A General Framework for Detecting Metaphorical Collocations

Marija Brkić Bakarić **Lucia Načinović Prskalo** **Maja Popović**

Faculy of Informatics and Digital Technologies, University of Rijeka
Radmile Matejcic 2, 51000 Rijeka, Croatia
{mbrkic, lnacinovic}@uniri.hr

ADAPT Centre, School of
Computing Dublin City University,
Ireland
maja.popovic@adaptcentre.ie

Abstract

This paper aims at identifying a specific set of collocations known under the term metaphorical collocations. In this type of collocations, a semantic shift has taken place in one of the components. Since the appropriate gold standard needs to be compiled prior to any serious endeavour to extract metaphorical collocations automatically, this paper first presents the steps taken to compile it, and then establishes appropriate evaluation framework. The process of compiling the gold standard is illustrated on one of the most frequent Croatian nouns, which resulted in the preliminary relation significance set. With the aim to investigate the possibility of facilitating the process, frequency, logDice, relation, and pretrained word embeddings are used as features in the classification task conducted on the logDice-based word sketch relation lists. Preliminary results are presented.

Keywords: metaphorical collocations, classification, gold standard, significant relations, evaluation framework

1. Introduction

This paper is concerned with defining a framework for detecting metaphorical collocations. Since manually annotating corpus is extremely time-consuming and tedious, a combination of computational-linguistic and theoretical-semantic approaches is applied. The aim is to explore different patterns involved in the formation of metaphorical collocations in Croatian and discover possibilities of their automatic extraction. The final goal of this research is to create multilingual inventories of metaphorical collocations extracted from comparable corpora.

In generic terms, collocations imply awareness of common, conventional use. Metaphorical collocations form a very specific subset of lexical collocations. They are interesting in terms of cross-language comparison, since an in-depth analysis might provide universal formation patterns.

In metaphorical collocations, the base, which is usually a noun, retains its basic meaning. The collocate, on the other hand, is used in its secondary meaning, which is a consequence of the lexicalized (not spontaneous, vanished) metaphor (Stojić & Košuta, 2021a). Idiosyncrasy that is present with collocations is even more present in the case of metaphorical collocations. If we compare equivalents in Croatian, English, and German regarding the concept of a "long-time bachelor", it is evident that the collocates are represented by different images, i.e., "time" in English, "bark" in Croatian (*okorjeli neženja*), and "carved in flesh" in German (*eingefleischther Junggeselle*). In English, a temporal dimension is present. In Croatian and German, on the other hand, a spatial dimension can be observed, i.e., its properties of thickness and depth, respectively (Geld & Stanojević, 2018). The same extra-linguistic reality is lexicalized in different ways, thus indicating arbitrariness. However, the lexicalization is driven by a metaphorical mechanism in both cases. This leads to a conclusion that the process of making a relation between the base and its collocate might be following the same pattern. In this paper we focus on the Croatian language formation patterns.

Manual or semi-automatic compilation of language resources is extremely time-demanding, and thus expensive. Each time a method is modified, or a new method is tested, a new round of evaluation has to be performed, resulting in a huge waste of resources.

This paper presents an approach to developing the gold standard of metaphorical collocations. The approach is described in detail in section 2, in which a general evaluation framework is also proposed. Section 3 describes the subset of the gold standard involving the most frequent noun in the Croatian language. The related work on the existing collocation extraction studies, with a particular focus on Croatian is presented in Section 4. Section 5 presents some preliminary results obtained by approaching the task as a classification task. Concluding remarks are given in the final section of this paper.

2. Framework

Prior research has shown that nouns usually form the base of metaphorical collocations and that they retain their meaning, while the change in meaning usually manifests itself in the collocate. Due to that, we first compile the list of the most frequent nouns. The manual processing is therefore done in order of frequency (Stojić & Košuta, 2021b). The procedure proposed for compiling the gold standard can be outlined by the following steps:

1. Precise specification of the task
2. Selection of a suitable source corpus
3. Profiling
    a. Establishing the collocation profile of the most frequent noun based on a selected metric
    b. Exhaustive search
4. Determining fertile grammatical relations.

After the selection of a suitable source corpus, the collocation profile of the most frequent noun is established, and fertile grammatical relations are determined based on an exhaustive search.

The semantic analysis of the collocates performed in the second phase of the third step gives insight into semantic shifts and reveals language formation patterns in the language of interest, which might eventually lead to accepting the hypothesis about the universality of the process.

Steps 3-4 are repeated until a predefined number of nouns has been processed, each time taking into account the next

Proceedings of the 18th Workshop on Multiword Expressions (MWE 2022) n@LREC2022, pages 3–8

most frequent noun. If convergence has not been reached, the predefined number of nouns is enlarged. The point of convergence is reached when there are no new grammatical relations added to the list of fertile relations. Since we aim at doing a cross-language comparison, as a follow-up, steps 2-4 are conducted separately for each language. In our case, these are defined on the basis of available linguists employed for the task, and include English, Croatian, German, and Italian. However, step 3a is adapted to allow for direct comparisons. The list of the most frequent nouns is therefore taken to be the intersection of the nouns that appear in all four lists. The rank is determined by our base language, which is taken to be Croatian, but the nouns found in these lists are mostly the same, with minor differences in their respective ranks. In this paper, the presented results are limited to the most frequent Croatian noun *godina* ("year") for which the required output from the linguists has been obtained.

The output of the procedure described above is a list of metaphorical collocations, which will be used as our gold standard in evaluating different automatic extraction methods. Under the limitations set by our gold standard, beside a potential linguistic filter, we introduce additional constraint related to filtering the obtained candidate lists based on the available, i.e., processed, nouns which represent the nodes or the base words of the metaphorical collocations. This will allow us to compute precision and recall. As an additional verification step, which is also used for enlarging the gold standard with new base words and their collocates, we propose extracting the list of candidates not found in the manually processed lists and asking linguists to check for metaphorical collocations. If new collocates are determined, they are added to the gold standard, and the evaluation procedure is re-run. This is done to make the gold standard unbiased towards the measure used for the preliminary extraction.

From the joint discussions in which linguistic experts for all four languages participated, it could be concluded that the task of determining metaphorical collocations is quite subjective. Therefore, the experts held several discussion sessions prior to performing analysis and compiling the final list of metaphorical collocations per each language, up until they felt confident enough that they could differentiate between different types of collocations and thus extract metaphorical collocations. Two linguists per language participated in the task and the final lists comprise only collocations for which both linguists agreed to be metaphorical.

3. Processing the most frequent Croatian noun

In this section we provide details on the procedure applied in analysing the most frequent Croatian noun.

3.1 Corpus

Since our base language for exploring different patterns involved in the formation of metaphorical collocations is Croatian, the first corpus we process is the Croatian Web Corpus (Ljubešić & Erjavec, 2011), which consists of texts collected from the Internet and contains over 1.2 billion words. The hrWaC corpus is PoS tagged with MULTEXT-East Croatian POS tagset version 5 (Erjavec & Ljubešić, 2016). Considering the source of the corpus, it comes as no surprise that misspellings or non-standard language variants are infiltrated into the word sketch results. Additionally, due to the statistical nature of the tools employed in the pre-processing phase, there are also cases of incorrect lemmas and incorrect part-of-speech (POS) tags.

3.2 Measure

A measure used for identifying collocations (step 3a that is concerned with establishing the collocation profile of the most frequent noun) that is used in this research is the measure logDice implemented in Sketch Engine[1]. More details about logDice can be found in (Rychlý, 2008), and about its Sketch Engine implementation in (Kilgarriff et al., 2015). It is based on the frequencies of the base word and its collocate, and on the frequency of the whole collocation (co-occurrence of the base and the collocate). Since logDice is not affected by the size of the corpus, it can be used to compare scores between different corpora. The equation for calculating the logDice score is given in (1).

$$logDice\left(w_1, R, w_2\right) = 14 + \log_2 \frac{2 \times |\left|w_1, R, w_2\right||)}{\left|\left|w_1, R, *\right|\right| + \left|\left|*, R, w_2\right|\right|} \quad (1)$$

3.3 Relations

Sketch Engine relies on the language-dependent pattern matching grammars defined within the system that allow the system to automatically identify possible relations of words to the keyword, in our case *godina*. This makes the relations highly likely to contain false positives, but also to miss some collocations. However, for the purpose of this research, we find all these issues to be minor, as the candidate lists undergo additional inspection by linguists. For the word *godina,* Sketch Engine generates a total of 21 grammatical relations: *kakav?, oba_u_genitivu, u_genitivu_n, a-koga-čega, n-koga-čega, koga-što, particip, prijedlog, infinitive, koga-čega, s_prilogom, a-koga-što, a-komu-čemu, komu-čemu, glagol_ispred_prijedloga, prijedlog-iza, veznik, koordinacija, imenica_iza_prijedloga, biti_kakav?* and *subjekt_od*. There are 1,747 unique collocates dispersed over different grammatical relations, out of a total of 5,019 collocation candidates. Since the focus of this research are lexical collocations, only those grammatical relations with auto-semantical lexemes are considered relevant, i.e., ***kakav?*** (descriptive)*, oba_u_genitivu* (an adjective and a noun both in genitive)*, u_genitivu-n* (a noun in genitive)*,* ***n-koga-čega*** (two nouns—one in genitive)*, a-koga-čega* (an adjective in nominative and a noun in genitive)*,* ***koga-što*** (accusative)*,* ***subjekt_od*** (subject of)*,* ***particip*** (participle)*,* ***biti_kakav?*** (be like what). Exact rules for the listed relations can be found in Sketch Engine. Approximate descriptions are given in brackets. The relations shown in bold are taken to form the final significance set (Stojić & Košuta, 2021b), as elaborated in more detail in the upcoming subsection.

[1] SketchEngine (https://www.sketchengine.eu/)

3.4 Annotation

During the annotation task, the annotators process relations one by one, by analysing the obtained collocations and, if necessary, corpus examples of its use (Stojić & Košuta, 2021b). They label whether a candidate is a collocation, and additionally, whether it is a metaphorical collocation. There is an additional field in which the annotators can leave comments. That field is mostly used for trying to distinguish between different concepts and processes involved in the formation of metaphorical collocations, such as terms, metonymy, lexicalized metaphor, and personification. Over 80% of the metaphorical collocations belonging to the relation *subject_od* are labelled as personification. Regarding the relation *n-koga-čega*, there is approximately equal ratio between terms and metaphors, with the number of terms slightly superior. The relations such as *kakav?*, *koga-što*, *particip*, and *biti_kakav* have over 60% of metaphorical collocations labelled as resulting from the metaphorization process. The relation *kakav* comprises also a substantial number of terms.

The total number of candidates processed is 673. Among these candidates, there are 202 collocations, while 194 of these collocations are labelled as metaphorical collocations. Around 25% of the collocations in the relations *kakav?* and *biti_kakav* overlap. Moreover, almost 100% of the collocations in the relations *kakav?* and *oba_u_genitivu* overlap, which is why the latter is excluded from the final relation significance set. In the relation *u_genitivu-n* the keyword is a collocate and not the base, so it is considered irrelevant. The relation *a-koga-čega* is also irrelevant because it does not reflect collocations but independent lexemes. Furthermore, there are 25 metaphorical collocations detected by chance while examining contexts in the relation *biti_kakav*[2]. The detailed statistics is shown in Table 1. The extracted significance set of relations consists of patterns comprising the base, which is a noun, and another noun (N), an adjective (A), or a verb (V). However, scatterplots show no discernible patterns which could be used for the identification of metaphorical collocations neither on the basis of their logDice scores nor on the basis of the collocation frequency.

Relation	# of cands	# of colls	# of m_colls	Ratio of m_colls
kakav?	99	54	54	55%
n-koga-čega	100	41	38	41%
koga-što	100	41	41	41%
particip	100	16	11	11%
subjekt_od	100	30	30	30%
biti_kakav?	74	20	20	55%
Total	673	202	194	29%

Table 1: The annotated dataset

4. Related work

To our knowledge, there are no studies on the extraction of metaphorical collocations. In this section we, therefore, tackle recent work on the extraction of collocations in general, and the related work for the language involved, namely Croatian.

The most extensive empirical evaluation which includes 84 automatic collocation extraction methods can be found in (Pecina, 2005). Another comprehensive evaluation of lexical association measures (AMs) and their combination is presented in (Pecina, 2010). Linear logistic regression, linear discriminant analysis, support vector machines and neural networks are used to learn a ranker based on 82 association scores and all perform better than the individual AMs. Principal component analysis shows that the number of model variables can be significantly reduced.

A more recent study covering 13 corpora, eight context sizes, four frequency thresholds, and 20 AMs against two different gold standards of lexical collocations is presented in (Evert et al., 2017). The results show that the optimal choice of an AM depends strongly on the particular gold standard used. With respect to the corpora, larger corpora of the same kind perform better, which is in line with the positive effects observed by (Pecina, 2010). However, the authors in (Evert et al., 2017) acknowledge that clean, balanced corpora are better than large, messy Web corpora of the same size. Additionally, they find that even measures that highly correlate sometimes achieve substantially different evaluation results.

Recently, approaches based on word embeddings have started to gain popularity. A comparison between a supervised machine learning approach and a heuristic-based approach is presented in (Ljubešić et al., 2021). Regarding the rankings of collocates, a supervised machine-learning approach produces more relevant results than the approach based on heuristics. Furthermore, the word embeddings approach, which encodes distributional semantics of words, is a more useful source of information for the ranking of candidates than logDice, which encodes frequency information. An approach for identifying candidates of monolingual collocations using syntactic dependencies followed by the process of creating bilingual word-embeddings and a strategy for discovering collocation equivalents between languages is shown in (Garcia et al., 2017). A distributional semantics-based model that classifies collocations with respect to broad semantic categories is proposed in (Wanner et al., 2017).

As far as Croatian is concerned, there are several papers dealing with collocation extraction in general. For example, (Petrovic et al., 2006) explore four different association measures (PMI, Dice coefficient, Chi-squared test and Log-likelihood ratio) on Croatian legal texts. They use a linguistic filter and take into account AN and NN for bigrams and ANN, AAN, NAN, NNN, NXN for trigrams, where A stands for adjectives, N for nouns, and X for others. The results show that PMI measure performs the best.

A language and collocation type independent genetic programming approach for evolving new association measures is presented in (Šnajder et al., 2008). An evolved measure performs at least as good as any AM included in the initial population. Most of the best evolved AMs take into account the POS information.

(Seljan & Gašpar, 2009) conduct automatic term and collocation extraction based on the parallel English-

[2] Duplicate candidates are excluded from the figures in Table 1.

Croatian corpus of legal texts using two statistically based tools and applying a post-processing linguistic filter. The frequency of syntactic patterns in the automatically obtained lists is in agreement with the manually compiled, and contains AN, NN and NPN

Authors in (Karan, Šnajder and Bašić 2012) are the first one to treat collocation extraction in Croatian as a classification problem. They apply several classification algorithms including decision trees, rule induction, Naive Bayes, neural networks, and Support Vector Machines (SVM). Features classes used include word frequencies, AMs (Dice, PMI, $\chi2$), and POS tags. SVM classifier performs the best on bigrams and the decision tree on trigrams. The features that contribute most to the overall performance are PMI, semantic relatedness, and features representing a subset of POS tags. Experiments are conducted on a manually annotated set of bigrams and trigrams sampled from a newspaper corpus. The results of F1 measure go up to 80%.

In (Hudeček & Mihaljević, 2020), collocation extraction is based on the use of the Sketch Engine Word Sketch tool on the Croatian Web Repository Online Corpus and Croatian Web Corpus corpora. The results are filtered to include only frequent collocations with a typical syntactic construction.

Similarly, in this research we start with the word sketches generated by Sketch Engine. Next, we analyse the performance of the selected classification algorithms in the task of making the resulting candidate list more meaningful. By applying a classifier to the resulting candidate lists, we can facilitate the process of manual analysis.

5. Preliminary results

Naïve Bayes (NB) is extremely fast classification algorithm and has shown to work quite well in some real-world situations despite its oversimplifying assumptions (Witten et al., 2017). Hence, we take it to be our baseline and compare it to a tree based C4.5, and to more complex Support Vector Machines (SVM) and MultiLayer Perceptron (MLP).

C4.5 algorithm (Quinlan, 1993) is a descendant of ID3. It is a classification algorithm in the form of decision tree in which a splitting criterion known as the gain ratio is used. Decision nodes specify tests carried out at individual attribute values, and contain one branch for each possible outcome, while leaf nodes indicate class. An instance is classified by starting at the root of the tree and moving downwards until a leaf is encountered.

The kernel-based SVMs (Vapnik, 1995) are among the most popular models in Natural Language Processing applications. SVMs capture all features and their interdependencies. In this paper we use the sequential minimal optimization algorithm for training a support vector classifier using polynomial kernels.

MLP is a classifier based on artificial neutral networks. We experiment with several configurations and present results obtained with three hidden layers with 5, 10, and 20 neurons, respectively, and with the learning rate set to 0.3, momentum rate to 0.2, the number of training epochs to 500, and the number of consecutive increases of error allowed for validation testing before training terminates to the value of 20.

In this paper the labelled instances are obtained by manually annotating word sketches from Sketch Engine. Each instance is represented by a vector of feature values. We perform experiments on two sets of features. The first one contains collocation frequency, logDice, and relation (f=3). The second one additionally contains pretrained word embeddings of collocates (Grave et al., 2018), making a total of 303 features (f=303). Word embeddings are added to capture both the semantic and syntactic meanings of words, since they are trained on large datasets. At this point, we do not take into account the word embedding of the base word *godina*, as no other base words have been processed.

Prior to running classification, we pre-process our dataset. We remove instances that do not have valid lemmas due to lemmatization errors. Additionally, we remove duplicate lemmas that are found across several grammatical relations and keep instances with the highest frequencies. However, we do this separately for positively and for negatively labelled instances.

We set 42 as the seed value for the random number generator and run a stratified 10-fold cross validation repeated 10 times. We test whether different algorithms perform significantly better or worse when the feature set is expanded with word embeddings.

Precision (the share of correctly classified positive instances among all positive instances in the system output) and recall (the share of correctly identified positive instances among all instances that should have been identified as positive) are used to evaluate the classification. We also report F-measure scores. Recall results are given in Table 2, precision scores in Table 3, and F-measure scores in Table 4.

When only three features are taken into account, NB is the best performing algorithm at 5% significance level regarding recall, and the worst regarding precision. At the same time, its recall score is severely affected by expanding the feature set by word embeddings. For the other three classifiers, there are no statistically significant differences between their individual recall scores on the two feature sets. The difference in the recall and precision scores between SVM and MLP with f=303 is statistically significant. Regarding F-measure, no statistically significant differences can be observed between the four algorithms when f=3. However, when f=303, NB is outperformed by the other three algorithms.

Recall	*f*=3	*f*=303
NB	**0.94***	0.40
C4.5	0.81	0.79
SVM	0.78	**0.80***
MLP	0.80	0.76

Table 2: Recall of the selected algorithms

Precision	*f*=3	*f*=303
NB	0.68	0.72
C4.5	0.73	0.77
SVM	0.74	0.74
MLP	0.75	**0.78***

Table 3: Precision of the selected algorithms

F-measure	*f*=3	*f*=303
NB	**0.78**	0.50
C4.5	0.77	**0.78**
SVM	0.76	0.77
MLP	0.77	0.77

Table 4: F-measure of the selected algorithms

If we take into account the fact that metaphorical collocations for the Croatian headword *year* ("year") account for barely 30% of the candidate list obtained through Sketch Engine based on the logDice score, we find these preliminary results promising. However, our current dataset only contains collocates of the most frequent noun. In what way these results will be affected when we expand the dataset remains to be seen.

6. Conclusion

Association measures such as logDice rely exclusively on co-occurrence statistics, which is hardly enough for collocations in the broad meaning, let alone for the subtype of metaphorical collocations. This work is done with the aim to determine a way to encode the relation that refers to collocates contributing the semantic feature to their respective base words, i.e., a metaphor.

In this research we propose a procedure for compiling the gold standard of metaphorical collocations and establish the general evaluation framework for our future work.

Manual processing of the base words and their candidate lists of collocates is extremely time-demanding. Up to this point, linguists have only completed the processing of the most frequent Croatian noun *godina*. Therefore, this paper presents work in progress. The analysis performed is done using the Word Sketch function of the Sketch Engine,, which is based on the logDice score. Through the analysis, six significant grammatical relations are determined. The final relation significance set might be updated as new base words and their collocates are added to the gold standard. The compilation of the gold standard will be performed for a predefined number of base words under the condition that the final relation significance set reached convergence. The relation significance set will allow us to introduce a meaningful linguistic filter to different extraction methods, either as a pre-processing or a post-processing step.

From the experiment presented in this paper, it is evident that collocate embeddings strongly affect the performance of NB in most metrics. However, regarding the other three algorithms, statistically significant differences are obtained only in precision and recall scores between SVM and MLP with *f*=303.

In the follow-up we plan to test different AMs and machine learning algorithms in order to detect methods that are most helpful in automating the procedure of extracting metaphorical collocations. Comparison between different methods might be beneficial for other Slavic languages.

The final goal of this research is to create parallel inventories of metaphorical collocations that are extracted from comparable corpora in Croatian, German, English, and Italian. Due to unpredictability inherent in collocations in general, tasks such as machine translation would highly benefit from such lists.

7. Acknowledgements

This work has been fully supported by Croatian Science Foundation under the project "Metaphorical collocations – Syntagmatic word combinations between semantics and pragmatic*s*" (IP-2020-02-6319).

8. Bibliographical References

Erjavec, T., & Ljubešić, N. (2016). MULTEXT-East Morphosyntactic Specifications, Version 5. Http://Nl.Ijs.Si/ME/Vault/V5/Msd/Html/Msd-Hr.Html.

Evert, S., Uhrig, P., Bartsch, S., & Proisl, T. (2017). E-VIEW-alation-a Large-scale Evaluation Study of Association Measures for Collocation Identification. In *Proceedings of ELex*, pages 531–549.

Garcia, M., García-Salido, M., & Alonso-Ramos, M. (2017). Using bilingual word-embeddings for multilingual collocation extraction. In *Proceedings of the 13th Workshop on Multiword Expression*s, pages 21–30. http://universaldependencies.org/

Geld, R., & Stanojević, M. M. (2018). Strateško konstruiranje značenja riječju i slikom : Konceptualna motivacija u ovladavanju jezikom. Srednja Europa.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning Word Vectors for 157 Languages. In *Proceedings of the International Conference on Language Resources and Evaluation* (LREC 2018), pages 3483–3487. https://fasttext.cc/

Hudeček, L., & Mihaljević, M. (2020). Collocations in the Croatian Web Dictionary - Mrežnik. *Slovenščina 2.0: Empirical, Applied and Interdisciplinary Research*, 8(2), 78–111. https://doi.org/10.4312/slo2.0.2020.2.78-111

Karan, M., Šnajder, J., & Bašić, B. D. (2012). Evaluation of Classification Algorithms and Features for Collocation Extraction in Croatian. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 657–662. http://www.lrec-conf.org/proceedings/lrec2012/pdf/796_Paper.pdf

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., & Rychlý, P. (2015). Statistics used in the Sketch Engine. https://doi.org/10.1007/s40607

Ljubešić, N., & Erjavec, T. (2011). hrWaC and slWac: Compiling Web Corpora for Croatian and Slovene. In *International Conference on Text, Speech and Dialogue*, pages 395–402. https://doi.org/10.1007/978-3-642-23538-2_50

Ljubešić, N., Logar, N., & Kosem, I. (2021). Collocation ranking: frequency vs semantics. *Slovenscina 2.0*, 9(2), 41–70. https://doi.org/10.4312/slo2.0.2021.2.41-70

Pecina, P. (2005). An Extensive Empirical Study of Collocation Extraction Methods. In *Proceedings of the ACL Student Research Workshop*, pages 13–18.

Pecina, P. (2010). Lexical association measures and collocation extraction. *Language Resources and Evaluation*, 44(1–2), 137–158. https://doi.org/10.1007/s10579-009-9101-4

Petrovic, S., Snajder, J., Dalbelo Basic, B., & Kolar, M. (2006). Comparison of Collocation Extraction Measures for Document Indexing. *Journal of Computing and Information Technology*, 14(4), 321. https://doi.org/10.2498/cit.2006.04.08

Rychlý, P. (2008). A Lexicographer-Friendly Association Score. In *Recent Advances in Slavonic Natural Language Processing*, pages 6–9.

Seljan, S., & Gašpar, A. (2009). First Steps in Term and Collocation Extraction from English-Croatian Corpus. In *Proceedings of 8th International Conference on Terminology and Artificial Intelligence*.

Šnajder, J., Dalbelo Bašić, B., Petrović, S., & Sikirić, I. (2008). Evolving new lexical association measures using genetic programming. In *Proceedings of ACL-08: HLT,* pages 181–184.

Stojić, A., & Košuta, N. (2021a). Istraživanje metaforičkih kolokacija - teorijska osnova i prijedlog modela opisa. *Linguistica,* 61(1), 81–91. https://doi.org/10.4312/linguistica.61.1.81-91

Stojić, A., & Košuta, N. (2021b). Izrada inventara metaforičkih kolokacija u hrvatskome jeziku i njihova obrada sa semantičkoga i pragmatičkoga aspekta. Fluminensia, 34(1) (u rukopisu).

Vapnik, V. N. (1995). The Nature of Statistical Learning Theory. Springer Sceience + Business Media, LLC.

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. https://doi.org/10.1093/ijl/ecw002