PyBossa does not include any quality control functions (e.g. inter-annotator agreement measures, automatic gold-standard comparison), so these were handled outside the platform using external custom-made Python scripts. PyBossa has already been used with great success in previous work, e.g. for annotating collocations for the Collocations Dictionary of Modern Slovene (Kosem et al., 2018).

3.3 Task Design and Crowdsourcer Recruitment

We designed a custom interface for the task (Figure 3). In each task, the crowdsourcer was presented with 10 headword-synonym pairs and three options to choose from: *Da* ‘Yes’ if the units were synonymous, *Ne* ‘No’ if they were not, and *Ne vem* ‘I don’t know’ if they were uncertain. The crowdsourcer had to tick all ten pairs to be able to finally click *Shrani* ‘Save’ and proceed to the next task.

Several measures were taken to reduce the number of mistakes during crowdsourcing. Radio buttons were used in order to reduce the number of misclicks (and the crowdsourcer could revise their annotation within each batch of 10 pairs as many times as they wanted). In addition, the user got a pop-up alert if they forgot to tick one of the buttons.

Six crowdsourcers⁵ were recruited for the task, all of them students of linguistics at the University of Ljubljana. They were familiarized with the Thesaurus and the goal of the task during an introductory briefing session. The guidelines were not overly specific, as the task is very similar to the voting system already enabled in the dictionary interface. The crowdsourcer’s main task was to provide their subjective judgment on whether the given headword-synonym pair would be useful in the Thesaurus.

⁴ <https://pybossa.com/> – DOI: 10.5281/zenodo.3239980

⁵ The targeted campaign is not a typical example of crowdsourcing (as it relies on a limited group of preselected crowdsourcers) and is actually a mock-up of a full-scale crowdsourcing campaign because it uses the same methodology which is independent of this specific project. The crowd used could be significantly larger and more diverse, and may be such in our future work. For the sake of simplicity, we refer to all these activities as crowdsourcing in the paper.

The image shows a screenshot of a web-based task interface. It contains two identical-looking task blocks stacked vertically. Each block has a rounded rectangular header containing a pair of terms: 'agitator || dekllica za vse' for the first and 'agitator || deček za vse' for the second. Below each header is a label 'Sopomenki:' followed by three radio button options: 'Da', 'Ne', and 'Ne vem'. A horizontal line separates the two task blocks. At the bottom right of the interface is a blue rectangular button with the white text 'Shrani'.

Figure 3. PyBossa task interface for annotating multi-word synonym pairs.

Based on the testing phase in which each synonym pair took approximately 7-8 seconds on average to evaluate (rounded up to 10 seconds), the crowdsourcing campaign was foreseen to take approximately 153 hours to complete. At the standard rate of the University of Ljubljana for student work (€7), it would cost approximately €1,071.

3.4 Results

In total, 56,745 responses were collected during the crowdsourcing task, three (in some rare cases four⁶) for each headword-synonym pair. The mean response time per synonym pair (after removing outliers above 15 seconds) was approximately 8.4 seconds, with 7.9 as the median value. The total time spent on tasks was approximately 92 hours (with a cost of cca. €650 in student work). In general, approximately 57% of the responses were positive and approximately 41% were negative, while only 1.6% were undecided. We tracked inter-annotator agreement for each crowdsourcer pair with two measures: the percentage of identical answers, and Cohen's kappa coefficient. The average percentage of identical answers between annotators was 71% (ranging from 63% to 79%), while Cohen's kappa coefficient ranged from 0.33 to 0.55 with an average of 0.42, which indicates fair to moderate agreement.

We also measured to what extent the crowdsourcers agreed on whether a given pair is adequate or inadequate by calculating information entropy for the responses of each

⁶ The number of responses was limited to three per synonym pair, but if a crowdsourcer started solving a particular task and another crowdsourcer was presented with it before the first crowdsourcer was done with it, both responses were registered, which sometimes resulted in four responses per synonym pair.

pair. An information entropy of 0 indicates perfect agreement (e.g. all responses are positive), while higher values indicate various degrees of disagreement (e.g. a value of 1.58 indicates a response combination akin to Yes-No-I Don't Know). The results are shown in Table 1.

Information Entropy	Frequency	Percentage
0	10,136	54.39
0.81	203	1.09
0.92	7,943	42.62
1	89	0.48
1.5	7	0.04
1.58	257	1.38

Table 1: Evaluated synonym pairs by information entropy.

More than half (54%) of the pairs featured complete agreement between crowdsourcers, while the majority of the rest featured a slightly mixed response (42%). A more detailed distribution is shown in Table 2, where response combinations are also grouped into categories by agreement.

A thorough manual categorization of the annotated data is beyond the limits of this paper, but we nevertheless show a brief overview of the results which indicate that several groups or categories can be formed based on different response combinations. We list illustrative examples for both the pairs with complete agreement and pairs with mixed responses in order to provide some insight into the results.

A large number of inconclusive or mixed responses was likely caused by unfamiliarity with the annotated synonym candidates (e.g. infrequent words, terminological units, or loanwords). On the other hand, disagreement occurs with pairs in which their semantic similarity is clear, but they are interchangeable only in specific language contexts.

Response Combination	Frequency	Percentage
Complete Agreement		
Yes, Yes, Yes	6,245	33.51
Yes, Yes, Yes, Yes	156	0.84
No, No, No	3,616	19.40
No, No, No, No	115	0.62
Mixed Response		
Yes, Yes, Yes, No	117	0.63
Yes, Yes, No	4,267	22.90
Yes, Yes, I don't know	228	1.22
Yes, No, No	3,163	16.97
Yes, No, No, No	81	0.43
No, No, No, I don't know	5	0.03
No, No, I don't know	200	1.07
Inconclusive Response		
Yes, Yes, No, I don't know	2	0.01
Yes, No, No, I don't know	4	0.02
Yes, No, I don't know	257	1.38
Yes, No, I don't know, I don't know	1	0.01
No, I don't know, I don't know	45	0.24
Yes, I don't know, I don't know	40	0.21
I don't know, I don't know, I don't know	4	0.02
Yes, Yes, No, No	89	0.48

Table 2. Evaluated synonym pairs by response combinations.

The examples with agreement that the multi-word unit is relevant for the Thesaurus include e.g. pairs in which **the multi-word unit is an explanation of a (frequently)**

single-word headword. This is often an explanation of loanwords or neologisms (*sendvičarna* ‘sandwich store’ – *trgovina s sendviči* ‘store with sandwiches’; *absentizem* ‘absenteeism’ – *odsotnost z dela* ‘absence from work’; *glosirati* ‘to gloss’ – *pojasniti v opombi* ‘to explain in a notation’), but explanations of more frequent vocabulary also occur (*zmagati* ‘to win’ – *priti na prvo mesto* ‘to get first place’; *zmleti* ‘to crush’ – *zdrobiti v prah* ‘to grind into dust’; *pravljichen* ‘fairytale-like’ – *kot iz pravljice* ‘like in a fairytale’). What is interesting to note is that the crowdsourcers did not see these examples as redundant and irrelevant, but want to keep them in the Thesaurus. The same is true for **multi-word pairs that differ only in a single word** and in which the two differing words are synonymous (e.g. *razdeliti na pokrajine* ‘to divide into regions’ – *razdeliti na province* ‘to divide into provinces’; *sprejem s koktejli* ‘cocktail reception’ – *zabava s koktejli* ‘cocktail party’; *obsoditi na pogubo* ‘to condemn to oblivion’ – *obsoditi na propad* ‘to condemn to downfall’). The third group includes examples which **differ in part-of-speech structure**, e.g. a pair of nominal phrases containing an adjectival or prepositional attribute (*cestni davek* ‘road tax’ – *davek za uporabo cest* ‘tax for road use’; *brivski pribor* ‘shaving kit’ – *pribor za britje* ‘kit for shaving’; *bralna očala* ‘reading glasses’ – *očala za branje* ‘glasses for reading’), or pairs containing instrumental-case and genitive-case phrases (*s spretnimi prsti* ‘with nimble-fingers’ – *spretnih prstov* ‘lit. [of] nimble fingers, nimble-fingered’). In some pairs, one of the units features a **semantically light verb** (*dati nižjo oceno* ‘to give a lower grade’ – *znižati oceno* ‘to lower the grade’; *dati novo ime* ‘to give a new name’ – *preimenovati v* ‘to rename [to]’). The most interesting examples (although rare) are the ones in which one or even both components of the pair are **idiomatic expressions** (e.g. *dati zeleno luč* ‘to greenlight’ – *uradno odobriti* ‘to officially approve’; *pogled resnici v oči* ‘lit. staring truth in the eye’ – *spust na realna tla* ‘lit. a descent to solid ground’).

On the other hand, examples that the crowdsourcers unanimously want removed from the databases include pairs that were semantically linked in the original bilingual dictionary but **are not themselves synonymous**, e.g. *trikrat tedensko* ‘three times per week’ – *vsake tri tedne* ‘once every three weeks’; *sod za olje* ‘oil barrel’ – *vinski sod* ‘wine barrel’; *speti v čop* ‘to put up [hair] in a ponytail’ – *splesti v kito* ‘to braid [hair]’. Similarly, in some examples **synonymy is limited to a specific context** (*ne sprejeti* ‘to not accept’ – *vreči na izpitu* ‘to fail [someone] on an exam’; *nizek* ‘low’ – *s plosko peto* ‘with a flat heel’). Other negatively evaluated examples include pairs in which one of the components contains a **semantically specific complement** (*zvižati se* ‘to squirm’ – *zvižati se kot črv* ‘to squirm like a worm’, *za kuhanje* ‘for cooking’ – *za kuhanje na visoki temperaturi* ‘for cooking on high heat’) or a **prepositional verb** (*iti do* ‘to go to’ – *spustiti se v* ‘to descend into’; *iti k* ‘to go to’ – *videvati se z* ‘to see [someone]’), as well as **inadequately paired masculine and feminine variants** occurring in more complex structures, which prevented them from being filtered out automatically (e.g. *študent medicine* ‘[male] student of medicine’ – *študentka medicinske fakultete* ‘[female] student of a Faculty of Medicine’). In rare cases, inadequate pairs stem from **errors during the automatic export** of synonym candidates (e.g. *official* – *na visokem položaju* ‘in high places’, *people* – *na visokem položaju* ‘in high places’).

Examples with mixed responses contain less clearly defined groups. Predominantly positively evaluated examples include pairs from the above-defined groups that are characterized by a certain degree of **semantic similarity**, but do not necessarily overlap in terms of synonymy in different contexts, e.g. *zaljubljen v gledališče* ‘in love with the theater’ – *zaljubljen v oder* ‘in love with the stage’; *zakopati v jamo* ‘to bury in a cave’ – *zakopati v luknjo* ‘to bury in a hole’). In some examples, the vocabulary is **infrequent or specialized**, and as such presumably not familiar to the crowdsourcers (e.g. *primogenitura* ‘primogeniture’ – *pravica prvorojenca do nasledstva* ‘the right of the firstborn to inheritance’; *kiras* ‘cuirass’ – *prsni del oklepa* ‘the chest part of armor’). Similar categories can be identified in predominantly negatively evaluated pairs: semantically similar pairs (but not similar enough), e.g. *zapreti v kletko* ‘to put in a cage’ – *zapreti v kurnik* ‘to put in a chicken coop’; *upodobiti na fotografiji* ‘to portray in a photograph’ – *upodobiti na sliki* ‘to portray in a portrait’) and pairs with infrequent or specialized vocabulary (*blindirati* – *obložiti s ploščicami* ‘to insulate with panels’).

The most interesting examples for further analyses are the ones found in the inconclusive responses. These predominantly consist of **infrequent, specialized vocabulary** likely unfamiliar to the crowdsourcers (*barg* ‘barge’ – *kanalski tovorni čoln* ‘channel cargo boat’; *alkova* ‘alcove’ – *okno v tinu* ‘window in a corner’), but also of **stylistically marked vocabulary** (*crkniti* ‘to drop dead’ – *zrušiti se od utrujenosti* ‘to collapse of exhaustion’), **loanwords** (*digest* – *zbirka izvlečkov iz člankov* ‘a collection of article excerpts’), or more complex examples of **masculine-feminine pairs** (*liftboy* – *uslužbenka pri dvigalu* ‘[female] employee at the elevator’). A separate category consists of examples in which the responses were certain, but divided (e.g. Yes, Yes, No, No). Besides the already mentioned pairs with **limited synonymy** (e.g. *torba za cunje* ‘bag for clothes’ – *vreča za cunje* ‘sack for clothes’; *medsebojno povezovanje* ‘interconnection’ – *navezovanje poslovnih stikov* ‘forming business contacts’), an interesting category are the examples with **phraseological components** that certain crowdsourcers may not have recognized (*iti zraven* ‘to come with’ – *priti v paketu* ‘to come with the package’; *nasloniti se nazaj* ‘to lean back’ – *sprostiti se* ‘to relax’). Disagreement can also be observed with phrases that express the **(im)perfectiveness of the action**, e.g. *razdražiti* ‘to irritate [perfective]’ – *iti na živce* ‘to go on [smn’s] nerves [imperfective]’; *nadeti si tančico* ‘to put on a veil’ – *nositi tančico* ‘to wear a veil’, *izbruhniti v smeh* ‘to burst into laughter’ – *pokati od smeha* ‘to be bursting of laughter’.

Among the already mentioned 54% of pairs with complete agreement, approximately 34% were evaluated as adequate and 20% as inadequate. While only 2.37% of all evaluated synonym pairs resulted in a response that was completely inconclusive, and the remaining 43% could be resolved through a majority vote at this stage, a seventh annotator was recruited to provide additional votes for pairs with mixed or inconclusive responses. Table 3 shows the results for the ambiguous pairs when taking into account the responses made by the seventh annotator.

Response Combination	Frequency	Percentage
Predominantly Positive		
Yes, Yes, Yes, Yes, No	63	0.77
Yes, Yes, Yes, No	2,373	28.97
Yes, Yes, Yes, No, No	87	1.06
Yes, Yes, Yes, No, I don't know	2	0.02
Yes, Yes, Yes, I don't know	135	1.65
Predominantly Negative		
Yes, No, No, No	2,103	25.68
Yes, No, No, No, No	56	0.68
Yes, No, No, No, I don't know	4	0.05
No, No, No, No, I don't know	4	0.05
No, No, No, I don't know	145	1.77
Yes, Yes, No, No, No	61	0.74
Inconclusive Response		
Yes, Yes, No, No	2,700	32.97
Yes, Yes, I don't know, I don't know	23	0.28
Yes, Yes, No, I don't know	204	2.49
Yes, No, No, I don't know	168	2.05
Yes, No, No, I don't know, I don't know	1	0.01
Yes, No, I don't know, I don't know	28	0.34
Yes, I don't know, I don't know, I don't know	1	0.01
No, No, I don't know, I don't know	30	0.37
No, I don't know, I don't know, I don't know	2	0.02

Table 3. Evaluated ambiguous synonym pairs with additional annotations.

With the addition of another vote, 2,660 synonym pairs start to converge more toward Yes and 2,373 toward No (with 60-80% of votes in favour of one or the other), while 3,157 still remain inconclusive (approximately 17% of all the multi-word synonym pairs included in the crowdsourcing task). Out of these, 2,700 examples keep conflicting responses (Yes, Yes, No, No). Many of these are pairs in which synonymy is limited to a specific context, while one component is semantically wider or inclusive, e.g. *redčiti se* ‘to thin’ – *začenjati dobivati plešo* ‘to begin to go bald’, *prečkanje puščave* ‘the crossing of the desert’ – *vožnja čez puščavo* ‘the drive across the desert’. In some examples, the multi-word units differ in a single word, but the substitution reduces the degree to which the units are interchangeable in use, e.g. *znova se sestati* ‘to have a reunion’ – *znova se zbrati* ‘to regroup’; *zelo smešna zgodba* ‘a very funny story’ – *zelo smešna šala* ‘a very funny joke’; *zbirati se v bazen* ‘to gather in a pool’ – *zbirati se v tolmun* ‘to gather in a pond’. Another category difficult to evaluate consists of examples with a semantically specific complement that is part of the original phrase, but is commonly left out in language use, e.g. *človek z dna* ‘a man from the bottom’ – *človek z dna družbene lestvice* ‘a man from the bottom of the social scale’; *dijak tretjega letnika* ‘third-year student’ – *dijak tretjega letnika srednje šole* ‘third-year high-school student’. The same is true of examples in which the synonym candidate is explanatory, but semantically too narrow or too vague (*čarodej* ‘magician’ – *praktikant črne magije* ‘practitioner of black magic’; *nož za sir* ‘cheese-cutting knife’ – *naprava za rezanje sira* ‘cheese cutter’). Some examples are problematic because of the (im)perfectiveness of the expressed action (*pihati od jeze* ‘to be seething with anger’ – *ujeziti se* ‘to become angry’), prepositional verbs (*vrniti se na* ‘to return to’ – *znova stopiti v* ‘to reenter into’), and (potential) phraseological units (*prinesti na krožniku* ‘to bring on a plate’ – *servirati na pladnju* ‘to serve on a platter’). On the other hand, at this stage there seem to be no more problematic examples with rare or specialized vocabulary, masculine-feminine pairs, paraphrases of part-of-speech structures, methodological extraction errors, or results that clearly are (or are not) synonymous.

This outcome is in line with our expectations based on previous findings: after the automatic compilation of the Thesaurus, an evaluation of the dataset was conducted by experts (linguists and lexicographers) on a random subset of headword-synonym pairs (not limited to multi-word synonym candidates). The goal of the task was to evaluate synonyms as either good, acceptable, or poor. The results of the expert evaluation are shown in Figure 4 (see Arhar Holdt et al., 2018 for more on the evaluation).

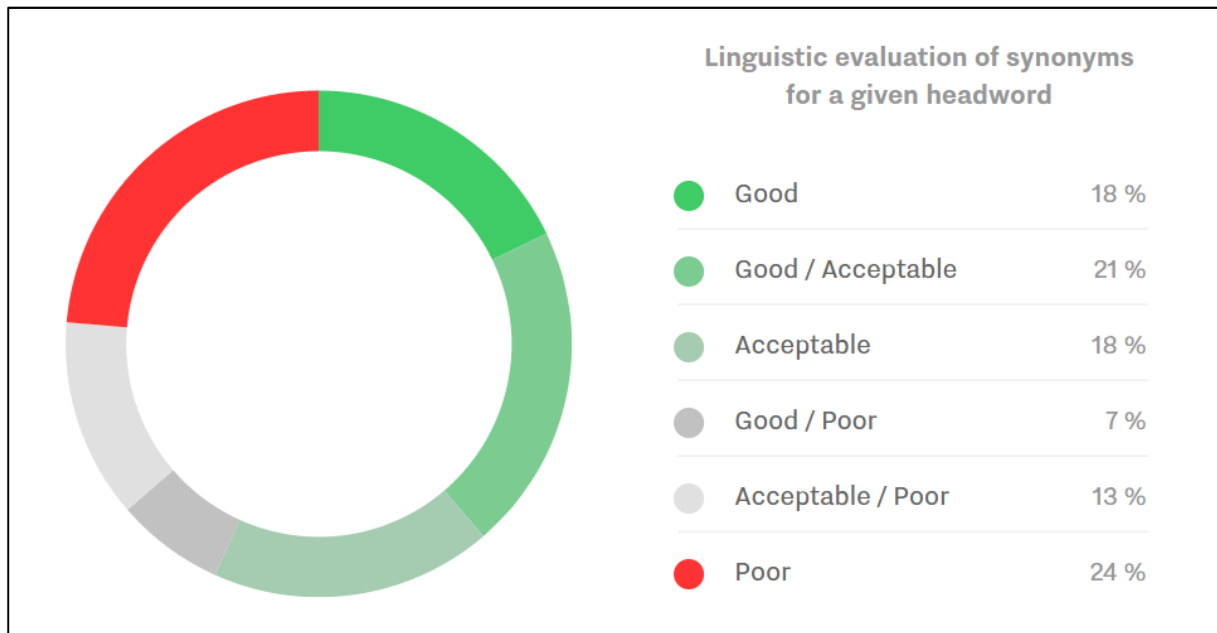


Figure 4: Expert evaluation of the synonyms in the Thesaurus of Modern Slovene.

Even with expert annotations, 20% of the evaluated synonym pairs showed considerable disagreement: 7% were simultaneously evaluated as good and poor, while 13% were rated as both acceptable and poor. This demonstrates that evaluating synonymy is not a trivial or one-dimensional task, and that some examples pose a challenge for both experts and non-experts. Synonymy is highly context-dependent, and to some extent subjective. With the Thesaurus of Modern Slovene, however, we took a greedier approach and opted to treat ambiguous synonyms as potentially adequate rather than exclude them, as the evaluation has shown that at least some users might find them useful.

4. Collecting user votes through the interface

Parallel to the targeted crowdsourcing campaign described in Section 3, user votes were also collected directly from the interface of the Thesaurus (for all synonym candidates, not just multi-word ones). In the period since the release of the Thesaurus (March 2018 – 3 June 2019), a total of 26,253 user votes was collected, 24,214 (92%) of which were upvotes and 2,039 (8%) were downvotes. The majority of votes (21,886, or 83%) was collected for the original Thesaurus synonyms, while a smaller portion (4,367, or 17%) was collected for the user-added synonyms. In this paper, we concentrate on the original Thesaurus synonyms.

A total of 17,904 headword-synonym pairs in the Thesaurus (5% of the entire dictionary) received at least one vote, the majority one vote (15,307 pairs) or two (2,035 pairs). 16,938 pairs received at least one upvote and 1,340 pairs received at least one downvote.

The results indicate that the users are positively inclined to the automatically compiled data (taking into account the headwords they have queried so far), with only 7.5% of the voted pairs having been downvoted, and even fewer (968, or 5.4%) having received only downvotes and no upvotes.

We list here several examples of user-voted synonym candidates, with the number of upvotes and downvotes in brackets. As can be expected, the most votes were collected for the example entries *zelen* ‘green’, *ideja* ‘idea’, and *spati* ‘to sleep’, which are often used during Thesaurus presentations as demonstrative examples (for voting as well); *zelen* – *mlad* ‘young’ (17+, 9-), *zelen* – *bled* ‘pale’ (11+, 8-). Disagreement in votes also occurs with terminological words or words of foreign origin (*splentitis* ‘spleentitis’ – *vnetje vranice* ‘spleen infection’, 6+, 6-; *hiša* ‘house’ – *polje* ‘field’, astrology, 1+, 4-; *izdajalec domovine* ‘traitor of the country’ – *kvisling* ‘quisling’, 2+, 3-) or stylistically marked words (vulgar: *drek* ‘shit’ – *en kurec* ‘piece of shit’, 2+, 3-). Pairs that have been more consistently downvoted include explicitly pejorative components (*teta* ‘auntie’ – *poženščen peder* ‘effeminate faggot’, 0+, 3-) or masculine-feminine pairs (*babica* ‘grandmother’ – *dedek* ‘grandfather’, 0+, 12-), as well as examples in which the vocabulary is general and stylistically neutral, but not synonymous enough (*domišljav* ‘pretentious’ – *zadovoljen* ‘satisfied’, 0+, 1-; *domišljav* ‘pretentious’ – *samozavesten* ‘confident’, 0+, 1-).

With positively voted candidates it is more difficult to pinpoint the reasons for the upvotes. It appears that users upvote very prototypical, unambiguous, and widely interchangeable synonyms (*lakomnost* ‘avarice’ – *pohlepnost* ‘greed’, 12+, 0-; *lep* ‘beautiful’ – *čeden* ‘handsome’, 12+, 2-; *groziti* ‘to threaten’ – *pretiti* ‘to menace’ 4+, 0-; *prebrisan* ‘ingenious’ – *premeten* ‘cunning’, 4+, 0-). Examples with single upvotes indicate that users vote for them systematically: presumably the same user votes for most of the synonym candidates in a headword (*abolirati* ‘to abolish’ – *ukiniti* ‘to cancel’, 1+, 0-; *abolirati* – *razveljaviti* ‘to cancel’, 1+, 0-).

An overview of the multi-word synonym candidates with votes (which this paper focuses on) shows that 868 examples with three or more words received user votes in the interface. Sixty of these include negative votes (e.g. pairs with redundantly repeated parts and/or masculine-feminine pairs (*paravojak* ‘[male] para-soldier’ – *pripadnica paravojaške organizacije* ‘[female] member of a paramilitary organization’, 0+, 2-), non-synonymous pairs (*človek* ‘human’ – *nabit z energijo* ‘full of energy’; 0+, 2-), and similar. As already mentioned, because of the methodology implemented in the compilation of the Thesaurus, multi-word units as headwords are often unusual (and are as such less queried), while the votes they receive are rather sporadic.

In the ideal scenario, the entire Thesaurus would be evaluated through user votes. A quick estimate reveals that this is not impossible: with cca. 184,000 headword-synonym pairs to be evaluated in the entire dictionary (not counting inverted pairs and user-added synonyms) and presupposing that four votes per pair would be enough to

distinguish the worst candidates from the best, a total of 736,000 thousand votes would have to be collected. So far, the Thesaurus has been accessed from approximately 38,000 different IP-addresses. If a quarter of these would vote on synonyms (taking into account that on average, a vote on a synonym pair takes 8 seconds), each individual would have to spend approximately 13 minutes (not necessarily all at once) voting on synonyms.

However, for this to be successful, user motivation is key. The Thesaurus already features Tasks of the Day (shown in Figure 5), a special subsection on the homepage that invites users to evaluate up to four headwords in which the synonyms have not yet received any votes. Additional features in a similar vein are planned.

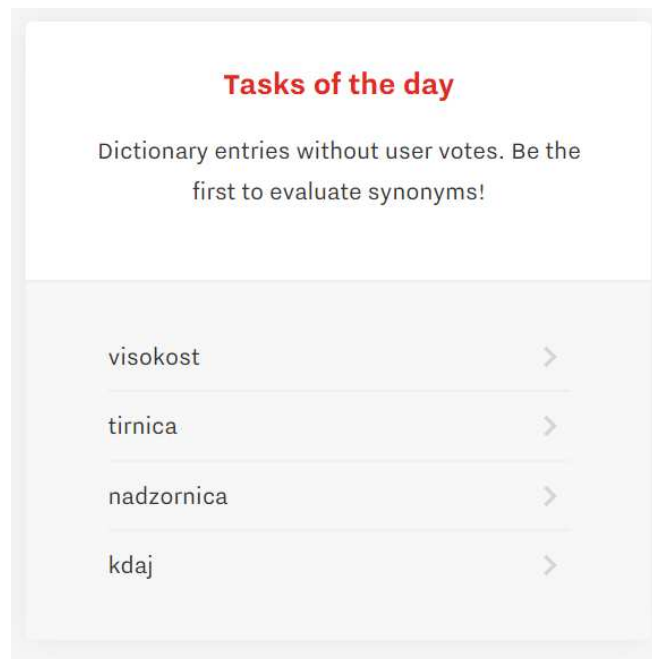


Figure 5: Tasks of the Day in the Thesaurus.

Furthermore, the Thesaurus logs show that only 48% of the Thesaurus has been queried so far (cca. 51,000 headwords out of 105,473), which means that half of the content has not even been seen by users yet. This suggests that new manners of presenting the content of the dictionary to the users are required, such as targeted crowdsourcing campaigns or gamification. This is part of our future work and we discuss it in more detail in the conclusion.

5. Conclusion and future work

In the paper, we presented two crowdsourcing activities aimed at cleaning up the first version of the Thesaurus of Modern Slovene. The targeted crowdsourcing campaign has processed 18,365 multi-word synonym candidates and resulted in a dataset of 5,882 negatively evaluated candidates (3,509 with complete agreement), 9,027 positively

evaluated candidates (6,367 with complete agreement), and 3,456 ambiguous candidates, which means the experiment cleaned up 81% of the multi-word synonym candidates included in the task. The votes collected through the Thesaurus interface resulted in a dataset of 17,904 synonym candidates with votes (5% of all synonyms candidates in the Thesaurus). Both datasets will be taken into account during the next Thesaurus upgrade in order to reduce the amount of irrelevant noise.

The results of both crowdsourcing activities have also produced a number of interesting findings. The results of crowdsourcing are similar to the results of the expert evaluation of the Thesaurus, which indicates that the method is indeed applicable to language resource compilation. The evaluation has also shed light on measures to be taken in the future: because of its automatic origin, the Thesaurus also consists of some unusual headwords that would otherwise not have been included (e.g. *zbirati se v bazen* ‘to gather in a pool’ – *zbirati se v tolmun* ‘to gather in a pond’). This also raises the question of the degree to which the Thesaurus is limited in terms of data, as it contains only the phrases used as translation equivalents for English headwords in The Oxford®-DZS Comprehensive English-Slovenian Dictionary. This calls for a more thorough analysis of the relevance of not only the synonym candidates, but the headwords as well, especially the ones that have not yet been queried by users (i.e. the ones not in the 48% of the Thesaurus queried so far). In the interface this issue might be addressed by providing users with the option to downvote headwords as well.

On the other hand, the voting system has provided votes for only 5% of the Thesaurus so far, which indicates that lexicographers would require more features to involve users in the cleanup. Even the most motivated users currently have no way of systematically contributing toward the improvement of the Thesaurus (other than by solving tasks of the day or by searching for random headwords). The evaluation presented in this paper has shown that while an evaluation of the entire Thesaurus is possible, user motivation is key. The inclusion of the Thesaurus data in a gamified environment would be an even more efficient and expedient manner of crowdsourcing user votes. To tackle this issue a mobile game is already in development as part of our future work. More targeted and short-term crowdsourcing campaigns extended to a larger crowd and aimed at solving specific problems would also be beneficial, particularly in combination with future updates that will add new, automatically extracted synonyms and evaluate user synonyms added through the interface. And last but not least, the annotated data we only briefly analysed in the paper offers great potential for linguistic studies on synonymy and the development of a concept of synonymy based on usefulness, as defined by the collective intuition of the language community.

6. Acknowledgements

The research presented in this paper was conducted within the project titled *The Thesaurus of Modern Slovene: By the Community for the Community* (2018–2019), which is financially supported by the Ministry of Culture of the Republic of Slovenia.

The authors acknowledge the financial support from the Slovenian Research Agency (research core funding No. P6-0411, Language Resources and Technologies for Slovene). The research was conducted within the framework of the CA160105 eNetCollect COST Action. The authors would also like to thank all the users of the Thesaurus of Modern Slovene and the crowdsourcers who participated in the synonym evaluation campaign: Haris Agović, Ajda Diaci, Zoran Fijavž, Barbara Gorišek, Tajda Liplin Šerbetar, Angelika Markič, and Jana Šter.

7. References

- Abel, A. & Meyer, C. (2013). The dynamics outside the paper: user contributions to online dictionaries. In I. Kosem et al. (eds.) *Proceedings of eLex 2013*, pp. 179–194.
- Arhar Holdt, Š., Čibej, J., Dobrovoljc, K., Gantar, A., Gorjanc, V., Klemenc, B., Kosem, I., Krek, S., Laskowski, C. & Robnik Šikonja, M. (2018). Thesaurus of Modern Slovene: By the Community for the Community. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (eds.) *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*. ISBN 978-961-06-0097-8). Ljubljana: Znanstvena založba Filozofske fakultete. 2018, pp. 401-410. <https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/118/211/3000-1>
- 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: proceedings of eLex 2015 Conference, 11-13 August 2015, Herstmonceux Castle, United Kingdom*. Ljubljana: Trojina, Institute for Applied Slovene Studies; Brighton: Lexical Computing. 2015, pp. 70-83. https://elex.link/elex2015/proceedings/eLex_2015_05_Cibej+Fiser+Kosem.pdf
- Fišer, D., Tavčar, D. & Erjavec, T. (2014). sloWCrowd: A crowdsourcing tool for lexicographic tasks. *Proceedings of LREC 2014*.
- Fort, K., Guillaume, B. & Chastant, H. (2014). *Creating ZombiLingo, a Game With A Purpose for dependency syntax annotation*. Proceedings of the Gamification for Information Retrieval (GamifIR'14) Workshop, Amsterdam, The Netherlands, April 2014.
- Kosem, I., Krek, S., Gantar, P., Arhar Holdt, Š., Čibej, J. & Laskowski, C. (2018). 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, Ljubljana University Press, Faculty of Arts, pp. 989-997. <https://e-knjige.ff.unilj.si/znanstvena-zalozba/catalog/view/118/211/2939-1.pdf> (25 August 2018).
- Krek, S, Laskowski, C., Robnik-Šikonja, M., Kosem, I., Arhar Holdt, Š., Gantar, P., Čibej, J., Gorjanc, V., Klemenc, B. & Dobrovoljc, K. (2018b). *Thesaurus of Modern Slovene 1.0*. <http://hdl.handle.net/11356/1166> (3 June 2019)

- Krek, S., Laskowski, C., Robnik-Šikonja, M., Kosem, I., Arhar Holdt, Š., Gantar, P., Čibej, J., Gorjanc, V., Klemenc, B. & Dobrovoljc, K. (2018a). *Sopomenke 1.0: Thesaurus of Modern Slovene*, <https://viri.cjvt.si/sopomenke> (17 June 2019).
- 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. 93-107. <https://elex.link/elex2017/wp-content/uploads/2017/09/paper05.pdf>.
- Lew, R. (2014). *User-generated content (UGC) in online English dictionaries*. OPAL - Online publizierte Arbeiten zur Linguistik 2014.4: 8-26. https://repozytorium.amu.edu.pl/bitstream/10593/5011/1/Lew_GAL.pdf
- Snow, R., O'Connor, B., Jurafsky, D. & Ng., A. Y. (2008). Cheap and Fast – But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, Honolulu, October 2008, pp. 254–263. <https://www.aclweb.org/anthology/D08-1027>.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>

