Up to No Good: Exploiting Word Embeddings for an Automatic Extraction of Candidates for a Lexicon of Slovene Taboo Language

Jaka Čibej

Centre for Language Resources and Technologies, University of Ljubljana
Faculty of Computer and Information Science, University of Ljubljana
Faculty of Arts, University of Ljubljana
Jožef Stefan Institute
E-mail: jaka.cibej@ff.uni-lj.si

Abstract

We present an approach to extracting candidates to be included in an open-access lexicon of Slovene taboo language by using word embeddings compiled from different Slovene corpora and a set of offensive and pejorative seed lexemes from the *Thesaurus of Modern Slovene 2.0*. While many studies on taboo language rely on surveys to collect data on taboo language and its use, our evaluation shows that the method with embeddings provides a good starting point for the compilation of a more comprehensive and empirically grounded taboo language lexicon. We describe the datasets used in the experiment, the process of extraction and its results, as well as the advantages and disadvantages of this method. From a set of approximately 120 Slovene seed lexemes, the initial analysis of extracted candidates resulted in 1,260 relevant lexemes. We briefly discuss potential future steps in the development of the lexicon in the context of other machine-readable Slovene language resources, such as the *Digital Dictionary Database of Slovene*. The extraction method is language-independent and can be directly applied to other languages.

Keywords: taboo language; automatic extraction; embeddings; corpora; Slovene

1. Introduction

Note: This paper contains language some may find offensive.

Lexicons of taboo language (i.e. language that some may find offensive) are useful language resources that can serve multiple purposes. In addition to their direct use either to automatically censor words deemed inappropriate for a given context (e.g., to help mitigate the problem of online hate speech or prevent malicious user contributions to digital dictionaries), they can also help filter out materials not suitable for educational purposes (see Zingano Kuhn et al., 2022) or games with a purpose (Arhar Holdt et al., 2020). They can also be used when training general large language models to remove pornographic and racist content from training data.

Taboo language lexicons can also be useful for linguistic analyses and contrastive translation studies as swearing and taboo language are frequently culturally specific. In addition, taboo language, particularly the section related to hate speech, needs to be well-documented in dictionaries as they are used as authoritative language resources (Gorjanc, 2005). Kaplan (2021), for instance, has emphasized the importance of lexicography in preventing the promotion of othering and oppression; lexicographers must avoid practices which actively provide problematic definitions or labels or omit important information on language use and its context (by not including all data on connotations of offensive lexemes or their context).

What is included in existing Slovene lexicographic resources, the largest being the *Dictionary of Slovenian Literary Language*[1] is either not available under an open-access license, is inaccurately or insufficiently represented (for instance, the only label used in the *Dictionary of Slovenian Literary Language* is *slabšalno* (pejorative), even though the context can be radically different in terms of intensity or tabooness: cf. *bedak* 'fool' vs. *peder* 'faggot'), or is limited in scope (e.g., the Thesaurus of Modern Slovene; Krek et al., 2023), with material originating mostly from corpora of standard Slovene, where the usage of offensive vocabulary is limited.

Machine-readable lexicons of taboo language have been compiled from existing language resources for other languages (e.g. van Huyssteen & Tiberius, 2023; Bond & Choo, 2021; more on this in Section 3), but no such lexicon exists yet for Slovene. We present the first step in constructing this lexicon by introducing an approach to compiling a list of Slovene taboo language candidates using the *fastText* embeddings trained on a number of Slovene corpora (including web crawls).

The paper is structured as follows: we provide a brief overview of the related work in Section 3 and describe the datasets used in our experiment in Section 3. In Section 4, we explain the method of extracting candidates, then annotate a portion of the extracted candidates to evaluate the effectiveness of the method (Section 5). We discuss the advantages and disadvantages of the method in Section 6 and conclude the paper with several suggestions for future work and steps for the future development of a lexicon of Slovene taboo language (Section 7).

2. Related Work

Many psycholinguistic studies of swearing and taboo words have already been carried out for English (see for instance Jay, 1992; McEnery, 2006; Jay & Janschewitz, 2008; Jay, 2020; a more comprehensive overview is also available in van Huyssteen & Tiberius, 2023), and the topic of taboo language has been frequently explored in recent years in natural language processing in the context of the automatic detection of hate speech or offensive content (e.g., Rosenthal et al., 2021; Schmidt & Wiegand, 2017) or removing problematic content from datasets used to train models: Qiu et al. (2024) perform several steps to remove unsafe content from the *Common Crawl* dataset, and Penedo et al. (2023) use a blacklist-approach to score page URLs in terms of the degree of unsafe content.

A recent psycholinguistic study on taboo words in multiple languages was conducted by Sulpizio et al. (2024), collecting taboo words for 13 different languages using surveys. Approximately 1,000 participants were involved in total (40-150 per language). For Slovene, a similar psycholinguistic study on taboo words was carried out by Kos (2024), while Klemenčič (2016) conducted a contrastive study of swearing in Slovene and Swedish. However, both studies focused on a limited set of expressions (either hand-picked or

[1] *The Dictionary of Slovenian Literary Language* (*Slovar slovenskega knjižnega jezika*) is available online at www.fran.si.

from surveys). The reliance of existing studies on surveys and hand-picked expressions demonstrates that there is a need for a comprehensive lexicon of Slovene taboo language in machine-readable format.

Similar projects aimed at providing such datasets have been undertaken for other languages: the Dutch lexicon *Taboelex* (van Huyssteen & Tiberius, 2023), for instance, is being compiled from existing taboo language resources for Dutch. Other examples include *Taboo WordNet* (Bond & Choo, 2021), a lexical resource with Japanese taboo language expressions, and *Hurtlex* (Bassignana et al., 2018), a collection of taboo words from approximately 50 languages.[2]

Our experiment with word embeddings is inspired by the use of embeddings in a lexicographic context as presented by Sørensen & Nimb (2018), who exploit semantic similarity to assist the editing of dictionary entries (by alerting the lexicographer about semantically similar units from the corpus that have not yet been included in the dictionary). We employ a similar approach to extracting taboo language candidates, as described in the following sections.

3. Data

The data we used in our experiment were extracted from the following language resources: the *Thesaurus of Modern Slovene 2.0* (Krek et al., 2023) and the *CLARIN.SI-embed.sl* word embeddings (Ljubešić & Erjavec, 2018; Terčon et al., 2023b). We describe the resources in more detail in the following subsections.

3.1 The Thesaurus of Modern Slovene 2.0

The *Thesaurus of Modern Slovene 2.0* (Krek et al., 2023) is an open-access dictionary of Slovene synonyms available in XML format under the Creative Commons CC-BY-SA 4.0 license. The Thesaurus was compiled automatically from an English-Slovene bilingual dictionary using co-occurrence graphs (Krek et al., 2017) and was designed as a fully digital responsive dictionary which features an interface in which users can add synonyms in a separate user-synonym section. Version 2.0 features some manual lexicographic work (described by Arhar Holdt et al., 2023) and forms an integral part of the *Digital Dictionary Database of Slovene* (DDDS; Kosem et al., 2021), a central dictionary database for Slovene. From the point of view of this paper, the most relevant improvement of the Thesaurus in version 2.0 is the addition of labels for several vulgar and extremely offensive entries (Arhar Holdt et al., 2022). These were added after a manual analysis of approximately 40,000 entries by three annotators (students of linguistics), and a final decision accepted by lexicographers. It should be noted, however, that the current version of the Thesaurus does not contain a comprehensive list of Slovene taboo-language entries; only the ones already included in the Thesaurus were assigned labels. The labeled vulgar and offensive entries were used as seed lexemes in our experiment (more on this in Section 3.3).

[2] Several ad hoc lists of offensive words also exist on various Github repositories, such as the *List of Dirty, Naughty, Obscene, and Otherwise Bad Words*; https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words

3.2 CLARIN.SI-embed.sl Word Embeddings

To extract more candidates for the lexicon of Slovene taboo language, we use embeddings based on the fastText architecture (Bojanowski et al., 2017). Although pre-trained fastText embeddings are openly available for 157 languages[3] (Grave et al., 2018), including Slovene, they were trained on texts from the *Common Crawl* dataset[4] and *Wikipedia.*[5] We use two versions of *CLARIN.SI-embed.sl*, a collection of pre-trained fastText embeddings trained using the skipgram model on a large collection of Slovene texts composed of existing corpora of Slovene such as the *GigaFida Corpus of Written Slovene* (Krek et al., 2020), the *JANES Corpus of Internet Slovene* (Erjavec et al., 2018), the *KAS Corpus of Academic Slovene* (Žagar et al., 2022), the *slWaC Web Corpus of Slovene* (Erjavec et al., 2011), and the *MaCoCu-sl Web Corpus of Slovene* (Bañón et al., 2023). The advantage of these corpora is that they have been annotated with state-of-the-art tools for morphosyntactic tagging and lemmatization of Slovene (Ljubešić et al., 2023; Terčon et al., 2023a; Terčon & Ljubešić, 2023). Similar to *Common Crawl*, *slWaC* and *MaCoCu-sl* also contain web texts in non-standard Slovene, which are less regulated and less edited than standard texts from newspapers, and are thus expected to contain more informal language use and more taboo language content. According to some studies, taboo language exhibits very low written frequency (Sulpizio et al., 2024), which is nevertheless slightly higher in texts like tweets (approximately 1% of the words we write on Twitter as adults are taboo words, as reported by Wang et al., 2014.). Unlike *Common Crawl*, the corpora have been more carefully curated to remove boilerplate texts, formatting errors, and texts misclassified as Slovene.

Version 1.0 (Ljubešić & Erjavec, 2018) contains two sets of embeddings; the first was compiled based on lower-case lemma forms (in combination with their morphosyntactic features), while the other is more robust and was compiled on lower-case word forms only. Version 2.0 (Terčon et al., 2023b) only contains embeddings compiled from lower-case word forms, but these were based on a larger corpus.

In this paper, we used the lemma embeddings from version 1.0 and the form embeddings from version 2.0. For clarity, we refer to these sets of embeddings as *CLARIN.SI-embed.sl-lemma 1.0* (with embeddings for 2,093,848 units) and *CLARIN.SI-embed.sl-form 2.0* (with embeddings for 3,471,054 units). The embeddings are 100-dimensional: two truncated examples are shown in Table 1, with the lemma embedding for *leto* 'year' as a common noun (*Nc* - noun, common - according to the MTE-6 morphosyntactic specifications) and the form embedding for the inflected form *leta* (which can be either genitive singular or nominative plural). We explain how these embeddings were used to extract candidates in Section 4.

3.3 Preparation of Seed Lexemes

Figure 1 shows an excerpt from the XML format of the *Thesaurus of Modern Slovene 2.0*, with the lexeme *imbecil* 'imbecile' (in teal color), its morphological features according to

[3] fastText word vectors for 157 languages: https://fasttext.cc/docs/en/crawl-vectors.html

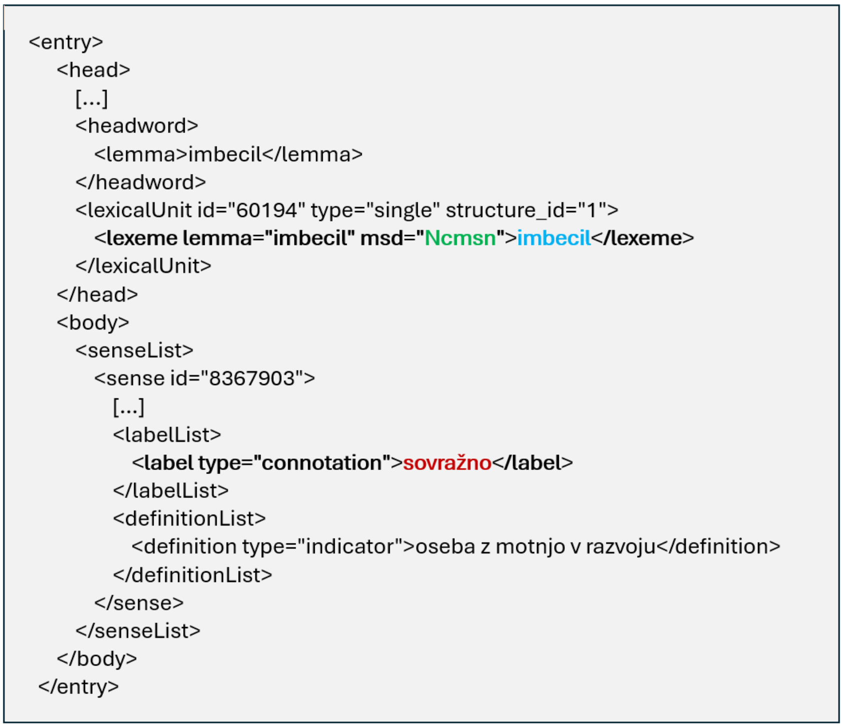
[4] Common Crawl: https://commoncrawl.org/

[5] Wikipedia: https://www.wikipedia.org/

Resource	Unit	Truncated Embedding
CLARIN.SI-embed.sl-lemma 1.0	leto#Nc	[-0.2615, ..., -0.093361, -0.21532]
CLARIN.SI-embed.sl-form 2.0	leta	[-0.36728, ..., -0.58205, -0.15223]

Table 1: Examples of (truncated) embeddings from *CLARIN.SI-embed.sl-lemma 1.0* and *CLARIN.SI-embed.sl-form 2.0.*

the Multext-East v6 morphosyntactic specifications for Slovene[6] (in green), and a label (in red) under a specific sense with a definition.



```
<entry>
  <head>
    [...]
    <headword>
      <lemma>imbecil</lemma>
    </headword>
    <lexicalUnit id="60194" type="single" structure_id="1">
      <lexeme lemma="imbecil" msd="Ncmsn">imbecil</lexeme>
    </lexicalUnit>
  </head>
  <body>
    <senseList>
      <sense id="8367903">
        [...]
        <labelList>
          <label type="connotation">sovražno</label>
        </labelList>
        <definitionList>
          <definition type="indicator">oseba z motnjo v razvoju</definition>
        </definitionList>
      </sense>
    </senseList>
  </body>
</entry>
```

Figure 1: Excerpt from the *Thesaurus of Modern Slovene 2.0* (XML format).

We first extracted all lexemes from the *Thesaurus of Modern Slovene 2.0* that contained at least one sense with *connotation* or *register* labels. Other labels (such as *domain*, *context*, *grammar*, *time*) were ignored. The *connotation* labels consisted of the following values: *grobo* (vulgar, coarse language), *sovražno* (hateful/offensive), *lahko izraža negativen odnos* (can express a negative attitude), and *izraža negativen odnos* (expresses a negative attitude). The only value of the *register* labels was *neformalno* (informal).

The extraction resulted in 181 distinct units (along with their MTE-6 morphosyntactic features). We filtered out the 25 multiword units as the *CLARIN.SI-embed.sl* embeddings only focus on single-word units. We then manually categorized the remaining 156 single-lexeme units according to a bottom-up categorization consisting of a total of 12 groups (as shown in Table 2).

[6] Multext-East v6 Morphosyntactic Specifications for Slovene: https://nl.ijs.si/ME/V6/msd/html/msd-sl.html

Category	**Number of Seed Lexemes**	**Examples**
Drugs and Alcohol Intoxication	11	*nažehtanost* ‘drunkenness (informal)’, *zadet* ‘high (informal)’, *nacejanje* ‘drinking (informal)’
Homophobia and LGBTIQA+	6	*peder*, ‘faggot’, *buzi* ‘fag’, *lezbača* ‘lesbo’
Misogyny	11	*baba* ‘hag’, *candra* ‘slut’, *kurba* ‘whore’
Racism	6	*črnuhinja* ‘negress’, *rumenokožec* ‘yellowskin’, *zamorec* ‘nigger’
Sexual References	33	*fukač* ‘fucker’, *kurac* ‘cock’, *natepavati* ‘to bang’
Political References	2	*rdečuh* (pejorative for a communist/leftist), *rdečuhinja* (pejorative for a female communist/leftist)
Other Vulgar and Pejorative Expressions	37	*crkniti* ‘to drop dead’, *lizun* ‘sycophant’, *razpizden* ‘pissed (angry)’, *zjebati* ‘to fuck up’
Xenophobia	3	*cigan* ‘gypsy’, *makaronar*, *polentar* (pejoratives for Italians)
Disability	5	*izrodek* ‘degenerate’, *kripelj* ‘cripple’, *nakaza* ‘freak’
Scatological References and Bodily Fluids	9	*drek* ‘shit’, *driska* ‘diarrhoea’, *usran* (adjective; covered in shit)
References to Cognitive Abilities	5	*debil* ‘idiot’, *debilen* ‘idiotic’, *imbecil* ‘imbecile’
Non-Taboo, but Non-Neutral Vocabulary	29	*biciklirati* ‘to cycle (informal), *žurirati* ‘to party (informal), *bordanje* ‘boarding (sport; informal)’

Table 2: Categorized seed lexemes from the *Thesaurus of Modern Slovene 2.0.*

We provide a brief description of each category and its contents. The category *Drugs and Alcohol Intoxication* consists of lexemes related to the use of alcohol, drugs and other illegal substances, as well as words related to people experiencing addiction. Although words from this category may not all be offensive or considered taboo in all contexts, we nevertheless treat them as relevant for the taboo lexicon because they can be problematic in educational contexts. The category *Homophobia and LGBTIQA+* contains slurs and pejorative words directed at the members of the LGBTIQA+ community, or pejorative words referring to non-heteronormative sexual activity. The category *Misogyny* contains pejorative words for women and girls that imply the female gender as a negative characteristic. The *Racism* category contains racial slurs that other people based on their skin pigmentation. Similarly, *Xenophobia* contains words that discriminate people based on their nationality or geographical origins. Among the most represented categories are *Sexual References*, with vulgar words referring to different types of sexual activities and participants in sexual activities. The category *Disability* contains pejorative words for people with disabilities, while *References to Cognitive Abilities* contains pejorative words that refer to a person's intellectual capacity or their mental health. *Political References* contained only two entries (*rdečuh* and *rdečuhinja*; pejorative expressions for left-leaning people or communists). *Scatological References* contained words related to excrement and other bodily fluids (vomit, spit, urine). A more general category *Other Vulgar and Pejorative Expressions* included all the other vulgar expressions that could not be classified into any of the other categories, such as the adjective *razpizden* (a vulgar word for 'angry'). The category *Non-Taboo, but Non-Neutral Vocabulary* was not used in the extraction of taboo candidates as it consisted of words that were not in any way offensive or problematic, but were nonetheless not neutral in terms of register (with informal and slang expressions).

It should be noted that these 11 categories do not constitute the final taxonomy that will be used to compile the lexicon of Slovene taboo language, as some notable categories (such as religious references or pejoratives used for body shaming) are missing from the set of seed lexemes extracted from the *Thesaurus of Modern Slovene*. A more comprehensive categorization will be made once more candidates are extracted and analyzed.

In some cases, a single unit could be categorized into multiple categories (depending on sense and context). For instance, *kurba* 'whore' could be classified as both a sexual reference or as an expression of misogyny. For the purposes of this experiment, we resolved these dilemmas by estimating the unit's semantic similarity to other candidates within each category and selecting the group which shared the most similar candidates. The assumption was that in the corpora the embeddings we use are based on, *kurba* would share more contexts with other words from the group on misogyny such as *psica* 'bitch' and *cipa* 'slut', as opposed to other references to sexual activities (such as *fukanje* 'fucking' or *tič* 'cock'). However, this is a purely pragmatic solution, which we adopted in this experiment because the approach is very robust and focuses on the embeddings of lexemes, not their senses.

In addition, it is important to note that there is a difference between the word's etymological origin and its sense; e.g. the adjective *zafukan* 'fucked up' is derived from the verb *zafukati* 'to fuck up', which is derived from the verb *fukati* 'to fuck'. Although the use of the word *zafukan* arguably evokes associations with sexual activity, its actual sense is more generally vulgar (i.e. it can be said of someone who is difficult to deal with). Similarly, the verb *napizditi* 'to scorn somebody' is derived from *pizditi* 'to speak angrily, to complain', which is derived from *pizda* 'cunt'. We categorized these into the most suitable semantic category based on their prevalent context from the *Gigafida Corpus of Standard Written Slovene* (Krek et al., 2020) and the *JANES Corpus of Internet Slovene* (Erjavec et al., 2018). The examples of *zafukan* and *napizditi*, for instance, were assigned to *Other Vulgar and Pejorative Expressions* despite the sexual connotations of the words they are derived from.

4. Candidate Extraction

The extraction method we use is based on the premise that words that are similar in terms of their offensiveness are used in similar contexts. In a large corpus, they should be at least contextually (if not necessarily semantically) close. By focusing on an embedding of a seed lexeme and searching for lexemes with similar embeddings, the method should yield additional candidates that can be included in the lexicon of Slovene taboo language. This is demonstrated in Figure 2, which shows the visualization of a selection of embeddings from *CLARIN.SI-embed.sl-lemma 1.0*.[7] The green units represent the embeddings of offensive lexemes from the *References to Cognitive Abilities* category. The blue dots represent the embeddings for *pes* 'dog' and *maček* 'cat', the red dots are two Slovene words for 'friend' (*prijatelj* and *kolega*), and the orange dots are three of the most frequent functional words in Slovene (*biti* 'to be', *in* 'and', and *v* 'in'). With the exception of *bedak*, most of the embeddings for the offensive lexemes are much closer to each other than to other candidates.

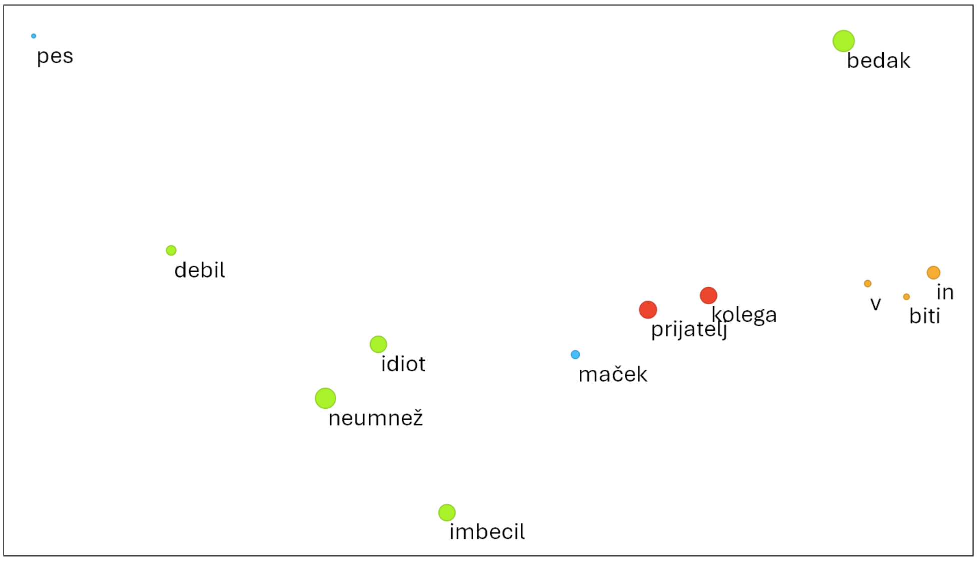


Figure 2: A selection of embeddings from *CLARIN.SI-embed.sl-lemma 1.0* visualized using Multi-Dimensional Scaling.

For each seed lexeme obtained from the *Thesaurus of Modern Slovene 2.0*, we extracted its embedding (first from *CLARIN.SI-embed.sl-lemma 1.0*, then from *CLARIN.SI-embed-sl-form 2.0*). This resulted in 122 seed embeddings for lemmas and 121 seed embeddings for forms. The difference can be attributed to the characteristics of three words: *uscane* (a

[7] The MDS visualization was made with *Orange Data Mining v3.38.0* (Demšar et al., 2013). The size of the dots corresponds to one of the embedding features with the greatest values among the offensive candidates in the selection.

pejorative noun for someone who has urinated over themselves; used to insult someone who is inexperienced or cowardly) is a homograph with the several inflected forms of the adjective *uscan* 'covered in urine'. All occurrences of *uscane* as a noun have probably been mislemmatized and mistagged as *uscan*, so the embedding is not present in the set of lemma embeddings. Similarly, the adjectives *joškat* and *joškast* (both meaning 'busty' in the context of women with large breasts) are missing from the form embeddings because they only occur in the feminine inflected forms (such as *joškata*, *joškasta*, *joškate*, *joškaste*).

We conducted two extractions of candidates: from CLARIN.SI-embed.sl-lemma 1.0 and CLARIN.SI-embed.sl-form 2.0. For each seed embedding (122 and 121, respectively), we compared it to each other embedding within the relevant set of embeddings by calculating the cosine similarity[8] for each pair:

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|}$$

In this way, approximately 255 million comparisons were made for the seed embeddings of lemmas (e.g., 122 seed embeddings × 2,093,848 total embeddings in *CLARIN.SI-embed.sl-lemma 1.0*) and approximately 420 million comparisons for form embeddings (121 seed embeddings × 3,471,054 total embeddings in *CLARIN.SI-embed.sl-form 2.0*). For the purposes of this experiment, we set the threshold for lowest cosine similarity at 0.75, which resulted in a large number of extracted candidates (as described below). Different numbers of candidates were extracted for different seed lexemes. The statistical distribution of extracted candidates is shown in Table 3.

Resource	**Mean**	**Median**	**Minimum**	**Maximum**	**Standard Deviation**
Lemmas	35,812.84	3,636	4	157,535	52,678.78
Forms	290.74	127	0	6,494	640.67

Table 3: Statistical distribution of extracted candidates (with a cosine similarity 0.75 or greater) for individual seed embeddings from *CLARIN.SI-embed.sl-lemma 1.0* and *CLARIN.SI-embed.sl-form 2.0*.

The lemma embeddings yielded a much larger quantity of candidates compared to form embeddings, with an average of approximately 36,000 candidates per seed embedding for lemmas (and a median of approximately 3,600) and only 291 candidates for forms (and a median of 127). This is expected, as form embeddings only take into account the context of one form, while the lemma embeddings focus on all occurrences of the lexeme in the corpus (with different inflected forms). It needs to be emphasized that not all of the extracted candidates are relevant – they simply represent words that occur in the same contextual radius (with the seed embedding as the centre) within the corpus that is the basis for the embeddings. We conduct a manual evaluation of the extraction method in Section 5. Because the lemma embeddings provided a much larger number of candidates,

[8] Cosine similarity takes a value between 0 and 1. A higher value indicates from contextual similarity between compared words.

we focused on those results in our evaluations due to the limited scope of this study, and leave the CLARIN.SI-embed-sl.form 2.0 results for future work.

We also analyzed the statistical distribution of extracted candidates by category. Table 4 shows the candidates by category for lemma embeddings. The highest number of candidates on average was extracted for lexemes from the categories *References to Cognitive Abilities*, *Homophobia and LGBTIQA+*, *Political References*, *Sexual References*, and *Scatological References and Bodily Fluids.* The offensive seed lexemes from these categories occur in a much more dense contextual space in Slovene corpora compared to other categories. A more systematic and in-depth study of the distribution of actual offensive lexemes is required, but the results possibly indicate that the most widespread manner of offending people in Slovene corpora is to insult their intelligence, shame them for their political views and negatively brand them as non-heteronormative.

Resource	**Mean**	**Median**	**Minimum**	**Maximum**	**Standard Deviation**
Cognitive	71,381.40	72,579	750	135,612	48,529.01
Disability	39,471	3,613	24	140,135	54,129.95
Drugs	863.55	130	8	3,825	1,348.58
LGBTIQA+	60,022.17	34,224	2,563	151,547	61,064.95
Misogyny	6,071.1	2,628	231	34,067	9,747.51
Political	58,700	58,700	740	116,660	57,960
Racism	11,302.17	1,784	138	51,651	18,472.65
Scatological	54,109.78	13,541	318	142,749	57,765.96
Sexual	44,029.97	1,749.5	135	157,535	61,953.49
Vulgar	34,211.6	5,171.0	4	147,906	48,629.51
Xenophobia	35,907.67	10,915	4,818	91,990	39,734.24

Table 4: Statistical distribution of extracted candidates by category from *CLARIN.SI-embed.sl-lemma 1.0.*

The extracted candidates are not unique to a single lexeme, however, and occur multiple times across lists for each individual seed embedding. Table 5 shows the distribution of only unique candidates within each category. The highest number of unique candidates was extracted in the *Sexual References* and *Other Vulgar and Pejorative Expressions* categories. The high number of unique candidates indicates that the threshold of 0.75 for cosine similarity could potentially be set higher. However, seed lexemes from the *Drugs and Alcohol Intoxication* category seem to occur in a much less dense contextual space, and the extraction yielded only approx. 5,700 candidates; with a higher threshold, the number could potentially be much lower. For discovering new candidates, a better approach is to cast a wider net in the beginning and truncate the extracted lists post-festum.

Table 6 shows an example of an extracted list of candidates (sorted by cosine similarity from highest to lowest) for the seed lexeme *debil* 'idiot' (from the *References to Cognitive Abilities* category). 11 of the 12 candidates are relevant for this category, while one (*pizdek*)

Category	**Number of Unique Candidates**
Drugs and Alcohol Intoxication	5,723
Homophobia and LGBTIQA+	157,130
Misogyny	44,651
Racism	56,064
Sexual References	190,737
Political References	117,071
Other Vulgar and Pejorative Expressions	196,444
Xenophobia	96,104
Disability	142,499
Scatological References and Bodily Fluids	165,204
References to Cognitive Abilities	144,006
Total	222,058

Table 5: Unique candidates by category extracted from *CLARIN.SI-embed.sl-lemma 1.0*.

is more generally vulgar. However, several lemmatization issues are apparent, such as *budal* (which should be *budala*), *pizdeki* (which should be *pizdek*), as well as *debili* and *debiliz* (both of which should be *debilizem*). In addition, the extracted list contains the candidate *butl*, which is the non-standard spelling of the word *butelj* 'idiot'.

Manually analyzing candidate lists (sorted by cosine similarity) for each seed embedding would involve a lot of redundant work, as the same candidates would have to be checked multiple times if they occurred in different lists. Cross-comparing each list with all the candidates previously analyzed would also be somewhat tedious, so we tested a more robust approach. For each of the 11 categories of seed embeddings, we collected all the unique candidates extracted within that category, then calculated the average cosine similarity (i.e. the sum of all cosine similarities of a specific candidate within that category divided by the number of seed embeddings in that category). In the aggregated list containing all candidates extracted within the category, the candidates with the highest overall average similarity scores would be ranked higher compared to the less relevant ones. We manually annotated a portion of these aggregated lists as described in Section 5.

5. Candidate Annotation and Results

We manually annotated a subset of candidates from each aggregated list in terms of the following: (a) whether the candidate is relevant for a lexicon of taboo language; (b) what category the candidate belongs in (as seen in Table 6 with the example *pizdek*, the extracted candidates do not necessarily belong in the same category as the seed lexeme); (c) is it primarily a taboo language candidate or is it only problematic in one of its senses, contexts or uses (e.g. *coprnica* 'witch', which can be used as an insult, but is not in and of itself offensive); (d) whether the candidate exhibits any non-standard spelling features

Candidate	Cosine Similarity
bebec	0,971
butl	0,966
idiot	0,961
kreten	0,958
drekač	0,954
budal	0,954
imbecil	0,951
bebo	0,951
pizdeki	0,949
debili	0,949
debiliz	0,949
bedak	0,948

Table 6: Top 12 units from the extracted list of candidates for the seed lexeme *debil* 'idiot' extracted from *CLARIN.SI-embed.sl-lemma 1.0.*

(as well as other spelling peculiarities such as self-censorship, e.g. *k*rac* instead of *kurac* 'cock') and, if so, what is its standardized lemma.

In terms of relevance, the candidates were annotated based on whether they could potentially be assigned one of the labels present in the *Thesaurus of Modern Slovene 2.0* (see Section 3.3) and whether they could be problematic in contexts such as educational materials or games with a purpose. In borderline or ambiguous cases, the annotation criteria were more inclusive as it is much easier to exclude irrelevant units from a limited pool later compared to reacquiring them from massive amounts of candidates. The annotation was done by a single linguist, whereas multiple annotations and inter-annotator agreement analyses will be done as part of our future work on the pool of identified candidates, relying on much stricter guidelines to classify the candidates into more fine-grained categories.

The analyzed candidates were checked in the *metaFida 1.0 Corpus of Slovene* (an aggregate corpus of different Slovene corpora; see Erjavec, 2023),[9] particularly in the case of ambiguous or incomprehensible lemmatization errors where the original form and its context were crucial to correctly interpret the candidate.

Table 7 shows the percentages of units that were annotated as relevant for the taboo lexicon from each of the aggregated lists (the numbers include repetitions across lists; for unique relevant candidates per category, see Table 8), with several examples from the category. It should be noted that not all the relevant candidates belonged in the same category that was the basis for the compilation of the list. For instance, the aggregated list for *Political References* included 68% of political candidates, 25% of vulgar candidates, and 1–2% each for *Cognitive References*, *LGBTIQA+*, *Scatological References*, and *Xenophobia.*

[9] metaFida 1.0 is available at: https://www.clarin.si/ske/#dashboard?corpname=mfida10

The lists were checked in the order listed in Table 7. After a manual analysis of each list, the next list was cross-compared with all the previously analyzed lists to automatically annotate all the previously manually analyzed candidates (to remove redundant work). The number of candidates analyzed was determined independently for each list, depending on the number of relevant material and whether the extracted candidates still yielded candidates within the list's category. More candidates will be analyzed in our future work, probably by recursively exporting more lists based on newly discovered candidates (see Section 7).

The extraction method has proven to be effective: 34–89% of the analyzed candidates were relevant (an average of 54%).

The total final count of unique relevant candidates by category is shown in Table 8. The analysis resulted in a total of 1,260 units (compared to the approximately 120 initial seed lexemes). The newly discovered units even include three new categories: *Body Shaming* (e.g., *debeluhar* 'fatso'), *Violence* (e.g., *umorček* 'murder (diminutive)'), and *Religion* (e.g., *muslič* (pejorative expression for Muslims)).

113 (9%) of the units are taboo only in a certain context/sense (e.g., *pujs* 'pig', *češplja* 'plum', can be used as an expression for female genitalia). 224 (18%) of the candidates feature special spelling characteristics, such as non-standard spelling (*yugovič* instead of *jugovič*, *rdecuh* instead of *rdečuh*; *zavraga* instead of *za vraga*), self-censorship (*pi*zda* instead of *pizda*), or even typos (*komunjzer* instead of *komunajzer*). Units with spelling peculiarities are still relevant for profanity filters, for instance, but not necessarily for lexicographic resources. In some cases, the spelling anomaly is relevant only for a single inflected form (or a small number of inflected forms, e.g. *norc*, a non-standard form for the nominative and accusative singular forms instead of *norec* 'madman'). In others, the spelling variation affects all inflected forms (*babnca* 'hag', *babnce*, *babnci*, etc.) or distinguishes between a single-lexeme variant and a multiword expression (*pizdumater* or *pizdu mater*).

6. Discussion

Figure 3 represents the entire workflow of extracting taboo language candidates from seed units to the final candidate list.

The method has shown several advantages compared to surveys which have been used in related work to collect taboo words for Slovene. Firstly, embeddings can detect non-standard forms or forms that are rare in the corpus and might not be detected through a frequency-based approach (including typos). Embeddings also reveal words used in offensive contexts even if they are not related to the field of initial seed lexemes or if they belong to a category not included in initial set.

The main disadvantage of this approach is its focus on single-word expressions and functions on the level of a lexeme, not necessarily its sense. The majority of the discovered units were primarily taboo, while the detection of problematic secondary senses seems more difficult for static embeddings, as was the case of *češplja* 'plum'; the candidates with the greatest cosine similarity to this had nothing to with its sexual sense ('vagina'), but mostly consisted of different types of fruit (e.g., *hruška* 'pear', *češnja* 'cherry'). While the approach with fastText embeddings is effective enough to cover the most obvious taboo

List	Total Analyzed	Relevant	Irrelevant	Relevant Examples from Category
Sexual	1.200	360 (30%)	840 (70%)	*lulek* 'willy', *fuk* 'fuck (noun)', *fafač* (male who performs oral sex on another man)
Xenophobia	223	76 (34%)	147 (66%)	*čapar*, *fureš*t, *špagetar* (expressions for foreigners of different nationalities)
Vulgar	500	329 (66%)	171 (34%)	*napizdevati* 'to scorn (vulgar)', *prifuknjen* 'fucked up', *jebemuvraga* (vulgar interjection)
Scatological	800	487 (61%)	313 (39%)	*posrati* 'to shit', *kozlati* 'to puke', *prdež* (person who farts)
Racial	300	103 (34%)	197 (66%)	*rdečekožec* 'redskin', *zamorklja* 'negress', *aziat* (expression for a person of Asian descent)
Political	300	265 (89%)	35 (11%)	*komunajzer* 'commie', *levakar* 'leftist (pejorative)', *desnuhar* 'right-winger (pejorative)'
Misogyny	404	247 (61%)	157 (39%)	*prasička* lit. 'sow' (diminutive and pejorative), *avša*, *babnica* (offensive expressions for women)
LGBTIQA+	400	239 (60%)	161 (40%)	*pederuh*, *pederajs*, *ritopikec* lit. 'ass-poker' (offensive expressions for gay men)
Drugs	100	58 (58%)	42 (42%)	*zakomirati* 'to fall into a stupor', *nažgan* 'drunk' (adjective, vulgar), *džanki* 'junkie'
Disability	100	63 (63%)	37 (37%)	*stvor* 'freak', *monstrum* 'monster'
Cognitive	100	46 (46%)	54 (54%)	*glup* 'stupid', *bedak* 'idiot', *trapa* 'idiot (feminine)'

Table 7: Percentages of relevant candidates in each aggregated list of extracted candidates from *CLARIN.SI-embed.sl-lemma 1.0*.

Category	**Number of Units**
Sexual	145
Xenophobia	53
Vulgar	564
Scatological	37
Racial	26
Political	164
Misogyny	68
LGBTIQA+	25
Drugs	15
Disability	6
Cognitive	87
Body Shaming	21
Violence	33
Religion	16
Total	**1,260**

Table 8: Final numbers of relevant units by category.

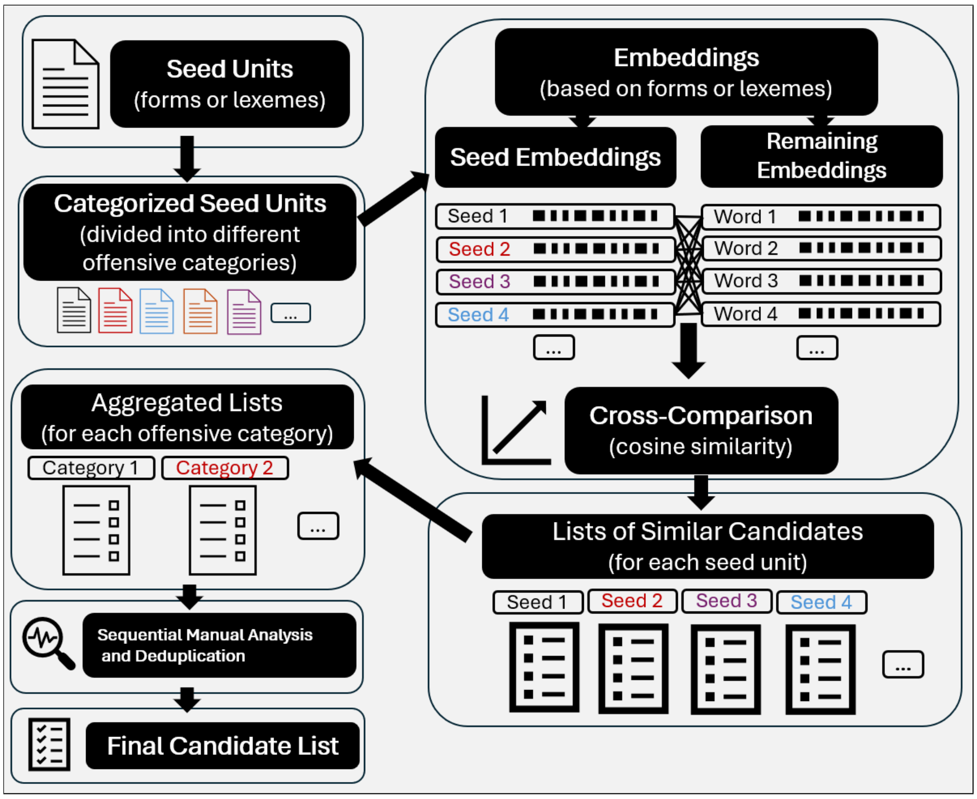


Figure 3: Depiction of the workflow for extracting taboo language candidates.

candidates, more advanced methods are warranted to detect more nuanced examples (see e.g. Jarquín-Vásquez et al., 2020, who leverage attention-based neural networks to take context into account and discriminate between offensive and non-offensive language use).

More candidates can also be discovered by recursively extracting the contextual radius of newly obtained candidates. When working with form embeddings, however, it is prudent to first generate all the inflected forms, then extract similar form embeddings for each form of the generated candidate, as form embeddings have been shown to yield much smaller amounts of candidates compared to lemma embeddings.

7. Conclusion

We have described an approach to extracting taboo language candidates from corpora using word embeddings. The evaluation shows that the method is effective and the initial analysis of candidates has resulted in a total of 1,260 units relevant for the lexicon (compared to the 122 seed lexemes). While we have not explicitly compared the proposed approach to the manual compilation of such lists through surveys or by browsing corpora (or frequency lists from corpora), it is reasonable to assume that this semi-automatic method requires less time and results in a much greater number of candidates, including potential non-standard spellings and authentic self-censored forms. In addition, the data acquired in this manner is based on real language use, not someone's perception of language use (if collecting data through surveys). The method is applicable to other languages, and *fastText* embeddings are available for numerous other languages (see Section 3.2.

The extracted units will be included in the *Digital Dictionary Database of Slovene* (DDDS; Kosem et al., 2021) for further lexicographic analysis; in order to make it compatible with the structure of *DDDS*, the discovered lemmas will be expanded with all inflected forms using *Pregibalnik*,[10] a custom-made open-access inflection software for lexicon expansion. Once the inflected forms are available, more candidates will be analyzed by taking into account form embeddings (as opposed to lemma embeddings analyzed in this paper) as well. This should provide additional units as *CLARIN.SI-embed.sl-form 2.0* was trained on a larger corpus. The embeddings can also be trained on corpora of spoken Slovene, as most of the taboo candidates we've extracted so far have come from web texts.

In addition, we will define a final taxonomy of categories based on related work and establish a system to include non-standard variants and other spelling anomalies (such as self-censorship and variants that contain both single lexemes and multiword expressions. The goal is to document units in the lexicon in such a way that the resource is useful for multiple purposes (e.g., as a filter for educational purposes; for instance, to improve GDEX extraction for Slovene (Kosem et al., 2018); as a comprehensive profanity filter, and as a basis for documenting taboo words in lexicographic resources). Future developments will also include annotation of the identified candidates with degrees of tabooness and arousal, similar to emotion lexicons such as *SloEmoLex* (Brglez et al., 2024). The lexicon will be made available under an open-access license (CC BY-SA 4.0).

[10] *Pregibalnik* is available as an API at: https://orodja.cjvt.si/pregibalnik/docs The code is available at: https://github.com/clarinsi/SloInflector

8. Acknowledgements

The research presented in this paper was conducted within the *LLM4DH* research project (*Large Language Models for Digital Humanities*; GC-0002) and the *PoVeJMo* research program (*Adaptive Natural Language Processing with Large Language Models*), funded within the *Recovery and Resilience Plan* (NOO; *Načrt za okrevanje in odpornost*) by the Slovenian Research and Innovation Agency (ARIS) and NextGenerationEU. The author acknowledges the financial support from the research program *Language Resources and Technologies for Slovene* (P6-0411) and the *CLARIN.SI Research Infrastructure* (I0-E004), both funded by the Slovenian Research and Innovation Agency (ARIS). The author also thanks the anonymous reviewers for their valuable comments.

Software

Arhar Holdt, Š., Gantar, P., Kosem, I., Pori, E., Logar, N., Gorjanc, V. & Krek, S. (2022). Sovražno in grobo besedišče v odzivnem Slovarju sopomenk sodobne slovenščine. In *Proceedings of the Conference on Language Technologies & Digital Humanities 2022*. URL https://elex.link/ojs/index.php/elex/article/view/36.

Arhar Holdt, Š., Gantar, P., Kosem, I., Pori, E., Robnik-Šikonja, M. & Krek, S. (2023). Thesaurus of Modern Slovene 2.0. In *Electronic lexicography in the 21st century. Proceedings of eLex 2023 conference.* URL https://elex.link/ojs/index.php/elex/article/view/36.

Arhar Holdt, Š., Logar, N., Pori, E. & Kosem, I. (2020). "Game of Words": Play the Game, Clean the Database. In Z. Gavriilidou, M. Mitsiaki & A. Fliatouras (eds.) *Lexicography for Inclusion: Proceedings of the 19th EURALEX International Congress, 7-9 September 2021, Alexandroupolis, Vol. 1.* Alexandroupolis: Democritus University of Thrace, pp. 41–49. URL https://www.euralex.org/elx_proceedings/Euralex2020-2021/EURALEX2020-2021_Vol1-p041-049.pdf.

Bañón, M., Chichirau, M., Esplà-Gomis, M., Forcada, M.L., Galiano-Jiménez, A., García-Romero, C., Kuzman, T., Ljubešić, N., van Noord, R., Pla Sempere, L., Ramírez-Sánchez, G., Rupnik, P., Suchomel, V., Toral, A. & Zaragoza-Bernabeu, J. (2023). Slovene web corpus MaCoCu-sl 2.0. URL http://hdl.handle.net/11356/1795. Slovenian language resource repository CLARIN.SI.

Bassignana, E., Basile, V. & Patti, V. (2018). Hurtlex: A Multilingual Lexicon of Words to Hurt. In E. Cabrio, A.M. Mazzei & F. Tamburini (eds.) *Proceedings of the Fifth Italian Conference on Computational Linguistics CLiC-it 2018.* Accademia University Press. URL https://doi.org/10.4000/books.aaccademia.3085.

Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. URL https://arxiv.org/abs/1607.04606. 1607.04606.

Bond, F. & Choo, M.Y.H. (2021). Taboo Wordnet. In P. Vossen & C. Fellbaum (eds.) *Proceedings of the 11th Global Wordnet Conference.* University of South Africa (UNISA): Global Wordnet Association, pp. 36–43. URL https://aclanthology.org/2021.gwc-1.5/.

Brglez, M., Caporusso, J., Hoogland, D., Koloski, B., Pollak, S. & Purver, M. (2024). Slovenian Emotion Dimension and Emotion Association Lexicon SloEmoLex 1.0. URL http://hdl.handle.net/11356/1875. Slovenian language resource repository CLARIN.SI.

Demšar, J., Curk, T., Erjavec, A., Črt Gorup, Hočevar, T., Milutinovič, M., Možina, M., Polajnar, M., Toplak, M., Starič, A., Štajdohar, M., Umek, L., Žagar, L., Žbontar, J., Žitnik, M. & Zupan, B. (2013). Orange: Data Mining Toolbox in Python. *Journal of*

Machine Learning Research, 14, pp. 2349–2353. URL http://jmlr.org/papers/v14/demsar13a.html.

Erjavec, T. (2023). Korpus metaFida 1.0. Poročilo projekta Razvoj slovenščine v digitalnem okolju. Aktivnost DS1.8. URL https://www.cjvt.si/rsdo/wp-content/uploads/sites/18/2023/03/RSDO_Kazalnik_Metakorpus.pdf.

Erjavec, T., Ljubešić, N. & Fišer, D. (2011). hrWaC and slWaC: Compiling web corpora for Croatian and Slovene. In *Text, Speech and Dialogue: 14th International Conference, TSD 2011.* Pilsen, Czech Republic: Springer, pp. 395–402.

Erjavec, T., Ljubešić, N. & Fišer, D. (2018). Korpus slovenskih spletnih uporabniških vsebin Janes. In D. Fišer (ed.) *Viri, orodja in metode za analizo spletne slovenščine.* Ljubljana, Slovenia: Znanstvena založba Filozofske fakultete Univerze v Ljubljani, pp. 16–43.

Gorjanc, V. (2005). Neposredno in posredno žaljiv govor v jezikovnih priročnikih: diskurz slovarjev slovenskega jezika. In *Družboslovne razprave 21(48).* Ljubljana, Slovenia, pp. 197–209. URL http://dk.fdv.uni-lj.si/dr/dr48Gorjanc.PDF.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A. & Mikolov, T. (2018). Learning Word Vectors for 157 Languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018).*

Jarquín-Vásquez, H.J., Montes-y Gómez, M. & Villaseñor-Pineda, L. (2020). Not All Swear Words Are Used Equal: Attention over Word n-grams for Abusive Language Identification. In K.M. Figueroa Mora, J. Anzurez Marín, J. Cerda, J.A. Carrasco-Ochoa, J.F. Martínez-Trinidad & J.A. Olvera-López (eds.) *Pattern Recognition.* Cham: Springer International Publishing, pp. 282–292.

Jay, T. (1992). *Cursing in America: A psycholinguistic study of dirty language in the courts, in the movies, in the schoolyards and on the streets.* Amsterdam: John Benjamins.

Jay, T. (2020). *2. Ten issues facing taboo word scholars.* Berlin, Boston: De Gruyter Mouton, pp. 37–52. URL https://doi.org/10.1515/9781501511202-002.

Jay, T. & Janschewitz, K. (2008). The pragmatics of swearing. *Journal of Politeness Research*, 4(2), pp. 267–288. URL https://doi.org/10.1515/JPLR.2008.013.

Kaplan, S.M. (2021). The role lexicographers can play in helping to vanquish insensitivity, brutality, othering, and wilful ignorance. In *25th International AFRILEX (African Association for Lexicography) Conference.* Stellenbosch, South Africa. URL https://hal.science/hal-03898094.

Klemenčič, I. (2016). *Fan, vad gör du?!: om svärande och svordomar i det svenska och slovenska språket. BA Thesis.* Ph.D. thesis, Faculty of Arts, University of Ljubljana.

Kos, J. (2024). *Psiholingvistični pristop k tabu besedam v slovenščini: diplomsko delo.* Ph.D. thesis, J. Kos. URL https://repozitorij.uni-lj.si/IzpisGradiva.php?lang=slv&id=158245.

Kosem, I., Koppel, K., Zingano Kuhn, T., Michelfeit, J. & Tiberius, C. (2018). Identification and automatic extraction of good dictionary examples: the case(s) of GDEX. *International Journal of Lexicography*, 32(2), pp. 119–137. URL https://doi.org/10.1093/ijl/ecy014. https://academic.oup.com/ijl/article-pdf/32/2/119/28858872/ecy014.pdf.

Kosem, I., Krek, S. & Gantar, P. (2021). Semantic data should no longer exist in isolation: the Digital Dictionary Database of Slovenian. In *9th EURALEX International Congress "Lexicography for Inclusion".* pp. 81–83. URL https://elex.is/wp-content/uploads/2021/09/Semantic-Data-should-no-longer-exist-in-isolation-the-Digital-Dictionary-Database-of-Slovenian_Kosem-Krek-Gantar_EURALEX2020.pdf.

Krek, S., Arhar Holdt, Š., Erjavec, T., Čibej, J., Repar, A., Gantar, P., Ljubešić, N., Kosem, I. & Dobrovoljc, K. (2020). Gigafida 2.0: The Reference Corpus of Written Standard Slovene. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck,

S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk & S. Piperidis (eds.) *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 3340–3345. URL https://aclanthology.org/2020.lrec-1.409/.

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 *Proceedings of eLex 2017: Lexicography from Scratch*. URL https://elex.link/elex2017/wp-content/uploads/2017/09/paper05.pdf.

Krek, S., Laskowski, C., Robnik-Šikonja, M., Kosem, I., Arhar Holdt, Š., Gantar, P., Čibej, J., Gorjanc, V., Klemenc, B., Dobrovoljc, K., Pori, E., Roblek, R. & Zgaga, K. (2023). Thesaurus of Modern Slovene 2.0. URL http://hdl.handle.net/11356/1916. Slovenian language resource repository CLARIN.SI.

Ljubešić, N. & Erjavec, T. (2018). Word embeddings CLARIN.SI-embed.sl 1.0. URL http://hdl.handle.net/11356/1204. Slovenian language resource repository CLARIN.SI.

Ljubešić, N., Terčon, L. & Čibej, J. (2023). The CLASSLA-Stanza model for morphosyntactic annotation of standard Slovenian 2.0. URL http://hdl.handle.net/11356/1767. Slovenian language resource repository CLARIN.SI.

McEnery, T. (2006). *Swearing in English: Bad Language, Purity and Power from 1586 to the Present*. London: Routledge.

Penedo, G., Malartic, Q., Hesslow, D., Cojocaru, R., Cappelli, A., Alobeidli, H., Pannier, B., Almazrouei, E. & Launay, J. (2023). The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only. URL https://arxiv.org/abs/2306.01116. 2306.01116.

Qiu, J., Lv, H., Jin, Z., Wang, R., Ning, W., Yu, J., Zhang, C., Li, Z., Chu, P., Qu, Y., Shi, J., Lu, L., Peng, R., Zeng, Z., Tang, H., Lei, Z., Hong, J., Chen, K., Fei, Z., Xu, R., Li, W., Tu, Z., Dahua, L., Qiao, Y., Yan, H. & He, C. (2024). WanJuan-CC: A Safe and High-Quality Open-sourced English Webtext Dataset. URL https://arxiv.org/abs/2402.19282. 2402.19282.

Rosenthal, S., Atanasova, P., Karadzhov, G., Zampieri, M. & Nakov, P. (2021). SOLID: A Large-Scale Semi-Supervised Dataset for Offensive Language Identification. URL https://arxiv.org/abs/2004.14454. 2004.14454.

Schmidt, A. & Wiegand, M. (2017). A Survey on Hate Speech Detection using Natural Language Processing. In L.W. Ku & C.T. Li (eds.) *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*. Valencia, Spain: Association for Computational Linguistics, pp. 1–10. URL https://aclanthology.org/W17-1101/.

Sulpizio, S., Günther, F., Badan, L., Basclain, B., Brysbaert, M., Chan, Y.L., Ciaccio, L.A., Dudschig, C., Duñabeitia, J.A., Fasoli, F., Ferrand, L., Filipović Đurđević, D., Ernesto Guerra, E., Hollis, G., Job, R., Jornkokgoud, K., Kahraman, H., Kgolo-Lotshwao, N., Kinoshita, S., Kos, J., Lee, L., Lee, N.H., Mackenzie, I.G., Manojlović, M., Manouilidou, C., Martinic, M., Ménder, M.d.C., Mišić, K., Chiangmai, N.N., Nikolaev, A., Oganyan, M., Rusconi, P., Samo, G., Tse, C.s., Westbury, C., Wongupparaj, P., Yap, M.J. & Marelli, M. (2024). Taboo language across the globe: A multi-lab study. In *Behavior Research Methods 56(4)*. Springer, p. 3794–3813. URL https://doi.org/10.3758/s13428-024-02376-6.

Sørensen, N.H. & Nimb, S. (2018). Word2Dict – Lemma Selection and Dictionary Editing Assisted by Word Embeddings. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (eds.) *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*. Ljubljana, Slovenia: Ljubljana University Press, Faculty of Arts, pp. 819–826.

Terčon, L., Čibej, J. & Ljubešić, N. (2023a). The CLASSLA-Stanza model for lemmatisation of standard Slovenian 2.0. URL http://hdl.handle.net/11356/1768. Slovenian language resource repository CLARIN.SI.

Terčon, L. & Ljubešić, N. (2023). The CLASSLA-Stanza model for morphosyntactic annotation of non-standard Slovenian 2.1. URL http://hdl.handle.net/11356/1786. Slovenian language resource repository CLARIN.SI.

Terčon, L., Ljubešić, N. & Erjavec, T. (2023b). Word embeddings CLARIN.SI-embed.sl 2.0. URL http://hdl.handle.net/11356/1791. Slovenian language resource repository CLARIN.SI.

van Huyssteen, G.B. & Tiberius, C. (2023). Towards a lexical database of Dutch taboo language. In *Electronic lexicography in the 21st century. Proceedings of eLex 2023 conference*. Ljubljana, Slovenia, pp. 197–209. URL https://elex.link/ojs/index.php/elex/article/view/13.

Wang, W., Chen, L., Thirunarayan, K. & Sheth, A.P. (2014). Cursing in English on twitter. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '14. New York, NY, USA: Association for Computing Machinery, p. 415–425. URL https://doi.org/10.1145/2531602.2531734.

Žagar, A., Kavaš, M., Robnik-Šikonja, M., Erjavec, T., Fišer, D., Ljubešić, N., Ferme, M., Borovič, M., Boškovič, B., Ojsteršek, M. & Hrovat, G. (2022). Corpus of academic Slovene KAS 2.0. URL http://hdl.handle.net/11356/1448. Slovenian language resource repository CLARIN.SI.

Zingano Kuhn, T., Arhar Holdt, Š., Kosem, I., Tiberius, C., Koppel, K. & Zviel-Girshin, R. (2022). Data preparation in crowdsourcing for pedagogical purposes: The case of the CrowLL game. *Slovenščina 2.0: empirical, applied and interdisciplinary research*, 10(2), pp. 62–100. URL https://journals.uni-lj.si/slovenscina2/article/view/11431.