UDK 81’322
81’373.612.2
Pregledni rad
Rukopis primljen 31. XII. 2021.
Prihvaćen za tisak 10. III. 2023.
https://doi.org/10.31724/rihjj.49.1.1

Marija Brkić Bakarić
Lucia Načinović Prskalo
Maja Matetić
Fakultet informatike i digitalnih tehnologija, Sveučilište u Rijeci
Radmile Matejčić 2, HR-51000 Rijeka
https://orcid.org/0000-0003-4079-4012
mbrkic@uniri.hr
https://orcid.org/0000-0002-8832-2527
lnacinovic@uniri.hr
https://orcid.org/0000-0003-4571-1546
majam@uniri.hr

INSIGHTS INTO AUTOMATIC EXTRACTION OF METAPHORICAL COLLOCATIONS

Collocations have been the subject of much scientific research over the years. The focus of this research is on a subset of collocations, namely metaphorical collocations. In metaphorical collocations, a semantic shift has taken place in one of the components, i.e., one of the components takes on a transferred meaning. The main goal of this paper is to review the existing literature and provide a systematic overview of the existing research on collocation extraction, as well as the overview of existing methods, measures, and resources. The existing research is classified according to the approach (statistical, hybrid, and distributional semantics) and presented in three separate sections. The insights gained from existing research serve as a first step in exploring the possibility of developing a method for automatic extraction of metaphorical collocations. The methods, tools, and resources that may prove useful for future work are highlighted.

1. Introduction

Multiword expressions (MWEs) or phraseological units constitute a significant part of the vocabulary of any language. The focus of this paper is on colloca-

tions as a category of phraseological units. Although there is no agreed definition of collocations, they are usually described as combinations of words whose components occur together in a short span of text more frequently than they would by chance. Seretan, Nerima and Wehrli (2004) assert that collocations imply awareness of common, conventional usage. They do not usually retain the meaning of some or even all of their components across languages, making them somewhat difficult for non-native speakers to learn. They are domain-dependent and language-dependent and differ in terms of length, syntactic patterns, and spacing between constituents (Thanopoulos et al. 2002).

In the absence of a universal definition, some studies use the term *collocations* in its broad sense, while others focus on a specific subtype of collocations. For example, Kita et al. (1994) refer to collocations as cohesive word clusters, which include idioms, frozen expressions, and compound words. Since the notion of typicality overlaps with stock phrases, technical terms, named entities, idioms, and supporting verb constructions, Antoch, Prchal and Sarda (2013) and Pecina (2010) take this even broader view. Strakatova et al. (2020) differentiate between semantically opaque idiomatic expressions and collocations that are not fully lexicalized. Some other authors such as Thanopoulos, Fakotakis and Kokkinakis (2002) place collocations in the subcategory of named entities.

A prerequisite to identifying and extracting collocations, both manually and automatically, is a clear definition of concepts and reliable tools and resources (Strakatova et al. 2020). Manually created language resources are reliable and seemingly error-free. However, their development is expensive and time-consuming. Therefore, many applications resort to the automatic extraction of collocation candidates.

In this paper, we focus on a subset of collocations known as metaphorical collocations, following Reder (2006). To the best of our knowledge, there are no studies on automatic extraction of metaphorical collocations to date. Metaphorical collocations belong to the category of lexical collocations. As Smadja (1993: 171) puts it, lexical collocations "roughly consist of syntagmatic affinities between open class words such as verbs, nouns, adjectives, and adverbs." Metaphorical collocations form a specific semantic subset of collocations in which the collocate is used figuratively, i.e., in its secondary meaning, which is a consequence of the lexicalized (not spontaneous, vanished) metaphor (Stojić and Košuta 2021).

Previous research has shown that the base, which is usually a noun, retains its basic meaning when collocated. The meaning of the collocate (usually verbs and adjectives) changes upon entering the relationship, resulting in polysemy, such as hr. *okorjeli neženja* 'long-time bachelor'. Strakatova et al. (2020) use the term *collocation* for such word pairs in which the meaning of the base is transparent, and the meaning of the collocate is not prototypical, whereas prototypical refers to the basic, most literal sense. The authors devise an annotation scheme presented in the form of a decision tree.

Many collocations are idiosyncratic, as pointed out by Lin (1998), in the sense that they are unpredictable by syntactic and semantic features. Idiosyncrasy is even more evident in different languages. This is especially true for metaphorical collocations. Take for example 'long-time bachelor', its Croatian equivalent *okorjeli neženja*, or its German equivalent *eingefleischter Junggeselle*. All three collocates are represented by different images – by 'time' in English, 'bark' in Croatian and 'carved in flesh' in German. However, in both Croatian and German the use of language is modelled on the basis of spatial dimension (through the property of thickness – *okorjeli neženja* – or depth – *eingefleischter Junggeselle*), but the same extra-linguistic reality is lexicalised in different ways, which seems to indicate arbitrariness. Nevertheless, the process of making a collocation compound seems to follow the same pattern using spatial dimension, i.e. the extra-linguistic reality was lexicalized in both languages by a metaphorical mechanism that motivated the meaning (Stojić and Košuta 2022). Patekar (2022) provides an extensive overview and discussion on the definition of metaphorical collocations and concludes with the definition we take on in this research. The author emphasises the need to differentiate between metaphorical expressions and metaphorical collocations. While none of the components of metaphorical expressions is used in its literal sense, in metaphorical collocations "the collocate is used figuratively, and the base literally, thus imbuing the collocation with metaphorical meaning" (Patekar 2022: 45).

As pointed out by Pecina (2010), the universally best method for extracting collocations does not exist. These tasks are highly dependent on the data, the language, and the notion of collocation itself. The main aim of this paper is to provide a systematic review of existing research on collocation extraction, existing methods, measures and resources in this field, with a particular interest in the

research involving Croatian. The motivation is to develop a procedure for automatic extraction of metaphorical collocations. The task of automatic extraction is challenging even for collocations, let alone the subset of metaphorical collocations. As Lin (1998) notes, collocations are recurrent but do not necessarily occur frequently in a corpus. This is especially true for metaphorical collocations. Following Walker (2011), we assume that much of the collocational behaviour of collocation components can be explained by the discovery of certain linguistic features that influence the way they are formed. The basic language of this study is Croatian, but the study also includes German, English and Italian. This opens the possibility to analyse the significant shifts in the established links across different languages. The main hypothesis we want to investigate is the existence of universal mechanisms in the formation of a large number of metaphorical collocations (across different languages). Although the choice of languages depends primarily on the availability of annotators, the inclusion of three different language families, namely Slavic (Croatian), Romance (Italian) and Germanic (English and German), adds strength to the potential conclusions regarding universal mechanisms.

In the remainder of this paper, we first list possible applications for the collocation extraction procedure and present related work on automatic collocation extraction. The section on related work is divided into the related work on association measures (AMs), hybrid approaches that consider linguistic information, and distributional semantics approaches. Next, we review available language tools and resources and discuss methods of interest. Finally, we outline plans and give suggestions for future work.

2. Possible applications

Church and Hanks (1990) were among the first to recognize the potential of automatic collocation identification for computational lexicography several decades ago. It remains true that one of the main benefits of collocation identification or extraction is to assist human lexicographers in compiling lexicographic information (e.g., identifying possible word senses, lexical preferences, usage examples, etc.). In addition to enriching traditional lexicons, it enables the crea-

tion of specialized collocation lexicons or even bilingual dictionaries and their use in translation studies. The translation of collocations is indeed not straightforward, since collocations differ from language to language.

The intended use of the collocation extraction method changed over time in line with advances in the field of natural language processing. It used to raise hopes in natural language generation and machine translation (Pearce 2001; Smadja 1993), but also in text simplification, since replacing collocates with simpler synonyms could greatly affect text flow (Pearce 2001). Ferret (2002) saw its potential in thematic segmentation or disambiguation of word sense. Regarding the latter, Walker (2011) has shown that it is possible to identify different meanings of a term based solely on different characteristic collocates. Regardless of all this, and perhaps most importantly, automatic extraction facilitates the process of manual annotation. The annotated datasets, on the other hand, can serve as material for data-driven approaches to collocation extraction and for various machine learning experiments, as noted by Strakatova et al. (2020).

3. Systematic literature review

In this section, we provide a chronological overview of the approaches from the existing literature, divided into three major lines of work. Existing research in the field of collocation extraction for Croatian is presented in a separate, fourth subsection.

The beginnings of collocation extraction are characterised by various efforts to find a suitable measure and to apply or adapt an existing measure. While some of these efforts make at least partial use of linguistic knowledge, others contain no linguistic information at all. In the last decade, distributional semantics models have gained the upper hand.

3.1. Statistical approaches

The first group of approaches to collocation extraction is based on checking typical collocation properties. These properties are formally described by math-

ematical formulas called association measures (AMs), which determine the degree of association between collocation components. The association value is calculated for each candidate collocation extracted from a corpus. It can be used for ranking or for a threshold-based classification (Pecina 2005).

Choueka, Klein and Neuwitz (1983) are among the first authors to deal with collocation extraction. Their approach is based solely on statistical and combinatorial properties of word distributions and is restricted to consecutive words only.

Church and Hanks (1990) propose a measure called *association ratio*, based on the concept of mutual information and restricted to two words that frequently occur together in a predefined window size. The measure differs from the mutual information measure in that it encodes a linear ranking and therefore rules out symmetry, meaning that the probability of word *x* and word *y* occurring together is not the same as the probability of word y and word x occurring together.

Dunning (1993) proposes a measure based on the *likelihood ratio*. It is defined as the logarithm of the ratio between the likelihoods of the hypotheses of dependence and independence, assuming that word occurrence follows a binomial distribution. The proposed measure is also suitable for text sets smaller than those required under the assumption of normal distribution.

The *cost criterion* proposed by Kita et al. (1994) refers to the processing cost for a sequence of words and quantitatively estimates the extent to which processing is reduced when the sequence is considered as a single unit. The approach relies heavily on absolute frequencies, and words with low frequencies have no chance of appearing as top candidates.

Smadja, Hatzivassiloglou and Mckeown (1996) use the *Dice coefficient* in iterative manner to find translations of source language collocations using parallel corpora. After finding individual words that highly correlate with a source language translation, these words are grouped into higher order combinations that can be labelled as rigid or flexible depending on the result of the corpus inspection. The Dice coefficient is calculated using maximum likelihood estimates for the conditional probabilities of one word appearing after a particular word.

Another purely statistical approach is presented by Shimohata, Sugio and Nagata (1997). Their approach is based on the idea that adjacent words are widely distributed when the string is meaningful, and localised when it is a substring

of a meaningful string. Therefore, the distribution of adjacent words is measured before and after the string, and the fragments are filtered using an *entropy* threshold.

The *mutual dependency (MD)* proposed by Thanopoulos, Fakotakis and Kokkinakis (2002) is based on the idea that dependency can be identified by subtracting information that the whole event carries from the Pointwise Mutual Information (PMI) score.

Walker (2011) extracts the most frequent collocates using *t-score* with respect to position and considering raw frequency. The initial set of base words is determined based on frequency. T-score is a statistical tool used to measure how much the distribution of something deviates from the norm. It reflects how frequently a particular combination occurs in the corpus. T-score exhibits strong bias towards frequency, which makes it incapable to identify rare collocations.

There is a line of research that is characterised by attempting to compare existing methods, and not combining or developing new methods (e.g., Krenn and Evert 2001; Thanopoulos et al. 2002; Pearce 2002; Pecina 2005). Pecina (2005) presents the most extensive empirical evaluation which includes 84 automatic collocation extraction methods.

Another line of research aims to combine different AMs. Any association measure can be used as a binary classifier by setting a threshold and treating phrases with scores above the threshold as collocations. Additionally, association measures can also be used as features for training classifiers.

Pecina (2005) proposes an approach that combines several basic methods and classifications. The presented study focuses on bigrams. The author uses logistic regression to evaluate subsets of attributes.

Pecina (2010) presents another comprehensive evaluation of lexical AMs and their combination. 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. Finally, the author also applies hierarchical clustering to obtain one representative for each cluster of highly correlated metrics, and then

progressively removes representatives when the resulting degradation in model performance is minimal.

Combining AMs using corresponding receiver operating characteristic (ROC) curves is presented by Antoch, Prchal and Sarda (2013). The authors combine representatives of clusters of equivalent AMs into more complex models to discover the "global collocation superclassifier." They treat collocations as successive words with a given meaning. Since the best-performing cluster includes representatives of statistical, linguistic, and information-theoretic AMs, the authors conclude that none of these theoretical approaches outperforms the others.

3.2. Hybrid approaches

While the above-referred research studies are concerned with finding appropriate measures for collocation detection and combining and evaluating them, the researchers described in this section consider one or more linguistic information or processes (e.g., part-of-speech tagging, stemming, parsing, etc.) in combination with the measures in automatic collocation detection.

Church and Hanks (1990) demonstrate the advantages of using part-of-speech (POS) tagged corpora. Soon after starting to use POS information in collocation extraction, various methods based on shallow or full parsing have been proposed.

Smadja (1993) proposes a three-step approach that allows for the integration of POS information. Two words co-occur if they occur in the same sentence and there are no more than 5 words between them, regardless of their order. The author extracts collocations of seemingly arbitrary length, i.e., from length 2 to 30. For each word in a sentence, collocates, position, POS tag, and frequency are recorded. The strength of the collocation is expressed by the z-score or the number of standard deviations from the mean (i.e., the difference between the frequency of a candidate collocation and the mean of the frequencies of all candidates divided by the standard deviation).

The first stage also uses the spread over positions or variance. The values of strength and spread are used in the subsequent filtering procedure. In that way bigrams with frequencies above a certain threshold that are used in a relatively

rigid way are extracted. In the second stage, the sentences containing the bigram are analysed, along with the distribution of words around the bigram and their POS tags, and again those above a certain threshold are retained, resulting in collocations with more than two words. All concordances in which two words occur within a certain distance and POS information are used for generating parses from which binary relations are extracted. Syntactic information in stage three is used for final filtering, thus increasing the accuracy from 40% to 80%. Dunning (1993) rebukes the use of the z-score as it significantly overestimates the importance of rare events.

Blaheta and Johnson (1997) use parsing and perform stemming to alleviate the ubiquitous problem of sparse data and focus only on verbs with particles.

Lin (1998) proposes using a parser to extract dependency triples from a corpus and applies mutual information for filtering.

Krenn (2000) applies a statistical POS tagger and a partial parser to extract collocation-specific syntactic constraints. Two models are evaluated – one using phrase entropy and another based on a lexicon using selected verbs as lexical keys.

Pearce (2001) presents an approach based on constraints on possible substitutions for synonyms within candidate phrases, using WordNet as a source of synonymic information. A pair of words is considered a collocation if one of the words clearly favours a particular lexical realisation of the concept that the other represents. The author proposes a corpus search. If the difference between the number of synonyms for a given word is above the threshold, a search of the World Wide Web and a dictionary is performed. A match in the number of occurrences indicates collocation. If dictionary information is missing, the pair is considered a potential collocation. The strength of collocation is expressed by the difference between the coincidence scores of the two most frequently occurring words with the word of interest. The score is normalised to the interval [0, 1] by dividing the result by the co-occurrence number of the most frequently occurring word.

Pearce (2002) calculates the difference between the probability of a phrase among all possible lexical realisations of the concept in terms of synonym sets of each word and the probability of those particular competing realisations in

the synonym set. The latter is computed by approximating the joint probability of independent trials with a maximum likelihood estimate. Any difference is converted into units of standard deviations.

Walker (2011) also advocates determining the collocation behaviour of a concept by comparing it to a synonym or close synonym.

Seretan, Nerima and Wehrli (2004) implement another hybrid approach. Syntactic analysis is used to select collocation candidates according to predefined collocation patterns, and the log-likelihood ratio (LLR) is applied to the sets of word co-occurrences (bigrams) obtained from the syntactic parser for each predefined pattern. The extracted bigrams are combined into larger n-grams to identify multi-word collocations. The method used is described in detail by Seretan, Nerima, and Wehrli (2003) and is particularly useful for identifying collocations whose terms are arbitrarily far apart due to syntactic processes.

The advantage that full parsing offers over a windowing method is highlighted by Seretan, Nerima, and Wehrli (2004) and Seretan and Wehrli (2006).

A collocation extraction system based on full parsing of source corpora, and therefore particularly suitable for finding collocation instances over long distances, is presented in Seretan and Wehrli (2009). Full parsing handles complex cases of extraposition such as passivisation, relativization, interrogation, apposition, etc., which are not handled by shallow parsing or window-based methods. The results show that even incorrect parsing leads to quality improvement.

Verma et al. (2016) propose general algorithms that are directional and work for n-grams of arbitrary order. The authors experiment with a POS unconstrained and a POS constrained variant, and use WordNet and the Web to check whether an n-gram is a collocation.

Bhalla and Klimcikova (2019) evaluate three collocation extraction tools (Sketch Engine, FLAX and Elia). The authors find that Elia, which uses dependency parsing, performs the best.

Garcia, Garcia Salido and Alonso-Ramos (2019) investigate the influence of 12 statistical measures on the automatic extraction of collocations in three different languages and find that the average performance of each association measure is similar irrespective of the language. Moreover, the results show that combining

dependency triples with raw frequency information performs equally well as the best AMs in most syntactic patterns and languages.

Strakatova et al. (2020) extract adjective-noun pairs with attributive dependency relation and tune thresholds for different AMs in order to perform a majority vote classification. Additionally, they use AMs as input features for a Support Vector Machine classifier and a feed forward neural network. The authors show that association measures either alone or combined are not able to detect collocations, possibly due to the fact that instances are biased towards logDice as that measure is used in the initial extraction.

3.3. Distributional semantics approaches

Distributional semantics is the leading approach to lexical meaning representation that has deeply changed in the last decades (Lenci et al. 2022). Distributional semantics represents lexical items with real-valued vectors that encode distribution. The vectors are commonly called embeddings. The traditional count models that build distributional vectors by recording co-occurrence frequencies were first replaced by the prediction models that learn vectors with shallow neural networks, followed by contextual models that use deep neural language models to generate contextualized vectors.

Static word embeddings such as those introduced by Mikolov et al. (2013) are based on the distribution of words in a language, but do not encode polysemy since all the possible word senses are represented by the same vector. Regardless, Strakatova et al. (2020) show that static embeddings detect both prototypical and non-prototypical meanings, as they tend to rely more on the noun. Capturing different meanings of words depending on context can be achieved by computing dynamic word representations conditioned on local context such as with BERT, which stands for Bidirectional Encoder Representations from Transformers (Devlin et al. 2019). BERT pre-trains deep bidirectional representations from unlabelled text by jointly conditioning on both left and right context in all layers. BERT follows a fine-tuning approach, applying pre-trained language representations to downstream tasks. Instead of using unidirectional language models to learn general language representations, BERT uses a masked language model pre-training objective and a next sentence prediction task that jointly pre-train

text-pair representations. Ljubešić and Lauc (2021) pre-train an Electra transformer language model BERTić on the texts crawled from the Croatian, Bosnian, Serbian, and Montenegrin web domains and present state-of-the-art results for four selected NLP tasks. Ulčar and Robnik-Šikonja (2020) pre-train CroSloEngual BERT, a trilingual model based on the Croatian, Slovenian and English corpora using the BERT architecture. Multilingual models such as BERT and CroSloEngual BERT enable cross-linguistic knowledge transfer among the languages on which they are trained.

A distributional semantics-based model that classifies collocations with respect to broad semantic categories is proposed by Wanner et al. (2017). 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 by Garcia et al. (2017). Strakatova et al. (2020) present a dataset of German adjective-noun collocations, which contains both positive and negative instances, and use it for evaluating different models in the task of automatic collocation identification. They experiment with different setups and show that static and contextualized word embeddings outperform the methods based on AMs. They also show that additional context and sense representations result in improvements but only for dynamic word embeddings. Ljubešić et al. (2021) show that the word embeddings approach, which encodes distributional semantics of words, is a more useful source of information for the ranking of candidates than logDice. Espinosa Anke, Codina-Filba and Wanner (Espinosa-Anke et al. 2021) examine language models for finding and categorising lexical collocations. The authors perform unsupervised collocation retrieval and supervised collocation classification in context.

The terms *non-compositional multi-word expression* (MWE) or *non-compositional phrases* may be to some extent related to the term *metaphorical collocation* if we consider the definition provided by Salehi et al. (2014): a combination of words with lexical, syntactic or semantic idiosyncrasy whose meaning is not predictable from the meaning of their constituents. Salehi et al. (2014) attempt to identify non-compositional MWE components using Wiktionary. The basic idea is that an expression is considered compositional if the definition of the expression contains the components of the expression. In the opposite case, the

expression is classified as non-compositional. In their later research, Salehi et al. (2015) use word embeddings to predict the compositionality of multi-word expressions. In predicting the non-compositionality of MWEs, they achieve better results with word embeddings than with traditional distributive vector representations. Yazdani et al. (2015) investigate distributive vector space models for semantic composition to detect non-compositionality for English noun compounds through unsupervised learning. The authors also evaluate the models and suggest additional methods to improve the results, such as using polynomial projection and enforcing sparsity. Hashimoto and Tsuruoka (2016) experiment with quantifying the compositionality level of phrases using the scoring function to adaptively weight compositional and non-compositional phrase embeddings. They propose a novel adaptive joint learning method for learning transitive verb phrase embeddings and verb-object compositionality.

3.4. Related work for the Croatian language

As far as the Croatian language is concerned, the number of researches on collocation extraction is modest, especially as far as metaphorical collocations are concerned, which are the focus of this research.

Interest in collocations as a specific linguistic phenomenon in Croatian linguistics dates back to the end of the 20th century (Ivir 1992; Borić 1998). However, collocations are tackled in terms of their definition, determination of collocation components, etc., and there is no mention of methods for extracting collocations automatically.

First experiments on collocation and terminology extraction are presented by Tadić and Šojat (2003). The authors show that PMI in combination with linguistic filters on non-lemmatized text yields poor results.

Petrović et al. (2006) compare four different AMs in the task of extracting collocations from Croatian legal texts for the purpose of document indexing and consider only bigrams and trigrams. The results show that PMI performs better than LLR, χ^2, and the Dice coefficient.

Seljan and Gašpar (2009) automatically extract terms and collocations from the English-Croatian corpus of legal texts using two statistically based tools –

MultiTerm Extract (MTE) and Lexterm (LT). The obtained results are further filtered using local regular grammars within the NooJ linguistic environment. The frequency of syntactic patterns in the lists obtained with both tools shows that the most frequently represented patterns are of the type AN, NN and NPN, where A stands for adjectives, N for nouns and P for prepositions.

Nine different co-occurrence measures for collocations in combination with POS filter and lemmatization are implemented in the tool TermeX (Delač et al. 2009).

Pinnis et al. (2012) present a workflow for the extraction of term candidates and for bilingual term mapping in comparable corpora. The candidates are filtered by a set of morphosyntactic patterns and a minimum frequency threshold, and then ranked using co-occurrence statistics. In the final phase a cut-off method is applied.

Karan, Šnajder and Bašić (Karan et al. 2012) evaluate classification algorithms and features in the task of collocation extraction for Croatian. They conduct a binary classification and use several classification algorithms including decision trees, rule induction, Naive Bayes, neural networks and Support Vector Machines (SVM). The features they use include word frequencies, AMs (Dice, PMI, χ^2), POS tags, and semantic word relatedness modelling in the form of latent semantic analysis. The authors conclude that the logistic regression classifier gives the best F1 score on bigrams and the decision tree on trigrams. The features that contribute most to the overall performance are PMI, semantic relatedness, and POS information.

Hudeček and Mihaljević (2020) describe the extraction of collocations for the Croatian web dictionary Mrežnik. Collocation extraction is based on the use of the Sketch Engine Word Sketch tool applied on the Croatian Web Repository Online Corpus and Croatian Web Corpus. The results are filtered on the basis of frequency and syntactic construction.

Ljubešić, Dobrovoljc, and Fišer (2015) use dependency syntactic patterns to identify MWE candidates in parse trees and build a resource called MWELex. The grammar they use is defined over morphosyntactic patterns, and then transformed to the corresponding dependency syntax level grammar. After extracting all the candidates from a parsed corpus, logDice is used for scoring and the

list is filtered based on frequency thresholds and morphological lexicons. The authors also inspect the possibility of applying the distributional approach for calculating semantic transparency of MWE candidates and obtain promising results.

Collocations have also been viewed in terms of the language of the profession, translation, and glottodidactics. For example, Miščin (2015) explores collocational competence of primary and secondary school students. Stojić and Košuta (2017) examine the method of using collocation connections in a foreign language. Blagus Bartolec (2017) analyses the frequency of verb collocations in administrative-functional texts of Croatian and discusses the possibilities of replacing verb collocations with a one-word verb. The focus of Šnjarić and Borucinsky (2020) is on the verb-noun collocations in scientific literature. The authors point out the lack of coverage of general scientific verb-noun collocations in existing general bilingual dictionaries with Croatian as the original language, which poses a difficulty for translators. Ordulj and Žauhar (2018) analyse the frequency and associative strength of 228 noun collocations in Croatian concluding that higher frequency collocations associate more strongly.

Metaphorical collocations are the focus of studies conducted by Brkić Bakarić, Načinović Prskalo and Popović (2022) and Načinović Prskalo and Brkić Bakarić (2022). Brkić Bakarić, Načinović Prskalo and Popović (2022) investigate the possibility of facilitating the creation of the gold standard by using frequency, logDice, relation, and pre-trained word embeddings as features in the classification task conducted on the logDice-based word sketch relation lists and present preliminary results for Croatian. A follow-up study by Načinović Prskalo and Brkić Bakarić (2022) extends the research to English and German.

4. Methods, measures, and resources

Since metaphorical collocations are the focus of this research, a corpus study is conducted to analyse metaphorical collocations in Croatian. The processing is based on the list of the most frequent Croatian nouns. In parallel, corpus research is conducted in three different languages (German, English and Italian) based on the list of translation equivalents of the most frequent Croatian nouns.

The aim of this method is not to identify the deviant structures, which previous research has already pointed out, but to create parallel inventories of metaphorical collocations in Croatian, German, English, and Italian. Besides suggesting the procedure for automatic extraction of metaphorical collocations, one of the main aims of future work is to investigate the hypothesis about the existence of universal mechanisms in the formation of a large number of metaphorical collocations.

Methods for extracting collocations are largely language-independent, but some language-specific tools are required for linguistic filtering of source corpora (e.g., POS taggers, lemmatizers, and syntactic parsers). Lemmatizers are used to detect all inflected forms of a lexical item, POS taggers are used to filter out specific word categories, while parsers are used to extract significant relations.

4.1. Lemmatization

Croatian is a morphologically extremely rich language, which is why morphological normalisation is performed as one of the preliminary steps in most procedures. The two basic procedures of morphological normalisation are stemming and lemmatization. Lemmatization involves finding a base form of a word, i.e., an entry form in a dictionary or a lemma. Stemming, on the other hand, is the process of removing affixes from different word forms to find a common stem for all forms. The use of lemmatization or stemming in automatic collocation extraction has its benefit in considering all word forms of a given collocational component. In all the above-mentioned researches on collocation extraction for Croatian, lemmatization was performed as one of the steps of corpus preprocessing. A lemmatizer and morphosyntactic tagger for Croatian is available at: https://reldi.spur.uzh.ch/blog/croatian-and-serbian-lemmatiser/ (Agić et al. 2013). A new lemmatizer for Croatian was published by Stanza as part of the CLASSLA fork for processing Slovenian, Croatian, Serbian, Macedonian, and Bulgarian (Ljubešić et al. 2019), available at: https://pypi.org/project/classla/. The reported F1 scores of both tools are very close to each other and are 98.

4.2. POS tagging

Another method used for corpus pre-processing is POS tagging. POS tagging refers to assigning adequate POS tags to words in a corpus. Given that some syntactic structures are more typical for the creation of collocations, it is clear that POS tagging plays an important role in the identification and extraction of collocations. According to Hudeček and Mihaljević (2020) and Stojić and Košuta (2022), four most typical syntactic structures of collocations in the Croatian web dictionary Mrežnik include verb + noun, adjective + noun, adverb + verb, and adverb + adjective.

One POS and morphosyntactic tagger for Croatian is available at: https://reldi.spur.uzh.ch/blog/tagger/ (Ljubešić et al. 2016). The morphosyntactic tagset used in the tagger is the revised MULTEXT-East V5 tagset for Croatian (Erjavec and Ljubešić 2016). Same as for lemmatizer, a new POS tagger is now available within the CLASSLA project (Ljubešić et al. 2019), available at https://pypi.org/project/classla/. The reported F1 score is 94.18. It uses MULTEXT-East V6 tagset (Erjavec 2019).

4.3. Parsing

Due to syntactic variability, in order to identify and extract collocations, one must take into account all syntactic contexts in which they can be realised, including long-distance dependencies (Seretan 2013). Parsing addresses this problem. This is especially important for languages with a freer word order, such as Croatian. Dependency grammars abstract from word order information and represent only the information necessary for parsing, which is beneficial for morphologically rich languages with free word order. Dependency parsing "describes the syntactic structure of a sentence in terms of directed binary grammatical relations between words" (Jurafsky and Martin 2009: chapter 15, 1). One dependency parser for Croatian is presented by Agić and Ljubešić (2015). A more recent tool for dependency parsing is also available as part of the CLASSLA project (Ljubešić et al. 2019), available at https://pypi.org/project/classla/.

4.4. Approaches

A variety of different AMs have been proposed to estimate lexical association based on corpus evidence. The main problem with such measures is that they are noisy and suffer from the problem of sparse data. They mostly come from mathematical statistics and range from those based on probabilities and linguistic contexts to purely heuristic ones. As the authors in (Thanopoulos et al. 2002) state, the simplest approach to collocation extraction would be to extract the most frequent word co-occurrences. However, since this does not take a priori word frequencies into account, the extracted sequences would be completely compositional and uninteresting.

Methods for selecting collocation candidates can be divided into window-based and syntactic methods. As Seretan and Wehrli (2006) point out, the former use the notion of linear proximity and the latter that of syntactic proximity. They also assert that linguistically uninformed methods are slower, less robust and less portable, as they do not need to pre-define syntactic configurations for collocation candidates. However, they need to be based on large corpora to achieve competitive results. In general, window-based systems cannot achieve the recall of parse-based systems because they do not recognise the "long-distance" pairs (Seretan and Wehrli 2006). More specifically, window-based systems yield more noise due to grammatically unrelated pairs within the collocation window. On the other hand, parser-based methods yield fewer but better pairs. However, the parser might also miss relevant pairs due to inherent analysis errors.

The related work research revealed numerous AMs applied to the task of collocation extraction. Roughly, they can be grouped into probability-based scores (e.g., PMI, MD, LLR, or statistical tests of independence such as Pearson's $\chi 2$, t-test, z-score), association coefficients (e.g., odds ratio, the Dice coefficient), and context measures (e.g., context entropy). In general, there is a trend of incorporating machine learning techniques and including long-distance relations, as well as incorporating static and contextualised word embeddings.

4.5. Evaluation

Since there is no general agreement on the definition of collocations, there is also no standard evaluation methodology. Collocations can be evaluated by a professional lexicographer, by native speakers, or by using a gold standard. Antoch, Prchal and Sarda (2013) claim that the use of a gold standard is crucial for empirical evaluation. Moreover, Krenn and Evert (2001) argue that a gold standard should be a reference set of collocations manually extracted from the full candidate data. Since a suitable resource is rarely available, some authors such as Thanopoulos, Fakotakis and Kokkinakis (2002) resort to WordNet. The authors acknowledge, however, that the use of WordNet is error-prone because it contains only the most frequently named entities and many WordNet entities are analytical descriptions of lexical entities rather than non-compositional multi-words of interest. In the absence of a suitable gold standard, Pearce (2002) performs an evaluation using multi-word information from a dictionary.

Performance can be measured by accuracy – the proportion of correct predictions or more commonly by precision – the proportion of correct positive predictions. Precision can be calculated using the manually identified true positives (TPs) extracted from the top of the significance list (Krenn and Evert 2001). However, this approach often suffers from low inter-annotator agreement. This is due to the fact that the notion of typicality overlaps with technical terms, proper names, idioms, etc. Even when annotators are instructed to count all these phenomena as collocations, the inter-annotator agreement is still quite low (Pecina 2010). Antoch, Prchal and Sarda (2013) also report a low inter-annotator agreement under the same scoring scheme. Only candidates recognized by all annotators as true collocations are usually included in the gold standard. With respect to the gold standard, Pearce (Pearce 2002) argues that evaluation with reference to a single standard is somewhat controversial, as there is no general agreement on the exact nature of collocations. The author concludes that it is necessary to evaluate against a set of gold standards, as well as to conduct native speaker and task-based evaluations.

Pecina (2010) uses the Average Precision (AP) measure. It can be defined as the expected value of precision for all possible values of recall (assuming a uniform distribution of recall), or even better, for recall in the interval from 10 to 90%.

The Mean Average Precision (MAP) can be defined as the mean value of average precision calculated for each data fold in the case of stratified cross-fold validation.

Visualizations usually include the percentage of correctly found pairs among the N best candidates. The baseline is often drawn by listing the candidates in the significant list in completely random order, or by ordering them according to the frequency of co-occurrence. The results can also be presented in different intervals defining the N-best percentage extraction lists. Seretan and Wehrli (2009) extract test sets of related pairs at four different levels of the significance list (1, 3, 5 and 10%).

When there is a gold standard, evaluation is usually done using precision-recall measures and comparing their precision-recall (PR) curves. By varying the threshold on the test scores, the PR curve plots classifier precision values on the y axis and classifier recall values on the x axis. When the threshold is low, every instance is labelled as positive, and the recall is 1. By increasing the threshold, the recall monotonically decreases as the number of true positives can either decrease or stay the same. Precision, on the other hand, can increase or decrease depending on the class of the instance that is included in the positive class by lowering the threshold. A thorough comparison of the precision-recall curves (PR) that visualise the quality of the classification – the higher up and to the right, the better – can circumvent the classification thresholds (Pecina 2010).

Pecina (2010) underlines the need for curve averaging since curves are only sample-based estimates of their true shape. Without precision-recall curves, MAP might not always be interpretable. Antoch, Prchal and Sarda (2013) use ROC curves. ROC curves visualise the probability of correctly classified entries versus the probability of misclassified entries for all thresholds of a selected evaluation measure.

4.6. Corpora

A corpus can be defined as "a large, principled collection of naturally occurring examples of language stored electronically" (Bennett 2010: 2). Since corpora provide rich models of language in terms of lexical, grammatical, morphologi-

cal, and semantic features, and collocation patterns, they are used by various groups of scholars such as linguists, social scientists, lexicographers, natural language processing experts, etc.

We start our research with the analysis of patterns of metaphorical collocations in the Croatian language. The corpus used for the task is the Croatian Web Corpus (Ljubešić and 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 and Ljubešić 2016). Although larger corpora may be available, such as cc100-hr (Conneau et al. 2020) or MaCoCu-hr 1.0 (Bañón et al., 2022) for Croatian, in this study the basis for manual annotation of metaphorical collocations are word sketches extracted from Sketch Engine (Kilgarriff et al. 2014), which constrains the selection of corpora.

The basis for the lexicalization of certain content is provided by perceiving extralinguistic reality through the mechanism of metaphorization and by establishing a link between language and reality through metaphor. The choice of metaphor used by a particular linguistic community is the result of cultural specificities, which lead to differences that become visible only in interlingual comparison (Stojić and Košuta 2020).

We assume that there are universal formation patterns in a large number of collocation relations. For this reason, in parallel to identifying metaphorical collocations in Croatian, we also identify metaphorical collocations in English, German and Italian. Therefore, we also use English, German, and Italian corpora. We opt for the English Web 2020 (enTenTen20), the German Web 2018 (deTenTen18) and the Italian Web 2016 (itTenTen16), as these are the largest, most comprehensive and up-to-date corpora for the respective languages available on Sketch Engine. They are composed of texts collected from the Internet. The corpus enTenTen20 is tagged with the TreeTagger tool using the English Web 2020 POS tagset, deTenTen18 is annotated with the RFTagger tool using the German RFTagger POS tagset, and itTenTen16 is annotated with the TreeTagger. Table 1 shows the total numbers of language units for each corpus.

Table 1. Total numbers of language units within each corpus

	hrWaC 2.2	enTenTen20	deTenTen18	itTenTen16
Tokens	1,405,794,913	44,968,996,152	6,382,147,542	5,864,495,700
Words	1,211,328,660	38,149,437,411	5,346,041,196	4,989,729,171
Sentences	67,403,219	2,099,033,556	342,730,929	227,944,684
Paragraphs	28,771,178	789,418,319	126,308,900	95,603,815
Documents	3,611,090	81,323,314	13,772,016	12,967,535

5. Discussion

Previous work has shown that the type of co-occurrence used for computing association measures has an impact on the quality of the collocation extraction and classification (Evert et al. 2017). Though feature representations based on AMs do not suffice for more semantically restricted classification task, as witnessed by Strakatova et al. (2020), AMs might be a good approach for extracting a number of collocation candidates.

This research study therefore starts with the collocate lists of the most frequent nouns obtained from lemmatized corpora, constrained by morphosyntactic patterns, and ranked by the logDice scores. Croatian is defined as the base language.

Since the ultimate goal is to extract and compare metaphorical collocations in Croatian, English, German, and Italian, it is worth noting that an AM that is suitable for a syntactic type in one language may be less suitable in other language, due to different lexical distribution (Seretan and Wehrli 2009). Moreover, Antoch, Prchal and Sarda (2013) point out the fact that different AMs detect different collocation types.

The choice of corpus is important. As Seretan and Wehrli (2006) point out, more data does not necessarily mean better results, for several reasons. First, it introduces more noise for the basic methods; second, some collocations systematically appear at large intervals when they favour passive constructions; and third, the overlooked cases affect the frequency profile of the discovered collocations.

Our future work involves a combination of two approaches – a computational linguistic approach and a pragmatic (theoretical-semantic) approach. It will be divided into two phases. The first phase will involve identifying the basic metaphorical collocations in four different languages and studying their composition. In the second phase, the translation equivalents will be identified and mapped to each other. The in-depth analysis will show whether all metaphorical collocations can be traced back to a motivating conceptual metaphor and whether a universal mechanism exists.

Overall, the methodological approach proposed for the first phase can be defined by the following four steps – a precise specification of the task, the selection of a suitable source corpus, the creation of a collocation profile for a selected set of high frequency nouns by determining fertile grammatical relations with respect to the selected corpus, and the comparison of the collocation behaviour of an item with the behaviour of its synonyms. The result of this first phase will be the gold standard that will be used in the evaluation of different automatic extraction procedures. The complete evaluation framework is presented by Brkić Bakarić, Načinović Prskalo and Popović (2022). The authors illustrate the process of compiling the gold standard on one of the most frequent Croatian nouns and present the preliminary relation significance set. The result of the second phase will be a list of translation equivalents.

The work presented in this and the related papers (Brkić Bakarić, Načinović Prskalo and Popović 2022; Načinović Prskalo and Brkić Bakarić 2022) is somewhat similar to that of Strakatova et al. (2020). Strakatova et al. (2020) also use a platform that gives word sketches and then build upon the knowledge about statistical properties of collocations in order to select a list of collocation candidates. Their approach differs in that they start with a list of adjectives and thus obtain co-occurring nouns for the subsequent manual annotation. While Brkić Bakarić, Načinović Prskalo and Popović (2022) conduct preliminary experiments with static word embeddings as additional features in the classification task, Strakatova et al. (2020) compare performance of non-linear classifiers when trained on static against dynamic word embeddings.

In general, the work presented in this paper is part of an ongoing project and there are many opportunities for future work. In addition to experimenting with different AMs and word embeddings as features in different classification con-

figurations, the identification of metaphorical collocations can be performed using CroSloEngual BERT (Ulčar and Robnik-Šikonja 2020) or BERTić (Ljubešić and Lauc 2021) trained on a dataset annotated with metaphorical collocations, i.e., using dynamic word embeddings.

6. Conclusion and future work

The focus of this paper is on metaphorical collocations, which form a special subset of collocations in which the collocate is used figuratively, i.e., in its secondary meaning. Since collocations are generally idiosyncratic, their automatic extraction poses a major challenge. Although there are many studies on automatic collocation extraction for widely used languages such as English, none of these studies deal with metaphorical collocations.

The main goal of this research is to provide a systematic literature review in the field of collocation extraction and to present existing methods, measures and resources. The motivation is to gain insight into the possibilities and make plans for the task of automatic extraction of metaphorical collocations.

The literature review is divided into three sections, depending on the approach used (statistical, hybrid, distributional semantics). A special section is devoted to the relevant works for Croatian. Overall, the conducted analysis shows a modest number of studies on the topic of automatic extraction of collocations for Croatian and none on the extraction of metaphorical collocations. Methods, tools and resources used in previous research and considered useful for future work are highlighted.

As for future work, there are two major research questions that will be addressed. The first relates to the study of whether there are universal mechanisms in the formation of a large number of collocation compounds, i.e. metaphorical collocations (in different languages). The second issue relates to investigating the performance of different methods for automatic extraction of metaphorical collocations and determining the best approach.

Acknowledgment

We would like to thank anonymous reviewers for their constructive comments and insightful suggestions that significantly improved the quality of the paper. This work has been fully supported by the Croatian Science Foundation under the project Metaphorical collocations – Syntagmatic word combinations between semantics and pragmatics (IP-2020-02-6319).

References

Agić, Željko; Ljubešić, Nikola. 2015. Universal Dependencies for Croatian (that Work for Serbian, too). *The 5th Workshop on Balto-Slavic Natural Language Processing*. INCOMA Ltd. Shoumen. 1–8. https://aclanthology.org/volumes/W15-53/.

Agić, Željko; Ljubešić, Nikola; Merkler, Danijela. 2013. Lemmatization and Morphosyntactic Tagging of Croatian and Serbiaň. *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing*. 48–57. https://aclanthology.org/W13-2400/.

Antoch, Jaromír; Prchal, Luboš; Sarda, Pascal. 2013. Combining Association Measures for Collocation Extraction Using Clustering of Receiver Operating Characteristic Curves. *Journal of Classification* 30/1. 100–123. https://doi.org/10.1007/s00357-013-9123-x.

Bañón, Marta; Esplà-Gomis, Miquel; Forcada, Mikel L.; García-Romero, Cristian; Kuzman, Taja; Ljubešić, Nikola; van Noord, Rik; Pla Sempere, Leopoldo; Ramírez-Sánchez, Gema; Rupnik, Peter; Suchomel, Vit; Toral, Antonio; van der Werff, Tobias; Zaragoza, Jaume. 2022. *Croatian web corpus MaCoCu-hr 1.0*. Slovenian Language Resource Repository. CLARIN.SI. http://hdl.handle.net/11356/1516.

Bennett, Gena R. 2010. *Using Corpora in the Language Learning Classroom: Corpus Linguistics for Teachers*. University of Michigan Press ELT. http://www.press.umich.edu/titleDetailDesc.do?id=371534.

Bhalla, Vishal; Klimcikova, Klara. 2019. Evaluation of automatic collocation extraction methods for language learning. *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics. Florence. 264–274. https://aclanthology.org/W19-4400/.

Blagus Bartolec, Goranka. 2017. Glagolske kolokacije u administrativnome funkcionalnom stilu. *Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje* 43/2. 285–309.

Blaheta, Don; Johnson, Mark. 1997. Unsupervised learning of multi-word verbs. *Proceedings of the 39th Annual Meeting of the ACL*. 54–60. https://aclanthology.org/volumes/P01-1/.

Borić, Neda. 1998. Semantički aspekt kolokacijskih odnosa s kontrastivnog stajališta. *Strani Jezici*. 27/2. 72–79.

Brkić Bakarić, Marija; Načinović Prskalo, Lucia; Popović, Maja. 2022. A General Framework for Detecting Metaphorical Collocations. *Proceedings of the 18th Workshop on Multiword Expressions*. Eds. Bhatia, Archna et al. European Language Resources Association. 3–8. https://aclanthology.org/2022.mwe-1.0/.

Choueka, Yaacov; Klein, Shmuel T.; Neuwitz, E. 1983. Automatic retrieval of frequent idiomatic and collocational expressions in a large corpus. *ALLC Journal* 4/1. 34–38.

Church, Kenneth Ward; Hanks, Patrick. 1990. Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics* 16/1. 22–29.

Conneau, Alexis; Khandelwal, Kartikay; Goyal, Naman; Chaudhary, Vishrav; Wenzek, Guillaume; Guzmán, Francisco; Grave, Edouard; Ott, Myle; Zettlemoyer, Luke; Stoyanov, Veselin. 2020. Unsupervised Cross-lingual Representation Learning at Scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics. 8440–8451.

Delač, Davor; Krleža, Zoran; Šnajder, Jan; Dalbelo Bašić, Bojana; Sarić, Frane. 2009. TermeX: A tool for collocation extraction. *Computational Linguistics and Intelligent Text Processing. CICLing 2009. Lecture Notes in Computer Science* 5449. Eds. Gelbukh, Alexander et al. Springer. Berlin – Heidelberg. 149–157. https://doi.org/10.1007/978-3-642-00382-0_12.

Devlin, Jacob; Chang, Ming-Wei; Lee, Kenton; Toutanova, Kristina. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT 2019*. Association for Computational Linguistics. 4171–4186. https://aclanthology.org/N19-1000/.

Dunning, Ted. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*. 19/1. 61–74.

Erjavec, Tomaž. 2019. *MULTEXT-East Morphosyntactic Specifications, Version 6*. http://nl.ijs.si/ME/V6/.

Erjavec, Tomaž; Ljubešić, Nikola. 2016. *MULTEXT-East Morphosyntactic Specifications*. https://nl.ijs.si/ME/Vault/V5/msd/html/.

Espinosa-Anke, Luis; Codina-Filbà, Joan; Wanner, Leo. 2021. Evaluating language models for the retrieval and categorization of lexical collocations. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 1406–1417. https://aclanthology.org/volumes/2021.eacl-main/.

Evert, Stefan; Uhrig, Peter; Bartsch, Sabine; Proisl, Thomas. 2017. E-VIEW-alation-a

Large-scale Evaluation Study of Association Measures for Collocation Identification. *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference*. Eds. Kosem, Iztok et al. Lexical Computing CZ s.r.o., Brno. Leiden. 531–549. https://elex.link/elex2017/proceedings-download/.

Ferret, Olivier. 2002. Using collocations for topic segmentation and link detection. *COLING 2002: The 19th International Conference on Computational Linguistics*. https://aclanthology.org/volumes/C02-1/.

Garcia, Marcos; García-Salido, Marcos; Alonso-Ramos, Margarita. 2017. Using bilingual word-embeddings for multilingual collocation extraction. *Proceedings of the 13th Workshop on Multiword Expressions*. Association for Computational Linguistics. 21–30. https://aclanthology.org/W17-1700/.

Garcia, Marcos; García-Salido, Marcos; Alonso-Ramos, Margarita. 2019. A comparison of statistical association measures for identifying dependency-based collocations in various languages. *Proceedings of the Joint Workshop on Multiword Expressions and WordNet*. Association for Computational Linguistics. 49–59. https://aclanthology.org/volumes/W19-51/.

Hashimoto, Kazuma; Tsuruoka, Yoshimasa. 2016. Adaptive Joint Learning of Compositional and Non-Compositional Phrase Embeddings. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. 205–215.

Hudeček, Lana; Mihaljević, Milica. 2020. Collocations in the croatian web dictionary – Mrežnik. *Slovenscina 2.0*. 8/2. 78–111. https://doi.org/10.4312/SLO2.0.2020.2.78-111.

Ivir, Vladimir. 1992. Kolokacije i leksičko značenje. *Filologija* 20/21. 181–189.

Jurafsky, Daniel; Martin, James H. 2009. *Speech and Language Processing: An Introduction to Natural Language, Computational Linguistics, and Speech Recognition* (2nd ed.). Pearson/Prentice Hall.

Karan, Mladen; Šnajder, Jan; Dalbelo Bašić, Bojana. 2012. Evaluation of Classification Algorithms and Features for Collocation Extraction in Croatian. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*. Eds. Calzolari, Nicoletta et al. European Language Resources Association. Istanbul. 657–662. https://aclanthology.org/volumes/L12-1/.

Kilgarriff, Adam; Baisa, Vít; Bušta, Jan; Jakubíček, Miloš; Kovář, Vojtěch; Michelfeit, Jan; Rychlý, Pavel; Suchomel, Vit. 2014. The Sketch Engine : Ten Years On. *Lexicography* 1/1. 7–36. https://doi.org/10.1007/s40607-014-0009-9.

Kita, Kenji; Kato, Yasuhiko; Omoto, Takashi; Yano, Yoneo. 1994. A comparative study of automatic extraction of collocations from corpora: Mutual information vs. cost criteria. *Journal of Natural Language Processing* 1/1. 21–33. https://doi.org/10.5715/jnlp.1.21.

Krenn, Brigitte. 2000. Collocation Mining: Exploiting Corpora for Collocation Iden-

tiication and Representation. *KONVENS 2000 / Sprachkommunikation, Vorträge der gemeinsamen Veranstaltung 5. Konferenz zur Verarbeitung natürlicher Sprache*. VDE Verlag. 209–214.

Krenn, Brigitte; Evert, Stefan. 2001. Can we do better than frequency? A case study on extracting PP-verb collocations. *Proceedings of the ACL Workshop on Collocations*. 39–46. http://www.collocations.de/EK/Articles/KrennEvert2001.pdf.

Lenci, Alessandro; Sahlgren, Magnus; Jeuniaux, Patrick; Cuba Gyllensten, Amaru; Miliani, Martina. 2022. A comparative evaluation and analysis of three generations of Distributional Semantic Models. *Language Resources and Evaluation*. 1–45. https://doi.org/10.1007/s10579-021-09575-z.

Lin, Dekang. 1998. Extracting Collocations from Text Corpora. *First Workshop on Computational Terminology*. 57–63.

Ljubešić, Nikola; Dobrovoljc, Kaja; Fišer, Darja. 2015. *MWELex-MWE Lexica of Croatian, Slovene and Serbian Extracted from Parsed Corpora. *Informatica* 39. 293–300.

Ljubešić, Nikola; Erjavec, Tomaž. 2011. hrWaC and slWac: Compiling Web Corpora for Croatian and Slovene. *Text, Speech and Dialogue. TSD 2011. Lecture Notes in Computer Science* 6836. Ur. Habernal, Ivan; Matousek, Vaclav. Springer. Berlin – Heidelberg. 395–402. https://doi.org/10.1007/978-3-642-23538-2_50.

Ljubešić, Nikola; Klubička, Filip; Agić, Željko; Jazbec, Ivo-Pavao. 2016. New Inflectional Lexicons and Training Corpora for Improved Morphosyntactic Annotation of Croatian and Serbian. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*. Eds. Calzolari, Nicoletta et al. European Language Resources Association. Portorož. 4264–4270. https://aclanthology.org/volumes/L16-1/.

Ljubešić, Nikola; Lauc, Davor. 2021. BERTić - The Transformer Language Model for Bosnian, Croatian, Montenegrin and Serbian. *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*. Eds. Babych, Bogdan et al. Association for Computational Linguistics. Kyiv. 37–42. https://aclanthology.org/volumes/2021.bsnlp-1/.

Ljubešić, Nikola; Logar Berginc, Nataša; Kosem, Iztok. 2021. Collocation Ranking: Frequency vs Semantics. *Slovenscina 2.0*. 9/2. 41–70. https://doi.org/10.4312/slo2.0.2021.2.41-70.

Ljubešić, Nikola; Dobrovoljc, Kaja. 2019. What does Neural Bring? Analysing Improvements in Morphosyntactic Annotation and Lemmatisation of Slovenian, Croatian and Serbian. *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*. Eds. Erjavec, Tomaž et al. Association for Computational Linguistics. Florence. 29–34. https://aclanthology.org/volumes/W19-37/.

Mikolov, Tomas; Chen, Kai; Corrado, Greg; Dean, Jeffrey. 2013. Efficient estimation

of word representations in vector space. *arXiv*:1301.3781. 1–12. https://doi.org/10.48550/arXiv.1301.3781.

Miščin, Evelina. 2015. Collocational competence of primary and secondary school students. *ExELL* 3/1. 8–25. https://doi.org/10.1515/exell-2016-0008.

Nacinovic Prskalo, Lucia; Brkic Bakaric, Marija. 2022. Identification of Metaphorical Collocations in Different Languages – Similarities and Differences. *Text, Speech, and Dialogue. TSD 2022. Lecture Notes in Computer Science* 13502. 102–112. https://doi.org/10.1007/978-3-031-16270-1_9.

Ordulj, Antonia; Žauhar, Valnea. 2018. Associative strength and frequency of 228 noun collocations in croatian. *Fluminensia* 30/2. 65–90. https://doi.org/10.31820/f.30.2.13.

Patekar, Jakob. 2022. What is a metaphorical collocation? *Fluminensia*, *34*(1), 31–49. https://doi.org/10.31820/F.34.1.5

Pearce, Darren. 2001. Synonymy in Collocation Extraction. *Proceedings of the Workshop on WordNet and Other Lexical Resources, Second Meeting of the North American Chapter of the Association for Computational Linguistics*. 41–46.

Pearce, Darren. 2002. A Comparative Evaluation of Collocation Extraction Techniques. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*. Eds. Calzolari, Nicoletta et al. European Language Resources Association. Portorož. 1530–1536. https://aclanthology.org/volumes/L16-1/.

Pecina, Pavel. 2005. An Extensive Empirical Study of Collocation Extraction Methods. *Proceedings of the ACL Student Research Workshop*. 13–18. https://aclanthology.org/volumes/P05-2/.

Pecina, Pavel. 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.

Petrović, Saša; Šnajder, Jan; Dalbelo Bašić, Bojana; Kolar, Mladen. 2006. Comparison of collocation extraction measures for document indexing. *Journal of Computing and Information Technology* 14/4. 321–327. https://doi.org/10.2498/cit.2006.04.08.

Pinnis, Mārcis; Ljubešić, Nikola; Stefanescu, Dan; Skadina, Inguna; Tadić, Marko; Gornostay, Tatiana. 2012. Term Extraction, Tagging, and Mapping Tools for Under-Resourced Languages. *Proceedings of the 10th Conference on Terminology and Knowledge Engineering (TKE 2012)*. 193–208. https://www.researchgate.net/publication/233807989.

Reder, Anna. 2006. Kollokationsforschung und Kollokationsdidaktik. *Linguistik Online* 28/3. 157–176.

Salehi, Bahar; Cook, Paul; Baldwin, Timothy. 2014. Detecting Non-compositional MWE Components using Wiktionary. *Proceedings of the 2014 Conference on Empiri-*

cal Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics. 1792–1797. https://aclanthology.org/volumes/D14-1/.

Salehi, Bahar; Cook, Paul; Baldwin, Timothy. 2015. A Word Embedding Approach to Predicting the Compositionality of Multiword Expressions. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics. 977–983. https://aclanthology.org/volumes/N15-1/.

Seljan, Sanja; Gašpar, Angelina. 2009. First Steps in Term and Collocation Extraction from English-Croatian Corpus. *Proceedings of 8th International Conference on Terminology and Artificial Intelligence*. http://www.irit.fr/TIA09/thekey/posters/seljan.pdf.

Seretan, Violeta. 2013. On collocations and their interaction with parsing and translation. *Informatics* 1/1. 11–31. https://doi.org/10.3390/informatics1010011.

Seretan, Violeta; Nerima, Luka; Wehrli, Eric. 2003. Extraction of Multi-Word Collocations Using Syntactic Bigram Composition. *Proceedings of the Fourth International Conference on Recent Advances in NLP (RANLP-2003)*. 424–431.

Seretan, Violeta; Nerima, Luka; Wehrli, Eric. 2004. A Tool for Multi-Word Collocation Extraction and Visualization in Multilingual Corpora. *Proceedings of the 11th EURALEX International Congress*. Eds. Williams, G.; Vessier, S. Université de Bretagne Sud. 755–766.

Seretan, Violeta; Wehrli, Eric. 2006. Accurate Collocation Extraction Using a Multilingual Parser. *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*. Eds. Calzolari, Nicoletta; Cardie, Claire; Isabelle, Pierre. Association for Computational Linguistics. 953–960.

Seretan, Violeta; Wehrli, Eric. 2009. Multilingual collocation extraction with a syntactic parser. *Language Resources and Evaluation* 43/1. 71–85. https://doi.org/10.1007/s10579-008-9075-7.

Shimohata, Sayori; Sugio, Toshiyuki; Nagata, Junji. 1997. Retrieving Collocations by Co-occurrences and Word Order Constraints. *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics. 476–481. https://aclanthology.org/volumes/P97-1/.

Smadja, Frank. 1993. Retrieving Collocations from Text: Xtract. *Computational Linguistics* 19/1. 143–178.

Smadja, Frank; Hatzivassiloglou, Vasileios; Mckeown, Kathleen R. 1996. Translating Collocations for Bilingual Lexicons: A Statistical Approach. *Computational Linguistics* 22/1. 1–38.

Šnjarić, Mirjana; Borucinsky, Mirjana. 2020. Glagolsko-imeničke kolokacije hrvatskoga, njemačkoga i engleskoga općeznanstvenog jezika u općoj dvojezičnoj e-leksikografiji. *Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje* 46/2). 1105–1127. https://doi.org/10.31724/rihjj.46.2.34.

Stojić, Aneta; Košuta, Nataša. 2017. Kolokacijske sveze u mentalnom leksikonu učenika stranog jezika. *Fluminensia* 29/2. https://doi.org/10.31820/f.29.2.9.

Stojić, Aneta; Košuta, Nataša. 2020. Collocations in L2 written text production. *Fluminensia* 32/2. 7–31. https://doi.org/10.31820/F.32.2.4.

Stojić, Aneta; Košuta, Nataša. 2021. Metaphorische Kollokationen – Einblicke in eine korpusbasierte Studie. *Linguistica (Slovenia)* 61/1. 81–91. https://doi.org/10.4312/linguistica.61.1.81-91.

Stojić, Aneta; Košuta, Nataša. 2022. Izrada inventara metaforičkih kolokacija u hrvatskome jeziku - na primjeru imenice godina. *Fluminensia* 34/1. 9–29. https://doi:10.31820/f.34.1.4.

Strakatova, Yana; Falk, Neele; Fuhrmann, Isabel; Hinrichs, Erhard; Rossmann, Daniela. 2020. All That Glitters is Not Gold: A Gold Standard of Adjective-Noun Collocations for German. *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020).* 11–16.

Tadić, Marko; Šojat, Krešimir. 2003. Finding multiword term candidates in Croatian. *Proceedings of IESL2003 Workshop.* 102–107.

Thanopoulos, Aristomenis; Fakotakis, Nikos; Kokkinakis, George. 2002. Comparative Evaluation of Collocation Extraction Metrics. *Third International Conference on Language Resources and Evaluation*. Eds. González Rodríguez, Manuel; Paz Suarez Araujo, Carmen. European Language Resources Association. Las Palmas. 620–625. https://aclanthology.org/L02-1000/.

Ulčar, Matej; Robnik-Šikonja, Marko. 2020. FinEst BERT and CroSloEngual BERT: less is more in multilingual models. *arXiv*:2006.07890. https://doi.org/10.48550/arXiv.2006.07890.

Verma, Rakesh; Vuppuluri, Vasanthi; Nguyen, An; Mukherjee, Arjun; Mammar, Ghita; Baki, Shahryar; Armstrong, Reed. 2016. Mining the Web for Collocations: IR Models of Term Associations. *Computational Linguistics and Intelligent Text Processing. CICLing 2016. Lecture Notes in Computer Science* 9623. Ed. Gelbukh, Alexander. 177–194.

Walker, Crayton Phillip. 2011. A Corpus-Based Study of the Linguistic Features and Processes Which Influence the Way Collocations Are Formed: Some Implications for the Learning of Collocations. *Source: TESOL Quarterly* 45/2. 291–312. https://doi.org/10.5054/tq.2011.247710.

Wanner, Leo; Ferraro, Gabriela; Moreno, Pol. 2017. Towards distributional semanti-

cs-based classification of collocations for collocation dictionaries. *International Journal of Lexicography* 30/2. 167–186. https://doi.org/10.1093/ijl/ecw002.

Yazdani, Majid; Farahmand, Megdad; Henderson, James. 2015. Learning Semantic Composition to Detect Non-compositionality of Multiword Expressions. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing.* Eds. Màrquez, Lluís; Callison-Burch, Chris; Su, Jian. Association for Computational Linguistics. Lisbon. 1733–1742. https://aclanthology.org/volumes/D15-1/.

Uvid u automatsko izlučivanje metaforičkih kolokacija

sažetak

Kolokacije su već dugi niz godina tema mnogih znanstvenih istraživanja. U fokusu ovoga istraživanja podskupina je kolokacija koju čine metaforičke kolokacije. Kod metaforičkih je kolokacija kod jedne od sastavnica došlo do semantičkoga pomaka, tj. jedna od sastavnica poprima preneseno značenje. Glavni su ciljevi ovoga rada istražiti postojeću literaturu te dati sustavan pregled postojećih istraživanja na temu izlučivanja kolokacija i postojećih metoda, mjera i resursa. Postojeća istraživanja opisana su i klasificirana prema različitim pristupima (statistički, hibridni i zasnovani na distribucijskoj semantici). Također su opisane različite asocijativne mjere i postojeći načini procjene rezultata automatskoga izlučivanja kolokacija. Metode, alati i resursi koji su korišteni u prethodnim istraživanjima, a mogli bi biti korisni za naš budući rad posebno su istaknuti. Stečeni uvidi u postojeća istraživanja čine prvi korak u razmatranju mogućnosti razvijanja postupka za automatsko izlučivanje metaforičkih kolokacija.

Keywords: metaphorical collocations, automatic extraction, association measures, hybrid approaches, distributional semantics approaches, evaluation measures

Ključne riječi: metaforičke kolokacije, automatsko izlučivanje, asocijativne mjere, hibridni pristupi, pristupi zasnovani na distribucijskoj semantici, mjere evaluacije