Faculty of Arts
University of Helsinki

COLLOCATIONS ON THE WAY.

HOW WORDS COME TOGETHER IN RUSSIAN

Daria Kormacheva

DOCTORAL DISSERTATION

To be presented for public discussion with the permission of the Faculty of Arts of the University of Helsinki, in Auditorium 2, Metsätalo, on June 15th, 2020, at 12 o'clock.

Helsinki 2020

The Faculty of Arts uses the Urkund system (plagiarism recognition) to examine all doctoral dissertations.

ISBN 978-951-51-6082-9 (paperback)
ISBN 978-951-51-6083-6 (PDF)

Unigrafia
Helsinki 2020

ABSTRACT

This thesis addresses the topic of collocations and their behavior based on Russian language data. In the course of four articles included in this compilation, I develop a better understanding of collocations that is based on a corpus-driven approach.

Chapter 1 defines the object of study and presents the definition of collocations that I developed during the course of this research: Collocations are defined as statistically significant co-occurrences of tokens or lexemes within a syntactic phrase that are extracted by statistics-based automatic analysis tools and are restricted to various extents: from semantically not-idiomatic to full idioms. I also discuss corpus linguistics as a theoretical framework, introduce the CoCoCo project within which this research was conducted, and provide an overview of research data and methodology that are applicable to the entire investigation.

Chapters 2 to 5 introduce the theoretical background for each of the four articles included in the thesis and discuss the major analytical findings.

Chapter 2 presents a discussion of the methods used to extract statistical collocations and provides results pertaining to the comparison of five metrics for extracting statistics-based collocations as well as the raw frequency for the Russian language. First, this research has demonstrated that the results of the discussed metrics are often correlated and, second, that the degree of idiomaticity of the extracted units varies significantly.

Chapter 3 offers a comparison of the empirical and phraseological perspectives on collocations and introduces research where I attempt to position empirical collocations within the scope of a phraseological theory. This research demonstrates that empirical collocations have different tendencies to form idiomatic lexical units and I reveal the shortcomings of describing the idiomaticity of expressions in terms of strict classes.

Chapter 4 examines grammatical profiling as a method used to define the optimal level of representation for collocations. I have demonstrated that collocations have different distributional preferences across the corpus. I have also analyzed the relationship between token and lexeme collocations based on the degree to which their grammatical profiles resemble the grammatical profiles of their headwords (although the border between the two types is not clear-cut). I also offered a plausible method of differentiating between these two collocation types.

Chapter 5 presents the main concepts of Construction Grammar and introduces the research where a substantial number of automatically extracted

collocations were demonstrated to form clusters of words that belong to the same semantic class, even when they are not idiomatic. Such constructional generalizations have shown that there is a more abstract level on which collocations can be stable as a class rather than on the level of single collocations.

Finally, in Chapter 6, I summarize all that has been accomplished during this research, and Chapter 7 provides an overview of the articles included in this thesis.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to everyone who supported me at any stage of my doctoral research process. I highly appreciated it.

My deepest gratitude is extended to my supervisor, Mikhail Kopotev, who encouraged me to begin and inspired me to complete my doctoral degree. No doubts that without him, this journey would not have been possible. I am grateful for his invitation to join the CoCoCo project in 2013 and for all our meaningful discussions, his time and his useful advice. But what is most important to me, was our friendly, positive and trusting atmosphere that was my greatest motivator and kept me going during my PhD studies. All this made my PhD experience positive no matter how difficult it was from time to time.

I am also thankful to my other colleagues from the University of Helsinki who contributed to this dissertation. I especially wish to thank my co-supervisor Ahti Nikunlassi for his help and advice and my co-author of the dissertation articles, Lidia Pivovarova, for our fruitful collaboration.

My work would not have been possible without the financial support from the Finnish Center for International Mobility (CIMO, currently EDUFI), who provided me with the opportunity to begin my research career in Finland. I also received support from the Research Foundation of the University of Helsinki, which granted me a position as a full-time PhD student, and the Doctoral School in Humanities and Social Sciences, which provided me with the dissertation completion grant.

Finally, and most importantly, I am infinitely grateful to my parents, Marina Kormacheva and Igor Gruzinskii, and to my other half, Vitaly Semenov, for always believing in me. Their love, encouragement and unconditional support always kept me going and were so much needed during this time of my life.

CONTENTS

1 INTRODUCTION

1.1 THE OBJECT OF THE STUDY: FROM WORDS TO COLLOCATIONS

Traditionally, words are considered to be basic units of language and are used as a departure point for the analysis of the major linguistic levels – semantic, morphological and phonological. However, the term *word* itself is rather vague, as no single definition of it is available and there are numerous examples that violate a naive understanding of it. Most of the contradictions are faced on the semantic level where the co-occurring words may form larger units that function as single words alone, as in *look up*, *in terms of*, *juridical person* or *young generation*. These types of expressions consist of more than one item and may be considered the basic units of language along with single words, as they represent holistic entities whose constituents are bound both semantically and syntactically.

However, the grouping of words is not restricted to idiomatic expressions. Frank et al. (2012) provide psycholinguistic evidence that frequent expressions are processed as whole chunks even when they are not idiomatic. These expressions form a substantial part of the lexical system of language and their number in one's vocabulary is assumed to have approximately the same order of magnitude as single words (Jackendoff 1997, 156; for an overview of the research supporting this claim based on different data, see Men (2018, 9–11)).

Such frequent expressions can be united under the term collocations defined by Sag et al. (2002, 9) as "any statistically significant cooccurrence, including all forms of multiword expressions <...> and compositional phrases which are predictably frequent." According to this definition, multiword expressions (MWEs) can be described generally as "sequences of words that act as a single unit at some level of linguistic analysis" and are characterized by the presence of the behavior such as:

1. "reduced syntactic and semantic transparency;
2. reduced or lack of compositionality;
3. more or less frozen or fixed status;
4. possible violation of some otherwise general syntactic patterns or rules;
5. a high degree of lexicalization (depending on pragmatic factors);
6. a high degree of conventionality" (Calzolari et al. 2002, 1934)

They include a wide range of linguistic phenomena but in general, no clear borders exist between different groups of MWEs, as was observed by Moon (1998) and Bartsch (2004, 33).

According to Sag et. al. (2002), collocations provide us with a means to automatically analyze statistically significant MWEs (such as *in short*, *part of speech*, *look up, traffic light* and *fresh air)* along with frequent compositional phrases without conventionalized forms, such as *selling a house* or *sell a house* where lexemes *to sell* and *house* remain together and form collocations only due to real-world facts and not due to the idiomatization of meaning or conventionalization of form.

Although the definition above somewhat satisfies the main objectives in my research, it is worth noting that many fields, such as natural language processing, lexicography, and second language acquisition, define collocations differently because they focus on different aspects of language. The lack of a single definition for the term *collocation* has created many controversies, and the vast number of studies on collocations may in fact address a wide range of different linguistic phenomena. Discussion of the terminological confusion on the term collocation that we encounter today is presented, for example, in Nesselhauf (2005, 11–18) or Evert (2005, 15–17). The main distinction lies between the distributional, or empirical, (i.e. frequency-based) and the phraseological (i.e. with a focus on lexicalization) approaches; I will return to this later in Chapter 3.

I adopt an empirical orientiation in my investigation and define collocations as the following:

> *Collocations are statistically significant co-occurrences of tokens or lexemes within a syntactic phrase that are extracted by statistics-based automatic analysis tools and are restricted to various extents: from semantically not-idiomatic expressions to full idioms.*

In terms of my analysis, idiomaticity is considered to be a facultative feature of collocations. The common characteristic of all the units under investigation is that their co-occurrence is distinguishable from the usage of the constituent collocates by themselves, as either one or both of the collocates are highly bound to each other. Any boundaries between different classes of collocations are considered to be vague – just as they are between MWEs. By adopting an empirical approach, I emphasize my interest in the full spectrum of potentially relevant linguistic units and focus on a practical corpus-driven application of collocational analysis–in particular collocation extraction–that can then be used in information retrieval, machine translation or any other task that requires automatic extraction of semantic information from texts.

1.2 CORPUS LINGUISTICS AS THE THEORETICAL FRAMEWORK

The definition that my study uses is consitent with the empirical approach adopted in corpus linguistics. Several central criteria that are common to the research within this paradigm are described well by Gries and Stefanowitsch in their introduction to *Corpora in Cognitive Linguistics* (2007), which are summarized below.

In corpus linguistics, an analysis of research problems is based on data from a balanced and representative collection of texts – a corpus. These data typically include various frequency lists, collocations and concordance lines, and the analysis utilizes statistical information. In particular, collocational analysis is based on token or lexeme co-occurrences and statistics on their frequencies. Analysis in corpus linguistics is systematic and exhaustive in the sense that it not only describes a single, usually the most frequent, examples in detail, but it also reveals a full spectrum of usage that includes the less frequent examples. The representativeness of a corpus is defined as "the extent to which a sample includes the full range of variability in population" (Biber 1993, 243). This is achieved by including a balanced range of genres in the corpus and by carefully selecting chunks of texts from these genres (McEnery et al. 2006, 13). As a result, a representative corpus is supposed to form a basis for generalizations about the language variety that it represents. For details on how to design a representative corpus and an overview of central issues of corpus linguistics, see Biber (1993), Chapter 3 in Tognini-Bonelli (2001) and Unit A2 in McEnery et al. (2006).

A large number of corpora are currently available that vary in their size, purpose, and level of annotation. For the Russian language, an overview of the resources is presented in Kopotev et al. (2017, 9–10). The existence of quality data sources has allowed us to use corpora in different types of research. Corpus linguistics methodology can be divided into two subcategories based on how a corpus is used: *Corpus-driven* and *corpus-based*.

Tognini-Bonelli (2001) explains the difference between the two by defining the *corpus-based approach* as a "methodology that avails itself of the corpus mainly to expound, test or exemplify theories and descriptions that were formulated before large corpora became available to inform language study" (2001, 65). In other words, the corpus is primarily used to provide evidence for an existing theory without leaving much theoretical and methodological space to formulate new theoretical generalizations (ibid. 66). The same approach is adopted by Wolf and Gibson (2005), who use a corpus to investigate and illustrate various discourse coherence structures, and

Dobrovolskij and Pöppel (2016), who consult parallel corpora to investigate idiomatic and compositional constructions.

In turn, the *corpus-driven approach* "aims to derive linguistic categories systematically from the recurrent patterns and the frequency distributions that emerge from language in context" (Tognini-Bonelli 2001, 87). This approach uses evidence from a corpus to formulate theoretical statements instead of illustrating pre-existing theories. The general methodological path begins from observation, proceeds to formulating a hypothesis, then to devising a generalization and finally to establishing a theoretical statement (ibid. 85). The understanding is thus based on what is attested in the corpus data as opposed to theoretically distinguishing certain categories first and then exemplifying them by consulting the corpus. However, this does not imply that linguistic prerequisites are excluded from this process. On the contrary, they may be incorporated into every stage. In fact, a deeper and more extensive analysis as a rule is achieved by incorporating annotation which is the process (and result) of "adding <…> interpretative, linguistic information to an electronic corpus of spoken and/or written language data" (Leech 1997, 2) (for example, part-of-speech tagging, more elaborate morphological information, and information on syntactic structures or semantic classes of lexemes). Examples of corpus-driven studies can be found, for example, in Biber (2009), who identifies the most common multi-word patterns in different university registers, Glynn and Fischer (2010), who present several studies on cognitive semantics, or Piperski (2016), who analyzed intra-speaker stress variation in Russian.

One of the mainstream topics that is often investigated within a corpus-driven framework is collocational analysis, and many useful tools have been developed based on major corpora. For the Russian language, they include tools such as:

- the SketchEngine service that utilizes several corpora, the largest being the Russian Web corpus ruTenTen corpus with more than 14 million words,
- basic collocation extraction tools from the University of Leeds based on ruWac corpus,
- the CoCoCo service for both collocation and colligation extraction (see Chapter 1.3 for details),
- the CoSyCo tool for syntactic collocation extraction (Klyshinsky et al., 2018).

Furthermore, both Sketch Engine (with its subproject RuSkELL (Apresjan et al., 2016)) as well as CoCoCo, were also developed to serve as platforms for language teaching.

1.3 CURRENT RESEARCH WITHIN THE COCOCO PROJECT

This research was conducted within the CoCoCo (*Collocations, colligations, and constructions;* http://cococo.cosyco.ru) project that is led by M. Kopotev at the University of Helsinki. This is one of the few online resources for the Russian language that helps to understand how native speakers of Russian use the language, and is simultaneously a helpful tool for language learners, teachers, and linguists (see Figure 1 for a screenshot of the interface). The CoCoCo project aims to provide a tool for the automatic extraction of selective preferences of either tokens or lexemes from Russian corpora. An analysis of co-occurrence data is conducted using statistical methods and a corpus-driven approach. Employment of statistical analysis implies utilizing large-scale corpora, and the system supports three of them. These are the disambiguated subcorpus of the Russian National Corpus, which is consulted for this research and includes approximately 6 million tokens (Plungjan 2005), the Russian Internet Corpus (I-Ru, Sharoff and Nivre 2011), which has approximately 140 million tokens, and a balanced part of the Taiga corpus, which includes 400 million tokens (Shavrina and Shapovalova 2017).

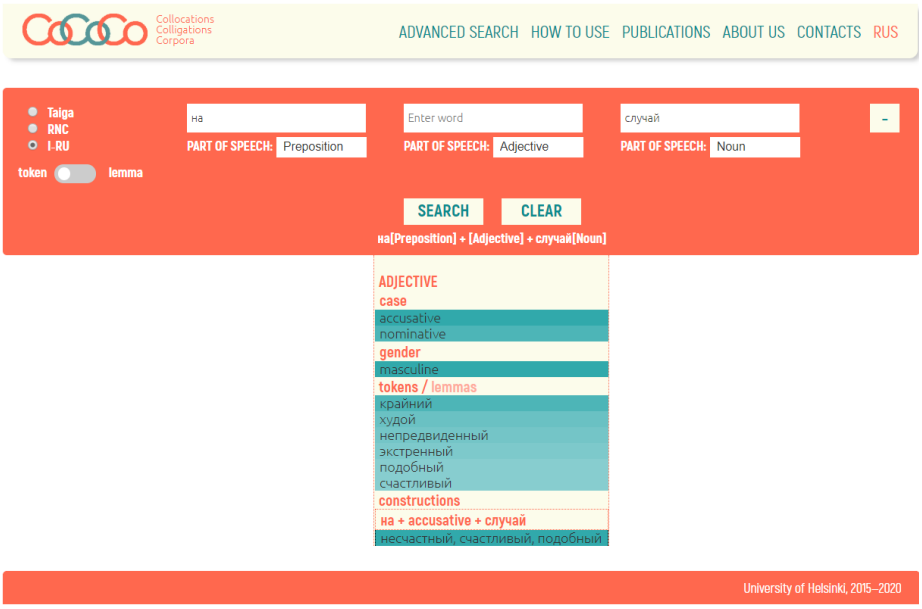


Figure 1 The interface of the CoCoCo that illustrates finding selectional preferences of the pattern *na* 'on, in, at' + x.Adj + *slučaj* 'case'

Our system, as described by Kopotev et al. (2015) adopts the concepts of Construction Grammar and analyzes the lexical and grammatical preferences of words equally. The aim of this approach is to determine the underlying cause for words to frequently co-occur "whether it is due to their morphological categories, or to lexical compatibility, or both" (ibid, 44). The information on the co-occurrence for a given word includes colligations (i.e., "co-occurrence of grammatical phenomena <…> with a word or phrase" (Sinclair 1998, 15)), collocations ("co-occurrences of words" (ibid., 15)), and collostructions (combinations of both). For example, this system can provide information that the verb *delit'sja* 'to share' is usually followed by a noun in the instrumentative case, that its most stable collocates include *opyt-om* 'experience-SG.INS' and *vpečatlenij-ami* 'impression-PL.INS', and that these collocates are, in turn, fillers in the collostruction *delit'sja* 'to share' + [*vpečatlenij-ami* 'impression-PL.INS', *opyt-om* 'experience.N.SG.INS', *vospominanij-ami* 'memory-PL.INS', *emocij-ami* 'emotion-PL.INS']. Yet another collostruction of the verb is *delit'sja* 'to share' + [*video-Ø* 'video-SG.INS', *fotografij-ami* 'photo-PL.INS', *zapis'-ju* 'record-SG.INS', *rolik-om* 'video-PL.INS']. An example of the application of this algorithm to grammatical and lexical patterns' extraction can be found in (Kopotev et al. 2013) and (Kormacheva et al. 2014).

While the entire project focuses on both grammatical and lexical data, my own research concerns only the part that is responsible for collocational analysis. In this article compilation, my intent was to examine collocations from different angles in order to create a comprehensive understanding of their behavior. The focal points that were investigated are automatic methods of collocation extraction, the complexity of collocational nature as illustrated by constructions, the degree of idiomaticity of the extracted units as well as their grammatical characteristics. The four articles that are included in this compilation attempt to provide answers to the following questions:

1. What measure(s) are best suited for automatic collocation extraction?
2. What is nature of the automatically extracted units if we look at them from a linguistic perspective that adopts a phraseological orientation?
3. What is difference between a token and lexeme representation of the collocations, and which is more plausible in the linguistic description?
4. Is there a method for formulating a generalized description of collocations that is based on how they cluster into bigger semantic classes?

The research findings were then used to improve our system in accordance with user needs and to provide a deeper understanding of what we are trying to model.

This research is a part of the CoCoCo project and it therefore inherits its corpus-driven methodological framework. Thus, a detailed semantic analysis of each collocation is not the main objective of this research; instead it focuses on a wider perspective where hundreds of analyzed expressions are meant to illustrate the thousands of expressions that were extracted from corpora. Our project aims to discover a method of narrowing an enormous volume of expressions that are included in corpora in order to serve those who are interested in conducting further analysis on a more detailed level, which is usually performed within a phraseological investigation. Thus, our approach does not contradict the concepts of a careful semantic analysis but aims to support it by providing the relevant preprocessed data. We therefore end where a deeper semantic analysis begins.

1.4 DATA AND METHODS

The consistency and interrelationship between the results from different experiments were achieved by consulting the same research data throughout the research. The analysis was conducted on the subset of the data used in the CoCoCo project and it was restricted to the n-gram corpus that had been extracted from the manually disambiguated subcorpus of the Russian National Corpus (Plungjan 2005). An n-gram corpus is a corpus of contiguous strings of *n* items (unigrams, bigrams, trigrams, etc.). For example, the sentence "The research was conducted on the n-gram corpus." consists of the following bigrams: *The research*, *research was*, *was conducted*, *conducted on*, *on the*, *the n-gram*, and the *n-gram corpus*. The given n-gram corpus provides comprehensive information on the grammatical characteristics of each token and the lexeme it belongs to, and this allows an in-depth analysis of a morphologically rich language. However, the one substantial drawback of this corpus is its size of approximately six million tokens, which is a lower number of tokens than for the other two corpora that are utilized in the CoCoCo project. Nevertheless, high-quality data that is fully reliable in terms of linguistic annotation and the necessity of making an accurate distinction between homonymous grammatical forms determined the selection of this disambiguated n-gram corpus as a data source.

The small size of the corpus affected this research in two ways. Firstly, we could not fully exclude low-frequency items from the evaluation. Statistical analysis requires a large amount of data. In fact, the final result is more reliable when more data is available. Words that exhibit a low frequency in the corpus

can significantly affect the outcome of the research because collocation extraction methods that are based on statistics can provide false-positive results when dealing with rare events. For example, a simple frequency ratio will yield the same value of 0.5 both when a word with a corpus frequency of 1 000 appears 500 times within a certain collocation and when a word with a corpus frequency of 2 appears once within a given collocation. In the first case, we may formulate some statistical generalizations, but the frequency ratio value from the second example is likely to be arbitrary. In this research, the words that occurred in the corpus only once were filtered out (for the article "*What do we get from extracting collocations? Linguistic analysis of automatically obtained Russian MWEs*" – fewer than 5 times). We would set this threshold higher if more data were available. Additionally, the small size of the corpus has also affected the scope of this research. For example, in *"Evaluation of collocation extraction methods for the Russian language,"* it was necessary to reduce the number of collocations from *A Russian-English Collocational Dictionary of the Human Body* that were used in the analysis because not all of them had a sufficient number of occurrences in the corpus.

Each article in this compilation focuses on Russian bigram collocations, and a number of collocational patterns were investigated. Each bigram under consideration constituted self-sufficient syntactic phrases that had meaning on their own. These bigrams were all continuous bigrams, that is, the first word of the bigram was directly followed by another without any intermediate tokens. The choice of the specific collocations that were to be examined was motivated by the requirement of having sufficient data for a reliable statistical analysis as well as by the objective to investigate the most frequently used lexemes as they constitute a basis for our everyday lexicon.

The article "*Choosing between lexeme vs. token in Russian collocations"* examines the 100 most frequent [Adjective + Noun] collocations in the corpus and none of the collocates were predefined. In the other three articles, one of the lexemes in a collocation was predefined and another belonged to a certain part of speech. For example, the pattern [vysokij 'high/tall' + x.Noun] is realized in collocations *vysokij tenor* 'high tenor', *vysokaja likvidnost'* 'high liquidity', *vysokaja častota* 'high frequency', etc.

In "*What do we get from extracting collocations? Linguistic analysis of automatically obtained Russian MWEs,*" I examine prepositions, which are a closed class of syntactic words, and used a weighted frequency ratio to extract the 100 most used collocations for the 25 Russian prepositions, which results in a total of 2 500 collocations. The weighted frequency ratio is a measure we proposed as "a ratio of a word frequency in a pattern to its frequency in the corpus multiplied by a logarithm of the word frequency in the general corpus" (Kopotev et al. 2017, 139). In "*Constructional generalization over Russian*

collocations," we examine collocates of the most frequent nouns, adjectives and verbs, again using a weighted frequency ratio to extract the 100 most relevant collocations for the 10 most frequent words within each class for a total of 3 000 analyzed collocations. We continue our discussion on noun collocates (both adjectival and verbal) in our article titled "*Evaluation of collocations extraction methods for the Russian language,*" and we examine twelve nouns that denote body parts that are entries in *A Russian-English Collocational Dictionary of the Human Body*. This article presents rankings that were obtained automatically for five computational methods of collocation extraction. We evaluated the collocations in these rankings both against the above-mentioned dictionary entries and a survey conducted with native speakers.

Thus, the focus of my research was on collocations that were extracted from the corpus using either their frequency characteristics or more elaborate statistical measures, such as their T-score (Church et al., 1991), MI (Church and Hanks, 1990), Log-likelihood (Dunning 1993), Dice (adopted from set theory, linguistic application discussed in, for example, Daudaravicius (2010)), and the weighted frequency ratio. An explanation for all these measures is provided by Kopotev et al. (2017, 138–139). By using these elaborate measures, it was possible to direct attention to collocations that were extracted from the corpus in an automated manner, which was the immediate topic of my interest and also to ensure that the results of the analysis were statistically significant. Another quantitative measure that was applied in this research is the Jenson-Shannon divergence (Lin 1991). In "*Choosing between lexeme vs. token in Russian collocations,"* this measure is proposed as a method to distinguish between the token and lexeme collocations by ranking collocations based on the difference between their grammatical profiles as compared to the grammatical profiles of their headword. Finally, the article on the evaluation of the collocation extraction methods assesses the extracted collocations against native Russian speakers' responses to the questionnaire that was used to measure their perception of the stability of these collocations. The same extracted collocations were also evaluated in terms of their presence or absence in a trustworthy dictionary. The aim was to demonstrate that different means of evaluation lead to the same result by triangulating the reference data sets that both assign a quantitative characteristic of stability to given collocations (a value on a scale in case of the informants and a binary value indicating presence or absence in case of the dictionary).

Nevertheless, the present compilation, and particularly the data analysis, utilizes both quantitative and qualitative methods. An overview of the methods used to combine or integrate the two methods is presented in Tashakkori and Creswell (2007).

The articles begin with a discussion of descriptive statistics on the data, which provides an overall, general perspective on the subject. The average member of a given sample is described in terms of mean and median values. The mean value is calculated by the sum of all data points divided by their number, whereas the median is calculated by ordering all data points along a continuum and then selecting the value in the middle (for details on the evaluation used for this ordering, see section 2.1). After providing the necessary context, I present a qualitative analysis of the obtained results. This is introduced as case studies, which include a linguistic analysis of the automatically obtained collocations, units that form constructions or the relation between the use of either tokens or lexemes in collocations. These case studies provide insight into the characteristics of empirical collocations as well as their relation to their lexical counterparts.

By combining quantitative and qualitative paradigms, my objective was to achieve a versatile and more comprehensive understanding of the investigated topic in question. While statistical information provides a basis for a generalized analysis, case studies allow for a deeper analysis of particular examples. As was pointed out previously by Fillmore in 1992, the most promising results are yielded by research that combines observations made through adopting the premises of corpus linguistics and the manual processing of examples adopted from traditional linguistics. Indeed, Filmore (1992, 35) observes the following:

> I don't think there can be any corpora, however large, that contain information about all of the areas of English lexicon and grammar that I want to explore. <...> (but) every corpus I have had the chance to examine, however small, has taught me facts I couldn't imagine finding out any other way.

2 AUTOMATIC EXTRACTION OF EMPIRICAL COLLOCATIONS

2.1 WAYS OF EVALUATING THE OBTAINED RESULTS

One of the main issues in the analysis of empirical collocations is to decide on a method to automatically extract them from a text and a procedure to evaluate the obtained results. The relatively open definition that was adopted earlier allows for a wide range of possible metrics to select empirical collocations from text data as well as ways to evaluate them.

Depending on what one intends to extract, a large number of statistical, and yet linguistically motivated, methods are available for collocation extraction. An overview of these methods is discussed in, for example, Wiechmann (2008) or Pecina and Schlesinger (2006). It is, nonetheless, common to use only a small subset of these metrics for both collocation extraction and for the assessment of the performance of newly developed metrics. The core set of measures that is typically used and compared include the t-score, χ^2-test, log-likelihood, and MI (Manning and Schütze, 1999, 151–187). It is important to note that adopting a large number of methods available does not mean that they can be used blindly, as the usage of these methods requires knowledge of what these methods are able to detect. Evert and Krenn (2005, 452) observe that a number of factors affect the usefulness of each individual method. These include the type of collocation to be extracted, ways of syntactic pre-processing, the approach to candidate extraction, the domain and size of the corpora, and the frequency thresholds used to filter out the data (Evert and Krenn 2005, 452). Many studies that focus on a comparison of different measures can be found and yet there is no single answer that is best. For example, Evert and Krenn (2001) report no significant differences between various measures for the German language, and Wermter and Hahn (2006, 791) argue that not only do all statistics-based measures display a similar performance, they also do not differ from the frequency of occurrence counts if no additional linguistic information is incorporated.

The performance of collocation measures is often evaluated by precision and recall measures. The description of these measures is, for example, presented by Manning and Schütze (1999). They define recall as the "proportion of the target items that the system selected" and precision as a "measure of the proportion of the selected items that the system got right" (ibid., 268). When applied to collocation extractions, recall describes the fraction of the collocates that have been retrieved among all the true

collocations, whereas precision describes how many of the extracted units are collocations. Another measure of evaluation applied in this research is non-interpolated average precision (Moirón and Tiedemann, 2006), which measures recall indirectly by taking into account the relative positions of the relevant items in a given ranking. Precision is therefore computed at each point of the ranking where a relevant entry is found, and then all precision points are averaged.

The question nonetheless remains open as to how we define a true positive example to be used in the evaluation, that is, how do we determine the formal criteria for categorizing a given expression as a collocation. Most of the evaluations are based on a researcher's own intuition or the knowledge of some other human expert. The goal remains the same – to obtain a manually extracted or annotated reference set of collocations. Among others, Smadja (1993, 166) argues that the evaluation of automatically retrieved collocations is best undertaken by a professional lexicographer. He employs this approach to evaluate his own technique for extracting collocations. The same approach is supported by Evert and Krenn (2001) for comparing results from different association measures. These evaluations, however, are not sufficient for acquiring a comprehensive understanding of the methods and they are not able to reveal the full picture of the usefulness of the methods (Evert and Krenn 2005, 451). An alternative method of evaluating collocation extraction techniques may be comparing them against a gold standard, if such standards are available. Pearce (2002) uses a list of collocations as a reference. This list was obtained from a dictionary that was reduced to include those collocations that occurred in the data at least once. Thanopoulos et al. (2002) in turn employ WordNet (Miller 1998), a resource that contains various semantic relationships between lexical items. Finally, assessments made by native speakers can also be used as evaluation technique that involves several people who provide intuitive judgments on how lexemes are used together in a target language, For example, this technique has been used by Blaheta and Johnson (2001). Yet it is controversial to assess the performance of collocation extraction measures in terms of any of these particular criteria and to fully evaluate the usefulness of the measures, it is necessary to evaluate them from several perspectives (Pearce 2002, 1535).

Khokhlova (2008, 2017) as well as Yagunova and Pivovarova (2010) discussed the topic of the comparison and/or evaluation of statistical measures for the Russian text data, but these works share the disadvantage of one-party evaluation that was described earlier. In automatic terminology detection, Braslavski and Sokolov (2006) used automatic evaluation based on the topic index of the books, and combined it with an expert evaluation of the extracted units. Our research group also investigated this topic on a smaller

scale. This was a pilot study for the article that is discussed in the next section and it was conducted on Russian prepositions (Kormacheva et al. 2014).

2.2 EVALUATION OF COLLOCATION EXTRACTION METHODS FOR THE RUSSIAN LANGUAGE

All attempts to organize an extensive evaluation of collocation extraction methods have failed because no common ground has been established to compare the results. In *"Evaluation of collocation extraction methods for the Russian language,"* my co-authors and my objective was to provide a systematic evaluation of Russian empirical collocations that were extracted by adopting five different measures for collocation extraction and using the raw frequency as a baseline. We compared them against survey responses from native speakers as well as an expert evaluation of a dictionary compilation. We only focused on precision and examined how many of the automatically extracted collocations actually overlap with linguistic idiomatic units. Our assumption was that the collocations that were considered fixed by native speakers as well as those listed in dictionaries should appear in higher positions on automatically extracted lists of collocations. Based on this assumption, we also aimed to provide a reliable cross-evaluation of these measures that would be able to determine the extent to which they correlate with each other and whether any of them could extract more idiomatic expressions from corpus data.

These research results demonstrate that the different measures used in automatic collocation extraction are often correlated with each other for the Russian language. The reason for this is that the performance of all these measures is inevitably affected by the collocational preferences of a given item. Some items participate in stable and idiomatic expressions more often than others. The measures expose the most similar behavior and results for expressions whose collocates have strong distributional preferences, as is the case for *skalit' zuby* 'to bare one's teeth', where the verb *skalit'* 'to bare one's teeth' is almost exclusively bound to the noun *zuby* 'teeth'. All the investigated measures are generally interchangeable to some degree and a choice between them should depend on the task in question. The similarities in cross-evaluation were demonstrated by the dictionary and survey evaluations, where we asked native speakers how regularly a pair of given words occurred together. The latter has also demonstrated that collocations with semantic shifts in their meaning are easier for native speakers to recognize as collocations than those whose semantic meaning is compositional. Automatic collocation extraction measures can, therefore, detect stable expressions that

are more difficult to distinguish for native speakers due to the compositional semantics.

The results of the collocation extraction in *"Evaluation of collocation extraction methods for the Russian language"* consistently reveal that for all measures, the degree of idiomaticity among the extracted units varies significantly and many of the retrieved expressions are stable but not completely idiomatic. This correlates with the results obtained in *"What do we get from extracting collocations? Linguistic analysis of automatically obtained Russian MWEs,"* where the same challenge was encountered when attempting to place automatically extracted collocations in the context of the linguistic theory. In the next chapter, we closely examine the difference in orientations adopted by the fields of corpus linguistics and phraseology and how they affect collocations.

3 THE CONTROVERSY IN DEFINITION: PHRASEOLOGY VERSUS CORPUS LINGUISTICS

3.1 COMMON PREREQUISITES

As I have mentioned previously, the notion of collocation is highly controversial in contemporary linguistics because there is no single approach to define it. All of the approahes are grounded in the premise that text generation is based on the interlacement of the open-choice principle as well as the idiom principle that was described by Sinclair (1991). The open-choice principle assumes that "words are treated as independent items of meaning" and that "each of them represents a separate choice" (ibid., 175). However, according to the idiom principle, at each point of text generation, "a language user has available to him a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analyzable into segments" (ibid., 110). These preconstructed phrases resemble words in that they can represent separate choices, and a language user alternates between these two principles when producing spoken or written text. The magnitude of this phenomenon was analyzed by Erman and Warren (2000), who calculated that English on average offers 71 choices for a text of 100 words. These choices consist of 45 single-word choices and 26 choices of prefabricated phrases, which are "combination(s) of at least two words favored by native speakers in preference to an alternative combination which could have been equivalent had there been no conventionalization" (2000, 31).

It follows that the key difference between the two approaches to collocations is then that in phraseology, the notion of collocations is based on a semantic shift in their meaning, while the empirical approach adopted in corpus linguistics defines collocations based on statistical information about the frequency of their usage and therefore includes a wider range of linguistic units. The study of collocations has a longer history in phraseology than the younger empirical approach has in corpus linguistics, but both have their own areas of application and deserve to be acknowledged. In fact, it would be instructive to introduce a new term to denote the newer meaning, however, and following other researchers, I adopt the term "empirical collocation," although it can be somewhat confusing.

3.2 COLLOCATIONS IN PHRASEOLOGY

In phraseology, collocations are investigated within the context of other phraseological units. There has been extensive research published on this topic and a historical overview of existing approaches is presented in (Penttilä, 2006) and for the Russian language, in Baranov and Dobrovolskij (2008). The latter also devised a detailed classification of phraseological units that is based on the two main properties that characterize phraseological units – idiomaticity and stableness. The degree to which each of these properties is present in each particular case defines the phraseological type of the expression. Baranov and Dobrovolskij (2008, 67) defined collocations as weakly idiomatic phraseological units that structurally represent a phrase where the main semantic component is used in its direct meaning. This suggestion of dividing collocations into further subclasses is based on lexical functions that were applied to collocational analysis as described by Mel'čuk. Mel'čuk (1995, 1998) proposed a classification where collocations as well as idioms, quasi-idioms, and pragmatemes are considered to be a subclass of lexical phrasemes and are defined as follows:

> A collocation AB of language L is a semantic phraseme of L such that its signified 'X' is constructed out of the signified of the one of its two constituent lexemes—say, of A—and a signified 'C' ['X' = 'A ⊕ C'] such that the lexeme B expresses 'C' contingent on A.
>
> (Mel'čuk 1998, 30)

One of the collocates in collocation is thus freely chosen because of its signified, while the other is contingent on the first one. All collocations can then be formally described with the help of lexical functions (LF).

> A Lexical Function f is a function that associates with a given lexical unit L, which is the argument, or keyword, of f, a set $\{L_i\}$ of (more or less) synonymous lexical expressions — the value of f — that are selected contingent on L to manifest the meaning corresponding to f: $f(L) = \{L_i\}$.
>
> (Mel'čuk 1998, 8)

A LF thus represents "a very general and abstract meaning <…> which can be lexically expressed in a large variety of ways depending on the lexical unit to which this meaning applies" (ibid, 8). For example, the LF *Magn* has the meaning "very; to a high degree; intense(ly)," as in *Magn(patience) = infinite*,

and the LF *Real2* has the meaning "realize; fulfill [the requirement of]" as in *Real2(exam) = to pass*.

An example of a collocation within the phraseological approach is *prolivnoj dozhd'* 'pouring rain', which we can describe with an LF in terms of Mel'čuk's theory: *Magn(dozhd') = prolivnoj*. While *dozhd'* 'rain' is used in its direct meaning and the whole expression cannot be considered an idiom, its collocate is highly bound to the headword. *Prolivnoj* 'pouring' occurs in the Russian National Corpus 912 times within the [Adjective + Noun] pattern, and 872 of these occurrences are with the lexeme *dožd'* 'rain'. Its remaining collocates include *doždik* 'rain.Diminutive', *grozy* 'thunderstorm.Pl', *osadki* 'precipitation.Pl', *liven'* 'downpour', *slezy* 'tear.Pl', *strui* 'stream.Pl' and several metaphoric usages. In addition, *dozhd'* 'rain' is also bound to *prolivnoj* 'pouring' as this adjective is the most frequent collocate for this noun. The second most frequent collocate of *dozhd'* 'rain' is *sil'nyj'* 'heavy' with 447 occurrences. Furthermore, while frequency alone in phraseology may not be a sufficient basis for considering an expression to be a collocation, corpus linguistics has another method of defining collocations and from this perspective, both *prolivnoj dozhd'* and *sil'nyj dozhd'* are considered collocations regardless.

3.3 COLLOCATIONS IN CORPUS LINGISTICS

In corpus linguistics, collocations are defined by focusing on the frequency of the expressions, and by using different statistical measures to refine the raw frequency. The same approach is utilized in the quantitative linguistic analysis that serves as a basis for computational linguistics with tasks such as automatic collocation extraction. In this paradigm, the subjective assessment of individual expressions in terms of their idiomaticity (in addition to stableness) is replaced by an overall picture of language usage that is recorded in corpora. Firth (1957) was the first to introduce the very general definition of collocations in an empirical sense with a focus on word co-occurrences and this definition was later elaborated on by other scholars. For example, Sinclair (1991, 170) describes collocation as "the occurrence of two or more words within a short space of each other in a text." Moon (1998, 26) also explicitly states that these co-occurrences must be frequently repeated or statistically significant and that there may or may not be "special semantic bonds between collocating items". Evert (2008, 4) follows Sinclair's definition by defining collocation as "a combination of two words that exhibit a tendency to occur near each other in natural language."

As empirical definitions of collocations are mainly frequency-based, all of them cover a heterogeneous group of linguistic expressions that range from

idiomatic expressions to rather free word combinations that tend to co-occur. Yet it has recently been discovered that frequency (and its various applications in statistical measures) is not the only descriptive feature of collocations. Kopotev and Steksova (2017, 59) assume that empirical collocations also possess a number of features that distinguish them from random word combinations and may become a basis for their future development into lexical collocations. These features include constructional restrictions, a unique grammatical profile and the possibility to insert other lexemes in a given collocation. These properties are, however, not obligatory and may be present to varying degrees.

The distinction between the different definitions of collocations in phraseological theory and in empirical studies was described by Evert (2008), who differentiates between *lexical* and *empirical* collocations. The former refers to the collocations in the phraseological meaning and constitute a part of a larger group, multiword expressions. In turn, empirical collocations refer to the "recurrent and predictable word combinations, which are a directly observable property of natural language" (Evert 2008, 3). In this compilation, my primary focus is on empirical collocations and I only utilize lexical collocations to refine the nature of empirical collocations. The term collocation is thus used alone to refer to empirical collocations.

3.4 WHAT DO WE GET FROM EXTRACTING COLLOCATIONS? LINGUISTIC ANALYSIS OF AUTOMATICALLY OBTAINED RUSSIAN MWES

The distinction between the theoretical and empirical approaches used to define collocations has motivated me to determine the degree to which the two concepts overlap. To achieve this, I attempted to place empirical collocations within the scope of a phraseological theory. In the article *"What do we get from extracting collocations? Linguistic analysis of automatically obtained Russian MWEs,"* prepositional phrases that were extracted in a fully automated manner were evaluated in terms of a linguistic classification proposed by Mel'čuk (1995). This theory was selected because it is operationalizable and should conceivably be able to account for all possible set phrases in a language as well as provide a comprehensive basis for their description.

For each of the 24 prepositions that were investigated, 100 collocations that were extracted using the weighted frequency ratio were manually annotated. The reason for assigning formal classes to the extracted collocations was that I attempted to determine the reliability of the results that were obtained by the

automatic collocation extraction from a theoretical point of view and also to ascertain whether quantitative methods are a useful and successful means of defining lexical relationships in raw text data. My objective was to investigate the shortcomings in describing empirical collocations in terms of strict classes and to demonstrate that the formal division of automatically extracted collocations into several non-overlapping classes is likely to be impossible.

The extracted stable expressions have displayed a tendency to form idiomatic lexical units, ranging from no participation in such expressions (for example, *nad* 'above' and *o* 'about') to frequent usage in expressions with varying degrees of idiomaticity (for example, *bez* 'without' as in *bez durakov* 'no kidding', lit. 'without fool.Gen.Pl' or *bez vesti* 'without a trace', lit. 'without piece of news.Gen.Sg'). In general, these units tend to occur in word combinations with no idiomatic meanings. This is accounted for by the limited overall number of Russian MWEs with prepositions. Consequently, algorithms will inevitably extract stable but not idiomatic word combinations when no idiomatic expressions are available.

The application of linguistic theory to the results is further complicated by a disparity between theory and practice: A clear division between different types of MWEs might technically be possible, but there are a large number of borderline cases. This limitation serves as motivation for considering empirical collocations as a continuum with space for both idiomatic and non-idiomatic units. Further analysis of this heterogeneous group of the extracted units has revealed that the latter may, in turn, form clusters. Members of such clusters, referred to as constructions, have similarities in their meaning, and these clusters reflect the distributional preferences of the headword. They are briefly examined in this article, and further elaborated on in the article titled *"Constructional generalization over Russian collocations."*

4 TOKENS VERSUS LEXEMES IN COLLOCATIONS

4.1 GRAMMATICAL PROFILING

Every collocation is a complex entity on the morphological level, and the optimal choice to represent a collocation involves the question of whether it is better to use tokens or lexemes and this is not a trivial undertaking. For single tokens, the choice between using tokens or lexemes as granular research units has been addressed by Newman (2008) as well as Newman and Rice (2006), who support analyzing inflected forms, as they have their own constructional and semantic properties (2006, 31). However, as Gries (2011) observes, the choice of an appropriate degree of granularity is challenging. He provides evidence based on an analysis of English argument constructions, suggesting that qualitative alterations in a token-based analysis are usually meaningless, although some of the quantitative characteristics are changed (2011, 249). Janda and Lyashevskaya (2011, 720) agree that "the appropriate level of granularity is determined by both the language and the linguistic phenomenon under analysis" but by citing Russian aspect as an example, they have also demonstrated that token-based analysis can provide useful insights for morphologically rich languages.

Gries (ibid.) as well as Janda et al. (2009, 2011) base their analysis of single tokens on the *grammatical profiles* of inflected forms. Janda and Lyashevskaja (2011, 719) define *grammatical profile* as "the relative frequency distribution of the inflected forms of a word in a corpus." This concept originates from the Behavioral Profile approach that was introduced by Gries and Divjak (2006). Their analytical orientation studies lexical relations based on broad context information that includes the morphological, syntactic and semantic characteristics of the expression that was gathered for all its usages. To summarize briefly, a behavioral profile provides a statistical analysis of a very detailed annotation of multiple linguistic dimensions (Gries 2010, 334), and a grammatical profile represents one of the dimensions of a behavioral profile.

A collocation often functions as a single unit. Therefore, the same issue of granularity also applies at this level, at least if we focus on syntactically well-formed units, such as those belonging to the [Adjective + Noun] pattern. Even though there is a clear consensus that one can investigate collocations on two different levels (for example, see Evert and Kermes 2003), the underlying motivations have not been discussed in detail and there are few arguments

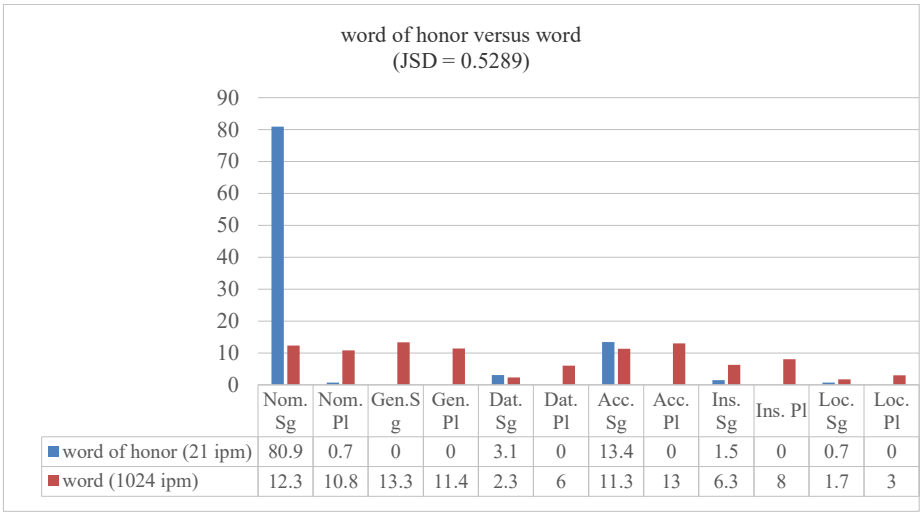
that advocate for one approach over another. Stubbs (2001, 69) notes that "different forms of lemmas may have quite different collocates," which means that an analysis of a whole lexeme may not suffice to cast light on particular usage patterns. Evert et al. (2004, 907) argue, however, that one reason for the analysis of lexeme-level collocations may be syncretic grammatical forms, which do not allow a one-to-one correspondence between surface forms and morphosyntactic features. Furthermore, in case of token-based analysis, the "frequency mass" of the collocations may spread over several different combinations of word forms. Either of the aforementioned reasons prevent a proper statistical analysis. As it relates to the Russian language, the importance of the basic unit for collocation analysis was touched on by Yagunova and Pivovarova (2014). Their article explores the classification of automatically extracted units on the scale between collocations and constructions; they compare lists of collocations extracted at the lexeme and token levels for the MI-measure. This comparison, while informative in terms of the potential options for analyses, does not discuss the underlying motivations in detail.

4.2 CHOOSING BETWEEN LEXEME VERSUS TOKEN IN RUSSIAN COLLOCATIONS

In the article "*Choosing between lexeme vs. token in Russian collocations,"* I aimed to systematically determine the optimal choice for representing collocations and whether it is better to use tokens or lexemes when describing them. This question was encountered each time a unit was selected for automatic collocation extraction. By *token collocations,* I refer to collocations that are only stable in certain forms and do not resemble the full grammatical profile of the headword. For example, according to my research, the collocation *čestnoe slovo* 'word of honor' was attested to be used in the nominative singular in approximately 81% of the occurrences, while its headword is used in this form in only approximately 12% of the occurrences. Their grammatical profiles are thus clearly different (see Figure 2). *Lexeme collocations*[1], in turn, follow the grammatical profile of their headwords. For example, the collocation *medicinskoe strachovanie* 'medical insurance' has

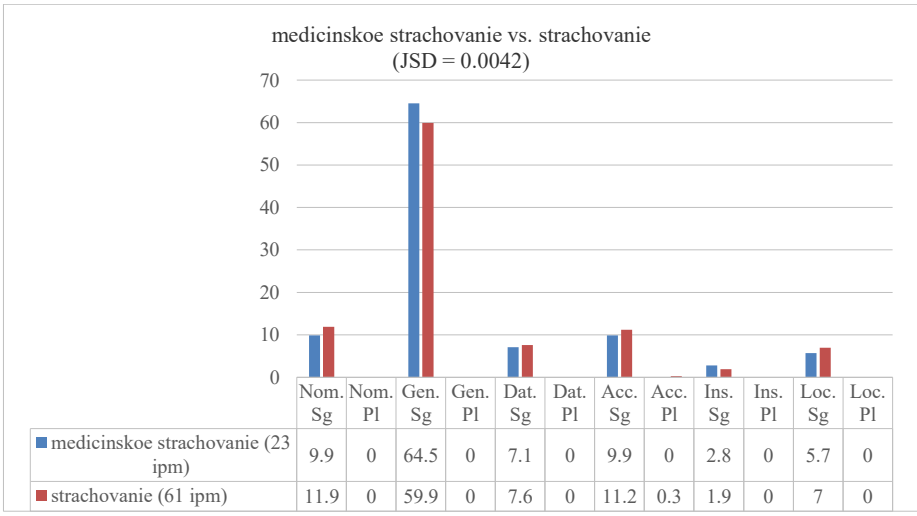
[1] It is important to note that in the earlier articles, the term *lemma collocation* was used to refer to *lexeme collocation*. Later, this term was changed to *lexeme collocation* because the latter reflects the essence of the phenomenon more accurately in that it emphasizes that the collocation is considered to be stable in the whole set of attested forms as opposed to single ones. As for the term *lemma collocation,* it appears to describe the representation of a collocation that can be stable by itself in any number of its forms.

approximately the same distribution of grammatical forms as *strachovanie* alone: The trends are the same and the maximum difference attested for the genitive singular is only about 5% (see Figure 3).



	Nom. Sg	Nom. Pl	Gen.Sg	Gen. Pl	Dat. Sg	Dat. Pl	Acc. Sg	Acc. Pl	Ins. Sg	Ins. Pl	Loc. Sg	Loc. Pl
word of honor (21 ipm)	80.9	0.7	0	0	3.1	0	13.4	0	1.5	0	0.7	0
word (1024 ipm)	12.3	10.8	13.3	11.4	2.3	6	11.3	13	6.3	8	1.7	3

Figure 2 Grammatical profiles for *čestnoe slovo* 'word of honor' and *slovo* 'word' (Kormacheva 2019, 14)



	Nom. Sg	Nom. Pl	Gen. Sg	Gen. Pl	Dat. Sg	Dat. Pl	Acc. Sg	Acc. Pl	Ins. Sg	Ins. Pl	Loc. Sg	Loc. Pl
medicinskoe strachovanie (23 ipm)	9.9	0	64.5	0	7.1	0	9.9	0	2.8	0	5.7	0
strachovanie (61 ipm)	11.9	0	59.9	0	7.6	0	11.2	0.3	1.9	0	7	0

Figure 3 Grammatical profiles for *medicinskoe strachovanie* 'medical insurance' and *strachovanie* 'insurance' (Kormacheva 2019, 14)

To provide a solid reason for choosing between token and lexeme collocations in the future, I consciously examined the distributional preferences for collocations as compared to those of their headwords and utilized grammatical profiling as a means of collocational analysis. I had two objectives in this phase of my research. First, I aimed to demonstrate that the distinction between token and lexeme collocations makes sense in terms of their different distributional preferences across the corpus. Second, I offered a plausible method for differentiating between the two types based on corpus data on the usage of the collocations. This method could be used further to explore the correlation between formal stability and semantic idiomaticity. In order to achieve this, a ranking is needed of a more comprehensive list of collocations according to their token/lexeme nature and this exceeds what had been planned within the current research. When applied to the 100 most frequent collocations, no clear correlation was found, as this list did not include much variability in the degree of idiomaticity and could not be a sufficient basis for this type of investigation.

This analysis demonstrates that both token and lexeme collocations are attested in corpus data and that neither is clearly predominant. Furthermore, in terms of grammatical profiling of collocations, they can again be placed on a scale between these two types. Therefore, depending on the particular task in question and the amount of data available, both of them can be used as a study unit when the choice is made consciously. In general, using token collocations, either alone or in bundles, appears to describe reality more accurately. In this case, both lexeme collocations and those that constitute an intermediate phenomenon between token and lexeme collocations (stable in several forms but not throughout the entire paradigm) can be presented as sets of token collocations. A measure that has been suggested for differentiating these two types of collocations places them on a scale according to their grammatical boundness, which is based on the differences in the grammatical profiles of a collocation and its headword. The obtained results can then be used either as they are or as a helpful basis for further manual analysis by a human expert, such as a lexicographer, as this would ease the task of providing lexical entries with relevant usage patterns.

5 CONSTRUCTIONS IN COLLOCATIONS

5.1 CONSTRUCTION GRAMMAR

When we focus on semantic, lexical and syntactic characteristics of the retrieved collocations as a constellation of properties, we follow the principles of Construction Grammar that were first formulated by Lakoff (1987) and Fillmore et al. (1988) and developed later by many other scholars, including a considerable contribution by Goldberg (1996, 2006). A detailed discussion of the main principles of this approach is presented in *The Oxford Handbook of Construction Grammar* (2013) or in the work of Fried and Östman (2004). The core presupposition within this theory is that no clear-cut division is made between lexicon and grammar, but instead there is a continuum, often called a *constructicon*. The units that this continuum consists of are constructions.

> A construction is <…> a pairing of form with meaning/use such that some aspect of the form or some aspect of the meaning/use is not strictly predictable from the component parts or from other constructions already established to exist in the language.
>
> (Goldberg 1996, 68)

The form-meaning pairing can be applied to all levels of grammatical description. Constructions thus differ in their formal and semantic complexity and include morphemes, words, idioms and abstract phrasal patterns, such as the English ditransitive subcategorization frame [S V O_i O_d], which is exemplified by "John gave Mary a book." (Gries and Stefanowitsch 2004, 212). Although constructions from the different sides of the continuum differ in their degree of idiomaticity (they range from idiosyncratic phenomena to productive patterns), all of them can be analyzed similarly. This serves as a good starting point for an account of all the automatically extracted empirical collocations in the same manner. Furthermore, Goldberg (2006, 64) claims in her later work that frequently used word sequences are constructions even when they are not idiosyncratic in meaning or form. Over time, these conventionalized sequences may develop special meanings. The frequency of usage that we observed to be an inherent property of empirical collocations may thus constitute a sufficient basis for a semantic shift in the meaning in the future.

Grammatical constructions also include abstract phrasal patterns, and these productive patterns may be instantiated in several particular instances

and thus account for the collocations that are not idiomatic alone but should rather be considered as a holistic phenomenon. This connection between constructions and lexemes that are drawn to each other was investigated by Gries and Stefanowitsch (2003) from the perspective of collostructions. They proposed a type of analysis that could account for these units: "Collostructional analysis always starts with a particular construction and investigates which lexemes are strongly attracted or repelled by a particular slot in the construction." (ibid., 214).

In other words, collostructions are the intermediate units that can account for the stability as well as idiomaticity (when present) of constructions when one member is a group of semantically close items. This way of formulating the problem directly correlates with the ideas of distributional semantics that were first formulated by Harris (1954, 156), who suggested that words that are similar in meaning tend to have a similar distribution. This distributional hypothesis was later supported by Firth's famous statement that "a word is characterized by the company it keeps" (1957, 11) and subsequently by many other scholars (for an overview, see Sahlgren 2008).

5.2 CONSTRUCTIONAL GENERALIZATION OVER RUSSIAN COLLOCATIONS

In the article *"Constructional generalization over Russian collocations,"* following Construction Grammar, we assume that the border between grammar and lexicon is vague, and that there are stable constructions that are formed by a group of lexical items that can fill them. We elaborate on observations about the diverse nature of the extracted expressions and the first constructional generalizations that were made in the previous articles. In this article, we approach the distributional hypothesis from a different angle: Instead of examining the contexts of several lexemes to define whether or not these lexemes are similar, we assess all the contexts of one particular lexeme in order to detect the common semantic features of these contexts. This produces a list of potential candidates for sharing a similar meaning because in this case, the headword itself forms a similar context for all its collocates. This list is then manually analyzed, and groups of semantically close lexemes are identified.

The aim of this article is thus to investigate the reasons for the co-occurrence of lexical items and to elaborate on the idea of *constructions* on the abstract level of generalization. Our objective was to perform a qualitative analysis of automatically extracted Russian collocations that could account for the highest-ranked items (for example, by *t-score*), which appear to be the

most stable ones lexically, and also for an extended list of the units that were extracted using a given measure. The aim was to determine the degree to which collocates of a given lexeme tend to share their semantic properties. In other words, we attempted to establish the proportion of collocates that participate in constructions in which lexical items are substitutable but still belong to the same semantic class, even when these collocates are not necessarily non-compositional on their own.

The results have shown that many automatically extracted empirical collocations are not idiomatic but can form clusters of words that belong to the same semantic class. For example, the collocation *goristoe mesto* 'mountainous place' is not fully idiomatic but it nevertheless constitutes a part of the construction [Adj'type of relief' + *mesto* 'place'], whose other variables include, *vozvyšennoe* 'high/elevated' or *nizmennoe* 'low-lying'. Fixed expressions can also represent elements of constructions, such as *žiznennyj vopros* lit. 'vital question' ('problem of life') is a part of the construction [Adj'of great importance' + *vopros* 'question/issue'], whose other variables include *složnyj* 'difficult', and *važnyj* 'important'. For the [Adjective + Noun] collocations, a total of 56% of the units extracted using the weighted frequency ratio can be accounted for by constructions. The number of idiomatic expressions among them is far more modest, and these expressions may or may not intersect with the expressions that participate in the constructions. The most productive pattern among the analyzed constructions was [*molo´doj* 'young' + Noun], whose top 100 collocates participated in 6 different constructions:

- [*molo´doj* 'young' + Noun'name/surname'] as in *molodoj pevec* 'young singer',
- [*molo´doj* 'young' + Noun'nationality'] as in *molodoj britanec* 'young Briton',
- [*molo´doj* 'young' + Noun'animal'] as in *molodoj l'venok* 'young lion',
- [*molo´doj* 'young' + Noun'bird'] as in *molodoj bekas* 'young snipe',
- [*molo´doj* 'young' + Noun'type of forest'] as in *molodoj osinnik* 'young aspen forest'.

Only 3 collocations out of 100 have non-generalizable meanings that are idiosyncratic and cannot be grouped together. Examples of these are *molodaja gvardija* 'young guard', *molodoe pokolenie* 'young generation' and *molodoj čelovek* 'young man'. Thus, interpreting the results of automatic collocation extraction can be considerably improved by assessing the information on participation in the constructions. The more theoretical outcome is that many idioms are not as unique as they may appear; a single idiom is often the most stable and developed representation of a whole construction that defines and supports the idiom.

6 DISCUSSION

The relationship between empirical and lexical collocations permeates current research, and my findings have provided insight into this subject. Empirical collocations were selected as a departure point because they are directly observable textual units and serve as a basis for many language processing tasks. Still, throughout the articles, I attempted to demonstrate that empirical collocations are not as different from lexical collocations as one might presume, and also that empirical collocations are often linguistically motivated. These two categories of collocations fulfil different purposes: The primary focus of empirical collocations is stability, while the essential and required property of lexical collocations is a certain degree of semantic idiomaticity and irregularity. Despite these differences, empirical and phraseological views on collocation are highly interrelated and their constituent units overlap substantially. When studying one of them, the other cannot be avoided. Many of the automatically extracted expressions are idiomatic to varying degrees and could be included in a dictionary. However, at the same time, many lexical collocations that are listed in dictionaries can be extracted from corpora with the same tools.

This research demonstrated that the two phenomena in question are related in several ways. Based on this research, I can conclude that one plausible explanation for the overlap of lexical and empirical collocations is that stability and idiomaticity tend to reflect each other. Empirical collocations undergo several stages on their way to becoming more lexicalized. High frequency of usage as well as the resulting stability allows us to perceive an expression as a complete entity. In time, these stable expressions may displace other synonymous expressions or even acquire additional meanings that are not directly deducible from their parts. At the same time, phraseological units, and herewith lexical collocations, have acquired special meaning, and this idiomatic means of expressing things often becomes a preferred one, gaining thus a stronger degree of stability. Still, idiomaticity is not a required characteristic for empirical collocations, and they should rather be placed on a stable-idiomatic continuum than classified in terms of strict classes. On one side of this continuum are expressions that are stable, but not idiomatic, while on the other are highly idiomatic expressions. Idiomaticity becomes a feature that varies to a degree. Indeed, automatic collocation extraction tools can only provide us with a means of positioning collocations on this scale. They cannot determine the exact threshold where either stableness or idiomaticity ends,

but they provide a useful basis for further insights during manual analysis and can certainly detect the most stable units that often will be idiomatic to some degree.

The existence of this stable-idiomatic continuum can be illustrated further by the fact that some of the automatically extracted collocations can be successfully unified under certain semantic classes and they can therefore participate in constructions. Constructions can thus offer an additional dimension to being stable, even if its members are not stable. Over time, some of them can become crystallized into lexical collocations, and this transformation occurs in the same way as with single empirical collocations where frequent usage coins special meaning as, for example, in *bez galstuka* (lit. 'without a neck-tie') that is also used idiomatically as in *vstreča bez galstuka*, 'an informal meeting'). Finally, I have found that although some lexical collocations tend to be used in a restricted number of all theoretically possible forms and many of the empirical collocations preserve the full grammatical profile of their headwords, in practice, the convergence of the two types of collocations is also reflected in the vague boundary between them. A large number of empirical collocations have only partly similar grammatical profiles of their headwords where not one but several forms are distinguishable.

The perception of empirical and lexical collocations as two interconnected phenomena that can be examined together instead of two separate entities, can have many important implications. This should increase the use of automatic methods in lexicographic practice, as information that is obtained empirically is a direct source of the speech/texts generated by language users. Electronic dictionaries potentially have no limit to the amount of information that can be included in them without affecting user experience negatively. Thus, it is of utmost importance to include full statistics on collocational, such as information their grammatical and lexical profiles, This information is especially useful for second language learners who do not have the same intuition for a language as native speakers have. The correct choice of collocates can indeed be a source of struggle for language learners and the development of tools that explain how words tend to co-occur will significantly facilitate the whole learning process.

This research has brought empirical and phraseological notions of collocations closer together. It has demonstrated that collocations are a complex phenomenon that is based on many characteristics. For this reason, it is not possible to define a set of properties that are inherent to all collocations because they are highly heterogeneous. Still, the possible properties, such as, stability, idiomaticity and the restrictiveness of the grammatical profile, can be combined in different ways, and the resulting collocations can then be

placed on the same plane but either in the core or periphery area. This tight bond that was discovered between empirical and lexical collocations is a direct indication that they should be analyzed together whenever possible and that any information available on one of them should be utilized and incorporated when studying the other.

7 LIST OF ORIGINAL PUBLICATIONS

This thesis consists of four articles that are listed below and they summarize the research I conducted within the CoCoCo project. The work includes reviewing relevant literature on the topic, discussing the structure of the co-authored articles, formulating goals and hypotheses, preparing and annotating the data, running the experiment, calculating statistics, analyzing the linguistic data, and evaluating the obtained results. My contributions to the co-authored articles are defined in parentheses below.

1. Kopotev, Mikhail, **Daria Kormacheva**, and Lidia Pivovarova. "Evaluation of collocation extraction methods for the Russian language." *Quantitative Approaches to the Russian Language*. Routledge, 2017, pp. 137–157.

(I was responsible for the evaluation of the performance of five measures and raw frequency against the dictionary data described in Chapter 3.1 of the article that included a comparison of two sources and analysis of the obtained results. I also contributed to preparing the experiment described in Chapter 3.2 (p. 146 in the published article), and I performed a qualitative analysis of the obtained experiment data (p. 150 in the published article). I also made major contributions to the introduction (Chapter 1) and the conclusion (Chapter 5) of the article.)

2. **Kormacheva, Daria**. "What do we get from extracting collocations? Linguistic analysis of automatically obtained Russian MWEs." *Journal of Research Design and Statistics in Linguistics and Communication Science* vol. 1, no. 2, 2015, pp. 169–189.

3. **Kormacheva, Daria**. "Choosing between lexeme vs. token in Russian collocations." *Scando-Slavica*, vol. 65, no. 1, 2019, pp. 77–93.

4. Kopotev, Mikhail, **Daria Kormacheva**, and Lidia Pivovarova. "Constructional generalization over Russian collocations." *Mémoires de la Société néophilologique de Helsinki*, 2016, pp. 121–140.

(I was responsible for the analysis of the 3 000 collocations that is described in Chapter 3 of the article. The task included manual annotation and linguistic analysis of these items in order to determine their idiomaticity and the proportion they constituted in the empirical collocation that participated in

constructions, as well as analysis of the obtained results. I also contributed extensively to the introduction (Chapter 1) and conclusion (Chapter 5) of the article.

REFERENCES

Apresjan, Valentina, Vít Baisa, Olga Buivolova, Olga Kultepina, Anna Maloletnjaja. "RuSkELL: online language learning tool for Russian language." *Proceedings of the XVII EURALEX International congress*, 2016, pp. 292–299.

Baranov, Anatolij, and Dmitrij Dobrovolskij. *Aspekty teorii frazeologii*. Moskva: Znak, 2008.

Bartsch, Sabine. *Structural and functional properties of collocations in English: A corpus study of lexical and pragmatic constraints on lexical co-occurrence*. Gunter Narr Verlag, 2004.

Biber, Douglas. "Representativeness in corpus design." *Literary and linguistic computing*, vol. 8, no. 4, 1993, pp. 243–257.

Biber, Douglas. "A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing." *International journal of corpus linguistics*, vol. 14. no. 3, 2009, pp. 275–311.

Blaheta, Don, and Mark Johnson. "Unsupervised learning of multi-word verbs." *39th Annual Meeting and 10th Conference of the European Chapter of the Association for Computational Linguistics*, 2001, pp. 54–60.

Braslavski, P., and E. Sokolov. "Sravnenie četyrekh metodov avtomatičeskogo izvlečenija dvuslovnykh terminov iz teksta." *Komp'juternaja lingvistika i intellektual'nye tekhnologii: Trudy meždunarodnoj konferencii Dialog*, 2006, pp. 88–94.

Calzolari, Nicoletta, Charles J. Fillmore, Ralph Grishman, Nancy Ide, Alessandro Lenci, Catherine MacLeod, and Antonio Zampolli. "Towards Best Practice for Multiword Expressions in Computational Lexicons." *LREC*, 2002, pp. 1934–1940.

Church, Kenneth Ward, and Patrick Hanks. "Word association norms, mutual information, and lexicography." *Computational linguistics*, vol. 16, no. 1, 1990, pp. 22–29.

Church, Kenneth, William Gale, Patrick Hanks, and Donald Hindle. "Using statistics in lexical analysis." *Lexical acquisition: exploiting on-line resources to build a lexicon*, 1991, pp. 115–164.

Daudaravicius, Vidas. "Automatic identification of lexical units." *Informatica*, vol. 34, no. 1, 2010, pp. 85–92.

Divjak, Dagmar, and Stefan Th. Gries. “Ways of trying in Russian: clustering behavioral profiles.” *Corpus Linguistics and Linguistic Theory*, vol. 2, no. 1, 2006, pp. 23–60.

Dobrovolskij, Dmitrij, and Ludmila Pöppel. “The discursive construction n v tom, čto and its parallels in other languages: a contrastive corpus study.” *Science for Education Today*, vol. 6, no. 6, 2016, pp. 164–175.

Dunning, Ted. “Accurate methods for the statistics of surprise and coincidence.” *Computational linguistics*, vol. 19, no. 1, 1993, pp. 61–74.

Erman, Britt, and Beatrice Warren. “The idiom principle and the open choice principle.” *Text-Interdisciplinary Journal for the Study of Discourse*, 2000, pp. 29–62.

Evert, Stefan, and Brigitte Krenn. “Methods for the qualitative evaluation of lexical association measures.” *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, 2001, pp. 188–195.

Evert, Stefan, and Brigitte Krenn. “Using small random samples for the manual evaluation of statistical association measures.” *Computer Speech & Language*, vol. 19, no. 4, 2005, pp. 450–466.

Evert, Stefan, and Hannah Kermes. “Experiments on candidate data for collocation extraction.” *Companion Volume to the Proceedings of the 10th Conference of The European Chapter of the Association for Computational Linguistics*, 2003, pp. 83–86.

Evert, Stefan, Ulrich Heid, and Krisitina Spranger. “Identifying morphosyntactic preferences in collocations.” *Proceedings of the 4th International Conference on Language Resources and Evaluation*, 2004, pp. 907–910.

Evert, Stefan. *The statistics of word cooccurrences: word pairs and collocations*, 2005.

Evert, Stefan. “Corpora and collocations.” *Corpus linguistics. An international handbook*, 2008, pp. 1212–1248.

Fillmore, Charles, Paul Kay, and Catherine O’Connor. “Regularity and Idiomaticity in Grammatical Constructions: The Case of let alone.” *Language*, vol. 64, 1988, pp. 501–538.

Fillmore, Charles J. “‘Corpus Linguistics’ or ‘Computer-aided armchair linguistics’.” *Directions in corpus linguistics. Proceedings of Nobel Symposium*, vol. 82, 1992, pp. 35–60.

Firth, John R. “A synopsis of linguistic theory 1930–1955.” *Studies in Linguistic Analysis*, 1957.

Frank, Stefan L., Rens Bod, and Morten H. Christiansen. "How hierarchical is language use?" *Proceedings of the Royal Society B: Biological Sciences* 279, no. 1747, 2012, pp. 4522–4531.

Fried, Mirjam, and Jan-Ola Östman, eds. *Construction grammar in a cross-language perspective*. Vol. 2. John Benjamins Publishing, 2004.

Glynn, Dylan, and Kerstin Fischer, eds. *Quantitative methods in cognitive semantics: Corpus-driven approaches*. Vol. 46. Walter de Gruyter, 2010.

Goldberg, Adele E. "Construction grammar." *Concise encyclopedia of syntactic theories*. Edited by Brown, Edward Keith, and James Edward Miller. Pergamon Press, 1996, pp. 401–437.

Goldberg, Adele E. *Constructions at work: The nature of generalization in language*. Oxford University Press on Demand, 2006.

Gries, Stefan Th., and Anatol Stefanowitsch. "Extending collostructional analysis: A corpus-based perspective on alternations." *International journal of corpus linguistics*, vol. 9, no. 1, 2004, pp. 97–129.

Gries, Stefan Th., and Anatol Stefanowitsch, eds. *Corpora in cognitive linguistics: corpus-based approaches to syntax and lexis*. Walter de Gruyter, vol. 172, 2007.

Gries, Stefan Th. "Behavioral profiles: A fine-grained and quantitative approach in corpus-based lexical semantics." *The Mental Lexicon*, vol. 5, no. 3, 2010, pp. 323–346.

Gries, Stefan Th. "Studying syntactic priming in corpora: implications of different levels of granularity." *Converging evidence: Methodological and theoretical issues for linguistic research*, 2011, pp. 143–165.

Harris, Zellig S. "Distributional structure." *Word*, vol. 10, no. 2-3, 1954, pp. 146–162.

Jackendoff, Ray. *The architecture of the language faculty*, no. 28. MIT Press, 1997.

Janda, Laura A., and Olga Lyashevskaya. "Grammatical profiles and the interaction of the lexicon with aspect, tense, and mood in Russian." *Cognitive linguistics*, vol. 22, no. 4, 2011, pp. 719–763.

Janda, Laura A., and Valery D. Solovyev. "What constructional profiles reveal about synonymy: A case study of Russian words for SADNESS and HAPPINESS." *Cognitive Linguistics*, vol. 20, no. 2, 2009, pp. 367–393.

Khokhlova, Maria. "Extracting collocations in Russian: Statistics vs. dictionary." *Proceedings of 9th International Conference on Textual Data statistical Analysis*, 2008, pp. 613–624.

Khokhlova, Maria. "Osobennosti statističeskih mer pri vydelenii bigramm." *Korpusnaja lingvistika–2017*, 2017, pp. 349–354.

Klyshinsky, E.S., N.Y. Lukashevich, and I.M. Kobozeva. “Creating a Corpus of syntactic co-occurrences for russian.” *Proceedings of the International Conference Dialogue 2018,* 2018, pp. 317–330.

Kopotev, Mikhail, Lidia Pivovarova, Natalia Kochetkova, and Roman Yangarber. “Automatic detection of stable grammatical features in n-grams.” *NAACL HLT 2013*, 2013, pp. 73–81.

Kopotev, Mikhail, Llorenc Escoter, Daria Kormacheva, Matthew Pierce, Lidia Pivovarova, and Roman Yangarber. “CoCoCo: Online Extraction of Russian Multiword Expressions.” *Proceedings of the 5th Workshop on Balto-Slavic Natural Language Processing*, 2015, pp. 43–45.

Kopotev, Mikhail, and Tatiana Steksova. *Isklučenie kak pravilo: Perehodnye edinicy v grammatike i slovare*. Litres, 2017.

Kopotev, Mikhail, Olga Lyashevskaya, and Arto Mustajoki. “Russian challenges for quantitative research.” *Quantitative Approaches to the Russian Language*, 2017, pp. 3–29.

Kormacheva, Daria, Lidia Pivovarova, and Mihail Kopotev. “Automatic collocation extraction and classification of automatically obtained bigrams.” *Proceedings Workshop on Computational, Cognitive, and Linguistic Approaches to the Analysis of Complex Words and Collocations*, 2014, pp. 27–33.

Lakoff, George. *Women, fire, and dangerous things. What Categories Reveal about the Mind*. University of Chicago press, 1987.

Leech, Geoffrey. “Introducing corpus annotation.” *Corpus annotation.* Routledge, 1997, pp. 11-28.

Lin, Jianhua. “Divergence measures based on the Shannon entropy.” *IEEE Transactions on Information theory*, vol. 37, no. 1, 1991, pp. 145–151.

Manning, Christopher D., and Hinrich Schütze. *Foundations of statistical natural language processing*. MIT press, 1999.

McEnery, Tony, Richard Xiao, and Yukio Tono. *Corpus-based language studies: An advanced resource book*. Taylor & Francis, 2006.

Mel’čuk, Igor. “Phrasemes in language and phraseology in linguistics.” *Idioms: Structural and psychological perspectives*, 1995, pp. 167–232.

Mel’čuk, Igor. “Collocations and lexical functions.” *Phraseology. Theory, Analysis, and Applications*, 1998, pp. 23–53.

Men, Haiyan. “The Notion of Collocation.” *Vocabulary Increase and Collocation Learning*, 2018, pp. 9–33

Miller, George A. *WordNet: An electronic lexical database*. MIT Press, 1998.

Moirón, Begona Villada, and Jörg Tiedemann. “Identifying idiomatic expressions using automatic word-alignment.” *Proceedings of the Workshop on Multi-word-expressions in a multilingual context*, 2006.

Moon, Rosamund. *Fixed expressions and idioms in English: A corpus-based approach*. Oxford University Press, 1998.

Newman, John, and Sally Rice. “Transitivity schemas of English EAT and DRINK in the BNC.” *Corpora in cognitive linguistics: Corpus-based approaches to syntax and lexis*, 2006, pp. 225–260.

Newman, John. “Aiming low in linguistics: Low-level generalizations in corpus-based research.” *Proceedings of the 11th International Symposium on Chinese Languages and Linguistics*, 2008.

Nesselhauf, Nadja. *Collocations in a learner corpus*. Vol. 14. Amsterdam: John Benjamins, 2005.

Pearce, Darren. “A Comparative Evaluation of Collocation Extraction Techniques.” *LREC*, 2002, pp. 1530–1536.

Pecina, Pavel, and Pavel Schlesinger. “Combining association measures for collocation extraction.” *Proceedings of the COLING/ACL on Main conference poster sessions*, 2006, pp. 651–658.

Penttilä, Esa. “It takes an age to do a Chomsky: Idiomaticity and verb phrase constructions in English.” *Unpublished doctoral dissertation, University of Joensuu*, 2006.

Piperski, A. “Intra-speaker stress variation in Russian: a corpus-driven study of Russian poetry.” *Komp’juternaja lingvistika i intellektual’nye tehnologii*, 2016, pp. 540–550.

Plungjan, V.A. (ed.) *Nacional’nyj korpus russkogo jazyka: 2003–2005. Rezul’taty i perspektivy*, 2005.

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. “Multiword expressions: A pain in the neck for NLP.” *Computational Linguistics and Intelligent Text Processing*, 2002, pp. 1–15.

Sahlgren, Magnus. “The distributional hypothesis.” *Italian Journal of Disability Studies*, vol. 20, 2008, pp. 33–53.

Sharoff, Serge, and Joakim Nivre. “The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge.” *Proc. Dialogue 2011, Russian Conference on Computational Linguistics*, 2011, pp. 657–670.

Shavrina, Tatiana, and Olga Shapovalova. “To the methodology of corpus construction for machine learning: “Taiga” syntax tree corpus and parser.”

Proceedings of the International Conference Corpus Linguistics–2017, 2017, pp. 78–84.

Sinclair, John. *Corpus, concordance, collocation*. Vol. 1. Oxford: Oxford University Press, 1991.

Sinclair, John. "The lexical item." *Amsterdam studies in the theory and history of linguistic science series*, vol. 4, 1998, pp. 1–24.

Smadja, Frank. "Retrieving Collocations from Text: Xtract." *Computational Linguistics*, vol. 19, no. 1, 1993, pp. 143–177.

Stefanowitsch, Anatol, and Stefan Th. Gries. "Collostructions: Investigating the interaction of words and constructions." *International journal of corpus linguistics*, vol. 8, no. 2, 2003, pp. 209–243.

Stubbs, Michael. *Words and phrases: Corpus studies of lexical semantics*. Blackwell Publishers, 2001.

Tashakkori, Abbas, and John W. Creswell. *The new era of mixed methods*, 2007, pp. 3–7.

Thanopoulos, Aristomenis, Nikos Fakotakis, and George Kokkinakis. "Comparative Evaluation of Collocation Extraction Metrics." *LREC*, vol. 2, 2002, pp. 620–625.

Tognini-Bonelli, Elena. *Corpus linguistics at work*. Vol. 6. John Benjamins Publishing, 2001.

Wiechmann, Daniel. "On the computation of collostruction strength: Testing measures of association as expressions of lexical bias." *Corpus Linguistics and Linguistic Theory* vol. 4, no. 2, 2008, pp. 253–290.

Wermter, Joachim, and Udo Hahn. "You can't beat frequency (unless you use linguistic knowledge): a qualitative evaluation of association measures for collocation and term extraction." *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, 2006, pp. 785–792.

Wolf, Florian, and Edward Gibson. "Representing discourse coherence: A corpus-based study." *Computational linguistics*, vol. 31, no. 2, 2005, pp. 249–287.

Yagunova, Elena, and Lidia Pivovarova. "The nature of collocations in the Russian language. The experience of automatic extraction and classification of the material of news texts." *Automatic Documentation and Mathematical Linguistics*, vol. 44, no. 3, 2010, pp. 164–175.

Yagunova, Elena, and Lidia Pivovarova. "Ot kollokacij k konstrukcijam." *Trudy instituta lingvističeskich issledovanij, 2. Russkij jazyk: grammatika konstrukcij i leksiko-semantičeskie podchody*, 2014, pp. 568–61