From Translation Equivalents to Synonyms: Creation of a Slovene Thesaurus Using Word co-occurrence Network Analysis

Simon Krek[1], Cyprian Laskowski[2], Marko Robnik-Šikonja[3]

[1] "Jožef Stefan" Institute, Artificial Intelligence Laboratory, Jamova 39, Ljubljana, Slovenia
&
University of Ljubljana, Centre for Language Resources and Technologies, Večna pot 113, Ljubljana, Slovenia
[2] University of Ljubljana, Faculty of Arts, Aškerčeva 2, Ljubljana, Slovenia
[3] University of Ljubljana, Faculty of Computer and Information Science, Večna pot 113, Ljubljana, Slovenia

E-mail: simon.krek@ijs.si, cyprian.laskowski@ff.uni-lj.si, marko.robnik@fri.uni-lj.si

Abstract

We describe an experiment in the semi-automatic creation of a new Slovene thesaurus from Slovene data available in a comprehensive English–Slovenian dictionary, a monolingual dictionary, and a corpus. We used a network analysis on the dictionary word co-occurrence graph. As the additional information, we used the distributional thesaurus data available as part of the Sketch Engine tool and extracted from the 1.2 billion word Gigafida corpus, as well as information on synonyms from a Slovene monolingual dictionary. The resulting database serves as a starting point for manual cleaning of the database with crowdsourcing techniques in a custom-made online visualisation and annotation tool.

Keywords: bilingual dictionary; translation equivalents; thesaurus; automated lexicography; network analysis

1. Introduction

Slovene as a language with approximately two million speakers in the Republic of Slovenia and an additional 0.5 million outside its borders who speak or understand it (cf. Krek, 2012: 44), has been a slow starter in relation to the availability of language reference books. The first, and to date the only, comprehensive monolingual dictionary was compiled and published in five volumes at the end of the 20th century, between 1970 to 1991; and until 2016 no thesaurus or similar reference book describing synonymy in Slovene had been available. In 2016, the Fran Ramovš Institute of Slovene Language, working under the umbrella of the Slovene Academy of Sciences and Arts, published a one-volume thesaurus on a little fewer than 1,300 pages, with its concept and data predominantly based on the already outdated monolingual dictionary. The academic thesaurus project started around 2002; therefore, it took 15 years to compile and publish the dictionary which is currently available only in printed form. The motivation to experiment with bilingual, corpus

and other types of data to create a thesaurus from scratch, originates from two basic deficiencies of the existing academic thesaurus: (1) it is available only in printed form and (2) it describes Slovene that was used in the middle of the 20th century and not the modern language.

In contrast, resources used to compile the thesaurus described in the paper were chosen to reflect primarily what is considered modern Slovene. Major changes in the political and economic system after 1991, when Slovenia became an independent state and abolished the post-WWII single-party system to introduce parliamentary democracy, also had a profound influence on language. Our source data originate from works created in the last 20 years and are explicitly and intentionally corpus-based. In this manner, the used data provide an accurate representation of the current state of the language.

The remainder of the paper is structured as follows. In Section 2, we describe the data sources: bilingual English–Slovene dictionary, the 1.2-billion-word Gigafida corpus of Slovene, and the Slovene monolingual dictionary. In Section 3, we first describe the procedure and algorithms used to automatically create the thesaurus data preprocessing, word co-occurrence graph and extraction of relevant synonyms with the Personal PageRank algorithm. We also present the evaluation of obtained synonymy and the final database. In Section 4, we discuss visualization of the thesaurus data, which we split into three parts: synonyms, collocations and good examples. Our visualization system includes a crowdsourcing component. In Section 5, we conclude the paper.

2. Source Data

2.1 Bilingual dictionary data

As the source dictionary for bilingual data, the Oxford-DZS Comprehensive English–Slovenian Dictionary (ODCESD, 2005–2006; Šorli et al., 2006) was used. Contrary to bi-directional bilingual dictionaries designed to serve for both encoding and decoding purposes for native speakers of languages involved, ODCESD is a mono-directional dictionary intended for Slovene native speakers decoding English texts. Consequently, the headword list is more extensive than usual (120,000), and senses receive a more in-depth treatment (specialised, archaic or rare). Organisation of senses is translation-based, meaning that all the senses which generate the same translation equivalent(s) are joined. While traditional bi-directional dictionaries generally avoid listing (near-)synonymous translations, ODCESD lists an exhaustive list of semantic and stylistic equivalents relying on the native speaker's ability to distinguish between nuances of meaning in translation equivalents. Close synonyms are separated by a comma, while words interchangeable in less than roughly 50% of contexts are separated by a semicolon. We use these strings of Slovene translation equivalents separated by commas or semicolons as a source of data on synonymy in Slovene.

2.2 Corpus data

The second source of information was the Gigafida corpus (Logar et al., 2012); in particular via the Thesaurus module in the Sketch Engine tool (Rychlý, 2016). Gigafida is a 1.2 billion-word corpus of some 40,000 texts of various genres. With its release in 2012, it represents the third iteration of the FIDA family of corpora, which is considered as the reference corpus series for Slovene, starting with the 100 million-word FIDA corpus in the year 2000, followed by the 620 million-word FidaPLUS corpus in 2006. In addition to the Sketch Engine tool, Gigafida is available in a custom-made web concordancer, together with its balanced 100 million-word subcorpus Kres.

2.3 Monolingual dictionary data

The third source of information was a monolingual Slovene dictionary (SSKJ, 2014), the data of which serving as an additional confirmation of associations between words. SSKJ provides the lexicographic description of Slovene from the second half of the 20th century in little more than 92,000 entries. Its first edition was compiled between 1970 and 1991, also representing the first, and to date the only, monolingual dictionary of modern Slovene. In 2014, the second edition was published with some 6,000 new entries, and the dictionary was partly updated. It is available online as part of the Fran dictionary portal and as an independent website.

3. Procedure

3.1 Preparation of data

The first step in building the database was the extraction of translation equivalents from the bilingual dictionary and normalisation of text where truncation devices were applied. The basis of data preparation was an XML version of the ODCESD, which had been stripped of information irrelevant for our purposes (including all English data). The main points of departure were the <tr> tags, which contained the translation(s) of a given headword in a given (sub)sense. For example:

```
<tr>zapustiti; opustiti; odpovedati se, odstopiti od</tr>
```

Here, the particular sense of a headword (abandon, v.) was given four translations, the last two of which are more similar to each other. The first two translations are considered as near synonyms separated by a semicolon, and the last two as core synonyms separated by a comma.

Our first step, however, was to expand two types of truncation devices used inside these tags: brackets and slashes. Brackets indicated shorter and longer versions of a translation, whether just a word or an entire phrase. We handled these by expanding the original text to both versions. For instance, "računalo, abak(us)" expanded to

"računalo, abak, abakus", and "mirovanje; začasni odlog (izvajanja)" became "mirovanje; začasni odlog, začasni odlog izvajanja".

Slashes indicated alternatives. There were two types of devices: if the slash was followed by a dash, it indicated alternative suffixes (e.g., gender variants), the first of which was identified by going back to the first instance of the letter after the dash. For instance, "zaveznik/-ica" became "zaveznik, zaveznica", and "bolj kot/od" expanded to "bolj kot, bolj od". Combinations of brackets and slashes also occurred, in which case the rules were combined. For instance, "(kavno/pšenično) zrno" became "zrno, kavno zrno, pšenično zrno".

Once these expansions were available, the <tr> contents were split at semicolons and commas, to generate a hierarchy of terms with <synch> and <syn> tags, respectively. Each term was placed in a <s> element, possibly with coded attributes which tracked the source truncation devices. Finally, any domain annotations (or labels) within the <tr>, represented by <la> tags, were also copied into the <synch>. For example:

```
<tr><la>knjiž.</la>prilagoditi (se), akomodirati (se);
prirediti; uskladiti</tr>
```

generated:

```
<synch>
  <la>knjiž.</la>
  <syn>
    <s p="1" t="o-vz">prilagoditi se</s>
    <s p="1" t="o-vz">prilagoditi</s>
    <s p="2" t="o-vz">akomodirati se</s>
    <s p="2" t="o-vz">akomodirati</s>
  </syn>
  <syn>
    <s>prirediti</s>
  </syn>
  <syn>
    <s>uskladiti</s>
  </syn>
</synch>
```

At this point, a large list of structured translation equivalents was available. The next key step was finding all the potential synonyms for each term, and generating a reorganised XML file, arranged by headword rather than translation chain (<synch>).

Therefore, a new file was generated with data organised by headwords, with counted frequencies of their co-occurrences with individual candidate synonyms.

For every unique string within the <s> tags, an entry was created with that string as the headword. Then, we generated a synonym candidate list for that headword by looking at all the other <s> strings which co-occurred within a <synch> with the headword. For each such candidate, we tabulated the "core" and "near" counts, by counting all the co-occurrences of the headword and candidate and checking whether they were in the same <syn> or not. In order to detect relationships between the candidates, we calculated these totals for every pair of candidates.

During this phase, we analysed the attributes and labels of the truncation mechanism and used them to filter the data. For instance, since the ODCESD truncation mechanism with a slash-hyphen combination was used mainly for separating female and male translations, we tracked this and filtered out terms with mismatching genders. This way "študent" (male student) and "študentka" (female student) would not be treated as synonyms. Similarly, we removed all the variants that derived from the bracket truncation mechanism, when it resulted in extra words so that only the longest string containing a shorter possible synonym was kept. We also filtered out some labels which were irrelevant for our purposes (e.g., American vs British English).

The result of this phase of data processing was an XML file with 135,073 headwords, organised into entries containing <direct_syns> and <indirect_syns>. The <direct_syns> contained the candidates and their core/near counts with respect to the headword, along with any labels that co-occurred with the combination. The <indirect_syns> contained all the pairs of candidates that co-occurred (with each other and the headword) and their core/near counts with respect to each other. For instance, the word "neobremenjenost" occurred in three <tr> tags in the ODCESD:

```
<tr><la>poet.</la>razpuščenost, zanesenost; sproščenost,
neobremenjenost</tr>
<tr>samozavestnost, neobremenjenost</tr>
<tr>brezskrbnost, neobremenjenost, sproščenost</tr>
```

This resulted in:

```
<entry>
  <head>
    <form>
      <headword>neobremenjenost</headword>
    </form>
  </head>
  <body>
    <syns>
      <direct_syns>
        <direct_syn core="1" near="0">
          <s>brezskrbnost</s>
        </direct_syn>
        <direct_syn core="0" near="1">
          <s>razpuščenost</s>
          <labels>
```

```
            <la>poet.</la>
          </labels>
        </direct_syn>
        <direct_syn core="1" near="0">
          <s>samozavestnost</s>
        </direct_syn>
        <direct_syn core="2" near="0">
          <s>sproščenost</s>
        </direct_syn>
        <direct_syn core="0" near="1">
          <s>zanesenost</s>
          <labels>
            <la>poet.</la>
          </labels>
        </direct_syn>
      </direct_syns>
      <indirect_syns>
        <indirect_syn core="1" near="0">
          <s>brezskrbnost</s>
          <s>sproščenost</s>
        </indirect_syn>
        <indirect_syn core="0" near="1">
          <s>razpuščenost</s>
          <s>sproščenost</s>
        </indirect_syn>
        <indirect_syn core="1" near="0">
          <s>razpuščenost</s>
          <s>zanesenost</s>
        </indirect_syn>
        <indirect_syn core="0" near="1">
          <s>sproščenost</s>
          <s>zanesenost</s>
        </indirect_syn>
      </indirect_syns>
    </syns>
  </body>
</entry>
```

The <indirect_syns> also helped us organise candidate synonyms into groups with a simple rule: if it is possible to reach one candidate from another through a sequence of one or more <indirect_syn> tags, then the candidates belong in the same group. With these data in place, we could set up a co-occurrence graph, as described in the next section.

3.2 Co-occurrence graph

The most important step in organising data according to word associations was the creation of a weighted co-occurrence graph. The graph contains frequencies of co-occurrence of translation equivalents from the whole database. We ran the Personal PageRank algorithm (Page et al., 1999) on this graph to rank the synonym list, separately for each synonym candidate. Having obtained lists of synonyms and near synonyms for each headword, we now pursued three goals: i) determine groups of

words with the same meaning, ii) rank the groups of words with the same meaning according to their semantic similarity with the headword, and iii) rank words within groups according to their frequency of use.

Graphs are a suitable formalism to model semantic relations. We created a word co-occurrence graph G=(V, E), where V is a set of nodes (each node represents one headword with its label if present), and E is a set of connections between nodes. The edge e_{ij} connects nodes i and j and has an associated weight $w_{ij} \in R^+$. The weight models the strength of semantic similarity between words i and j. The larger the weight, the stronger the association between words i and j. Value $w_{ij} = 0$ means that there is no synonymy between words i and j. We organise values w_{ij} into a matrix W, called an adjacency matrix as its values contain degrees of adjacency between nodes. We calculate each cell of the matrix as a weighted sum of synonymy information from our three sources. The primary source of information is the core and near counts for headword-candidate or candidate-candidate combinations (core counts were given twice the weight of near counts). In addition, data from the Thesaurus module in Sketch Engine (Rychlý, 2016) and a monolingual Slovene dictionary, were included as additional information. The Figure 1 below shows a graphical representation of core and near synonyms for the headword *hiša*. Words in rectangles form groups with the same meaning e.g., *bivališče* and *domovanje*. The groups are subgraphs connected with <indirect_syn> tags, as described above. In the actual graph, these words are all connected to the headword but we excluded the connections from the graph to avoid clutter.

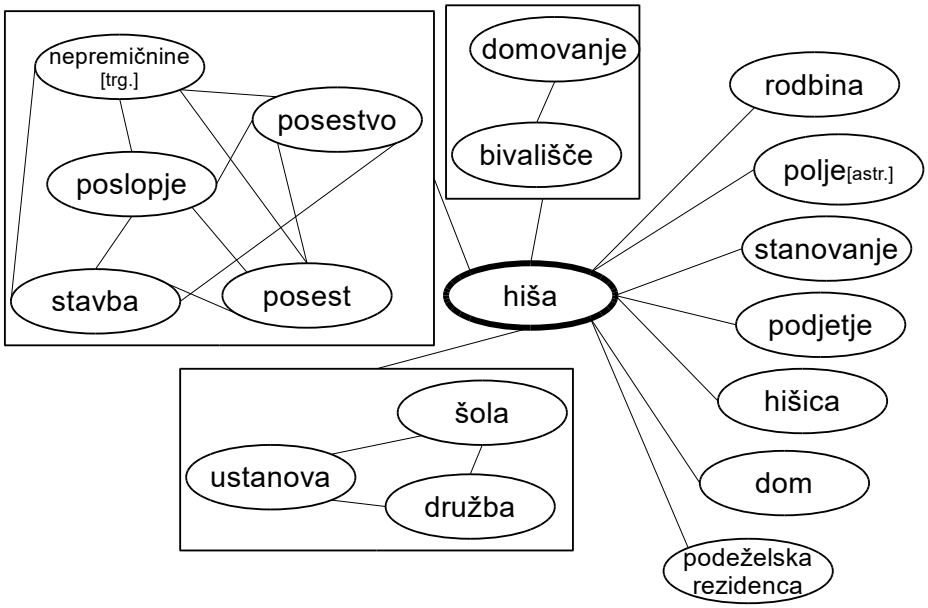

Figure 1: Co-occurrence graph ('hiša' – house)

We set the weight of each connection as the linear combination of contributing factors (core or near synonym, association score from the Sketch Engine tool and confirmation from the monolingual dictionary).

$$W = \text{coreWeight} \cdot \text{coreCount} + \text{nearWeight} \cdot \text{nearCount} + \text{sskjWeight} \cdot \text{sskjScore} + \text{sketchWeight} \cdot \text{sketchScore}$$

Here coreWeight, nearWeight, sskjWeight, and sketchWeight are weights given to each contributing factor. We used a preliminary evaluation on a small number of different headword categories to set sensible default values, namely coreWeight=2, nearWeight=1, sskjWeight=3, and sketchWeight=1. The most important factors are coreCount and nearCount, which are determined as the number of joint occurrences of connected words as core synonyms and near synonyms, respectively. The sskjScore and sketchScore are auxiliary factors with values between 0 and 1 (note that coreCount and nearCount mostly have a far greater range), which strengthen information from the bilingual dictionary. The sskjScore for headword i and connected word j is 0 if the SSKJ dictionary does not contain word j in the description of headword i. If the dictionary description contains the word j then the sskjScore is a value between 0 and 1 depending on the frequency of word j in the Gigafida corpus. If the corpus contains more than 50 instances of word j, the value of sskjScore is 1, if the frequency of a word is less than or equal to 3, sskjScore=0, otherwise it linearly depends on the frequency of j in Gigafida: sskjScore = (frequency - 3) / (50 - 3). The sketchScore is actually the logDice score (Rychlý, 2008) and is reported by the Sketch Engine as the default word association score. It is based on co-occurrence of two words in a corpus of documents, in our case in Gigafida.

Ranking of nodes is one of the frequently used tasks in the analysis of network properties and several so-called node centrality measures exist to assess the influence of a given node in the graph. The objective of ranking is to assess the relevance of a given node either globally (with regard to the whole graph) or locally (relative to some node in the graph). A well-known ranking method is PageRank (Page et al., 1999), which was used in the initial Google search engine. For a given network with the adjacency matrix W, the score of the i-th node returned by the PageRank algorithm is equal to the i-th component of the dominant eigenvector of W'^{T}, where W' is the matrix W with rows normalised so that they sum to 1. This can be interpreted in two ways. The first interpretation is the 'random walker' approach: a random walker starts walking from a random vertex v of the network and in each step walks to one of the neighbouring vertices with a probability proportional to the weight of the edge traversed. The PageRank of a vertex is then the expected proportion of time the walker spends in the vertex, or, equivalently, the probability that the walker is in the particular vertex after a long time. The second interpretation of PageRank is the view of score propagation. The PageRank of a vertex is its score, which it passes to the neighbouring vertices. A vertex v_i with a score PR(i) transfers its score to all its neighbours. Each neighbour receives a share of the score proportional to the strength of the edge between itself and v_i. This view explains the PageRank algorithm with the principle that in order for a vertex to be highly ranked, it must be pointed to by many highly ranked vertices. Other methods for ranking include Personalized-PageRank (Page et al., 1999), frequently abbreviated

as P-PR, that calculates the vertex score locally to a given network vertex, SimRank (Jeh & Widom, 2002), diffusion kernels (Kondor & Laerty, 2002), hubs and authorities (Kleinberg, 1999) and spreading activation (Crestani, 1997).

As we need a locally sensitive ranking of the nodes (i.e., the importance for each headword separately), we applied the P-PR algorithm to each headword (graph node) to rank the synonyms according to the strength of their relationship with the headword.

3.3 Evaluation

We evaluated the obtained results on a set of 50 randomly chosen headwords from different categories by comparing the returned ranks with those manually assigned by two annotators and language experts (denoted as E1 and E2). The selection of headwords by part-of-speech was controlled (25 nouns, 10 verbs, 10 adjectives, 5 adverbs) and some filters were applied, such as the minimum number of synonyms. The overall number of synonyms in 50 entries was 550 (302 core and 248 near synonyms). They were evaluated against two criteria: (1) each synonym was scored according to a three-point numerical scale, with 2 indicating a valid synonym, 1 an acceptable synonym and 0 not a synonym; and (2) the placement of synonyms to groups indicating semantic distinctions was manually evaluated. The main conclusions from the evaluation process were:

1. The results of the procedure are highly useful with regards to both the obtained sets of synonyms and their ranking.

Table 1 below presents the distribution of scores (0–2) for core and near synonyms and both evaluators.

	Evaluator E1						**Evaluator E2**					
score	**core**	%	**near**	%	**sum**	%	**core**	%	**near**	%	**sum**	%
2	164	54.3	58	23.4	222	40.4	107	35.4	20	8.1	127	23.1
1	64	21.2	84	33.9	148	26.9	110	36.4	125	50.4	235	42.7
1+2	228	**75.5**	142	**57.3**	370	**67.3**	217	**71.9**	145	**58.5**	362	**64.0**
0	74	24.5	106	42.7	180	32.7	85	28.1	103	41.5	188	34.2
sum	**302**	100	**248**	100	**550**	100	**302**	100	**248**	100	**550**	100

Table 1: Distribution of scores

The results show that around three quarters of the core synonyms are either valid or acceptable (E1 75.5%, E2 71.9%), as well as more than half of the near synonyms (E1 57.3%, E2 58.5%), which indicates that both core and near synonyms should be taken into account in the final database, although with a clear distinction between the two groups. Together around two-thirds of the obtained results are useful (E1 67.3%, E2 64%).

The agreement between the annotators is shown in Table 2.

Scores (E1-E2)	No. (synonyms)	Sum	%	Sum (1-2)	%
2-2 (valid)	98				
1-1 (acceptable)	99				
0-0 (not synonym)	130	327	59,5		
1-2 (acceptable – valid)	115			442	80,4
0-1 (not synonym – acceptable)	70				
0-2 (not synonym – valid)	38				
Sum	550				

Table 2: Inter-annotator agreement

Annotators completely agreed in 59.5% of cases, or in 80.4% if the difference between valid and acceptable synonyms is ignored, which is less important than between a synonym and non-synonym. The critical disagreement is between valid synonyms and non-synonyms (38 cases). The qualitative analysis shows that in the majority of cases the disagreement originates from a different strategy of evaluators in relation to a limited number of specific situations. For example, a multi-word synonym contains a single-word synonym which is already included in the list (headword: *videti* 'to see', single-word synonym: *srečati* 'to meet', multi-word synonym: *večkrat srečati* 'to meet repeatedly'), or some synonyms are semantically valid but differ in register, style, etc. (headword: *sodnik* 'judge', synonym: *rihtar* 'judge' (old informal expression)). In general, the bias should be towards inclusion: if at least one evaluator found the synonym acceptable, it is likely that some dictionary users would be interested in seeing it in the dictionary. The percentage of synonyms with at least one positive score is 76.4%.

2. Sense groups can only be considered as an indication of sense distribution and cannot be explicitly used as separate dictionary senses.

Table 3 in Appendix 1 shows the results of the evaluation of the sense groups. We used the Adjusted Rand Index (ARI) (Hubert & Arabie, 1985) to calculate agreement between the sense groupings constructed in the automatic procedure (A) or created by the evaluators (E1 or E2). ARI counts pairs of items that are in the same (or different) group in one grouping and at the same time are part of the same (or different) group in the other grouping. It yields a score of 1 for identical groupings and 0 for matching of groups expected by chance (so it is also possible to have negative ARI score). Unsurprisingly, the overlap is high in the case of monosemous headwords with a small number of synonyms, but in the case of polysemous headwords with a higher number of synonyms, the numbers show that the results are not good enough to use groups directly as separate senses. We note that the overlap between the two human evaluators is not very high which shows that the task is fairly difficult and there can be multiple solutions (presumably due to different granularities of categorisation).

Some other important qualitative observations from the evaluation are:

- a selected set of labels should be kept, particularly those indicating terminology (field labels);
- synonym candidates resembling definitions should be removed;
- in (rare) cases when a headword can belong to multiple parts-of-speech, synonyms are mixed and should be separated according to the different parts-of-speech;

We considered the results satisfactory and therefore did not test other network centrality measures (besides P-PR) as this would require another evaluation round. This may be an interesting further work.

3.4 Final database

The result was an automatically created thesaurus with 78,276 headwords where at least one core or near synonym is available. Of them, 41,555 are single-word and 36,721 multi-word headwords. Headwords with the highest number of synonyms are adjectives *hud* (angry, bad), with 110 as the total number synonyms (core + near) and *oster* (sharp) with the highest number of 62 core synonyms. The distribution of synonyms is uneven, with a long tail: 27,136 headwords have only one synonym (core or near), with the next 14,451 headwords having two synonyms. In Figure 2 we show a graphic representation of the number of synonyms per entry/headword.

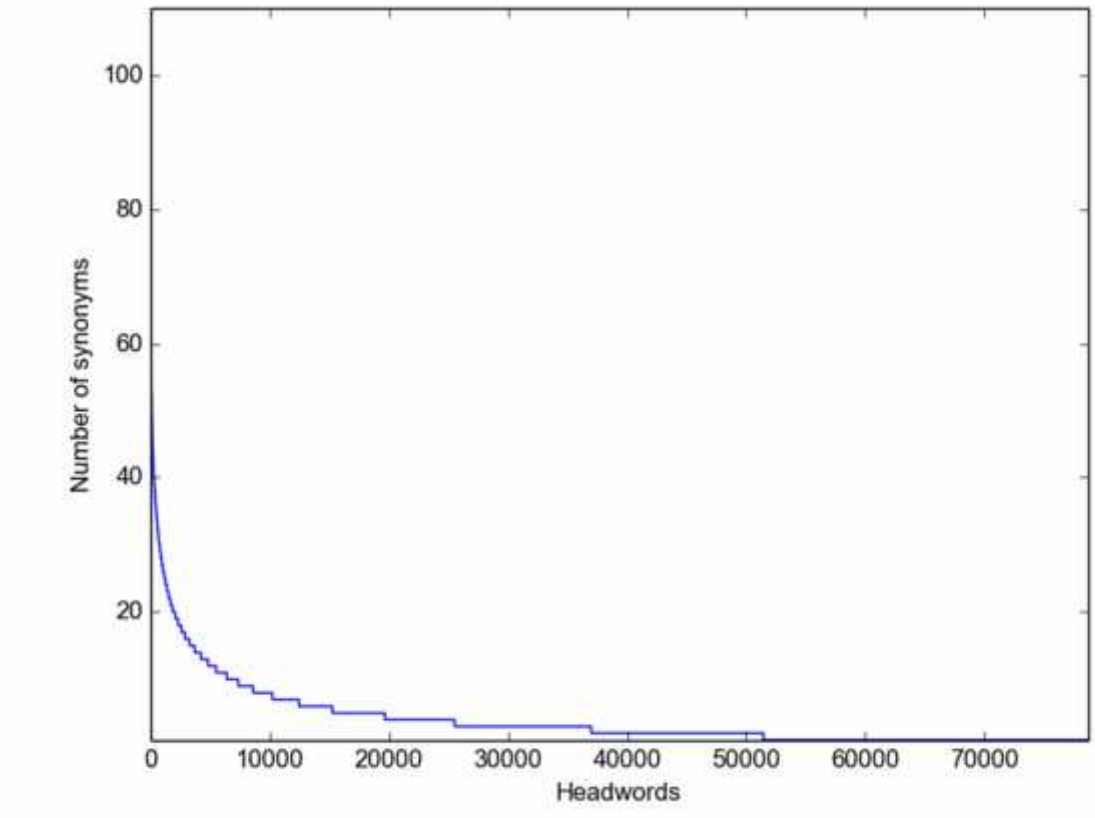

Figure 2: Number of synonyms per entry/headword

Finally, two types of additional information were added to all headwords and synonyms: (1) frequency of occurrence from the Gigafida corpus was added to single-word headwords and synonyms, and (2) part-of-speech category was attributed to headwords. This data set in XML format represents the result of the automatic procedure which was used in the visualisation and crowdsourcing tool described in the next section.

4. Visualisation and Crowdsourcing

From the very beginning of the experiment, automated generation of a thesaurus from the dictionary and corpus data was not the final goal. Crowdsourcing as a language resource creation method is also gaining ground in "traditional" lexicography (Čibej et al., 2015), and the ultimate goal was, in fact, to enable experimenting with crowdsourcing in creating a thesaurus for Slovene.[1] For this purpose, an online visualisation and annotation tool has been developed that enables users of the automatically created data to give feedback on the quality of generated synonyms. The tool was specifically designed to help users in recognising the differences between the two synonyms and enable scoring on the basis of contextual data. For this purpose, we use data extracted from the Gigafida corpus as produced by the Sketch-Diff module in the Sketch Engine tool. Users of the online thesaurus visualisation platform can advance through three consecutive pages when consulting each entry, and through linking to the previously available web concordancer, end up in the Gigafida corpus as the final consultation tool. The three pages are (1) synonyms page, (2) collocations page, and (3) good examples page.

4.1 Synonyms page

The initial page shows sets of synonyms in two groups: core and near synonyms. The two groups can be listed according to four criteria: ranking, alphabet, word length and user scores. We also use frequency data from the Gigafida corpus to show which synonyms are more frequent in the language in general, with a four-degree slider consecutively dimming less frequent synonyms. Finally, the most important features of the initial page are evaluation buttons accompanying each synonym enabling users to score the synonym.

4.2 Collocations page

Provided that relevant data can be found in the Gigafida corpus, we show collocations of both the headword and the synonym on the collocations page. For this purpose, we use data extracted from the Gigafida corpus taking advantage of a combination of Word Sketch and Thesaurus modules in the Sketch-Diff part of the Sketch Engine tool. Sketch-diffs show which collocations appear more frequently with

[1] The crowdsourcing phase has not begun at the time of the submission of the article, therefore we cannot provide any numbers on the success of crowdsourcing.

either the headword or the synonym, or equally with both. Collocations are selected on the basis of the "sketch grammar", a formalism identifying statistically relevant words appearing in specific grammatical relations in the corpus, such as an adjective preceding a noun. The grammar was developed for automatic extraction of data for a new Slovene collocations dictionary (Krek et al., 2016), which was used also in the online thesaurus tool. By clicking on a synonym on the initial page, the user is presented with collocations in four grammatical relations, in three different groups: those that are equally distributed in the context of both the headword and the synonym, and those that appear only with either the headword or the synonym, thus providing the means to the user to understand subtle semantic and contextual differences of synonyms.

4.3 Good examples page

As part of the collocations dictionary project, a tool producing "good dictionary examples" or GDEX tool (cf. Kosem, 2016) was also developed, and relevant dictionary examples were extracted from the Gigafida corpus. If good examples exist in the extracted database for a particular collocation, they are shown on the last page in the visualisation tool. In addition, each combination of the headword/synonym + collocate is equipped with a direct link to the Gigafida web concordancer producing all available concordances for that particular combination.

5. Conclusion

We described the automatic creation of a Slovene thesaurus from bilingual dictionary data, monolingual dictionary data, and corpus data. The methodology that was used in generating the thesaurus is general and can be applied to other dictionaries and languages. The step specific to our case is preprocessing, where the original dictionary data are transformed into XML form, suitable for our purposes. Other steps, such as the creation of word co-occurrence graph or ranking of nodes with P-PR, etc., are more general and can be reused if the input data are available in a suitable format. The methodology is flexible and allows integration of different sources of word association information. Furthermore, we described how we intend to use crowdsourcing in lexicography, which is a new trend, particularly in relation to cleaning the automatically generated data or data extracted from a corpus. Conceptually, visualisation and crowdsourcing solutions are also language independent and, given that similar types of data (synonyms, collocations, good examples) are used, solutions in the interface can be used in other dictionary portals.[2] We are convinced that this trend will grow and that other similar initiatives will follow in the coming years. We believe that our experiment on an automatic generation of a thesaurus represents a step in this direction.

[2] All relevant information will be available on the CJVT website (http://www.cjvt.si) after the lauch of the synonym dicitonary portal which is expected in October 2017.

6. Acknowledgements

We would like to thank DZS, the publisher of the *Oxford-DZS comprehensive English–Slovenian dictionary*, for their permission to conduct the experiment with the Slovene translation equivalents part of the bilingual dictionary database. Jan Kralj kindly provided the Python source code of P-PR algorithm. We would also like to thank the evaluation team: Špela Arhar Holdt, Jaka Čibej, Polona Gantar and Iztok Kosem.

7. References

Čibej, J., Fišer, D. & Kosem, I. (2015). The role of crowdsourcing in lexicography. In I. Kosem, M. Jakubíček, J. Kallas & S. Krek (eds.) *Proceedings of the eLex 2015 conference, 11-13 August 2015, Herstmonceux Castle, United Kingdom*. Ljubljana/Brighton: Trojina, Institute for Applied Slovene Studies/Lexical Computing, pp. 70-83.

Crestani, F. (1997). Application of spreading activation techniques in information retrieval. *Artificial Intelligence Review*, 11(6), pp. 453-482.

Hubert, L. & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2(1), pp. 193–218.

Jeh, G. & Widom, J. (2002). SimRank: A measure of structural-context similarity. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 538-543. ACM.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. (2014). The Sketch Engine: ten years on. In *Lexicography*, vol 1. issue 1. Springer, pp. 7–36.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5), pp. 604-632.

Kondor, R. I. & Laerty, J. D. (2002). Diffusion kernels on graphs and other discrete input spaces. In *Proceedings of the 19th International Conference on Machine Learning*, pp. 315-322.

Kosem, I. (2016). Interrogating a corpus. In P. Durkin (ed.) *The Oxford Handbook of Lexicography*. Oxford: Oxford University Press, pp. 76-93.

Krek, S. (2012). *The Slovene Language in the Digital Age / Slovenski jezik v digitalni dobi*. META-NET White Paper Series "Europe's Languages in the Digital Age", Springer, Heidelberg.

Krek, S., Gantar, P., Kosem, I., Gorjanc, V. & Laskowski, C. (2016). Baza kolokacijskega slovarja slovenskega jezika. *Proceedings of the Conference on Language Technologies & Digital Humanities, Faculty of Arts, University of Ljubljana*. Ljubljana, Slovenia. pp. 101-105.

Logar Berginc, N., Grčar, M., Brakus, M., Erjavec, T., Arhar Holdt, Š. & Krek, S. (2012). *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba*. Ljubljana: Trojina, zavod za uporabno slovenistiko; Fakulteta za družbene vede.

Page, L., Brin, S., Motwani, R. & Winograd, T. (1999). *The PageRank Citation Ranking: Bringing Order to the Web*. Technical Report. Stanford InfoLab.

Rychlý, P. (2016). Evaluation of the Sketch Engine Thesaurus on Analogy Queries. In *Tenth Workshop on Recent Advances in Slavonic Natural Language Processing*, RASLAN. pp. 147–15.

Rychlý, P. (2008). A lexicographer-friendly association score. In *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN*, pp. 6–9.

Šorli, M., Grabnar, K., Krek, S., Košir, T. (2006). Oxford-DZS comprehensive English-Slovenian dictionary. In *Proceedings XII EURALEX international congress*. Edizioni dell`Orso: Universita di Torino: Academia della Crusca, pp. 631-637.

Dictionaries:

ODCESD: *The Oxford®-DZS Comprehensive English-Slovenian Dictionary*. (2005-2006). Ljubljana: DZS.

SSKJ: *Slovar slovenskega knjižnega jezika (Dictionary of Literary Slovene)*. (2014). 2nd edition. Ljubljana: ZRC SAZU.

Appendix 1

Table 3: The comparison of synonym groups produced automatically (A) or by two human evaluators (E1 or E2).ARI score of 0 corresponds to an agreement between randomly assigned groups and 1 to perfectly matching groups. Note that the agreement expressed by ARI score may not be linear.

headword	POS	E1-E2	A-E1	A-E2	groups	synonyms	core	near	SSKJ
pritrditi	v	0.39	0.16	0.06	20	46	21	25	3
ugotoviti	v	0.14	0.04	0.21	20	51	13	38	1
vodilo	n	0.32	0.21	0.28	15	34	16	15	4
prepričati	v	0.21	0.03	0.04	14	28	8	20	1
kraj	n	0.48	0.42	0.27	12	19	12	7	7
preklic	n	0.48	0.12	0	12	26	11	15	5
sodnik	n	0.53	-0.03	-0.05	11	18	6	12	3
zrasti	v	0.34	0.25	0.41	11	25	14	10	10
hudodelstvo	n	0.55	0.17	0	9	12	3	9	1
drgnjenje	n	0.26	0.22	0.20	9	15	6	9	3
miniaturen	adj	0.26	0.14	0.34	9	18	14	4	2
temačno	adv	0.55	0.22	0.42	8	23	10	13	3
videti	v	0.34	0.33	0.42	8	14	4	10	12
analiza	n	0.26	0.09	0.26	8	18	11	7	1
pretepač	n	0.65	0.04	-0.02	7	11	6	5	1
krutost	n	0.37	0.29	0.15	6	18	11	7	1
žvižganje	n	0.01	0.14	0.14	6	10	4	6	5
zmršen	adj	1	0.66	0.66	5	10	4	6	1
kliše	n	1	0.41	0.41	5	7	7	0	2
prileżnica	n	0.64	0.26	0.13	5	6	5	1	1
pretepanje	n	-0.08	0.35	0.24	5	7	5	2	1
odveza	n	1	0.62	0.62	4	4	4	0	2
prasica	n	0.37	0.53	0.24	4	12	10	2	3
grotesken	adj	0.15	0.15	1	4	10	3	7	2
razsipništvo	n	0.12	0.30	0.02	4	7	1	6	1
somišljenik	n	0.63	0.06	-0.10	3	5	2	3	1
izpodbijati	v	0.55	1	0.55	3	5	2	3	2
odmevno	adv	0	0.33	0.57	3	4	3	1	-
spenjati	v	0	0	0	3	3	3	0	2
bivalen	adj	1	1	1	2	4	3	1	1
pravokotnica	n	1	1	1	2	3	1	2	1
zalust	n	1	1	1	2	3	1	2	1
povratno	adv	1	0.33	0.33	2	4	4	0	2
nagajivka	n	0.51	-0.17	-0.18	2	10	10	0	2

tarnanje	n	0.22	-0.06	-0.10	2	15	12	3	1
gensko	adv	0	0	1	2	2	1	1	-
prevohati	v	0	0	1	2	2	2	0	2
despotski	adj	-0.11	0.32	-0.22	2	9	9	0	1
ponazarjati	v	-0.20	0.57	-0.29	2	4	4	0	1
spodjesti	v	1	1	1	1	4	2	2	3
predelan	adj	1	1	1	1	2	1	1	1
ultramoderen	adj	1	1	1	1	2	2	0	1
paleolitik	n	1	1	1	1	2	2	0	1
investicija	n	1	1	1	1	1	1	0	2
konjunkcija	n	1	1	1	1	1	1	0	2
nevedno	adv	1	1	1	1	1	1	0	1
vsemogočnost	n	1	1	1	1	1	1	0	1
naturen	adj	1	1	1	1	1	1	0	1
popoten	adj	1	1	1	1	1	1	0	1
vrvičen	adj	0	0	1	1	2	2	0	1

E1-E2 – ARI between groups of Evaluator 1 and Evaluator 2
A-E1 – ARI between automatically assigned groups and Evaluator 1
A-E2 – ARI between automatically assigned groups and Evaluator 2
groups – number of groups created by the P-PR algorithm
synonyms – number of synonyms
core – number of core synonyms
near – number of near synonyms
SSKJ – number of senses in the monolingual Slovene dictionary