How Lexicographers Evaluate User Contributions in The Thesaurus of Modern Slovene in Comparison to Dictionary Users

Magdalena Gapsa[1], Špela Arhar Holdt[1,2]

[1] University of Ljubljana, Faculty of Arts, Aškerčeva cesta 2, 1000 Ljubljana, Slovenia

[2] University of Ljubljana, Faculty of Computer and Information Science, Večna pot 113, 1000 Ljubljana, Slovenia

E-mail: Magdalena.Gapsa@ff.uni-lj.si, Spela.ArharHoldt@fri.uni-lj.si

Abstract

User involvement can be a valuable asset in expediting the process of language resource development, given that a thoughtful methodology is implemented. A successful example is the Thesaurus of Modern Slovene, which incorporates user participation to improve its automatically generated content. To shed light on the otherwise invisible lexicographic decision-making processes and to develop editorial protocols based on the needs of dictionary users, we investigated how differently lexicographers evaluate user-suggested synonyms compared to other user groups. We conducted an evaluation of nearly 1,000 user-suggested synonyms, assessed by a total of 42 evaluators from 7 user groups, and tested four hypotheses about lexicographers as evaluators. After evaluation, the Inter-Annotator Agreement (IAA) in all groups was calculated using Krippendorff's alpha and entropy, the evaluators' comments were classified into bottom-up categories, and the data were statistically analysed. In accordance with our assumptions, the lexicographers provided the most detailed arguments and identified the highest number of potential shortcomings of the suggested synonyms. However, they also scored the second lowest IAA among all groups and were more opposed to discarding user suggestions. We discuss the possible reasons for these results and emphasise their value for the further development of responsive dictionaries.

Keywords: user involvement; responsive dictionary; synonyms; user evaluation; lexicographers

1. Introduction

The Thesaurus of Modern Slovene is a state-of-the-art example of a digitally-born dictionary created automatically from pre-existing openly available language resources (Krek et al., 2017).[1] It was prepared to address the lack of openly available synonym data for modern Slovene, and it serves as a benchmark for data reusability and user involvement for other languages facing similar issues. The development of the Thesaurus is based on a responsive dictionary model (Arhar Holdt et al., 2018), where the initial version of the resource is generated automatically and made available to the public under an open licence as soon as it is deemed useful. The data is then gradually revised, with the help of users, to ensure ongoing improvement. This iterative process is vital due to the presence of noise and the absence of certain types of essential lexical

[1] Thesaurus of Modern Slovene is available in the interface at https://viri.cjvt.si/sopomenke/eng/ and as a database at http://hdl.handle.net/11356/1166.

information in the automatically generated database.[2]

In the Thesaurus, the users are allowed to suggest new synonym candidates and evaluate existing ones. The possibilities for user participation, as well as many other novelties introduced by the responsive dictionary model, were positively rated and well accepted by the user community (Arhar Holdt, 2020). In practice, allowing the option of suggesting new synonyms has proven especially fruitful, as the number of collected synonym candidates is high: 60,976 at the time of writing. To ease participation, user suggestions are displayed in the dictionary interface immediately and without editorial intervention. However, a lexicographic review and approval process is required before suggestions can be included in the openly accessible dictionary database.

Although a preliminary study by Arhar Holdt and Čibej (2020) suggested that a very limited number of user inputs were malicious, there is currently no large-scale study on the content and relevance of user-suggested data. Conducting such a study would enable an assessment of the quality of user contributions and identification of potential problems that could be addressed to enhance user participation. To address this gap, we carried out an evaluation campaign utilising almost 1,000 user-suggested synonyms from the Thesaurus of Modern Slovene. A total of 42 evaluators, chosen based on their profession or interests, participated in the study.[3] In Gapsa (2023), a summative analysis of the results was presented, while this paper focuses specifically on how lexicographers evaluated user-suggested synonyms in comparison to other user groups, such as language editors, translators, and teachers.

2. Related work

The present study belongs to the field of lexicographic user research and builds upon established methodological frameworks (a comprehensive overview of existing methodologies is provided in Welker, 2013a, 2013b). Lexicographic user research emphasises the importance of user-centred design in the development and evaluation of lexicographic products. It has a tradition reaching back to the 1960s (e.g. Barnhart, 1962; Householder, 1967), but the research area was firmly established later in the 1980s and 1990s (e.g. Tomaszczyk, 1979; Hartman, 1987; Atkins, 1998; Nesi, 2000). The emergence of the digital medium in the 2000s offered a vast array of new methodological possibilities (e.g. Bergenholtz and Johnsen, 2013; Müller-Spitzer, 2014; Lew and De Schryver, 2014). In the last decade, existing approaches were also critically evaluated and surpassed (Bogaards, 2003; Tarp, 2009; Lew, 2015; Kosem et al., 2018):

[2] The data published in Thesaurus 1.0 was not lexicographically post-processed. The entries and synonym candidates were presented in a form of lemmata (without part-of-speech or other metadata that would help disambiguate between forms), semantic descriptions were replaced by automatically obtained semantic clusters, and the data also lacked dictionary labels, apart from domain ones. Version 2.0, currently undergoing testing, aims to address some of these issues, as outlined by Arhar Holdt et al. (In press).

[3] The gathered data are available in the Repository of the University of Ljubljana: http://hdl.handle.net/20.500.12556/RUL-144064

older studies have most often been criticised for having too few participants or for being too homogeneous (students were the most likely group to participate, as they are the easiest for researchers to access).

In our case, the participants in the study represent dictionary users, while at the same time serve as evaluators of user-suggested synonyms. Previous studies, mainly from the field of NLP, have shown that non-experts are capable of successfully performing tasks of assessing synonymy or word similarity. Crowdsourced evaluations of synonyms have been applied in various contexts, such as evaluating the degree of similarity between words (Schnabel et al., 2015) and creating gold standards for evaluation and training tasks (e.g. Hill et al., 2015; Schneidermann et al., 2020). Human annotations of similarity have been used as evaluation methods in Word-in-Context and SemEval tasks (e.g. Pilehvar and Camacho-Collados, 2019; Breit et al., 2021; Armendariz et al., 2020), and crowdsourcing-oriented tools have been developed for different wordnets to detect and correct errors (e.g. Braslavski et al., 2014; Fišer et al., 2014; Rambousek et al., 2018).

3. Methodology

3.1 Preparation of the dataset

Similar to intrinsic evaluations in NLP tasks (see e.g. Schnabel et al., 2015 and Schneidermann et al., 2020), where pre-selected inventories of word pairs are used, we used a list of 546 Slovene nouns occurring as headwords (or headword-like units) in various openly available language resources for modern Slovene: the Thesaurus of Modern Slovene 1.0 database (Krek et al., 2018), the sloWNet 3.1 database (Fišer, 2015), the Lexical Database for Slovene (Gantar et al., 2013), the Comprehensive Slovenian-Hungarian Dictionary (Kosem et al., 2021), and the database of nouns labelled with semantic types (Kosem and Pori, 2021).[4] We then extracted user-suggested synonyms for these nouns from the Thesaurus of Modern Slovene 1.0 interface using a custom made script, prepared specifically to track user contributions. The number of suggestions varied for each noun, and not all nouns had suggestions. In total, we extracted 972 synonyms for 307 nouns.

3.2 Selection of user groups

We selected the desired user groups based on the typology of potential dictionary users by Arhar Holdt et al. (2016, pp. 181-184) and the results of a study on user attitudes towards the lexicographic novelties introduced by the Thesaurus (Arhar Holdt, 2020, p. 477). On the one hand, the typology provided a theoretical overview of the user

[4] This work is part of a larger study in a PhD research project aiming to improve the connectivity and reusability of Slovene synonym data in the digital environment. Certain decisions, e.g. the selection of headwords for the evaluation data, were made with other research objectives in mind.

groups according to the main situations of dictionary use (in the educational process, for professional purposes or for leisure activities). On the other hand, the user study indicated which user groups were most represented among the participating active users of the Thesaurus.

Combining both pieces of information as well as our research questions, we have selected 7 user groups, as presented in Figure 1: Lexicographers (L), Translators (T), Language Editors (LEd), Marketers (M), Teachers of Slovene (ToS), Language Enthusiasts (LEn), and Students (S) of linguistic studies. Our aim was to cover all three scenarios of dictionary usage. We included lexicographers in the study due to their critical role in the editorial process of evaluating synonyms. In addition to representing the educational aspect of the study, we also included students to pilot the research before its wider implementation (Gapsa, 2022).

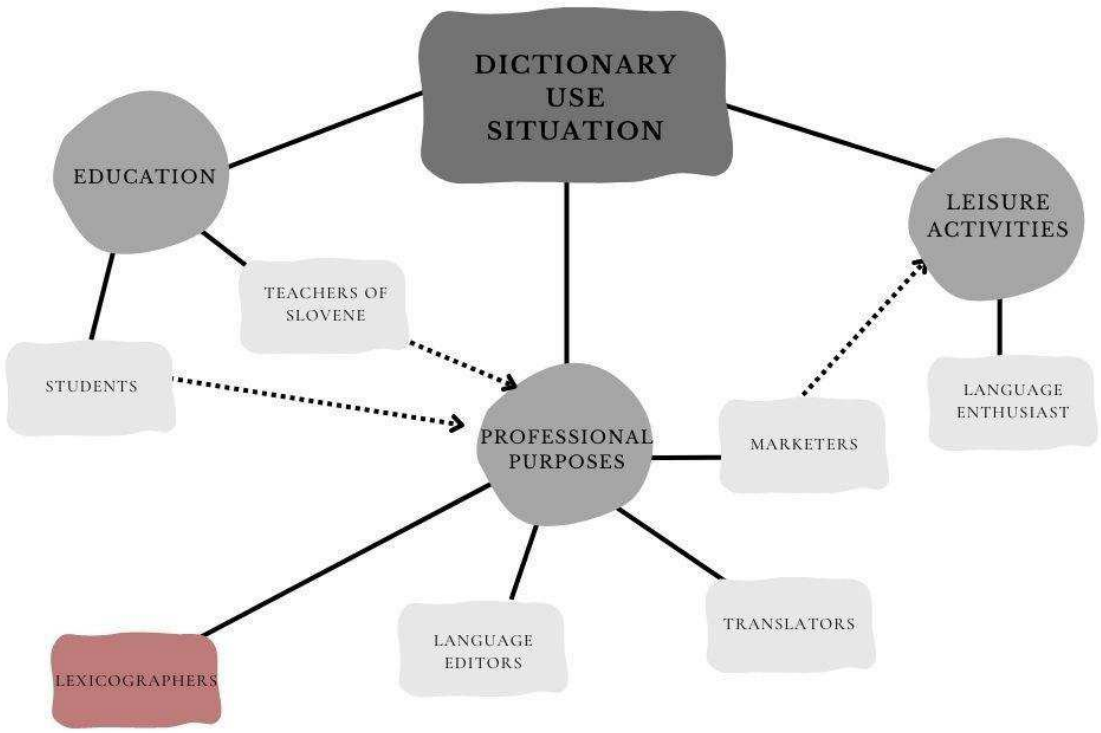

Figure 1: Overview of the selected user groups based on three main dictionary use situations

3.3 Recruiting participants

Considering the cautionary notes against qualitative user studies with a too limited number of participants (Tarp, 2009, 290), and taking into account the resources available for our study, we opted to include six evaluators per group, for a total of 42

evaluators.

The first groups recruited were Students, who had at the time participated in the development of the Thesaurus from 1.0 to 2.0.[5] They already knew the Thesaurus and had experience in analysing linguistic data and could help test the evaluation process, tools and guidelines, as well as estimate the time needed for the task and set a financial compensation for the participants. Secondly, the group of Lexicographers was assembled under the umbrella of the same project. Recruitment of representatives from other user groups took place in several rounds. The call for applications for Teachers of Slovene, Translators and Language Editors was published, first via the CJVT newsletter and then via the CJVT Facebook profile. A call for applications for Language Enthusiasts, which was also answered by Marketers, was posted in two Facebook groups, which serve as a forum for asking and answering language-related questions: 'For at least approximately correct use of the Slovene language' and 'Association of Amateur Orthographers AND Grammarians'.[6] The call briefly presented the task and the conditions of participation, including the payment.

3.4 Data evaluation

The participant data was prepared in separate Google Sheets spreadsheets,[7] where we listed all 972 user-suggested synonyms and their corresponding headwords. Each participant was asked to evaluate whether the words in each pair were synonyms or not by answering the question, "Are the words in the pair synonyms?" for all 972 pairs. Table 1 presents the four possible answers and their suggested uses. In cases where participants answered "CONDITIONAL YES,' it was mandatory for them to explain the specific issues they identified. While comments were encouraged for the other three answer options, they were not mandatory.

Answer	**When to use**
YES	If you believe that the words in the pair are synonyms.
NO	If you believe that the words in the pair are not synonyms or in the case of obvious errors, typos, etc.

[5] Project Synonyms and Collocations 2.0 – SoKol, Upgrading fundamental dictionary resources and databases of CJVT UL.

[6] In Slovene: Za vsaj približno pravilno rabo slovenščine and Društvo ljubiteljskih pravopisarjev IN slovničarjev.

[7] Google Sheets was used due to its accessibility, popularity, cost-effectiveness and option for continuous editing and saving of the answers.

CONDITIONAL YES	If you believe that the words in the pair can be synonyms, but at the same time you see limitations or have doubts, e.g. because the words are synonyms only in a certain meaning or context, one or both words are marked, etc.
NOT SURE/DON'T KNOW	If you are not sure whether the words are synonyms, you do not know one or both of the words in the pair or the meaning of one or both of the words in the pair, or you have difficulty deciding.

Table 1: Overview of possible answers in the guidelines for evaluators

The objective was to test the evaluators' understanding of relevant synonymous data. The guidelines provided to participants were intentionally general, without defining synonymy or providing examples of potential synonym pairs (as opposed to e.g. Hill et al., 2015, where a brief definition of similarity was provided together with examples of similar word pairs to better illustrate the difference between similarity and association or relatedness) or suggesting where borderline cases should be classified to avoid influencing the participants' answers. Similarly to Hill et al., 2015, we wanted the participants to rely on their language intuition (thus we discouraged them from consulting other language resources like dictionaries, corpora, etc.) and presented them with context-free word pairs (which is also an experience users get when browsing Thesaurus 1.0, as synonym candidates are listed without sense disambiguation or examples of use).

To ensure quality control of the evaluation process, participants also completed a brief questionnaire using the online survey tool 1ka.[8] The questionnaire was designed to collect background information about the evaluators and confirm their placement in the designated user groups. It also enabled participants to provide feedback about potential problems with the evaluation process.

3.5 Research Hypotheses

For this study, we tested 4 hypotheses about Lexicographers as an evaluator group:[9]

[8] Online survey tool 1ka: https://www.1ka.si/d/en

[9] The formulation of the four hypotheses was driven by the aim of ensuring quality control in the user participation aspect of the dictionary-making process. In our workflows, lexicographers, who serve as the editors of the dictionary and possess first-hand experience in organising synonyms in the Thesaurus, undertake the evaluation of user contributions on behalf of the participating community. For this study, it is crucial to establish the lexicographers' evaluations as a gold standard and explore the divergence of their decisions from those made by other participating groups. Consequently, we are also testing hypotheses that may appear obvious or counterintuitive from this particular standpoint.

- H1: Lexicographers' evaluation would be more consistent and their Inter-Annotator Agreement would be higher than in other groups.
- H2: Lexicographers would argue their decisions in more detail than other groups.
- H3: Lexicographers would make statistically different decisions about (un)acceptability of user-suggested synonyms and identify more potential problems with user-suggested synonyms than other groups.
- H4: Lexicographers would be more reserved to include user-suggested synonyms than other groups.

3.6 Data Analysis

To address the hypotheses, different approaches were used.

Firstly, Inter-Annotator Agreement (IAA) between the evaluators was calculated using Krippendorff's alpha (Krippendorff, 1970).[10] Calculations were made for each of the synonym pairs within each user group to facilitate clustering of IAA levels (as opposed to manually identifying all possible IAA levels) and to make the data more comparable between groups. The total number of answers received was 40,801, as a total of 23 answers were missing. Since the possible answers were nominal categories and not a scale, entropy[11] was calculated to determine the distribution of possible answers.

Secondly, evaluators' comments were manually categorised according to their content. The categories were created bottom-up, based on the material analysed. The final list of categories comprised 11 possible categories, some of which allowed for further sub-categories, notably the category Other. Multi-layered categorisation was used because some of the comments, although coming from a single commentator, contained multiple pieces of information that could be classified into different categories, e. g. "dialectal and calque"[12], "a stylistic label would be needed, in one of the meanings", etc. The categories and their definitions, as well as selected examples of categorised comments, can be found in Table 2.

[10] The IAA is usually calculated using Fleiss' Kappa (see M. Vila et al. 2015, p. 85), however, Krippendorff's alpha (Krippendorff, 1970) was used here because of rare cases of missing answers.

[11] Both calculations are very sensitive to the subtlest differences in answers, therefore both were used as a filtering tool to facilitate the analysis and comparison of the results.

[12] Translations are approximate and may not cover all specifics. Slovene headwords and suggestions are provided with English translations. Evaluators' comments are presented in English. Translations aim to aid understanding and fluency of reading.

Category name	**Definition**	**Example of evaluators' comments**	**Synonym pair**
limited context or certain sense(s) of the word(s)	context or certain sense; limited usage; other senses or a need for sense disambiguation	Synonyms only in one meaning.	žoga – podaja ('a ball' - 'a pass')
insufficient sense	additional qualifiers seem to be a necessary component of the meaning	A piece of fabric intended for cleaning can be a cloth, let's say.	blago – krpa ('a fabric' - 'a cloth')
semantic discrepancy	semantically related but not necessarily always interchangeable words ; related but different concepts	The customer is not necessarily the subscriber. It can be a random customer or just a visitor to the shop/store etc.	stranka – naročnik ('a client' - 'a subscriber')
alternative semantic relation	other semantic relationship (e.g. hyper-/hyponymy, mero-/holonymy)	The suggested synonym is a hypernym of the headword.	hotel – prenočišče ('a hotel' - 'an accommodation')
unknown word or sense	unfamiliarity with word or suggested sense	I do not know the second word.	izseljenec – ezul ('an emigrant' - 'an exile')
definition	explanation, definition or description	The suggested synonym sounds more like a definition to me.	anatomija – veda o telesni zgradbi ('an anatomy' - 'science of body structure')
incomplete word units	multi-word expressions suggested as single words	In the form of *imeti pogum* - *imeti jajca*.	pogum – jajca ('a courage - 'balls')
opinionizing	evaluators opinion on the	Perhaps a little	elita – creme de la crème

	suggested synonym	bit too French.	('an elite' - 'crème de la crème')
foreign words	loanword, foreignism, calque or non-standard loanword	Merely as a literal translation of a foreign word from Latin.	aplikacija – namestitev ('an application' - 'an installation')
marked	marked word or a qualifier or tag needed, sometimes very specific, e.g. dialectal, pejorative	Colloquially.	cigareta – dim ('a cigarette' - 'a smoke')
other	remarks that do not fall into the above categories	Consider singular-plural.	pošta – maili ('a post' - 'mails')

Table 2: Comment categories with definitions and examples of use

It was also possible to identify certain problems that occurred with the user-suggested synonyms, but which were not frequent enough to be included in a separate category. Such comments were subcategorised within main categories. This was particularly the case with e.g. phraseological units or metaphorical senses, which created subcategories within main category *Limited context or certain sense(s) of the word(s)*, cases of meronymy, which were put under main category *Alternative semantic relation* or specific semantic labels that were mentioned with comments regarding a headword or user-suggested synonym being *Marked*.

Thirdly, to determine possible dependencies between the user groups and their most frequent comments, statistical tests were carried out, i.e. contingency tables were prepared and a chi-square test was run, followed by calculations of Pearson residuals to determine whether there were statistically significant differences between the groups. Pearson residuals below -1.92 or above 1.92 indicate a statistically significant difference. In the following chapter, we present the results of the study.

4. Results

4.1 Consistency and Inter-Annotator Agreement

Our first hypothesis was that Lexicographers would be the group with the highest IAA of all groups, which would indicate that their answers are more inherently consistent than those of the other groups. The hypothesis is based on the presumption that lexicographers evaluate user-suggested synonyms on the basis of common and

comparable expert knowledge and experience, which would facilitate higher consistency.

To test the hypothesis, we compared: "full IAA", where all evaluators within the group chose the same answer; "very high IAA", where 5 out of 6 evaluators chose the same answer; "high IAA", where 4 out of 6 evaluators chose the same answer; and "moderate IAA", where 3 out of 6 evaluators chose the same answer. Here, we distinguished "tied answers": the instances where 3 evaluators agreed on one answer and the remaining 3 evaluators agreed on another answer. Figures in Table 3 show that, on average, evaluators scored at least *high IAA* on almost 60% of the whole evaluation set and *moderate IAA* on 33% of the set.

User group	**Full IAA**	**Very high IAA**	**High IAA**	**TOTAL at least high IAA**	**Moderate IAA**	**Tied answers**
Lexicographers	28 (3 %)	136 (14 %)	**341 (35 %)**	505 (52 %)	395 (41 %)	**130 (13 %)**
Language Editors	139 (14 %)	222 (23 %)	286 (29 %)	647 (67 %)	271 (28 %)	58 (6 %)
Language Enthusiasts	52 (5 %)	149 (15 %)	336 (35 %)	537 (55 %)	359 (37 %)	109 (11 %)
Marketers	**188 (19 %)**	**256 (26 %)**	272 (28 %)	**716 (74 %)**	219 (23 %)	59 (6 %)
Translators	46 (5 %)	195 (20 %)	300 (31 %)	541 (56 %)	349 (36 %)	32 (3 %)
Students	34 (3 %)	140 (14 %)	263 (27 %)	437 (45 %)	**396 (41 %)**	72 (7 %)
Teachers of Slovene	165 (17 %)	209 (22 %)	285 (29 %)	658 (68 %)	255 (26 %)	55 (6 %)

AVERAGE	**93 (10 %)**	**187 (19 %)**	**298 (31 %)**	**577 (59 %)**	**321 (33 %)**	**74 (8 %)**

Table 3: Distribution of number of pairs with at least high IAA between groups

Lexicographers achieved the second lowest *at least high IAA* among all groups (the only group that scored lower were Students, see Discussion). Their *full* and *very high IAA* was the lowest of all the evaluator groups, at only 3% and 14% respectively (again, a similar percentage was achieved by the Student group). On the other hand, their *high IAA* (35%) was the highest of all groups, followed by Language Enthusiasts. Lexicographers also scored the second highest number of pairs with *moderate IAA*, closely after Students. Finally, they scored the highest number of pairs with tied answers. These results reject the first hypothesis: data shows that Lexicographers were below average in terms of IAA, their answers within the group were less consistent and most often tied in comparison to other groups.

4.2 Detailed argumentation of the decisions

The second hypothesis assumed that the Lexicographers would give a more detailed argumentation of their decisions indicating that they were better informed about the potential problems of the data than other evaluator groups. To test this assumption, we compared the number of comments made and categorised between the different groups and the number of categorised comments for each category within the groups. The numbers are shown in Table 4.

User group	**L**	**LEd**	**LEn**	**M**	**T**	**S**	**ToS**	**TOTAL**	**AVG.**
Comments made	**2,717**	363	783	640	1,234	2,593	252	**8,582**	1,226
Comments categorised[13]	1,802	388	708	609	1,249	**1,845**	246	**6,846**	978

[13] Repeating comments were deduplicated – if multiple evaluators in the same group made comments that fell into the same category, it was only counted once.

limited context or certain sense(s) of the word(s)	**625**	51	121	65	166	435	18	**1,481**	212
insufficient sense	5	28	40	31	**89**	60	35	**288**	41
semantic discrepancy	36	56	57	92	**200**	188	35	**664**	95
alternative semantic relation	75	44	35	28	80	**190**	19	**471**	67
unknown word or sense	247	53	115	194	166	**276**	83	**1,134**	162
definition	**93**	0	17	1	22	65	0	**198**	28
incomplete word units	23	1	9	9	**58**	17	5	**122**	17
opinionizing	6	17	9	11	11	**22**	1	**77**	11
foreign words	0	19	15	**43**	36	22	0	**135**	19
marked	425	92	247	84	279	**426**	27	**1580**	226
other	**267**	27	43	51	142	144	23	**697**	100

Table 4: Number of comments made and categorised per user group and the distribution of the comment categories among the user groups. The abbreviations are: L –Lexicographers, LEd – Language Editors, LEn – Language Enthusiasts, M – Marketers, T – Translators, S – Students, ToS – Teachers of Slovene, AVG. – average

As the figures in Table 4 show, the Lexicographers indeed made the highest number of comments of all the evaluator groups. When comparing the number of categorised comments, Lexicographers scored second highest. The group that behaved most similarly to Lexicographers were again Students.

As mentioned in Section 3.6, some categories were further divided, particularly the category *Other*. Not only did Lexicographers contribute the most comments to this category, their comments also generated most subcategories: about 30 subcategories

compared to 10-15 subcategories[14] in the other evaluator groups. The subcategories most frequently observed in the Lexicographers group were:

- coined synonyms - the comments indicated that this vocabulary is probably characteristic of the suggester's idiolect, and therefore hardly understood or used by the wider community, e.g. *klitoris* 'a clitoris' – *gumbek* 'a button', *menstruacija* 'a menstruation' – *rdeča armada* 'red army',

- terminological correctness - the comments indicated that it needed to be checked whether the suggested synonym can be used in a terminological sense of the headword, e.g. *epidemija* 'an epidemic' – *pandemija* 'a pandemic', *mandarina* 'a mandarine' – *klementina* 'a clementine',

- collocations - the comments indicated that the suggested synonym might be collocative or part of a collocation, e.g. *avtoriteta* 'an authority' – *spoštovan strokovnjak* 'a respected professional', *babica* 'a granny' – *starejša gospa* ' an elderly lady',

- alternative spellings - the comments indicated that a word has no standard written form or that different spellings are possible, e.g. *bonbon* – *bombon* 'a candy', *parfum* – *parfem* 'a perfume',

- doubts on actual use - the comments indicated that it needed to be checked whether the user-suggested synonym is confirmed in modern language, e.g. *alkohol* 'alcohol' – *veselje* 'a joy', *ogrlica* 'a necklace' – *kolje* 'a necklace', *smrad* 'a stench' – *zaudarek* 'a reek',

- doubts on the frequency of use - the comments indicated that it needed to be checked whether the user-suggested synonym is frequent enough in the modern language, e.g. *avtoriteta* 'an authority' – *veščak* 'an expert', *izseljenec* 'an emigrant' – *ezul* 'an exile'.

Overall, Lexicographers made more comments in total and those categorised as *Other* than other groups. Moreover, their comments revealed more subcategories, especially within the category *Other*. These subcategories reflect issues identified and commented on more often or typically by Lexicographers. Both facts support the hypothesis that Lexicographers would give more detailed and informed argumentation of their answers and decisions.

4.3 Focus on different problems

The third hypothesis assumed that Lexicographers' decisions about (un)acceptability of users suggestions would be statistically different from decisions of other groups, as

[14] Except for Students, whose comments contained ample explanations that could be sorted into nearly 30 subcategories.

lexicographers are likely to identify different potential problems than other evaluator groups. To test this assumption, contingency tables were prepared and a chi-square test of independence was performed to finally calculate the Pearson's residuals. The calculations of the Pearson's residuals are shown in Table 5.

Category	**L**	**LEd**	**LEn**	**M**	**T**	**S**	**ToS**
limited context or certain sense(s) of the word(s)	11,915	-3,594	-2,597	-5,814	-6,337	1,798	-4,827
insufficient sense	-8,132	2,891	1,873	1,064	5,031	-1,998	7,664
semantic discrepancy	-10,496	2,995	-1,407	4,286	7,167	0,679	2,282
alternative semantic relation	-4,397	3,351	-1,964	-2,146	-0,638	5,600	0,505
unknown word or sense	-2,978	-1,405	-0,209	9,274	-2,841	-1,692	6,620
definition	5,664	-3,350	-0,768	-3,958	-2,349	1,594	-2,667
incomplete word units	-1,607	-2,249	-1,018	-0,562	7,577	-2,769	0,295
opinionizing	-3,169	6,050	0,368	1,586	-0,813	0,275	-1,062
foreign words	-5,961	4,104	0,279	8,944	2,292	-2,384	-2,202
marked	0,450	0,261	6,542	-4,769	-0,543	0,012	-3,951
other	6,170	-1,988	-3,424	-1,396	1,318	-3,197	-0,408

Table 5: Pearson residuals of the distribution of the comment categories among the user groups. The abbreviations are: L –Lexicographers, LEd – Language Editors, LEn – Language Enthusiasts, M – Marketers, T – Translators, S – Students, ToS – Teachers of Slovene

The group of Lexicographers was the one that most frequently commented on the need for sense disambiguation, while other groups were less concerned about it. Secondly, different evaluator groups frequently commented that the suggestion lacked an essential sense component to be considered synonymous while Lexicographers rarely made such comments. Thirdly, Lexicographers rarely commented on semantic discrepancies between the headword and the user-suggested synonyms, while other groups reported such cases quite frequently. Furthermore, they also reported cases of alternative semantic relations less frequently than other groups. The data also show that Lexicographers were less likely to report cases of unknown word(s) or meaning(s). On the other hand, they were more likely than other groups to comment that the suggestion is a "definition" or "description" rather than a synonym. There were no significant differences between Lexicographers and other evaluators in reporting cases of incomplete word units.

The data presented in Table 5 also clearly show that Lexicographers were less inclined to comment on the foreign origin of word(s), while other groups (with the exception of the Teachers of Slovene) emphasised this relatively frequently. They were also somewhat less likely than other groups to provide comments that had no other value but to express opinions. Marked vocabulary was commented on by the Lexicographers at approximately the same rate as within other groups. They did, as already mentioned, contribute more comments that were categorised as *Other* than the remaining groups.

If we summarise the above results and the data from the previous section, we can conclude that the third hypothesis is true. Lexicographers did indeed focus on other issues. Possible explanations for these findings are addressed in the Discussion.

4.4 Rigour and reserve in incorporating user suggestions

The fourth hypothesis assumed that Lexicographers are more rigorous in their decisions and more reserved to accept user suggestions and consequentially include them in the Thesaurus database. To test this assumption, we compared the total number of NO and CONDITIONAL YES answers within each evaluator group and the distribution of answers chosen by the evaluators in the *full, very high* and *high IAA* cases. Table 6 shows the total number of answers given by each group. The highest values for each answer are underlined and in bold.

User group	TOTAL given answers[15]	YES	NO	CONDITIONAL YES	NOT SURE/DON'T KNOW
Lexicographers	5,829	2,720	492	**1,956**	661
Language Editors	5,823	3,009	1,908	467	439
Language Enthusiasts	5,828	2,916	**1,924**	611	377
Marketers	**5,832**	**3,590**	1,404	300	538
Translators	5,831	2,614	1,687	742	788
Students	5,831	1,797	1,187	1,940	**907**
Teachers of Slovene	5,827	3,383	1,556	407	481
AVERAGE	**5,829**	**2,861**	**1,451**	**918**	**599**

Table 6: Total number of answers given per evaluators group.

As the figures in Table 6 show, Lexicographers gave the answer CONDITIONAL YES more frequently than other evaluators groups. Students achieved an almost identical total number of CONDITIONAL YES answers, while other evaluators gave this answer much less frequently. The total number of CONDITIONAL YES answers supports the assumption that Lexicographers would be more cautious and reserved to include user-suggested synonyms as they were suggested. However, the total number of NO answers proves that the assumption that Lexicographers would reject more data was wrong, as

[15] Occasional missing answers were noted, therefore the numbers given in column 2 vary between groups and rarely equals the total number of possible answers in a group (6 evaluators x 972 pairs = 5,832 possible answers).

Lexicographers gave the NO answer significantly less often than other groups.

Similar results can be observed when looking at the distribution of answers in pairs with *at least high IAA*, which is shown by Figure 2. It shows the summarised number of pairs with each of the possible answers per evaluator group and the average distribution of answers in the case of *full, very high* and *high IAA*.

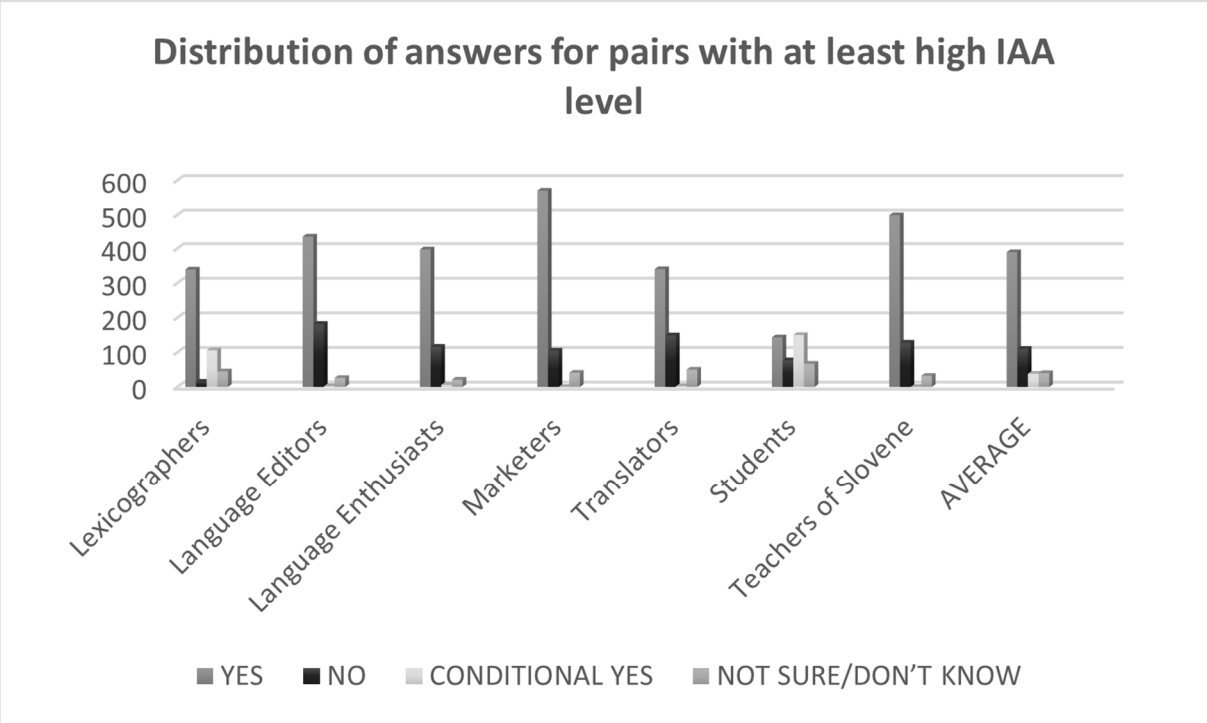

Figure 2: Distribution of number of answers per IAA level.

As the data in Figure 2 show, the two groups that chose the answer CONDITIONAL YES more often than other groups and at the same time achieved at least a high IAA were Students and Lexicographers, suggesting that they made more comments explaining their scruples about the synonym pair, but were also less decisive than other groups who tended to answer YES or NO. The strictest group that rated most pairs as unsuitable were the Language Editors, while Lexicographers turned out to be the least strict and rigorous group of all evaluators.

5. Discussion

The results yielded valuable information about Lexicographers as evaluators. Out of four hypotheses concerning Lexicographers and their decisions when evaluating synonymy, only two were corroborated. The data revealed that Lexicographers were the least consistent group, with the second lowest overall Inter-Annotator Agreement

(IAA) score (factoring in *full, very high*, and *high IAA* cases) and the highest number of tied responses. Furthermore, they were the least rigorous, deeming only a small proportion of data unsuitable for the Thesaurus. However, Lexicographers demonstrated a broader perspective than other groups, frequently selecting CONDITIONAL YES as their answer and offering insights into issues and problems that other evaluators addressed less frequently. The results also indicate that Lexicographers prioritised different issues than other evaluator groups.

Initially, the Lexicographers' answers were meant to serve as a benchmark for evaluation. It was assumed that the lexicographic team's expertise would uniformly reflect the main problems and needs of Thesaurus users and that other evaluator groups would validate this. However, the presented analysis of the Lexicographers' answers revealed that this would not be possible. While the low inter-annotator agreement (IAA) among evaluators was partially due to the four possible decisions allowed, it was surprising that the Lexicographers scored below average on IAA and were more indecisive than other groups. The only group with a lower *at least high IAA* was the Students, however, their performance may have been influenced by imperfect guidelines and a poorer understanding of the task since they were simultaneously evaluating the data and testing the evaluation design (see Gapsa, 2022).

We had expected the Lexicographers to identify both more and different issues with the user-suggested content, while also covering the most common problems and limitations of the Thesaurus and its data. We were surprised to find that they placed disproportionate emphasis on certain issues, which highlights the fact that not all evaluator groups have a universal opinion of the Thesaurus's limitations. Lexicographers focused more on the lack of sense disambiguation and cases of definition instead of actual synonymy, while semantic discrepancies, insufficient senses, or foreign origin of vocabulary were issues raised more frequently by other evaluators. It is possible that the Lexicographers were biased by previous attempts to identify user needs and develop updating solutions, leading them to identify such cases more frequently than other groups. They also operated with more precise terminology, which can explain some of the differences.[16]

The design of the research itself may have influenced the Lexicographers' responses. The evaluators were not limited to binary YES-NO choices, but could also select a NOT SURE/DON'T KNOW response or a CONDITIONAL YES response. Lexicographers, in particular, were more likely to choose the latter option than other evaluator groups (with the exception of Students). From a lexicographic perspective, the difference between YES and CONDITIONAL YES responses, especially when combined with comments, is significant. It indicates that either the suggestion or the

[16] Lexicographers' familiarity with "dictionary definitions" facilitated their recognition, but some of the user-suggested synonyms identified by Lexicographers as definitions were actually between descriptions or hypernyms, which other evaluators considered as alternative semantic relations.

headword requires further review and editing, which should be prioritised due to the inadequacy of the current data. Interestingly, Lexicographers were less likely to give a NO response than other groups, perhaps due to their desire to preserve as many synonym candidates as possible and thus provide Thesaurus users with multiple options to choose from. To assist users in making their choice, Lexicographers wanted to ensure that the suggested synonyms were accompanied by semantic information, labels, usage examples, and so on, rather than simply discarding imperfect data. Additionally, Lexicographers did not hesitate to acknowledge that they were unfamiliar with some of the vocabulary. However, the total number of such responses in the Lexicographers group was only slightly higher than average.

The Students group and the Lexicographers shared some interesting similarities. The Inter-Annotator Agreement and number of comments made were almost identical in both groups. Notably, the Students also provided detailed comments, particularly those that were further subcategorized under the “Other” category. They also emphasised alternative spellings, terminological correctness, and issues related to the frequency of use or actual usage of vocabulary. Both groups displayed a greater awareness of the Thesaurus' limitations and had a better understanding of how to name and address them. They were also involved in the updating process and understood the tools and technologies available to facilitate lexicographic review processes, such as verifying data with corpora. Additionally, both groups appeared to take the task more seriously than the other groups, as evidenced by the considerable number of comments as well as the lack of humorous remarks. This could potentially explain the other similarities observed between them.

6. Future work

In this paper, we aimed to explore the differences in how synonymy is perceived and evaluated by Lexicographers, who are experts in the field and typically viewed as the primary evaluators of user-suggested data, and six other groups representing a broader community of dictionary users with diverse professions and interests in language data. The results of the evaluation campaign not only provide a basis for future studies but also have practical implications. They will serve as a guide for drafting editorial protocols, prioritising tasks, and improving the Thesaurus of Modern Slovene. The findings clearly indicate the need for detailed lexicographic guidelines that define appropriate data and the types of additional information pertaining to user suggestions. The guidelines should be based on the priorities identified in the study and supported by empirical data from corpora, as evidenced by the Lexicographers' comments in the "Other" category. The comments highlighted issues such as alternative spellings, frequency of use, and evidence of use in specific meanings, which must be considered in the editorial protocols for future Thesaurus updates. An application-oriented approach would be to add new types of information to the Thesaurus, such as semantic disambiguation, labels, and metadata. Some of these solutions have already been incorporated in the updated version of Thesaurus 2.0.

This paper provides insights into the development of similar online language resources for other languages, based on the involvement of users as collaborators. The study shows that users can offer relevant and useful synonym candidates, but it is also important to involve them as evaluators. The significant differences in the evaluation of synonymy between Lexicographers and other evaluator groups highlight the ongoing need to monitor community priorities and needs and to address them to ensure the actual responsiveness of the responsive lexical resources.

7. Acknowledgements

The authors acknowledge that the project Empirical foundations for digitally-supported development of writing skills (J7-3159) and the programme Language Resources and Technologies for Slovene (P6-0411) were financially supported by the Slovenian Research Agency. The Agency funds the first Author PhD research fund within the programme P6-0411, from which the majority of the evaluators were compensated. Additional funding source for evaluators was the project Upgrading fundamental dictionary resources and databases of CJVT UL, funded by the Ministry of Culture of the Republic of Slovenia in the period 2021–2022.

8. References

Arhar Holdt, Š. (2020). How Users Responded to a Responsive Dictionary: the Case of the Thesaurus of Modern Slovene. *Rasprave Instituta za hrvatski jezik i jezikoslovlje*, 46(2), pp. 465–482. https://doi:10.31724/rihjj.46.2.1

Arhar Holdt, Š., & Čibej, J. (2020). Rezultati projekta "Slovar sopomenk sodobne slovenščine: Od skupnosti za skupnost". In D. Fišer, & T. Erjavec (eds.) *Zbornik konference Jezikovne tehnologije in digitalna humanistika, 24. – 25. september 2020, Ljubljana, Slovenija*, pp. 3–9. Available at: http://nl.ijs.si/jtdh20/pdf/JT-DH_2020_Arhar-Holdt-et-al_Rezultati-projekta_Slovar-sopomenk-sodobne-slovenscine.pdf

Arhar Holdt, Š., Čibej, J., Dobrovoljc, K., Gantar, P., Gorjanc, V., Klemenc, B., . . . 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, 17-21 July 2018, Ljubljana*, pp. 401–410). Available at: https://euralex.org/wp-content/themes/euralex/proceedings/Euralex%202018/118-4-2991-1-10-20180820.pdf

Arhar Holdt, Š., Gantar, P., Kosem, I., Pori, E., Robnik Šikonja, M. & Krek, S. (In press). Thesaurus of Modern Slovene 2.0. *Proceedings of the Eighth Biennial Conference eLex 2023, Brno, Czech Republic, 27–29 June 2023.*

Arhar Holdt, Š., Kosem, I., & Gantar, P. (2016). Dictionary User Typology: The Slovenian Case. In T. Margalitadze, & G. Meladze (eds.) *Proceedings of the XVII EURALEX International Congress: Lexicography and Linguistic Diversity*, pp. 179-187. Available at https://euralex.org/wp-content/themes/euralex/proceedings/Euralex%202016/euralex_2016_015_p179

.pdf

Armendariz, C.S., Purver, M., Pollak, S., Ljubešić, N., Ulčar, M., Vulić, I., & Pilehvar, M.T. (2020). SemEval-2020 Task 3: Graded Word Similarity in Context. In A. Herbelot, X. Zhu, A. Palmer, N. Schneider, J. May, & E. Shutova (eds.) *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pp. 36–49. https://doi:10.18653/v1/2020.semeval-1.3

Atkins, S.B.T. (ed.). (1998). *Using Dictionaries: Studies of Dictionary Use by Language Learners and Translators*. Tübingen: Max Niemeyer Verlag.

Barnhart, C.L. (1962). Problems in editing commercial monolingual dictionaries. *International Journal of American Linguistics*, 28(2), pp. 161–181.

Bergenholtz, H. & Johnsen, M. (2005). Log files as a tool for improving Internet dictionaries. *Hermes. Journal of Linguistics*, 34, pp. 117–141.

Bogaards, P. (2003). Uses and users of dictionaries. In P. Van Sterkenburg (ed.) *A Practical Guide to Lexicography*. Amsterdam and Philadelphia: John Benjamins, pp. 26–33.

Braslavski, P., Ustalov, D., & Mukhin, M. (2014). A Spinning Wheel for YARN: User Interface for a Crowdsourced Thesaurus. In S. Wintner, M. Tadić, & B. Babych (eds.) *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 101–104. https://doi:10.3115/v1/E14-2026

Breit, A., Revenko, A., Rezaee, K., Pilehvar, M.T., & Camacho-Collados, J. (2021). WiC-TSV: An Evaluation Benchmark for Target Sense Verification of Words in Context. In P. Merlo, J. Tiedemann, & R. Tsarfaty (eds.) *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 1635–1645. Available at: https://aclanthology.org/2021.eacl-main.140/

Fišer, D., Tavčar, A., & Erjavec, T. (2014). sloWCrowd: A crowdsourcing tool for lexicographic tasks. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, . . . S. Piperidis (eds.) *Proceedings of the Ninth International Conference on Language Resources and Evaluation. LREC'14, May 26-31, 2014, Reykjavik, Iceland*, pp. 3471–75. Available at: http://www.lrec-conf.org/proceedings/lrec2014/pdf/1106_Paper.pdf

Gapsa, M. (2022). Ocenjevanje uporabniško dodanih sopomenk v Slovarju sopomenk sodobne slovenščine – pilotna študija. In D. Fišer, & T. Erjavec (eds.) *Proceedings of the Conference on Language Technologies and Digital Humanities, September 15th – 16th 2022, Ljubljana, Slovenia*, pp. 308–316. Available at: https://nl.ijs.si/jtdh22/pdf/JTDH2022_Gapsa_Ocenjevanje-uporabnisko-dodanih-sopomenk-v-Slovarju-sopomenk-sodobne-slovenscine.pdf

Gapsa, M. (2023). *"But why??" Evaluation of user-suggested synonyms in the Thesaurus of Modern Slovene*. Research Square. https://doi.org/10.21203/rs.3.rs-2775161/v1

Hartmann, R.R.K. (1987). Four perspectives on dictionary use: a critical review of research methods. In A.P. Cowie (ed.) *The Dictionary and the Language Learner*.

Tübingen: Niemeyer, pp. 11–28.

Hill, F., Reichart, R., & Korhonen, A. (2015). SimLex-999: Evaluating Semantic Models With (Genuine) Similarity Estimation. *Computational Linguistics*, 41(1), pp. 665–695. https://doi:10.1162/COLI_a_00237

Householder, F.W. (1967). Summary report. In F.W. Householder and S. Saporta (eds.) *Problems in lexicography*. Bloomington: Indiana University Publications, pp. 279–282.

Kosem, I., & Pori, E. (2021). Slovenske ontologije semantičnih tipov: samostalniki. In I. Kosem (ed.) *Kolokacije v slovenščini*, pp. 159–202. https://doi:10.4312/9789610605379

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, 17-21 July 2018, Ljubljana*, pp. 989–997. Available at: https://euralex.org/publications/collocations-dictionary-of-modern-slovene

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, C. Tiberius, M. Jakubíček, J. Kallas, S. Krek & V. Baisa (eds.) *Proceedings of eLex 2017: Lexicography from Scratch, 19-21 September 2017, Leiden, Netherlands*, pp. 93-109. Available at: https://elex.link/elex2017/wp-content/uploads/2017/09/paper05.pdf

Krippendorff, K. (1970). Estimating the reliability, systematic error, and random error. *Educational and Psychological Measurement*, 30(1), pp. 61–70.

Lew, R. & de Schryver, G.-M. (2014). Dictionary users in the digital revolution. *International Journal of Lexicography*, 27(4), pp. 341–359.

Lew, R. (2015) Opportunities and limitations of user studies. In C. Tiberius & C. Müller-Spitzer (eds.) *Research into dictionary use. Wörterbuchbenutzungsforschung. 5. Arbeitsbericht des wissenschaftlichen Netzwerks „Internetlexikografie“*. Mannheim: Institut für deutsche Sprache, pp. 6–16.

Müller-Spitzer, C. (2014). *Using Online Dictionaries*. Berlin – Boston: De Gruyter Mouton.

Nesi, H. (2000). *The Use and Abuse of EFL Dictionaries. How learners of English as a foreign language read and interpret dictionary entries*. Tübingen: Max Niemeyer Verlag.

Pilehvar, M.T., & Camacho-Collados, J. (2019). WiC: the Word-in-Context Dataset for Evaluating Context-Sensitive Meaning Representations. In J. Burstein, C. Doran, & T. Solorio (eds.) *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 1267–1273. https://doi:10.18653/v1/N19-1128

Rambousek, A., Horák, A., & Pala, K. (2018). Sustainable long-term WordNet development and maintenance: Case study of the Czech WordNet. *Cognitive*

Studies/Études cognitives, 18. https://doi:10.11649/cs.1715

Schnabel, T., Labutov, I., Mimno, D., & Joachims, T. (2015). Evaluation methods for unsupervised word embeddings. In L. Màrquez, C. Callison-Burch, & J. Su (eds.) *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 298–307. https://doi:10.18653/v1/D15-1036

Schneidermann, N., Hvingelby, R., & Pedersen, B. (2020). Towards a Gold Standard for Evaluating Danish Word Embeddings. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, . . . S. Piperidis (eds.) *Proceedings of the 12th Language Resources and Evaluation Conference (LREC) 2020, Marseille, France, 13th-15th May 2020*, pp. 4754–4763.

Tarp, S. (2009). Reflections on lexicographical user research. *Lexikos*, 19(1), pp. 275–296. https://doi:10.5788/19-0-440

Tomaszczyk, J. (1979). Dictionaries: users and uses. *Glottodidactica* 12, pp. 103–119.

Vila, M., Bertran, M., Martí, M. A., & Rodríguez, H. (2015). Corpus Annotation with Paraphrase Types: New Annotation Scheme and Inter-annotator Agreement Measures. *Language Resources and Evaluation*, 49, pp. 77–105. https://doi.org/10.1007/s10579-014-9272-5

Welker, H.A. (2013a). Methods in Research of Dictionary Use. In R.H. Gouws, U. Heid, W. Schweickard & H.E. Wiegand (eds.) *Dictionaries. An International Encyclopedia of Lexicography: Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography*. Berlin, New York: Walter de Gruyter, pp. 540–547.

Welker, H.A. (2013b). Empirical Research into Dictionary Use since 1990. In R.H. Gouws, U. Heid, W. Schweickard & H E. Wiegand (eds.) *Dictionaries. An International Encyclopedia of Lexicography: Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography*. Berlin, New York: Walter de Gruyter, pp. 531–540.

Databases:

Fišer. D. (2015). *Semantic lexicon of Slovene sloWNet 3.1*. Slovenian language resource repository CLARIN.SI. Available at: http://hdl.handle.net/11356/1026

Gantar, P., Krek, S., Kosem, I., Šorli, M., Kocjančič, P., Grabnar, K. … Nina Drstvenšek, N. (2013). *Leksikalna baza za slovenščino 1.0*. Slovenian language resource repository CLARIN.SI. Available at: http://hdl.handle.net/11356/1030

Kosem, I., Bálint Čeh, J., Ponikvar, P., Zaranšek, P., Kamenšek, U., Koša, P. … Gorjanc, V. (2021). *Comprehensive Slovenian-Hungarian Dictionary 1.0*. Slovenian language resource repository CLARIN.SI. Available at: http://hdl.handle.net/11356/1453.

Krek, S., Laskowski, C., Robnik-Šikonja, M., Kosem, I., Arhar Holdt, Š., Gantar, P. … Dobrovoljc, K. (2018). *Thesaurus of Modern Slovene 1.0*. Slovenian language resource repository CLARIN.SI. Available at: http://hdl.handle.net/11356/1166