COLLOCATION RANKING: FREQUENCY VS SEMANTICS

Nikola LJUBEŠIĆ
Jožef Stefan Institute; Faculty of Computer and Information Science, University of Ljubljana

Nataša LOGAR
Faculty of Social Sciences, University of Ljubljana

Iztok KOSEM
Faculty of Arts, University of Ljubljana; Jožef Stefan Institute

Ljubešić, N., Logar, N., Kosem, I.: Collocation ranking: frequency vs semantics. Slovenščina 2.0, 9(2): 41–70.

DOI: https://doi.org/10.4312/slo2.0.2021.2.41-70

Collocations play a very important role in language description, especially in identifying meanings of words. Modern lexicography's inevitable part of meaning deduction are lists of collocates ranked by some statistical measurement. In the paper, we present a comparison between two approaches to the ranking of collocates: (a) the logDice method, which is dominantly used and frequency-based, and (b) the fastText word embeddings method, which is new and semantic-based. The comparison was made on two Slovene datasets, one representing general language headwords and their collocates, and the other representing headwords and their collocates extracted from a language for special purposes corpus. In the experiment, two methods were used: for the quantitative part of the evaluation, we used supervised machine learning with the area-under-the-curve (AUC) ROC score and support-vector machines (SVMs) algorithm, and in the qualitative part the ranking results of the two methods were evaluated by lexicographers. The results were somewhat inconsistent; while the quantitative evaluation confirmed that the machine-learning-based approach produced better collocate ranking results than the frequency-based one, lexicographers in most cases considered the listings of collocates of both methods very similar.

Keywords: collocations, word embeddings, logDice, general language, academic language

1 INTRODUCTION

The importance of the notion of collocation has been acknowledged by linguists for a long time, ever since J. R. Firth's famous statement: "You shall know a word by the company it keeps" (Firth, 1957). In fact, collocations themselves are considered by many as lexical units with different levels of semantic transparency (Singleton, 2000). As a result, even transparent collocations (and not only idioms, phrases and other more fixed multiword units) have started to receive more attention in dictionaries.

Collocation identification requires a computational approach. Several statistics for measuring collocation have been proposed in the past decades, for example t-score, MI, MI3, the log-likelihood ratio, the Dice coefficient, etc. (see Manning and Schütze, 1999, for an overview). In fact, collocation has been the pervasive driving force behind the development of tools for analysing and describing language in general. However, with progress also new challenges arose. Problematic aspects of different statistical approaches for measuring collocation have often been discussed (cf. Kilgarriff and Kosem, 2012), which led to the proposals of new measures such as logDice (Rychlý, 2008), which has been developed with lexicographic use in mind, and has been used by a large number of dictionary projects.

Nowadays, new, non-statistical methods are slowly finding their way into dictionary-making (and language) analysis. We thus decided to test one popular and up-to-date language modelling technique, namely word embeddings (Levy and Goldberg, 2014; Li and Jurafsky, 2015; Camacho-Collados and Pilehvar, 2018; etc.).

1.1 The aim and the scope of the paper

As Levy and Goldberg (2014, p. 302) explain, in the embeddings, distributional semantics word embeddings are vector representations of all the contexts in which a word occurred, and "enable efficient computation of word similarities through low-dimensional matrix operations". Recent uses of word embeddings for identifying collocations are well recorded (cf. Section 2). Various experiments proved the method to be moderately to highly successful in various tasks. We decided to find out how well it performs when given one other task, that is a task of collocate ranking. Since our research was lexicographically

oriented, we were especially interested in how well the method performs in comparison to the lexicographically highly popular logDice metric (Rychlý, 2008), which uses heuristics (i.e. a set of fixed rules).

Broadly speaking, we also wanted to find out whether a dictionary-making process (in our case a Slovene dictionary-making process) could become less time consuming and more efficient, if complemented with collocate ranking data acquired by the semantic-based method of word embeddings.

In order to establish how well word embeddings tackle the task of collocate ranking for lexicographic purposes we set a two-part experiment. It consisted of:

1. the quantitative analysis of
    a) heuristic-based vs machine-learning-based approach to collocate ranking, and
    b) frequency-based vs semantics-based machine-learning approach to collocate ranking;
2. the qualitative analysis of different collocate ranking results, which was performed by lexicographers.

In both analyses, two datasets were used:

- a general Slovene language dataset named KOLOS (Kosem et al., 2018), and
- a Slovene for special purposes (LSP) dataset named KAS (Erjavec et al., 2020).

Namely, we also wanted to draw some initial conclusions about the two approaches to collocation ranking with regards to differences in text type, and monosemy/polysemy of words.

All in all, the experiment arose from an actual dictionary-making process, and is described here with the purpose of bringing possible benefits to similar endeavours elsewhere as well.

2 MEASURING COLLOCATIONS: ASSOCIATION MEASURES, AND MORE RECENT – WORD EMBEDDINGS

An extensive body of research exists on measuring collocation strength or collocativity (e.g. Berry Rogghe, 1973; Church and Hanks, 1990; Church et al., 1991; Biber, 1993; Manning and Schütze, 1999; Evert, 2004; Gries, 2013), and different statistical methods (i.e. association measures) have been used up to this day. Association measures have also been regularly compared, and new ones proposed. Two good overviews of association measures are Wiechmann (2008) who compared 47 different association measures, and Pecina (2009), who conducted a comparison of more than 80 measures for collocation extraction. General observations of the majority of such studies were aptly summarized by Evert (2009), namely that "different association measures will produce entirely different rankings of the collocates" (ibid., p. 1218) and that "there is no ideal association measure for all purposes" (ibid., p. 1236).

A recent study by Evert et al. (2017) inspected the role of variables such as corpus size, context span, and frequency threshold in collocation identification. Using two different dictionaries as gold standards, it proved that "very large Web corpora and small co-occurrence contexts produce the best results" (ibid., 543). Moreover, in terms of co-occurrence span, researchers concluded that syntactic dependency was the best choice in most cases.

There is some literature on association measures used on Slovene corpus data as well (e.g. Gorjanc and Vintar, 2000; Gorjanc and Fišer, 2010), however there are no studies that would comprehensively compare the effectiveness of various association measures for identifying collocations in Slovene. As far as language description is concerned, in recent years most Slovene lexicographical and terminological projects have started using the Sketch Engine (Kilgarriff et al., 2004) and rely on association measures provided by this tool, especially logDice which is used by the well-known Word sketch function. However, as Gantar et al. (2015) and Gantar et al. (2016) observed, logDice often misses, or attributes very low ranking to certain important collocates, which is why researchers started combining logDice and raw frequency rankings when extracting and analysing collocates for dictionary purposes.

All association measures have one shortcoming in common: even if they are limited by predefined syntactic relations (such as in word sketches), they rely solely on co-occurrence frequencies and do not consider semantic aspects of words. And precisely this type of information is contained in word embeddings.

Word embeddings have been used extensively in the field of natural language processing (NLP) in the last decade. For example, Rodríguez-Fernández et al. (2016a) followed the well-known association approach early identified in Mikolov et al. (2013), where *king* to *man* is the same as *queen* to *woman*. They applied the same technique to collocation extraction, hoping to obtain the proper headword for the collocate *suggestion*, related to the known *take a walk* collocation. In their approach they hoped to be able to remove the *walk* information from *take* and add the *suggestion* information, ending up with *make* being a near-neighbour of the resulting vector, that is they calculated *vec(take) – vec(walk) + vec(suggestion)* with the goal of the result being close to *vec(make)*. This approach, evaluated in follow-up work, obtained a mean reciprocal rank (MRR) score between 0.01 and 0.47.[1]

Another piece of work by the same group of authors (Rodríguez-Fernández et al., 2016b) found that a linear transformation of the headword embedding can be used to predict the optimal collocate word embedding, learning this transformation per Mel'cuk semantic typologies (Mel'cuk, 1996). They did not compare this approach to the basic frequency-based one, nevertheless they achieved promising, but varying results, with the mean reciprocal rank (MRR) of the best-performing system between 0.3 and 0.9. This methodology was followed by Enikeeva and Mitrofanova (2017), who applied it to Russian data. They reported slightly higher MRR scores, ranging from 0.48 to 0.9. Again, they did not compare their results to the traditional frequency-based methods.

Liu and Huang (2017) showed that using the cosine distance between the distributional word representations of headwords and collocates as a function for

1 Mean reciprocal rank (MRR) is a relative score meant for ranked results that calculates the average of the inverse of the ranks at which the first positive instance occurs. MRR ranges between 0 and 1, and an MRR of 1 is obtained if in each ranking the positive instance is ranked in the first position, an MRR of 0.5 is obtained if in each ranking the positive instance occurs in second position, 0.33 for the third position and so on.

ranking collocation candidates yielded just slightly better results measured by F1 than the chi-square and mutual information co-occurrence statistics. Additionaly, Wanner et al. (2017) used distributional word representation to classify collocations into semantic classes, and Garcia et al. (2017) used multilingual word embeddings to find collocation translations in other languages.

Examining related literature, we can conclude that regardless of the fact that word embeddings are a very popular source of semantic information and that their usage as input features for making predictions in NLP has been considered a standard approach for years now, they have not yet been tested in a supervised learning setting on the task of general collocation ranking.

3 RESEARCH

3.1 Methodology

3.1.1 Research questions

In order to establish how well word embeddings tackle the task of collocate ranking for lexicographic purposes in the case of Slovene, we compared the embeddings results to the results obtained using the logDice method. The comparisons were made in a quantitative and qualitative way and were led by the following three research questions:

Q1: Which approach produces lexicographically more relevant rankings of collocates: the one that uses machine learning over manually annotated data, or the one that uses heuristics?

Q2: Which approach is a more useful source of information for the rankings of collocates: the word embeddings approach, which encodes distributional semantics of words, or the logDice approach, which encodes frequency information?

Q3: Which ranking of collocates is preferred by lexicographers: the embeddings ranking, or the logDice ranking?

As questions imply, we wanted to know whether the currently still dominant approach of using heuristics for collocate ranking is really better than the machine learning approach, which implicitly learns the underlying rules from examples. The second question was aimed at comparing two sources of

information – frequency, which is used in a heuristic way in logDice, and distributional semantics, which is exploited from word embeddings via machine learning. Finally, our third question put potential users of the two compared approaches (i.e. lexicographers) into focus and examined their preferences in actual cases.

3.1.2 Collocation datasets

3.1.2.1 KOLOS dataset

The KOLOS dataset contained a carefully selected set of 333 headwords, consisting of 154 nouns, 73 verbs, 81 adjectives, and 25 adverbs. The selected headwords were as heterogeneous as possible in terms of word class subcategories (e.g. plural nouns, countable nouns, transitive vs intransitive verbs etc.), corpus frequency, level of polysemy (number of different meanings), semantic characteristics (e.g. abstract vs concrete senses; qualitative vs classifying adjectives), etc. For each headword, we used collocations extracted for the purposes of the Collocations Dictionary of Modern Slovene (Kosem et al., 2018; Kosem et al., 2019). It should be noted that we already had a set of validated collocations from the Slovene Lexical Database (Gantar et al., 2016), and in order to devise a training dataset of good and bad collocation candidates, we decided to annotate only new ones (i.e. not yet validated collocations). This meant that we were often annotating the collocations slightly lower down the logDice-ordered list for each grammatical relation.

In the annotation task, the annotators were presented with a collocation, the information of its grammatical relation, and a corpus example of its use. The annotation of collocations was conducted in the Pybossa tool,[2] with each collocation being annotated by three annotators-linguists. The examples were extracted with the GDEX tool (Kosem et al., 2013) in the Sketch Engine (Kilgarriff et al., 2008), using the Slovenian configuration. The annotators were presented with three main answer groups – YES ('yes, this is a valid collocation'), NO ('no, this is not a collocation') and I DON'T KNOW ('I don't know if this is a collocation or not') (the YES and NO groups had additional sub-options, but they were not used in this experiment).

2 https://mnozicenje.cjvt.si/

Taking YES, NO and I DON'T KNOW answers, the agreement was analysed and the final decision for the training dataset, which could only be YES or NO, was made on the basis of the agreement (e.g. total agreement was YES or NO), while in borderline cases the final decision was made by making additional annotation or after joint discussion by the annotators.

The whole KOLOS dataset consisted of 17,540 collocation candidates belonging to 260 different grammatical relations. For the experiments performed in this paper we organised collocation candidates under 7,460 headwords (those being any of the two lexical parts of a bidirectional grammatical relation, so for *take a walk* we would have two collocations, once under the headword *take*, once under the headword *walk*). Experiments were done only on headwords that (1) had at least 10 collocation candidates for a specific grammatical relation as our evaluation was headword-based (this was the only data organisation that allowed evaluation of frequency-based statistics), and that (2) covered both the positive and the negative class so that discriminative machine learning (distinguishing between good and bad examples) can be performed. With these selection criteria the KOLOS dataset was shrunk to the most frequent 8 grammatical relations (actually 4 bidirectional relations), 212 headwords and 2,671 collocation candidates.

3.1.2.2 KAS dataset

The KAS dataset is a set of academic Slovene headwords, such as *analiza* (*analyses*), *tabela* (*table*), *razlikovati* (*to distinguish*), *relativno* (*relatively*), accompanied by collocations and examples of use (Logar et al., 2019). The set was built from a one-billion-word corpus KAS (Erjavec et al., 2020). The corpus was harvested from the Open Science Portal of Slovenia (2000–2015). For the most part (71% of tokens), it consists of BSc and BA theses, followed by MSc and MA theses (20%), and PhD theses (4%). Firstly, the initial list of candidates for the vocabulary of academic headwords was built by using the method of frequency profiling (Rayson and Garside, 2000). With this method we extracted lemmas that most differentiated the KAS corpus from a fiction part of the general corpus Kres (Logar et al., 2012, p. 79–97). Secondly, we inspected each lemma on the list in the KAS corpus concordances, and also checked its typical context in the Sketch Engine tool. In this manner we

determined whether the word in question belonged to a common expert discourse or not (the latter were excluded as it meant they were either grammatical words or technical terms). And thirdly, the final list of 463 headwords identified as typical of academic Slovene was supplemented by collocations and three examples of use for each collocation. The extraction of data was automatic; we used the same methodology as in the case of the KOLOS dataset (Kosem et al., 2011; Krek, 2012; Gantar et al., 2015; Kilgarriff and Kosem, 2012; Logar et al., 2014).

Automatically extracted data was then reviewed. We corrected the most obvious tagger performance mistakes, rearranged not ideally semantically grouped collocates, and deleted personal proper names, deixis, modal verbs and verbs with very broad meaning (e.g. *to be*, *to be about (sth)*). Nevertheless, all deletions remained part of the dataset, but were labelled as NEGATIVE collocation candidates.

Content-wise, the KAS dataset was heterogeneous with regards to its meaning and text function, but was either obviously or indirectly related to three roughly defined segments (Logar and Erjavec, 2019, p. 212–213): (a) the formal structure and the writing of academic texts (e.g. in English *bibliography*, *introduction*, *conclusions*; *empirical*, *defined*, *mentioned*; *to define*, *to cite*); (b) the methodology of academic texts (e.g. *method*, *hypothesis*, *respondent*; *to analyse*, *to identify*, *to classify*); or (c) the presentation and interpretation of the research data (e.g. *number*, *portion*, *dependence*; *measured*, *calculated*, *accurate*; *to result from*, *to indicate*, *to cause*; *subsequently*, *relatively*, *successfully*). With regard to word class, out of 463 headwords 226 were nouns, 119 adjectives, 86 verbs, and 32 adverbs (Logar et al., 2019). As far as the use in the KAS corpus is concerned, all words in the KAS dataset were monosemous.

In total, the KAS dataset consisted of 70,254 collocation candidates belonging to 342 different grammatical relations, organised under 5,220 headwords. By applying the same selection criteria as on the KOLOS dataset, our final KAS dataset on which we performed experiments shrunk to 8 grammatical relations (gramrels hereafter), 525 headwords and 14,722 collocation candidates.

3.1.3 Corpus information

The frequency and semantic information for our collocation candidates was obtained from the Gigafida 2.0 corpus (Krek et al., 2020). For calculating the frequency and logDice information as representatives of the frequency signal we used the Sketch Engine API. For calculating the (head)word embeddings as representatives of the semantic signal we used the fastText tool (Bojanowski et al., 2016) – in skip-gram mode with default parameters – and the lemma and part-of-speech annotations present in Gigafida 2.0, KAS and other large corpora of Slovene (Ljubešić and Erjavec, 2018).

3.2 Experiment

As explained in the Introduction section, our experiment consisted of two main parts:

1. the quantitative analysis, and
2. the qualitative analysis.

In both, we compared two approaches to collocate ranging, i.e. the logDice method and the word embeddings method. In the quantitative analysis, we performed two parts of the experiment, and in the qualitative part one more followed. Each of the three parts of our experiment was directly related to one of the research questions formulated at the beginning of the research.

3.2.1 Quantitative analysis

3.2.1.1 Experimental setup

In the quantitative part of the experiment, our goal was to compare traditional statistic-based approaches to collocate ranking with approaches based on machine learning. Since the only organisation that we can obtain through traditional approaches are ranked results (collocation candidates with higher frequency or higher logDice score are ranked higher), we set up our machine-learning experiments also in the way that enabled us to obtain ranked results. To evaluate traditional methods in their regular usage scenario, we performed evaluation on a per-gramrel and per-headword basis.

For our evaluation metric, we used the AUC (area-under-the-curve) ROC (receiver operating characteristic) score, which is considered to be the go-to

evaluation metric for ranking results, especially if the classes (positive and negative collocation candidates) are not balanced. Precisely this was the case in our datasets as in our original KOLOS dataset we had 13,812 positive candidates and 3,728 negative ones. The situation in the KAS dataset was similar, with 53,150 positive and 8,811 negative collocation candidates.

The AUC ROC score quantifies the quality of a ranking result, with the worst-possible ranking (all negative collocation candidates being ranked higher than all positive candidates) obtaining the result of 0.0, a perfect ranking (all positive collocation candidates being ranked higher than all negative collocation candidates) obtaining the result of 1.0, and a random ranking (positive and negative candidates being randomly mixed) obtaining the result of 0.5.

For performing supervised machine learning experiments, we used support-vector machines (SVMs), a regular go-to algorithm in traditional machine learning. We did not use more recent neural-network approaches as (1) their parameters are harder to interpret, and (2) initial experiments on our datasets had shown very similar results regardless of the machine-learning approach used. We had to be able to predict continuous values to be used for ranking candidates, thus we trained SVM regressors. All our implementations are written in the scikit-learn toolkit (Pedregosa et al., 2011).

Given that we obtained AUC ROC scores per each ranking (i.e. for each gramrel and headword we got a score), we had to set up a way to average all scores on some defined level. We aimed at averaging on the gramrel and overall level. As (1) different headwords under specific gramrels had a different number of candidates, and (2) different gramrels had a different number of candidates, we decided to normalise our results given the number of candidates, that is each collocation candidate would have the same impact on the final score of a method.

Supervised machine learning required two sets of data: training data (the data the model is built on) and testing data (the data the built model is evaluated on). Therefore, we performed a five-fold cross-validation, that is we split our training data into five groups, running five iterations of using four groups for training and one group for testing. By doing so we managed to evaluate the

model on each data point available, which is directly comparable to the output of the statistic-based ranking methods where we do not require training data. Furthermore, we made sure that headwords were sampled into groups, so that there was no spillage between training and testing data (e.g. training on some collocations of a headword and testing on other collocations of that headword). This makes the machine-learning approach quite challenging and measures to what extent the model can generalise regularities on the gramrel level, but not on the level of specific headwords present in our dataset.

3.2.1.2 Results

As explained, we obtained results on two datasets, KOLOS and KAS, by comparing four different approaches to collocation candidate ranking:

- freq: ordering via decreasing frequency of the collocations;
- logDice: ordering via decreasing logDice statistic of the collocations (using the frequencies of the headword, collocate and collocation);
- SVM_freq: machine learning the ranking from the frequency of the collocation, the headword, the collocate and the logDice statistic (all frequencies being represented on the logarithm scale);
- SVM_emb: machine learning the ranking from the embeddings of the headword, the collocate, and a sum of the two embeddings (to represent in a basic fashion the interaction between the two embeddings).

In Table 1, we present our results on the KOLOS dataset, together with the statistics on the size of the dataset for each gramrel. In Table 2, we give a similar description and results on the KAS dataset. Focusing first on the overall results on each dataset (the TOTAL row), the depicted picture is quite simple. The answer to our first research question, namely whether machine learning approach produces more relevant rankings of collocates than the approach based on heuristics, is positive. On the KOLOS dataset the two statistic-based approaches yielded scores of 0.52 and 0.47, while the two machine-learning-based approaches obtained scores of 0.58 and 0.71. On the KAS dataset the statistic-based approaches achieved scores of 0.58 and 0.63, while the machine-learning-based approaches obtained scores of 0.76 and 0.87.

With our second research question regarding the usefulness of both ambeddings approach and the logDice approach we again favoured the former. On the KOLOS dataset the frequency-based learning obtained the score of 0.58, while the semantic-based approach achieved the score of 0.71. On the KAS dataset the numbers obtained were 0.76 and 0.87, aiming at the same conclusion. Even more, there was only one gramrel (among 16) on which the machine-learning approach based on semantic information did not score the best results among the four approaches evaluated here (namely, the logDice score 0.65 for the *VERB + noun (accusative)* gramrel, see italics in Table 2).

An interesting, if not troubling observation is that ranking results via heuristics are quite close to the random baseline, with an average result on the KOLOS dataset of around 0.5 and on the KAS dataset of around 0.6. This suggests that their ranking is actually quite incapable of pushing the negative candidates as far down as possible. However, it still might be that the overall order of candidates via these two heuristics is useful for human use. In our experiments, we were aware only of the positive vs negative collocation candidate distinction and not of all subtle differences that collocations bring in a ranking scenario.

Table 1: *KOLOS dataset: the ranking results of the machine learning approach**

gramrel	# heads	# collos	freq	logDice	SVM_freq	SVM_emb
adjective + NOUN	38	576	0.526	0.405	0.56	**0.653**
ADJECTIVE + noun	54	983	0.503	0.463	0.534	**0.692**
NOUN + noun (genitive)	22	481	0.698	0.353	0.712	**0.78**
noun + NOUN (genitive)	47	967	0.517	0.501	0.631	**0.723**
VERB + noun (accusative)	13	231	0.468	0.443	0.432	**0.64**
verb + NOUN (accusative)	13	242	0.444	0.405	0.472	**0.737**
ADVERB + adjective	12	261	0.368	0.677	0.602	**0.802**
adverb + ADJECTIVE	13	221	0.584	0.62	0.515	**0.669**
TOTAL	212	3962	**0.523**	**0.469**	**0.577**	**0.71**

* Capital items: the headword and the starting point of the collocation (also from here forward, i.e. in Table 2, Table 4, etc.).

Table 2: *KAS dataset: the ranking results of the machine learning approach*

gramrel	# heads	# collos	freq	logDice	SVM_freq	SVM_emb
ADJECTIVE + noun	53	1737	0.537	0.563	0.665	**0.738**
adjective + NOUN	118	3045	0.58	0.689	0.8	**0.932**
NOUN + noun (genitive)	46	1677	0.559	0.534	0.603	**0.866**
noun + NOUN (genitive)	72	1999	0.565	0.556	0.623	**0.878**
VERB + noun (accusative)	18	828	0.619	*0.651*	0.59	0.556
verb + NOUN (accusative)	77	1947	0.632	0.597	0.913	**0.922**
ADVERB + adjective	52	1468	0.745	0.709	0.802	**0.894**
adverb + ADJECTIVE	89	2021	0.431	0.706	0.915	**0.954**
TOTAL	525	14722	**0.576**	**0.628**	**0.757**	**0.871**

For the different gramrels we also performed a correlation analysis to measure to what degree the results through gramrels and applied methods are stable between the two datasets. We calculated the Pearson correlation coefficient between the 8 results for each of the four methods on the KOLOS and on the KAS dataset. For the frequency method, we obtained a significant ($p = 0.043$) strong negative result ($r = -0.722$), and for the logDice method we again obtained a significant ($p = 0.029$), but strong positive result ($r = 0.758$). For the SVM_freq method our result was not significant ($p = 0.36$) and was moderately negative ($r = -0.375$), while for the SVM_emb method the result was also not significant ($p = 0.183$), but was moderately positive ($r = 0.524$). These results show that in the machine learning scenario achievements on specific grammatical relations differ quite a lot between datasets, while the logDice method was similarly (un-)successful on different gramrels. Nevertheless, the samples we obtained these calculations on are very small and one should take these results with caution. The only claim that could be made here is that in most cases the per-gramrel results are quite inconsistent.

3.2.2 Qualitative analysis

3.2.2.1 Experimental setup

We expected that lexicographers, too, would prefer the machine-learning results to those of heuristics, hence we tested our third hypothesis by presenting

them with two side-by-side columns for each headword in a specific grammatical relation, one column representing logDice ranking and one column representing embeddings ranking of collocates (see an example in Table 3). Lexicographers were asked to evaluate which column was more informative to them (column A or B), but they could also choose an answer *Both columns are similarly (un)informative*. This meant that either (a) both measures were equally informative or useful, or that (b) none of the measures was informative or useful. In addition, participants were alerted to the fact that they were evaluating results of the two aforementioned collocation extraction methods, but did not know which column was the result of which method. We also instructed them to pay more attention to top halves of lists in both columns. No other instructions for the evaluation process were given.

Table 3: *KOLOS dataset: headword* belina *(whiteness), grammatical relation: NOUN + noun (genitive) (the whiteness of ___)*

ranking	logDice (A)	embeddings (B)
1.	**zob (*tooth*)****	stena (*wall (interior)*)
2.	**sneg (*snow*)**	pokrajina (*landscape*)
3.	marmor (*marble*)	oblačilo (*clothes*)
4.	**polt (*complexion*)**	perilo (*washing*)
5.	perilo (*washing*)	kamen (*stone*)
6.	platno (*linen*)	marmor (*marble*)
7.	**papir (*paper*)**	obleka (*dress*)
8.	stena (*wall (interior)*)	koža (*skin*)
9.	zid (*wall*)	platno (*linen*)
10.	kamen (*stone*)	zid (*wall*)
11.	nebo (*sky*)	**sneg (*snow*)**
12.	obleka (*dress*)	nebo (*sky*)
13.	oblačilo (*clothes*)	**papir (*paper*)**
14.	pokrajina (*landscape*)	**polt (*complexion*)**
15.	koža (*skin*)	**zob (*tooth*)**
16.	obraz (*face*)	obraz (*face*)

** Bold print = in the case of the embeddings method, a noticeable drop in the ranking; underlined words = in the case of the embeddings method, a noticeable increase in the ranking.

This part of the experiment was partially done via a set of .txt documents and partially via an online survey. First, a preliminary evaluation on a smaller set

of .txt documents was performed by two lexicographers; one familiar with the KAS database and the other familiar with the KOLOS database. During this phase, the lexicographer evaluating the KAS database favoured logDice as having better ranking results, while the second lexicographer in some cases preferred the embeddings and noticed that the performance of this method might have been gramrel dependent. Since preliminary evaluation was inconclusive, seven other lexicographers were later invited to participate in the study (that is the online survey part of it).

The questionnaire of the online survey only included headwords from the KOLOS dataset, while the KAS dataset was further inspected only by the lexicographer who conducted the preliminary analysis. The reason for this decision was that all lexicographers invited to the online survey had experience with general dictionary and general dictionary-like resources and they were all involved in the KOLOS project, while only one lexicographer participated in the KAS project, that is the part that focused on general academic discourse vocabulary. Since we wanted to keep the expectations and initial positions of all of the lexicographers homogeneous, we kept them separate, as well as the datasets they evaluated.

Further KAS dataset analysis that was performed, as mentioned, by one lexicographer was done on eight randomly chosen headwords in ten different grammatical relations (i.e. 80 headwords: 24 nouns, 8 adjectives, 32 verbs, 16 adverbs), which in total summed up to 2,095 collocations repeated in two columns. On average, this meant 26 collocates per headword in a specific gramrel (with the smallest number of 10 and the largest number of 93 collocates per headword). In this second phase of the evaluation, the lexicographer evaluating the KAS dataset paid a closer attention to top halves of collocate columns, as did the online survey participants.

The online survey consisted of 63 headwords (34 nouns, 18 adjectives, 11 verbs) and their collocates in seven different grammatical relations. Because we wanted to broaden the number of gramrels, only three of them were the same in both datasets. The survey was divided into seven separate grammatical relation subsurveys, which meant that each grammatical relation had its own survey link. This was done to keep the cognitive load manageable for participants (they could complete the survey for one grammatical relation

and continue with the next one on another day), and to facilitate the analyses. In total, there were 146 pairs of collocate lists (i.e. questions in the survey; see Table 4). It should be noted that due to various reasons (time constraints etc.) not all the participants completed all seven grammatical relation surveys.

Table 4: *Online surveys: number of headwords and number of lexicographers participating*

gramrel	number of headwords	number of participants
VERB + noun (accusative)	12	6
verb + NOUN (accusative)	26	8
ADJECTIVE + noun	19	6
adjective + NOUN	30	6
adverb + ADJECTIVE	11	7
NOUN + noun (genitive)	19	6
noun + NOUN (genitive)	29	6

3.2.2.2 Results

3.2.2.2.1 KAS collocates ranking

As Table 5 and Figure 1 show, the lexicographer evaluating the KAS database in the second phase of the study again did not find the embeddings rankings better than the logDice rankings. In almost two thirds of cases (51/80), she decided that both columns were very similar, and in almost all of the rest of them (26/80), in her opinion, the embeddings performed worse. Thus, a small number of only three cases of embeddings performing better can be perceived as exceptions.

A closer look at grammatical relations reveals that the success of both ranking methods differs according to the lexicographers' judgments. Collocate ranking according to logDice was preferred in grammatical relations *NOUN + noun (genitive)* and *VERB + noun (accusative)*, while the ranking results of both methods were very similar in four relations (right side of Figure 1): *NOUN + "for" + noun (accusative)*, *VERB + "and_or" + verb*, *ADVERB + adjective*, and *ADVERB + verb*.

Table 5: *KAS dataset: logDice ranking vs embeddings ranking of collocates per grammatical relation (in absolute numbers and percentage)*

gramrel	logDice better: number and (%)	embeddings better: number and (%)	very similar: number and (%)
ADJECTIVE + noun	4 (50)		4 (50)
VERB + adverb	3 (37)		5 (63)
NOUN + noun (genitive)	5 (63)	1 (2)	2 (25)
VERB + noun (genitive)	3 (38)	1 (2)	4 (50)
VERB + noun (accusative)	5 (63)		3 (37)
NOUN + preposition *v* (*in*) + noun (locative)	4 (50)		4 (50)
NOUN + preposition *za* (*for*) + noun (accusative)	1 (2)		7 (98)
VERB + conjunction *in_ali* (*and_or*) + verb			8 (100)
ADVERB + adjective	1 (2)		7 (98)
ADVERB + verb		1 (2)	7 (98)
TOTAL	**26 (32)**	**3 (4)**	**51 (64)**

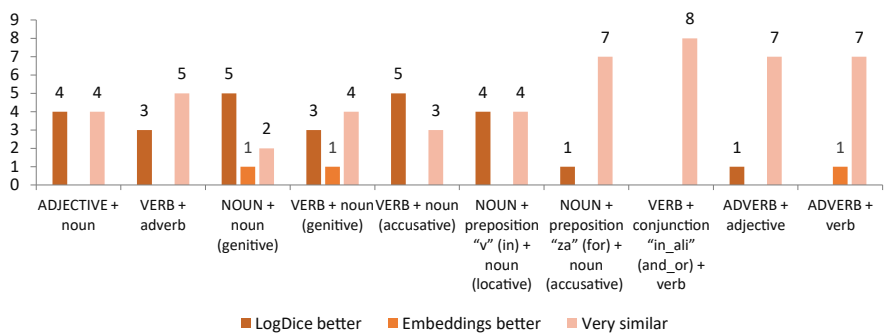

Figure 1: KAS dataset: logDice ranking vs embeddings ranking of collocates per grammatical relation (in absolute numbers).

3.2.2.2.2 KOLOS collocate ranking

Overall, the most popular answer in the online survey was *Both columns are similarly (un)informative* (45% of the answers, Table 6), which indicates that the participants having a general dictionary-like resource in mind did not, almost half of the time, consider one ranking better than the other.

Table 6: *KOLOS dataset: logDice ranking vs embeddings ranking of collocates per grammatical relation (in absolute numbers and percentage)*

gramrel	logDice better: number and (%)	embeddings better: number and (%)	very similar: number and (%)	TOTAL ANSWERS: number
VERB + noun (accusative)	16 (24)	22 (33)	28 (42)	66
verb + NOUN (accusative)	74 (37)	24 (12)	102 (51)	200
ADJECTIVE + noun	33 (31)	31 (29)	44 (41)	108
adjective + NOUN	73 (42)	21 (12)	80 (46)	174
adverb + ADJECTIVE	27 (39)	8 (11)	35 (50)	70
NOUN + noun (genitive)	23 (21)	36 (33)	50 (46)	109
noun + NOUN (genitive)	44 (26)	62 (37)	62 (37)	168
TOTAL	**290 (32)**	**204 (23)**	**401 (45)**	**895**

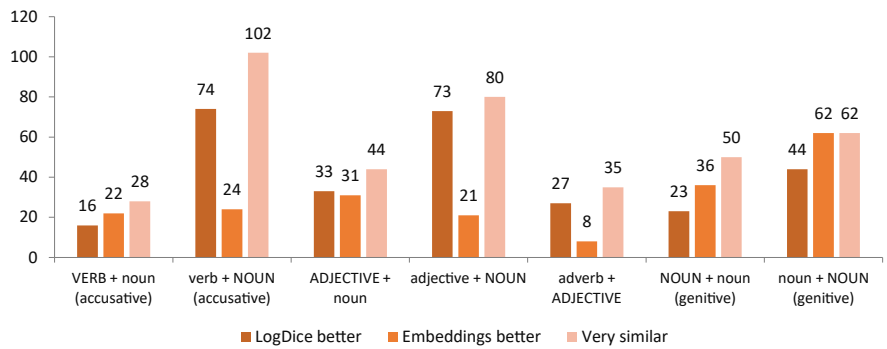

Figure 2: KOLOS dataset: logDice ranking vs embeddings ranking of collocates per grammatical relation (in absolute numbers).

Of the two measures, logDice was considered better more frequently than embeddings, with 32% vs 23% answers selected respectively. However, as Table 6 and Figure 2 show, this ratio between the two measures varied considerably according to the grammatical relation. Ranking of collocates according to logDice was more preferred in grammatical relations *verb + NOUN (accusative)*, *adjective + NOUN*, and *adverb + ADJECTIVE*. On the other hand,

embeddings ranking was preferred in *VERB + noun (accusative)*, *NOUN + noun (genitive)*, and *noun + NOUN (genitive)* grammatical relation.

We also searched for patterns in the results on a headword level, especially for headwords that featured in at least two different grammatical relations. We wanted to establish whether certain headwords prefer one of the measures across different grammatical relations. Similar to above mentioned findings, logDice was again preferred more often than embeddings, with the participants preferring it at 26 headwords in different grammatical relations, while embeddings results were preferred at only 14 headwords (for the remaining headwords no considerable differences in preferences were observed). There were also no clear patterns that the headwords identified had in common.

At the end of both evaluations, we made a numerical comparison of the results in total for both datasets (Figure 3).

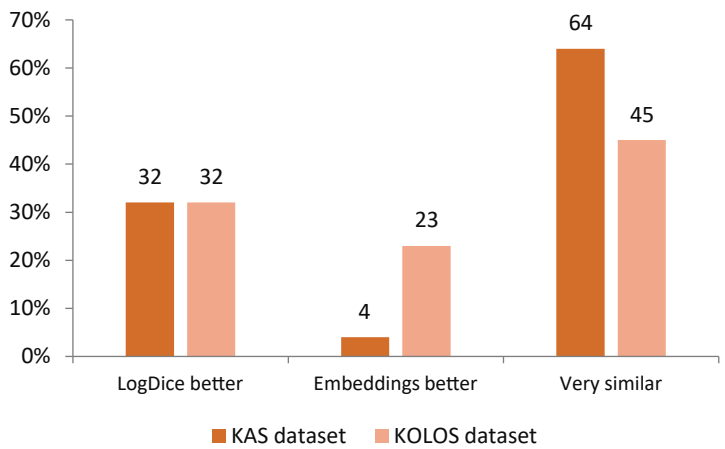

Figure 3: KAS and KOLOS dataset: logDice ranking vs embeddings ranking of collocates – both evaluations in total (in percentage).

Even though our deduction is limited due to the fact that only one lexicographer examined the KAS dataset, one feature in Figure 3 stands out: to a noticeably larger extent (23%) the embeddings rankings of collocates of the KOLOS dataset were recognised as more informative than those of the KAS dataset (4%). It is possible that this is a consequence of the KOLOS being

much more polysemous. If that is the case, at least this part of our qualitative analysis favours the semantic-based method to logDice metrics. Nevertheless, as a whole our third research question, namely which ranking of collocates is prefered by lexicographers, must be answered in the following way: lexicographers prefer the logDice ranking.

4 DISCUSSION

The main point that needs to be discussed is the difference between the results of quantitative and qualitative analyses. With the results of the quantitative analysis so convincingly in favour of the embeddings approach, it was somewhat surprising to learn that the lexicographers did not confirm this finding. In this section, we present some possible explanations for this discrepancy.

But first, let us turn our attention to the fact that in comparison to the KOLOS dataset, higher scores of the machine-learning-based approaches were consistently obtained on the data from KAS. It seems like this was influenced by two features: the (non)specialised content of the two corpora, and the monosemy or polysemy of selected headwords. As mentioned, all headwords in the KAS dataset are monosemous (but not technical), and secondly, the KAS corpus is domain- and genre-specific; on the other hand, more than half of the KOLOS headwords were polysemous and therefore used in various contexts, but they (and their collocates) also originated from a general, domain and genre diverse corpus of Slovene. The latter very probably limited the machine-learning process, while the first enhanced it. It is our belief this should be kept in mind in follow-up testings of the embeddings method and its use in dictionary-making projects.

When answering our first and second question using the AUC ROC score and the SVM learning algorithm, the machine-learning-based approaches ranked better than statistic-based ones (KOLOS scores on frequency information: 0.52 vs 0.58), and the semantic information given through word embeddings was more useful than frequency information (KOLOS scores on using machine learning on frequency and embeddings: 0.58 vs 0.71). Yet lexicographers' most frequent evaluation was a non-decisive one: to them in half or more cases (45% for KOLOS and 64% for KAS) both rankings of collocates seemed very similar. In fact, the survey participants' comments suggest that

the task of deciding which ranking was better even proved frustrating at times. Many KOLOS survey participants mentioned that they often deliberated on monosemous or polysemous characteristic of the headword, similarity of collocates and their broad meaning, while the lexicographer evaluating KAS dataset disfavoured columns that had too general or too technical words among approximately top ten collocates. Nevertheless, the votes of all of them were given with considerable uncertainty and were very diverse.

Our survey instructions were intentionally non-explicit, in other words: instruction-wise, we did not address the aforementioned differences. We wanted to learn in general, whether the semantic nature of the embeddings collocation extraction method could be recognised and found advantageous for lexicographic work. Unfortunately, our conclusions suggest that higher algorithm scores, though numerically significant, were in most part not obvious to humans. Just one segment of KOLOS vs KAS evaluation results confirmed that there is indeed some potential in the semantic nature of the embeddings collocate rankings; namely 23% of the much more polysemic KOLOS dataset was recognised as more informative than the logDice ranking, while this was the case for only 4% of the KAS database. However, since KAS data was evaluated solely by one lexicographer, further studies should examine this indication in more detail.

With regard to the embeddings method being gramrel dependent, i.e. that it is more successful for some grammatical relations, but not the others, nothing can be concluded. By choosing a set of 17 various relations (KAS: 10, KOLOS: 7), with only three of them overlapping, gramrel-wise we were able to get a broader view, but the number of headwords per each grammatical relation was thus reduced (in total KAS: 80, KOLOS: 63). Subsequently, none of the relations was analysed comprehensively. Even with gramrels that overlapped in the datasets (*ADJECTIVE + noun*; *NOUN + noun (genitive)*; *VERB + noun (accusative)*), the survey results were not uniform and do not allow for any obvious inference. The question of gramrel importance for the task of embedding-based collocation extraction is in fact rather questionable as initial experiments on training one single model for collocation extraction on all gramrels showed very similar results to those of training separate models for each gramrel. For the sake of a better control over the process and a more

interesting analysis, in this research we opted for keeping gramrel data and gramrel experiments separate, but other scenarios are, of course, possible for future fine tunings of the method.

Finally, we must consider the part human intuition, or rather lexicographers' knowledge, experience, and past and present project involvement played in our experiment. Lexicographers' evaluation, though an expert one, played a crucial role not once, but twice. Firstly, during the annotation of collocations before the quantiative part of the experiment; and secondly, after it in the form of lexicographers' judgments of the informativeness of the collocate rankings. Machine learning was, of course, performed on the pre-annotation dataset taken as a kind of gold standard, which actually meant that the lexicographers' preferences in the post-ranking phase primarily reflected annotators' preceding decisions. Here, it is important to stress that both groups of experts consisted of almost the same people, though the time that passed between the two phases of the experiment was about five months. Also, since the pre-treatment of the KAS datasets was not identical to the pre-treatment of the KOLOS dataset, and the same goes for the evaluation part of the experiment, the comparison between the results of both datasets is far from optimal. In this respect, our conclusions need to be treated as just preliminary.

5 CONCLUSIONS

Recent trends in lexicography have focused on automating certain aspects of language description, especially those related to collocations and examples (e.g. Kilgarriff and Rychlý, 2010; Rundell and Kilgarriff, 2011). As Cook et al. (2013, p. 50) point out, a "striking outcome of the work done so far in this area is that automation not only delivers efficiency savings but also leads to improvements in quality".

Lexicographers are used to inspecting long lists of collocates, separating the wheat from the chaff, but when automatically produced language resources are in question, different results of different extraction tools matter, and improvements in quality are always possible. In our research, we used a supervised machine-learning approach to collocation extraction and ranking with the aim of establishing how advantageous it is when compared to heuristic frequency-based logDice metrics. We found that while supervised approaches

do improve over the unsupervised baseline in an automation setting, in most cases the lexicographers did not appreciate this "improvement".

Nevertheless, the results are not discouraging. They prove (and confirm) that, ideally, a good collocation extraction tool is one that combines computational measurements and lexicographers' input. Obviously, modern lexicography is still an inherently multidisciplinary endeavour with the never justly answered question of how to measure what is informative, relevant, and significant – this seems even more so for language resources of the digital era.

Acknowledgments

The research was conducted as part of the project Collocation as a basis for language description: semantic and temporal perspectives (J6-8255), funded by the Slovenian Research Agency, and within the national research programme Slovene language – basic, contrastive, and applied studies (P6-0215), and the national research programme Language resources and technologies for Slovene language (P6-0411), also funded by the Slovenian Research Agency.

REFERENCES

Berry-Rogghe, G. L. (1973). The Computation of Collocations and their Relevance in Lexical Studies. In A. J. Aitken, R. W. Bailey, and N. Hamilton-Smith (Eds.), *The Computer and Literal Studies* (pp. 103–112). Edinburgh, New York: University Press.

Biber, D. (1993). Representativeness in Corpus Design. *Literary and Linguistic Computing*, *8*(4), 243–57.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching Word Vectors with Subword Information. In H. Schütze (Ed.), *Transactions of the Association for Computational Linguistics 5* (pp. 135–146).

Camacho-Collados, J., & Pilehvar, M. T. (2018). From Word to Sense Embeddings: A Survey on Vector Representations of Meaning. *Journal of Artificial Intelligence Research 63*, 743–788.

Church, K. W., Gale, W., Hanks, P., & Hindle, D. (1991). Using Statistics in Lexical Analysis. In U. Zernik (Ed.), *Lexical Acquisition: Exploiting On-line Resources to Build a* Lexicon (pp. 116–164). Erlbaum, Hillsdale, NJ.

Church, K., & Hanks, P. (1990). Word Association Norms, Mutual Information and Lexicography. *Computational Linguistics*, *6*(1), 22–29.

Cook, P., Lau, J. H., Rundell, M., McCarthy, D., & Baldwin, T. (2013). A Lexicographic Appraisal of an Automatic Approach for Detecting New Word Senses. In I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets & M. Tuulik (Eds.), *Electronic Lexicography in the 21st Century: Thinking Outside the Paper, Proceedings of the eLex 2013 Conference, 17–19 October 2013, Tallinn, Estonia* (pp. 49–65). Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut.

Enikeeva, E. V., & Mitrofanova, O. A. (2017). Russian Collocation Extraction Based on Word Embeddings. In V. Selegey et al. (Eds.), *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2017"* (pp. 52–64). Moscow: The Computational Linguistics and Intellectual Technologies.

Erjavec, T., Fišer, D., & Ljubešić, N. (2020). The KAS Corpus of Slovenian Academic Writing. *Language Resources & Evaluation 55*, 551–583.

Evert, S. (2004). The Statistics of Word Cooccurrences: Word Pairs and Collocations, PhD Thesis. University of Stuttgart.

Evert, S. (2009). Corpora and Collocations. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics: An International Handbook 2* (pp. 1212–1248). Berlin/New York: Mouton de Gruyter.

Evert, S., Uhrig, P., Bartsch, S., & Proisl, T. (2017). E-VIEW-alation – a Large-scale Evaluation Study of Association Measures for Collocation Identification. In I. Kosem, C. Tiberius, M. Jakubíček, J. Kallas, S. Krek & V. Baisa (Eds.), *Electronic Lexicography in the 21st Century, Proceedings of eLex 2017 Conference* (pp. 531–549). Leiden, Netherlands/Brno: Lexical Computing CZ s.r.o.

Firth, J. R. (1957). *Modes of Meaning: Papers in Linguistics*: 1934–1951. London: Oxford University Press.

Gantar, P., Kosem, I., & Krek, S. (2016). Discovering Automated Lexicography: the Case of the Slovene Lexical Database. *International Journal of Lexicography*, *29*(2), 200–225.

Gantar, P., Krek, S., Kosem, I., & Gorjanc, V. (2015). Collocation Dictionary for Slovene: Challenge for Automatic Extraction of Data and Crowdsourcing. In

G. Corpas Pastor, M. Buendía Castro & R. Guttiérrez Florido (Eds.), *Computerised and Corpus-based Approaches to Phraseology: Monolingual and Multilingual Perspectives (Fraseologı´a computacional y basada en corpus: perspectivas monolingu¨es y multilingu¨es), Europhras, 2015* (pp. 84–86). Malaga: Lexytrad, Research Group in Lexicography and Translation.

Garcia, M., García-Salido, M., & Alonso-Ramos, M. (2017). Using Bilingual Word-embeddings for Multilingual Collocation Extraction. In S. Markantonatou, C. Ramisch, A. Savary & V. Vincze (Eds.), *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)* (pp. 21–30). Valencia: Association for Computational Linguistics.

Gorjanc, V., & Fišer, D. (2010). *Korpusna analiza*. Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani.

Gorjanc, V., & Vintar, Š. (2000). Iskanja po korpusu slovenskega jezika FIDA. In T. Erjavec & J. Gros (Eds.), *Jezikovne tehnologije: Zbornik konference* (pp. 20–27). Ljubljana: Institut Jožef Stefan.

Gries, S. (2013). 50-something Years of Work on Collocations. *International Journal of Corpus Linguistics*, *18*(1), 137–165.

Kilgarriff, A., Rychlý, P., Smrz, P., & Tugwell, D. (2004). The Sketch Engine. In G. Williams & S. Vessier (Eds.), *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004 Lorient, France July 6–10, 2004* (pp. 105–116). Lorient: Université de Bretagne – sud.

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M., & Rychlý, P. (2008). GDEX: Automatically Finding Good Dictionary Examples in a Corpus. In E. Bernal & J. DeCesaris (Eds.), *Proceedings of the Thirteenth EURALEX International Congress* (pp. 425–432). Barcelona, Spain: Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra.

Kilgarriff, A., & Rychlý, P. (2010). Semi-automatic Dictionary Drafting. In G.-M. de Schryver (Ed.), *A Way with Words: A Festschrift for Patrick Hanks* (pp. 299–312). Kampala: Menha Publishers.

Kilgarriff, A., & Kosem, I. (2012). Corpus Tools for Lexicographers. In S. Granger & M. Paquot (Eds.), *Electronic Lexicography* (pp. 31–56). Oxford: Oxford University Press.

Kosem, I., Gantar, P., & Krek, S. (2013). Automation of Lexicographic Work: an Opportunity for Both Lexicographers and Crowd-sourcing. In I. Kosem,

J. Kallas, P. Gantar, S. Krek, M. Langemets & M. Tuulik (Eds.), *Electronic Lexicography in the 21st century: Thinking Outside the Paper, Proceedings of the eLex 2013 Conference, 17–19 October 2013, Tallinn, Estonia* (pp. 32–48). Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut.

Kosem, I., Husak, M., & McCarthy, D. (2011). GDEX For Slovene. In I. Kosem & K. Kosem (Eds.), *Electronic Lexicography in the 21st century: New Applications for New Users, Proceedings of eLex 2011, 10–12 November 2011, Bled, Slovenia* (pp. 150–159). Ljubljana: Trojina, Institute for Applied Slovene Studies.

Kosem, I., Krek, S., Gantar, P., Arhar Holdt, Š., Čibej, J., & Laskowski, C. (2018). Collocations Dictionary of Modern Slovene. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (Eds.), *Proceedings of the 18th EURALEX International Congress: Lexicography in Global Contexts, 17–21 July 2018, Ljubljana* (pp. 989–997). Ljubljana: Ljubljana University Press, Faculty of Arts.

Kosem, I., Gantar, P., Krek, S., Arhar Holdt, Š., Čibej, J., Laskowski, C., Pori, E., Klemenc, B., Dobrovoljc, K., Gorjanc, V., & Ljubešić, N. (2019). *Collocations Dictionary of Modern Slovene KSSS 1.0*. Ljubljana: Slovenian Language Resource Repository CLARIN.SI. Retrieved from http://hdl.handle.net/11356/1250 (26. 8. 2021)

Krek, S. (2012). New Slovene Sketch Grammar for Automatic Extraction of Lexical Data: Presentation given at SKEW3, Brno, Czech Republic, 21–22 March 2012. Retrieved from https://trac.sketchengine.co.uk/attachment/wiki/SKEW-3/Program/Krek_SKEW-3.pdf?format=raw (26. 8. 2021)

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 (Ed.), *LREC 2020: Twelfth International Conference on Language Resources and Evaluation: May 11–16, 2020, Palais du Pharo, Marseille, France, Conference Proceedings* (pp. 3340–3345). Paris: ELRA – European Language Resources Association.

Levy, O., & Goldberg, Y. (2014). Neural Word Embedding as Implicit Matrix Factorization. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 27 (NIPS 2014)* (pp. 1–9).

Li, J., & Jurafsky, D. (2015). Do Multi-sense Embeddings Improve Natural Language Understanding?. In L. Màrquez, C. Callison-Burch & J. Su (Eds.), *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 1722–1732). Lisbon: Association for Computational Linguistics.

Liu, X., & Huang, D. (2017). Translation Oriented Sentence Level Collocation Identification and Extraction. In D. Wong & D. Xiong (Eds.), *Machine Translation, CWMT 2017: Communications in Computer and Information Science 787* (pp. 78–89). Singapore: Springer.

Ljubešić, N., & Erjavec, T. (2018). *Word Embeddings CLARIN.SI-embed.sl 1.0*. Ljubljana: Slovenian Language Resource Repository CLARIN.SI. Retrieved from http://hdl.handle.net/11356/1204 (26. 8. 2021)

Logar, 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.

Logar, N., Gantar, P., & Kosem, I. (2014). Collocations and Examples of Use: a Lexical-semantic Approach to Terminology. *Slovenščina 2.0*, *2*(1), 41–61.

Logar, N., & Erjavec, T. (2019). Slovene Academic Writing: a Corpus Approach to Lexical Analysis. In I. Simonnæs (Ed.), *New Challenges for Research on Language for Special Purposes: Selected Proceedings from the 21st LSP-Conference, 28–30 June 2017, Bergen, Norway* (pp. 205–217). Berlin: Frank & Timme.

Logar, N., Kosem, I., & Erjavec, T. (2019). *Collocation Lexicon of Slovene Academic Discourse Aleks*. Ljubljana: Slovenian Language Resource Repository CLARIN.SI. Retrieved from http://hdl.handle.net/11356/1245 (26. 8. 2021)

Manning, C. D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing, Chap. 5: Collocations*. Cambridge, Massachusetts: The MIT Press.

Mel'cuk, I. (1996). Lexical Functions: a Tool for the Description of Lexical Relations in a Lexicon. *Lexical Functions in Lexicography and Natural Language Processing, 31*, 37–102.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. Retrieved from https://arxiv.org/abs/1301.3781 (26. 8. 2021)

Pecina, P. (2009). Lexical Association Measures and Collocation extraction. *Language Resources and Evaluation*, *44*(1–2), 137–158.

Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Rayson, P., & Garside, R. (2000). Comparing Corpora using Frequency Profiling. In *WCC'00, Proceedings of the Workshop on Comparing Corpora, 9*, 1–6.

Rodríguez-Fernández, S., Carlini, R., Espinosa Anke, L., & Wanner, L. (2016a). Example-based Acquisition of Fine-grained Collocation Resources. In N. Calzolari et al. (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* (pp. 2317–2322). Portorož: ELRA.

Rodríguez-Fernández, S., Carlini, R., Espinosa Anke, L., & Wanner, L. (2016b). Semantics-driven Recognition of Collocations Using Word Embeddings. In K. Erk & N. A. Smith (Eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 499–505). Berlin: Association for Computational Linguistics.

Rundell, M., & Kilgarriff, A. (2011). Automating the Creation of Dictionaries: Where Will It All End?. In F. Meunier, G. Gilquin & M. Paquot (Eds.), *A Taste for Corpora: in Honour of Sylviane Granger* (pp. 257–282). Amsterdam: John Benjamins.

Rychlý, P. (2008). A Lexicographer-Friendly Association Score. In P. Sojka & A. Horák (Eds.), *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2008* (pp. 6–9). Brno: Masaryk University.

Singleton, D. (2000). *Language and the Lexicon: an Introduction*. New York: Oxford University Press.

Wanner, L., Ferraro, G., & Moreno, P. (2017). Towards Distributional Semantics-Based Classification of Collocations for Collocation Dictionaries. *International Journal of Lexicography*, *30*(2), 167–186.

Wiechmann, D. (2008). On the Computation of Collostruction Strength. *Corpus Linguistics and Linguistic Theory, 42*, 253–290.

RAZVRŠČANJE KOLOKATORJEV V SEZNAM: POGOSTOST *PROTI* SEMANTIKI

Kolokacije imajo v opisu jezika zelo pomembno vlogo. Še zlasti to velja za prepoznavanje pomena besed. Zato so postali v moderni leksikografiji neobhoden del pomenske členitve prav seznami kolokatorjev, razvrščeni po eni od statističnih mer povezovalnosti. Prispevek prikazuje primerjavo med dvema pristopoma k razvrščanju kolokatorjev: (a) metodo logDice, ki je zelo uveljavljena in temelji na pogostosti, ter (b) metodo besednih vložitev, ki je nova in temelji na strojnem učenju ter besedni semantiki. Primerjavo med rezultati obeh pristopov smo naredili na dveh zbirkah podatkov za slovenščino, eno z iztočnicami in njihovimi kolokacijami iz splošnega jezika, drugo z iztočnicami in njihovimi kolokacijami iz strokovno-znanstvenega jezika. Pri ocenjevanju rezultatov smo uporabili dve metodi: v kvantitativnem delu preizkusa smo izvedli nadzorovano strojno učenje z AUC ROC evalvacijo algoritma podpornih vektorjev (SVM); v kvalitativnem delu pa so rezultate obeh pristopov k razvrščanju kolokatorjev ocenili še leksikografi. Ugotovitve niso enoznačne; medtem ko je kvantitativno ocenjevanje pokazalo, da je pristop s strojnim učenjem in semantično razpršenostjo dal boljše razvrstitve kolokatorjev kot pristop, ki izhaja iz pogostosti, pa so leksikografi večinoma ocenili, da so seznami kolokatorjev obeh pristopov med sabo zelo podobni.

Ključne besede: kolokacije, besedne vložitve, logDice, splošni jezik, strokovno-znanstveni jezik