7 Evaluation of collocation extraction methods for the Russian language

Lidia Pivovarova, Daria Kormacheva, and Mikhail Kopotev

1. Introduction

Studies of collocations can be traced back to classical works by John Firth, Michael Halliday, John Sinclair, and Igor Mel’čuk, to mention but a few. Today, the term “collocation” is used in linguistic literature in many different ways. Stefan Evert (2008, pp. 3–4) proposed a distinction between *lexical* and *empirical* collocations: the former is an object of linguistic analysis, while the latter is an output of computational algorithms. In this paper, we focus on empirical collocations, which we understand as co-occurrences that 1) are frequent enough to be extracted automatically and 2) may be semantically and/or syntactically bounded to various extents. Defined in this way, empirical collocations largely overlap with multiword expressions in general, which are defined as a “sequence of words that acts as a single unit at some level of linguistic analysis” (Calzolari et al., 2002, p. 1934) and represent a complex notion that includes several phenomena, such as multiword lexemes, terminological expressions, phrasal verbs, idioms, and collocations. To make a clear distinction between multiword expressions and empirical collocations, the latter are understood as “recurrent and predictable word combinations, which are a directly observable property of natural language” (Evert, 2008, p. 3); thus, they are extractable statistically without deeper linguistic analysis. The questions that arise are, first, how well do different computational methods extract *lexical* collocations, and, second, how do the extracted data correspond to each other, as well as to expert evaluation and speakers’ intuition? Thus, our main focus here is to determine to what extent each method is reliable beyond its statistical significance *per se*.

In this study, we examine closely the computational methods for empirical collocation extractions that are widely used in corpus-based studies, sometimes without proven efficiency. Our study takes a look at five window-based methods, i.e., methods based on strings of words and ignoring syntactic relations (Evert and Krenn, 2001), and evaluates their reliability with Russian data. Transferring methods that work for one language (primarily English) to other languages may not preserve the same level of accuracy for a number of reasons, one of which is the larger number of tokens for each lemma in a morphologically richer language, such as Russian. Thus, when a method is applied to new language data, it should be validated for this language by comparing it to other known methods and by evaluation by experts and/or ordinary native speakers.

In this paper, we test two hypotheses:

1. Lexical collocations listed in a dictionary (i.e., one considered fixed by experts in the field) should have a higher ranking in automatically extracted lists of empirical collocations.
2. Lexical collocations considered more plausible by native speakers tend to have higher rankings in automatically extracted lists.

The article is divided into four parts. First, we present a short overview of the methods used in automatic collocational analysis. Then, we evaluate these methods against two independent opinions: professional linguists and respondents who are native speakers of Russian. Finally, we discuss the advantages and disadvantages of each method and draw some conclusions.

2. Overview of existing measures for collocation extraction

A great deal of effort has been invested in automatic collocation extraction from corpora. Numerous tools[1] and methods have been used; for example, Wiechmann (2008) surveys 47 statistical methods, while Pecina and Schlesinger (2006) review 82. Even these surveys do not supply a complete list, since newer methods are constantly being developed (see, for example, the program for further research in Gries, 2013). As a result, there is no consensus on which method is the most suitable for collocation extraction. Depending on the data and the ultimate goal of a research project, one or another measure may be best suited to a particular case. However, despite the variety of methods, there is a core set that is *de facto* applied more often. The classical introduction to the topic (Manning and Schütze, 1999, pp. 151–187) describes four methods (t-score, χ^2-test, log-likelihood, MI). Evert and Krenn (2001) list five measures in their earlier paper (MI, log-likelihood, t-score, χ^2-test, and raw co-occurrence frequency), while in his later work, Evert (2008) uses six, somewhat different methods (MI, MI^k, local-MI, z-score, t-score, and simple-ll [=log-likelihood]).

In this paper, we evaluate five statistical metrics as well as raw **frequency**, starting with a brief introduction and an explanation of the basic ideas underlying each (more detailed descriptions can be found in Manning and Schütze, 1999 and Evert, 2005).

1. **t-score** (Church et al., 1991), $t-score(p,w)=\sqrt{f(p,w)}-\dfrac{f(p)\times f(w)}{\sqrt{f(p,w)}}$,

   where p is a pattern, w is a filler of this pattern, *f(p)* and *f(w)* are their frequencies in the corpus, and *f(p,w)* is a frequency of their co-occurrence in the corpus. The **t-score** is a measure based on a statistical Student test. In a way, the first component here is ranking by simple collocation frequency; the second is an adjustment component that decreases the score value for excessively frequent words. A **t-score** ranking is usually similar to ranking by raw frequency, but with the most frequent words filtered out (Stubbs, 1995).

2. **Log-likelihood or LL** measures how likely it is that two words will appear in a corpus independent of each other, given the observed data: $LL = 2\sum_{ij} Oij\, log \frac{Oij}{Eij}$, where i = [p, p] (where p means not p, i.e., all other words in the corpus), j = [w, w], O_{ij} is an observed frequency of ij combination in a corpus, and $Eij = f(i) \times f(j)$ is an expected frequency of this combination, calculated as a product of collocate frequencies. There are some varieties in likelihood calculation; we use the most standard representation of the formula.[2]
3. Pointwise mutual information, **MI** (Church and Hanks, 1990): $MI(p,w) = log \frac{f(p,w)}{f(p) \times f(w)}$. **MI** is a measure of *surprisal* adopted from information theory, which measures how likely it is to see the second collocate, given the first one (and vice versa, since the measure is symmetrical). A known problem with **MI** is that it prefers infrequent collocations; consequently, it is highly sensitive to any noise in the data and should always be used together with frequency filtering (Yagunova and Pivovarova, 2010).
4. The **Dice** score is adopted from set theory, where the overlap between two sets is measured as the size of the interception divided by the size of the union. For collocation extraction, the **Dice** score is usually used in the form $Dice(p,w) = \frac{2 \times f(p,w)}{f(p) + f(w)}$. Ranking by **Dice** is in general similar to that of **MI**, but **Dice** is less sensitive to infrequent collocations, which is why it is better in most practical applications (Daudaravicius, 2010).
5. The weighted frequency ratio, **wFR**, is 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: $wFR(p,w) = \frac{f(w)}{f(p,w)} \times \log(f(w))$. This measure is intended to identify features that appear in the pattern more frequently than in the general corpus while simultaneously giving preference to words that are more frequent in general. In our previous work (Kopotev et al., 2016), **wFR** has proven to be the best means for extracting colligations, i.e., syntactic word preferences. We were able to show that **wFR** works slightly better than raw frequency, since it is less sensitive to the corpus noise. Its efficiency for collocation extractions is under investigation in this paper.

It is not always possible to give a clear interpretation of the numerical values obtained using these measures. The measures are based on different principles; not all of them have clear statistical justification. As a consequence, it is not possible to tell whether the differences between numerical values are significant or not.

Comparison across the measures is even more difficult since the measures are not normalized. Thus, we do not compare the specific values of the measures. Instead, we use each measure to produce *a ranking* in order to compare and evaluate these rankings.

3. Evaluation of automatically obtained collocations

Evaluating collocation extraction – comparing it to morphological taggers or syntactic parsers – is a challenging task for many reasons. First, any lexicon is much larger than a repertoire of grammatical features. Even in Russian, which is a morphologically rich language, there are only 156 morphosyntactic features (e.g., Masculine gender or Past tense), whereas the number of lemmas is a thousand-fold. The second issue is the natural vagueness of lexical connections. While grammatical rules are mainly obligatory, lexical links are probabilistic, and variability is often possible even within full idioms, not to mention collocations. All this makes it difficult to design a proper evaluation of collocations (cf. Evert and Krenn, 2001). Most of the previous evaluations of Russian data were based on the intuition of the evaluators or/and the available dictionaries (see Khokhlova, 2008 for MI, t-score, log-likelihood; Braslavsky and Sokolov, 2006 for freq, χ^2, MI, log-likelihood; Mitrofanova et al., 2008 for MI; Toldova et al., 2013 for MI). As far as we are aware, our research is the first in which five collocation extraction measures and the raw frequency have been tested on a large corpus of data and systematically evaluated both against dictionary compilations and native speakers' responses.

The evaluation is organized in two ways. First (Section 3.1), the automatically obtained rankings were compared to data collected by experts in *A Russian-English Collocational Dictionary of the Human Body* in order to determine the portion of collocations in our corpus that appears in the *Dictionary* and the relative position in the rankings. The second evaluation (Section 3.2) was carried out with a survey designed for native speakers of Russian. We investigated which of the extracted empirical collocations were identified as plausible by native speakers, who generally do not rely on any particular theory but have a feeling for the language they speak. Both of the above-described approaches have their limitations. Although a dictionary is a good example of expert knowledge in the field, it is generally admitted that this kind of source is often not comprehensive and has the disadvantages of being personalized and outdated after a lapse of time. Experiments with native speakers give insight into the current state of a language, but are much more difficult to conduct. The experiments have to be carefully planned, and there are always limitations on the number of examples that can be presented to the participants. However, we believe that a double evaluation allows us to eliminate the weaknesses of each approach and evaluate our results in the most comprehensive way.

In this paper, we focus on nouns denoting parts of the body (e.g., *nos* 'nose,' *serdce* 'heart'), and we investigate collocations that match two patterns: [x.ADJ + NOUN] and [x.VERB + NOUN], where each Noun is from a list of body parts, and the second collocates X are taken from all the combinations found in the corpus. Only lemma collocations are taken into consideration in this research;

i.e., all morphological representations of a given word are summarized here into one lemma. For example, both token collocations *zagoreloe lico* 'tanned face' and *zagorelye lica* 'tanned faces' represent the same lemma collocation *ZAGORELOE LICO* 'tanned face.' All calculations and evaluations are made on bigrams, extracted from the annotated and grammatically disambiguated sub-corpus of the RNC (approximately six million running words). Two groups were manually excluded from the evaluation: bigrams that are part of bigger n-grams (since we do not analyze extended context here), and, for semantically ambiguous nouns, collocations that have a non-body-part meaning (e.g., *fizičeskoe lico* 'juridical person'; Cf. *blednoe lico* 'pale face').

For each word, we collected all of the bigrams that matched the pattern and then ranked them using **Dice**, **t-score**, **log-likelihood**, **MI**, **wFR**, and **frequency**. In that way, we obtained six rankings and compared them to the collocations from *A Russian-English Collocational Dictionary* or from the survey responses. The comparisons of different rankings and evaluations of their effectiveness were carried out using un-interpolated average precision (UAP; Manning and Schütze, 1999, pp. 535–536). The more relevant the results that concentrated at the top of the list, the higher the UAP value, which means that UAP indirectly measures recall. UAP is calculated as follows: at each point *c* of the ranking *r* where a relevant result S_c (here, a dictionary entry or a collocate according to native speakers) is found, the precision $P(S_1 \ldots S_c)$, i.e., the percentage of relevant tokens, is computed and all precision points are then averaged: $UAP(r) = \frac{\sum P(S_1 - S_c)}{|S_c|}$

3.1. *Comparison with dictionary data*

In this section, in order to obtain an overall picture, we evaluate the performance of five measures and the raw frequency against the expertise of highly professional linguists by comparing the corpus-driven results to *A Russian-English Collocational Dictionary of the Human Body*[3] (Iordanskaia et al., 1996). There are several collocational dictionaries for Russian (e.g., Denisov et al., 1978; Bratus et al., 1979), but this one was chosen because it is a practical realization of the Meaning-Text Theory, whereby lexical connections (called "lexical functions" there) are given undivided attention. The foundation of this theory was proposed by Mel'čuk as a systematic description of lexical relations (Mel'čuk, 1995, 2012–2015). An important point where we follow the *Dictionary* and the theory behind it is a classification of fixed expressions, or *phrasemes* (see Mel'čuk, 1995). In particular, we rely on the definition of lexical collocation developed by Mel'čuk in both our *Dictionary* and our native speakers' evaluation:

> A collocation AB of 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' [so that 'X' = 'A ⊕ C')] such that the lexeme B expresses 'C' contingent on A.
>
> (Mel'čuk, 1995, p. 182)

A Russian-English Collocational Dictionary was intended to be one of the most ambitious practical realizations of this theory, which allowed us to rely on its relatively high accuracy in comparing its contents to other sources. At the same time, since the group of words denoting body parts tends to form many expressions in the Russian language, we expected that a full dictionary on this specific topic would cover these expressions satisfactorily.

Overall, only 43 percent of the *Dictionary*'s verbal expressions are represented in our corpus data and only 56 percent of its adjectival expressions. Among the bigrams found in the corpus, approximately 40 percent are found in the *Dictionary*. This would not be a problem in itself, since most of the corpus bigrams are not supposed to be lexical collocations, while the *Dictionary* is meant to provide them. However, there is one substantial limitation of the *Dictionary* data for our research: Because it concentrates on the human body, the *Dictionary* focuses on physical characteristics (such as shape, size, color, aesthetics) and medical conditions and does not cover collocations in which "the body part meaning of the word does not show up in the meaning of the idiom" (Iordanskaia et al., 1996, p. viii). As an illustration, the well-known collocation *zavoevat' serdce* 'to win someone's heart' is not found in the *Dictionary*. Keeping in mind that the *Dictionary*'s coverage is somewhat limited, we are aware that the UAP values are *a priori* lower than they would be if a more complete dictionary were available. Nevertheless, the dictionary comparison has a considerable advantage in that we can rank *all* collocations matching the pattern and then analyze complete rankings. This type of evaluation is impossible with native speakers, because they cannot annotate the complete list of the corpus collocations without losing concentration.

The results of comparing the six rankings with the *Dictionary* are presented in Tables 7.1 and 7.2. As mentioned above, the UAP values are based on precision and take into account the overall ranking of the collocations. The more *Dictionary* entries concentrated at the top of the list, the higher the UAP.

For the [x.ADJ + NOUN] pattern, the most efficient measure for collocation extraction with average UAP=0.60 is the raw **frequency**. It is followed by **t-score**, **Dice**, and **log-likelihood**. For the [x.VERB + NOUN] pattern, the **t-score** demonstrated the best average UAP=0.56. However, as opposed to the first pattern, the **t-score** does not drastically differ from **frequency**, **Dice**, and **log-likelihood**, which show roughly the same performance, only a few hundredths behind. **wFR** and especially **MI** show poorer results for both [x.ADJ + NOUN] and [x.VERB + NOUN] patterns.

Moreover, this straightforward conclusion does not focus on another crucial point, namely, that words have different tendencies to participate in collocations. If we take a closer look at the query [ADJ + *serdce* 'heart'], we see that, regardless of the measure, the UAP is very low: 0.23 for the best measure (**t-score**) and only 0.06 for the least efficient score (**MI**). There are only ten expressions for this pattern found in the *Dictionary*, and five of them appeared in our data: *iskustvennoe serdce* 'artificial heart,' *bol'noe serdce* 'diseased heart/heart in bad condition,' *slaboe serdce* 'weak heart,' *zdorovoe serdce* 'healthy heart,' and *horošee serdce* 'good heart.' In general, our corpus contains 41 bigrams for the main word *serdce*

Table 7.1 The UAP for the [x.ADJ + NOUN] patterns according to A Russian-English Collocational Dictionary of the Human Body

WORD	*DICE*	*T-SCORE*	*LOG-L*	*MI*	*WFR*	*FREQ*
boroda 'beard'	0.64	0.69	0.69	0.54	0.59	0.72
glaz 'eye'	0.56	0.53	0.48	0.27	0.33	0.59
golos 'voice'	0.63	0.61	0.58	0.28	0.41	0.58
lico 'face'	0.46	0.43	0.40	0.28	0.32	0.47
noga 'leg'	0.62	0.67	0.59	0.25	0.34	0.70
nos 'nose'	0.46	0.73	0.60	0.33	0.37	0.82
ruka 'arm/hand'	0.50	0.46	0.40	0.23	0.26	0.52
serdce 'heart'	0.22	0.23	0.22	0.06	0.07	0.22
sleza 'tear'	0.38	0.43	0.40	0.24	0.31	0.41
volos 'hair'	0.73	0.76	0.72	0.47	0.55	0.74
zub 'tooth'	0.58	0.64	0.60	0.45	0.47	0.70
šeja 'neck'	0.49	0.69	0.59	0.40	0.44	0.75
AVG	0.52	0.57	0.52	0.32	0.37	0.60

Table 7.2 The UAP for the [x.VERB + NOUN] patterns according to A Russian-English Collocational Dictionary of the Human Body

WORD	*DICE*	*T-SCORE*	*LOG-L*	*MI*	*WFR*	*FREQ*
glaz 'eye'	0.60	0.60	0.60	0.42	0.51	0.63
golos 'voice'	0.47	0.42	0.45	0.28	0.36	0.35
krov' 'blood'	0.38	0.45	0.41	0.29	0.29	0.42
lico 'face'	0.37	0.33	0.34	0.15	0.24	0.27
noga 'leg'	0.50	0.53	0.48	0.35	0.42	0.52
ruka 'arm/hand'	0.60	0.59	0.56	0.40	0.50	0.57
serdce 'heart'	0.57	0.64	0.58	0.22	0.39	0.62
sleza 'tear'	0.75	0.81	0.78	0.63	0.67	0.81
zub 'tooth'	0.60	0.64	0.60	0.44	0.51	0.61
AVG	0.54	0.56	0.53	0.35	0.43	0.53

'heart,' and it is no surprise that the UAP values are so low, since the fraction of collocations is as low as 0.12 here. Additionally, since the word *serdce* 'heart' is frequently used and refers to more than health and physical conditions, many other expressions are missed in the *Dictionary*, although they appear in our corpus data, e.g., *čistoe serdce* 'pure heart,' *kamennoe serdce* 'heart of stone,' *razbitoe serdce* 'broken heart,' and so on.

Other patterns, for example, [VERB + *sleza* 'tear'], [ADJ + *šeja* 'neck'], [ADJ + *nos* 'nose'], and [ADJ + *noga* 'leg'], tend to form more fixed expressions, as reflected in the higher UAP values – over 0.70 for the best measure (**frequency**) and relatively high as well for the other measures. It is worth noting that the UAP of a ranking for one measure often correlates with the rankings by other measures. In other words, a lower **frequency** or **t-score** also predicts lower **Dice, MI**, or **log-likelihood** scores for the same pattern. This is especially obvious for the [x.VERB + NOUN] patterns (Table 7.2). Thus, we believe that the performance of the measure depends on the collocational preferences of a given word, i.e., on its tendency to participate in fixed expressions.

It is also worth noting that different measures are not equally sensitive to the distributional preferences of words. For example, the empirical collocations *skalit' zuby* 'to bare one's teeth' and *pokazat' zuby* 'to show teeth' are both lexical collocations, but their distributional preferences differ drastically since their meanings are different. *Pokazat'* 'to show' is a light verb in the sense that it has less semantic content and is thus used with many nouns, while *skalit'* 'to bare' is much more restricted in its semantics and is used almost exclusively before the word *zuby* 'teeth' and rarely with its near synonyms (e.g., *klyk* 'fang,' *čeljust'* 'jaw'). Accordingly, the first collocation – *skalit' zuby* 'to bare one's teeth' – has the first or second position in the rankings for all the measures (**Dice, frequency, t-score, log-likelihood, MI**, and **wFR)**, while the other expression – *pokazat' zuby* 'to show teeth' – is in 73rd position with **MI**, 72nd with **Dice**, 68th with **log-likelihood**, and 74th with **wFR** (out of 82). Once again, the **t-score** and **frequency** perform better in this case; in their lists, the collocation *pokazat' zuby* 'to show teeth' is in the 27th and 18th positions, respectively. This demonstrates that **MI**, **Dice**, **log-likelihood**, and **wFR** work well if the collocations have strong distributional preferences. But if both parts of the collocations have broader distribution, they receive quite a low ranking; the situation with the **t-score** is slightly better.

Let us now take a closer look at two case studies. First, adjective collocates for *boroda* 'beard,' which has high values for all the measures, and, second, adjective collocates for *serdce* 'heart,' which has the lowest results of all the measures. In both case studies, we present collocates for the three best measures arranged in descending order.

Case study [ADJ + *boroda* 'beard']

- **Frequency** (UAP=0.72): *černyj* 'black,' ***sedoj* 'gray-haired'**, *okladistyj* **'full,'** *zelenyj* 'green,' *belyj* 'white,' *sinij* 'blue,' ***kurčavyj* 'curly,'** *ognennyj* 'fiery red,' *gustoj* 'thick,' *krasnyj* 'red.'
- **t-score** (UAP=0.69): *černyj* 'black,' ***sedoj* 'gray-haired,'** ***okladistyj* 'full,'** *zelenyj* 'green,' *belyj* 'white,' *sinij* 'blue,' ***kurčavyj* 'curly,'** *ognennyj* 'fiery red,' *gustoj* 'thick,' *krasnyj* 'red.'
- **Dice** (UAP=0.64): ***okladistyj* 'full,'** ***sedoj* 'gray-haired,'** ***kurčavyj* 'curly,'** *ognenno-krasnyj* 'fiery red,' *klinovidnyj* 'wedge-shaped,' *svetlo-rusyj* 'light blond,' *šelkovistyj* 'silky,' *sivyj* 'ash-gray,' *popovskij* 'priest,' *obledenelyj* 'ice-covered.'
- **Log-likelihood** (UAP=0.69): ***okladistyj* 'full,'** ***sedoj* 'gray-haired,'** *krasnyj* 'red,' *serebrjanyj* 'silver,' *dostojnyj* 'worthy,' *svetlyj* 'bright,' *želtyj* 'yellow,' ***kurčavyj* 'curly,'** *zelenyj* 'green,' *ognennyj* 'fiery red.'

Listed above are the top ten items extracted by the **frequency**, **t-score**, **Dice**, and **log-likelihood** (the intersections are shown in bold). This case demonstrates distinct lists for each of the three measures, since only three collocates – *okladistyj* 'full,' *sedoj* 'gray-haired,' and *kurčavyj* 'curly' – are found in all three lists. These three words are also found in the *Russian-English Collocational Dictionary*, but not all of the other items appear there. Given the rankings, **Dice** seems to produce more relevant lists of collocations, which, in Stefan Evert's terms, are lexical, since it ranks at the top of the expressions ranked by both of the other two measures and the *Dictionary*. The ranking by **t-score** is equal to the ranking done with the raw **frequency** (if we consider the top ten collocates). Thus, both of these measures are less applicable if the bigram has a low query frequency. **Dice** allows the extraction of less frequent but more predictable words, for example, *klinovidnyj* 'wedge-shaped,' which is almost exclusively reserved for describing a beard and thus has a strong connection to this noun. The third best measure – **log-likelihood** – shows the same advantage of being less dependent on frequency. However, like the **t-score** and **frequency**, this measure extracts less fixed collocations than does **Dice**.

Case study [ADJ + *serdce* 'heart']

- **Frequency** (UAP=0.22): ***čelovečeskij* 'human,'** ***dobryj* 'good/kind,'** ***bol'noj* 'diseased,'** ***čistyj* 'pure,'** ***ženskij* 'woman,'** *sobstvennyj* 'own,' ***slabyj* 'weak,'** *spokojnyj* 'calm,' *milyj* 'dear,' *tjaželyj* 'heavy,' *russkij* 'Russian.'
- t-score (UAP=0.23): *čelovečeskij* 'human,' *dobryj* 'good/kind,' *bol'noj* 'diseased,' *čistyj* 'pure,' *ženskij* 'woman,' *slabyj* 'weak,' *sobstvennyj* 'own,' *spokojnyj* 'calm,' *milyj* 'dear,' *razbityj* 'broken.'
- **Dice** (UAP=0.22): *razbityj* 'broken,' *donorskij* 'donor,' ***čelovečeskij* 'human,'** ***bol'noj* 'diseased,'** *iskusstvennyj* 'artificial,' ***dobryj* 'good/kind,'** ***ženskij* 'woman,'** *čuvstvitel'nyj* 'sensible,' ***čistyj* 'pure,'** ***slabyj* 'weak.'**
- Log-likelihood (UAP=0.22): *čelovečeskij* 'human,' *dobryj* 'good/kind,' *bol'noj* 'diseased,' *donorskij* 'donor,' *čistyj* 'pure,' *ženskij* 'woman,' *razbityj* 'broken,' *iskusstvennyj* 'artificial,' *slabyj* 'weak,' *spokojnyj* 'calm.'

In this case study, we see that the lists largely overlap: six of ten words are found on all lists. Especially similar are the results for the **frequency**, **t-score**, and **log-likelihood**, where the word rankings are almost the same. In the case of the **t-score**, the ranking is very similar to that of the raw **frequency**. However, for **log-likelihood** the list slightly differs, since it does not correlate that closely with the raw **frequency**. For example, the words *donorskij* 'donor' and *iskusstvennyj* 'artificial,' despite their low query frequency, occupy a higher position on the list (and are not at the top of the **t-score** list). **Dice** introduces some degree of variety as well. Although the set of words is again almost the same, especially compared to the **log-likelihood**, the set is ranked quite differently. We suppose that the similarities in the extracted top lists are due to the nature of collocates. As opposed to the collocates of *boroda* 'beard,' most of the collocates of *serdce* 'heart' are relatively frequent in the corpus overall (for example, for *čelovečeskij* 'human,' ipm=216; for *dobryj* 'good/kind,' ipm=211; and for *čistyj* 'pure,' ipm=181). Since all of the measures more or less take into account differences between corpus and query frequencies, this might lower the sensitivity to the query values.

To summarize, at this point we have established that, while the raw **frequency** for adjectives and the **t-score** for verbs perform slightly better than the other measures, they all provide similar results, and it may be more plausible to suppose that different measures are intended to identify different kinds of collocations.

3.2. Evaluation by native speakers

In an attempt to narrow down the evaluation procedure, we created a questiuonary aimed at evaluating combinations of words, which have a tendency to co-occur together. The components in these expressions are rarely fully predictable and within a given phrase, the choice between several lexical items is possible, e.g., *to apply for a [job/position/presidency]*. When choice is available, *what* is chosen is probabilistic: for instance, in the expressions *to apply for a X*, the word *job* is five times more likely than the word *position* and nine times more probable than *presidency* (according to the Corpus of Contemporary American English, COCA). The evaluation by native speakers was used to explore their intuition about what they consider to be a collocation, which was explained as a set of words that regularly co-occur regardless of underlying grounds – idiomatic or otherwise – for their co-occurrence. Participants were asked, “Is the given collocation plausible or not?” and instructed to rate the collocations on a five-grade scale “plausible – rather plausible – uncertain – rather unplausible – unplausible.”

For the questionnaire, we used the same data that have been analyzed in the *Dictionary* evaluation. This allowed us not only to focus on what is considered lexical collocations *per se* by native speakers, but also to analyze the ranks of the extracted items. In the survey, 20 automatically extracted collocates for each word and two distractors (added to control the quality of the output) were randomly presented for evaluation. The distractors were chosen from the bottom of the extracted collocation lists. They usually had the query frequency of 1 and a considerably higher corpus frequency; e.g., *temnye nogi* ‘dark legs’ (query frequency = 1, corpus frequency of the collocate = 1155), or *nemeckie volosy* ‘German hair’ (query frequency = 1, corpus frequency of the collocate = 966). To avoid misinterpretation, only the most natural and frequent collocation form was taken as a default representation. Practically speaking, in most cases either the infinitive or the third person singular was used for verbal patterns (e.g., *zakryt’ glaza* ‘to close one’s eyes’ or *donessja golos* ‘N’s voice was heard’), with the nominative case used for the adjective patterns (e.g., *blednoe lico* ‘pale face’ or *umelye ruki* ‘skillful hands’). In rare cases, other default representations were chosen; e.g., the instrumental case for the collocation *golymi rukami* ‘barehanded’, because it is more frequently used than the nominative case. The decision as to whether we present a nominal or another form for the evaluation was based on the frequency of these collocations in the corpus; if there were no prevailing forms, then the nominal one was presented. All participants were native Russian speakers, most of them female (around 90 percent in each experiment). Most of the respondents had (or were at that moment working on) at least a bachelor's degree, with about half of the participants having linguistics as a major subject (Table 7.3). The experiment was organized mainly among students of the School of Linguistics at the Higher School of Economics, Moscow, and through social media. Responses that gave a positive answer to more than three distractors were filtered out. Finally, from 23 to 52 responses were examined for collocations, depending on the questionnaire to which a particular collocation belonged.

The inter-annotator agreement was measured with both Fleiss Kappa (Fleiss, 1971) and Krippendorff's Alpha (Krippendorff, 2004). Fleiss Kappa measures a pairwise agreement between raters for nominal data, which means that all disagreements are treated equally. Krippendorff's Alpha allows measurement of disagreement between multiple raters for ordinal data and thus is more suitable for our case. Another advantage of Krippendorff's Alpha is that it is applicable to incomplete responses, i.e., as when some answers are missed. Thus, in computing Alpha, we interpreted the answer "uncertain" as a missed value. However, since Kappa is more frequently reported, both coefficients are provided in this article. More details on these measures and their interpretations can be found in Artstein and Poesio (2008).

Table 7.4 presents the inter-annotator agreement for several experiment settings, where "5 grades" mean that the full scale of answers is used, while "3 grades" mean that the scale is reduced to three grades and responses are broken down into three groups (plausible – uncertain – unplausible). The Kappa values between 0 and 0.2 can be interpreted as a slight agreement and those between 0.2 and 0.4, as a fair agreement (Artstein and Poesio, 2008, p. 576). Krippendorff's Alpha ranges between 0 and 1, where 0 means that participants are unable to complete the task and therefore respond randomly. Since all the results in the table are statistically significant, the null hypothesis, namely, that the respondents made a random choice, can be rejected. At the same time, the inter-rater agreement is rather moderate, much less than is usually considered suitable for constructing

Table 7.3 Educational background of participants

High school graduate	2%
Undergraduates, no degree	29%
Bachelor/Master's degree	69%

Table 7.4 Inter-annotator agreement

		5 grades	*3 grades*
[x.ADJ + NOUN]	Fleiss *K*	0.17	0.30
	Krippendorff's Alpha	0.40	0.38
[x.VERB + NOUN]	Fleiss *K*	0.14	0.30
	Krippendorff's Alpha	0.36	0.39

a gold-standard dataset in corpus linguistics (Artstein and Poesio, 2008, p. 591). In fact, thresholds in the inter-rater agreement, especially for non-trivial tasks, are extensively discussed (e.g., Feinstein and Cicchetti, 1990; Cicchetti and Feinstein, 1990; Teufel and Moens, 2002; Di Eugenio and Glass, 2004; Antoine et al., 2014). An important methodological point was formulated by Artstein and Poesio (2008, p. 591):

> [s]etting a specific agreement threshold should not be a prerequisite for publication. Instead . . . researchers should report in detail on the methodology that was followed in collecting the reliability data . . . whether agreement was statistically significant, and provide a confusion matrix or agreement table so that readers can find out whether overall figures of agreement hide disagreements on less common categories.

Since we do not split stimuli into any preset groups, we cannot break the inter-agreement values into corresponding classes; instead, we plot the response standard deviation against an average response value. This demonstrates that, for fixed idiomatic expressions, on the one hand, and for free word combinations, on the other, agreements are much higher than they are for expressions that are neither this nor that. Thus, based on this principle, we can conclude that, although native speakers have some general notion of collocations, in many cases, they find it difficult to determine the degree of their collocability.

An average standard deviation in the responses was 0.58, which is quite high. However, there was a great variety of responses to some questions. In Figure 7.1, all of the bigram collocations used in the survey are located as points on the 2D-space, where the X-axis shows average responses and the Y-axis shows the standard deviation. As can be seen from the plot, the higher the absolute value of an average response, the less was the variety among the participants. The largest deviation, i.e., the greatest variety of responses, is typical of the collocations, with an average response around zero. The green vertical line on the plot separates bigrams that have an average response of more than 1 and are considered plausible in our evaluation. For these collocations, the standard deviation in responses is smaller than for other bigrams (avg=0.50). This plot may be interpreted as if the participants distinguished clearly between collocations that are either plausible or not, but had difficulty with borderline expressions. The crossed-out dots in the plot represent data points and show the highest and lowest degrees of variety among the participants: *uderžat' slezy* ‘hold back tears’ (sd = 0.861) and *zolotye ruki* ‘gifted hands’ (sd = 0; all participants rated it as plausible).

Below, we present the results of the survey. Table 7.5 illustrates the number of lexical collocations among the top twenty collocations. The original 5-grade scale was coded in the range -2 to 2, and the threshold of 1 was used for the evaluation. This scale implies that a collocation with an average response of more than 1 is considered plausible to some degree by a vast majority of the participants, but it is possible that some were uncertain about their choice. The evaluation itself was done in two ways: first, by using precision, i.e., the percentage of collocation

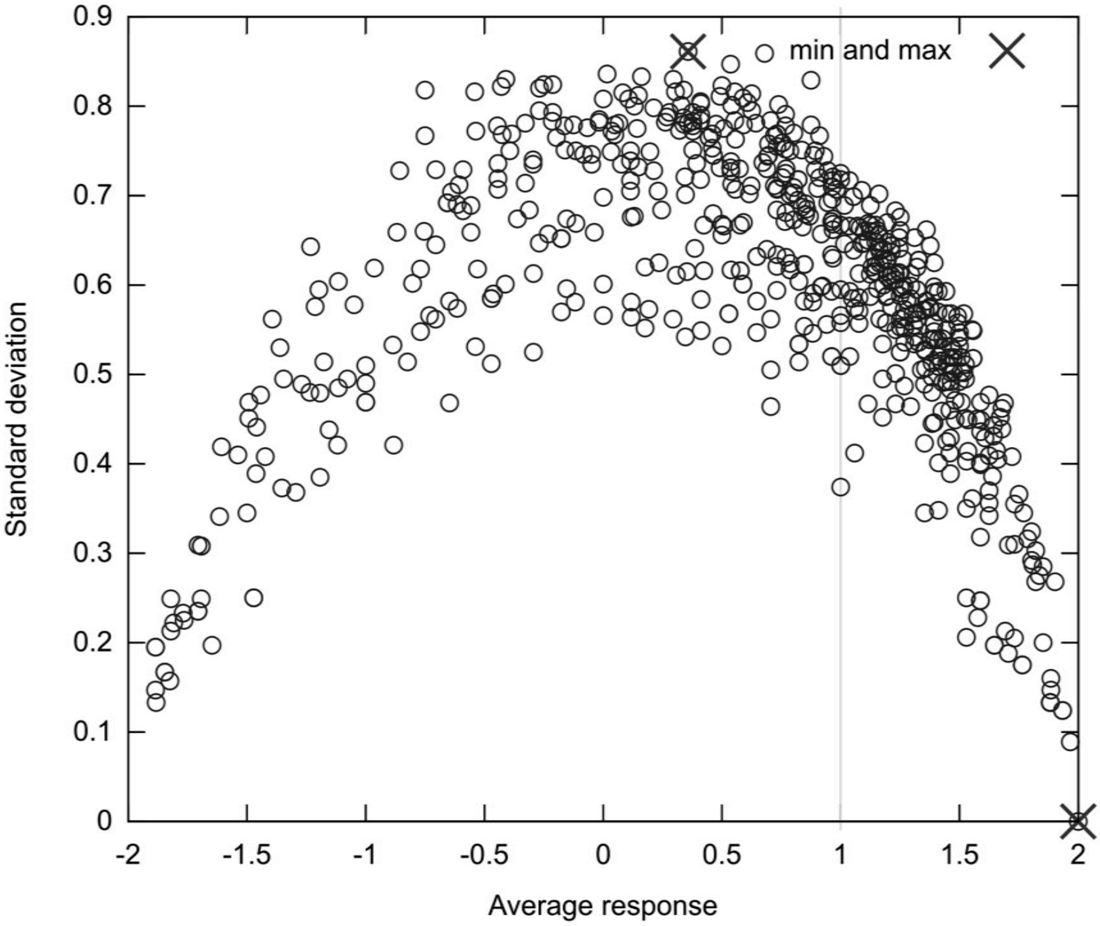


Figure 7.1 Average standard deviation in the responses.

Table 7.5 The number of lexical collocations among the top ten empirical collocations according to native speakers

MEASURE	*PREC*	*UAP*
DICE	0.62	0.81
T-SCORE	0.61	0.84
LOG-L	0.62	0.85
MI	0.48	0.62
WFR	0.52	0.70

defined as plausible according to the threshold, and, second, by using UAP, which reflects the successfulness of the rankings.

Roughly from 48 to 62 percent of collocations extracted by each measure was considered plausible by our participants. However, the UAP values were systematically higher than the precision values, which means that these collocations tend to be concentrated around the higher positions on the list. Once again, **log-likelihood, t-score**, and **Dice** did not differ drastically from each other and had nearly the same values. **wFR** was in fourth place, with **MI** performing slightly worse (judging by UAP) and thus the least efficient.

The nature of the extracted expressions varies considerably with all measures. As we pointed out above, some collocations were easy for all native speakers to define their plausibility, while others caused more difficulty in the interpretations. For some bigrams, the participants found it difficult to define not only the degree, but also whether an expression was plausible at all. In the following examples, we discuss briefly some typical cases in which agreement could not be reached on the collocations' plausibility (the numbers in parentheses show how many participants rated the collocation in each of five answers from plausible to unplausible). First, there were expressions without any semantic shift and yet considered plausible by a substantial number of the participants; for example, *morščinistaja šeja* 'wrinkled neck' (15—7—3—12—15) or *nežnaja šeja* 'delicate/tender neck' (16—9—8—7—12). Second, difficulties were caused by collocations with figurative meanings. These expressions describe an emotional state/reaction of a person with a physical comparison that cannot be measured fairly by any means: *ščemit' serdce* 'the heart ached/bled' (9—5—2—14—18) and *poholodelo serdce* 'the heart grew cold' (9—7—7—9—16). Finally, there were examples such as *uderžat' slezy* 'to hold back tears' (12—9—4—8—15), which are considered collocations in the *Dictionary*, but were not described as such by half of the participants. The uncertainty in this case might be explained by a similar expression *sderživat' slezy* 'to hold back tears' (see below), which is much more frequently used – almost twice as often, according to the RNC.

Nevertheless, there are collocations that were clearly defined as plausible by the vast majority of the participants, such as the following examples:

- *lebedinaja šeja* 'swan-like neck' (0—1—2—2—47)
- *zaščemilo serdce* 'the heart ached/bled' (1—1—4—10—32)
- *eknulo serdce* 'the heart skipped a beat' (0—1—1—5—41)
- *podstupajut slezy* 'tears well up' (1—2—1—16—28)
- *pustit' slezu* 'to shed a tear' (0—1—1—9—37)
- *sderživat' slezy* 'to hold back tears' (1—3—0—15—29)

All of the collocations considered plausible by the overwhelming majority of participants implied at the very least some restricted meaning (most often metaphorical) and are all listed in *A Russian-English Collocational Dictionary of the Human Body.*

To sum up, native speakers were not always able to distinguish between idioms, lexical and frequency-based collocations. They tended to define as plausible those expressions which involve a certain semantic shift, in other words, those where the meaning of the whole expression is non-compositional. In turn, marking a frequently used but semantically compositional bigram as plausible is a much more challenging task, reflected in greater deviations from the average response. Thus, although frequently used, semantically transparent collocations can be captured with statistical measures, and a speaker does not easily recognize such items when asked to do so. Probably other methods, such as the self-pacing reading test or eye-tracking control, could provide more reliable results than just a survey if these methods are available.

Nevertheless, considering our results for the measures shown in Table 7.5, we can say that the ranking by **t-score**, **log-likelihood**, and **Dice** includes more collocations than the ranking using the other methods, and, what is especially important, the ranking itself is more convincing. These results correlate to some extent with the results obtained from the *Dictionary* comparison (Tables 7.1 and 7.2). Thus, we are sure that this kind of evaluation was worth adding to our assessment of the *Dictionary* data.

4. Discussion

One of the main difficulties in dealing with collocation extraction is that collocability has different realizations in a language: some collocations emerged because they are frequently used, even if both collocates have an open distribution and may be frequently combined with other words. Others appear because one collocate has a strict selectional preference, even though the collocation itself is not that frequent. In this paper, we have demonstrated that 1) collocations with stronger selectional preferences are easier to extract than those consisting of frequently used collocates, and 2) different statistical measures suit the extraction of these two types to different extents.

It follows from what has been said that there is no single best method for dealing with collocation extraction. All standard methods are close to one another and produce intersecting results. First, this means that for non-crucial tasks, any method will work to some degree. However, in order to obtain more accurate results, using several methods seems to be the most reliable approach (cf. Evert and Krenn, 2001, p. 8). Although **t-score**, **log-likelihood**, **Dice**, and **frequency** showed the best results in our experiments, we cannot claim that any one of them is reasonably better than the others, nor can we state that **MI** and **wFR** are completely unacceptable in dealing with the Russian language. The choice of method depends on the goals a researcher wants to achieve.

A t-score extracts the collocations used most frequently in a language; e.g., the adjective *černaja* 'black' is not specifically used with the noun *boroda* 'beard,' yet the collocation *černaja boroda* 'blackbeard' is frequently used. The **t-score** ranking in our experiments was very similar to the **frequency** ranking, though few differences in the rankings were crucial, since the very frequent words are those that often match a pattern by accident. For example, according to our data, the verb *byt'* 'to be' usually appears at the top of the frequency list for almost any [x.VERB + NOUN] query, but collocations with *byt'* 'to be' usually have been filtered out by the **t-score**.

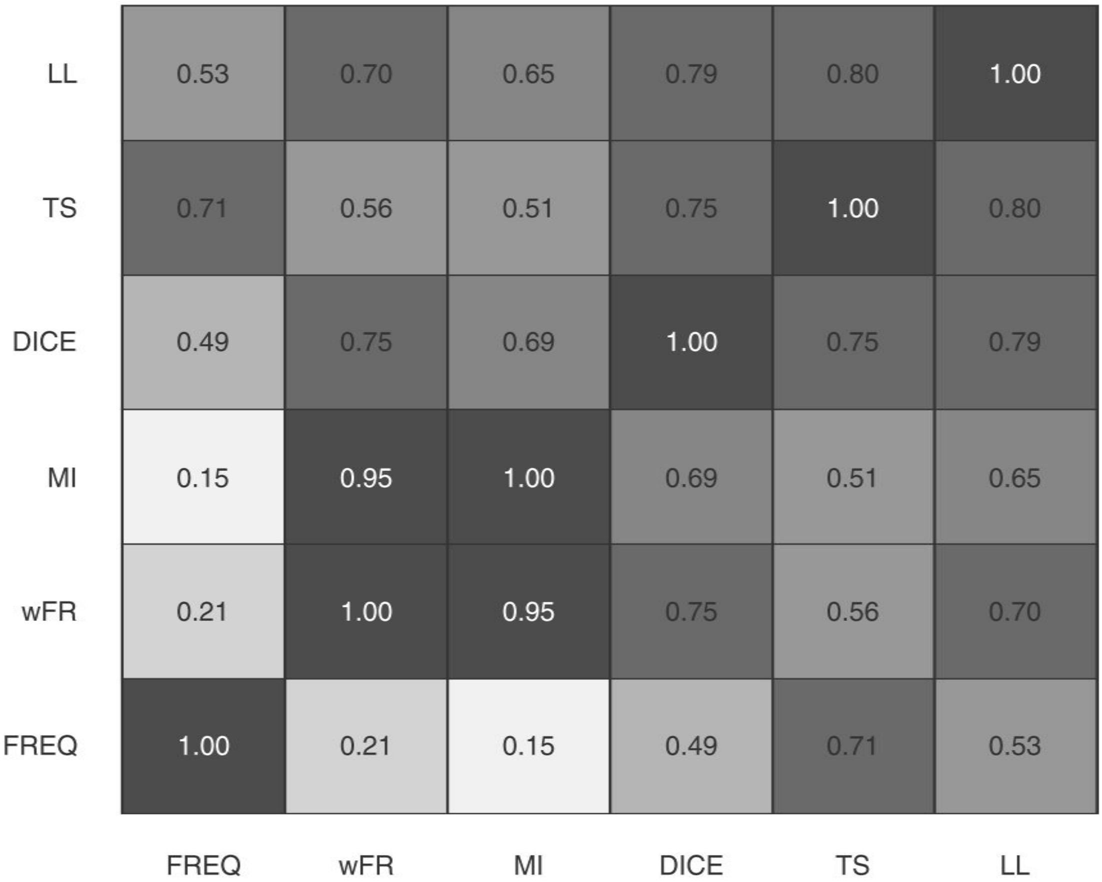
The **MI** measure is grounded in information theory and measures the level of uncertainty in finding a second collocate, given the first one. For some idiomatic expressions, the level of uncertainty is very low; thus, **MI** places such collocations at the top in its ranking. Since **MI** refers to infrequent collocations, it is highly sensitive to any noise in the data and should always be used with frequency filtering. For example, for the query [ADJ + *šeja* 'neck'], the first bigram in the **MI** ranking is *bleklaja šeja* 'faded neck'. The word *bleklyj* 'faded' appears only ten times in the corpus and once in this query. Statistically, in our data this word is strongly attached to *šeja* 'neck' (just because the corpus is rather small), and, consequently, this collocation occupies the first position on the list. Frequency filtering is a commonly used pre-processing stage when the object of study is a corpus as a whole and extracted collocations are intended to describe a particular genre, author, and so on. In our study, the objects themselves are fixed expressions. Fixedness does not necessarily imply frequency; even in our relatively large data, some expressions that are identified as plausible by native speakers occur only one or two times. Thus, we cannot use too aggressive a frequency cut-off, and, consequently, **MI** is not the most appropriate measure for this kind of study.

In turn, **Dice** extracts collocations that are fixed in a narrow sense: the collocates tend to occur with each other. Instead of predictability in the information-theoretical sense, **Dice** measures a simple ratio between collocation frequency and the sum of the collocates' frequencies. Unlike **MI**, this measure does not give preference to infrequent collocations. This can be seen in the above-mentioned example *bleklaja šeja* 'faded neck.' In the **Dice** measure, this bigram ranks only twelfth, since **Dice** takes into account the frequency of the word *šeja* 'neck' itself.

Log-likelihood can also be understood as a cross-entropy (or average mutual information; see Evert, 2005) between the observed and the expected distribution of a random variable, which may take four values: the collocation in question, collocations that include the first word only, collocations that include the second word only, and all other collocations. The greater the difference between the observed and the expected distributions, the more significant it is when collocates co-occur in the data. This measure is useful in extracting collocations with strong selectional preferences similar to **MI**. At the same time, **log-likelihood** rankings can be very different from **MI** rankings, since these measures use slightly different information: the **log-likelihood** formula includes frequencies for all the other collocations in the corpus, while the local **MI** focuses on the given collocation and the frequencies of its collocates.

wFR, even though it was the best measure for the colligation extraction task (Kopotev et al., 2013), turned out to be unsuitable for collocation extraction. **wFR** is highly correlated with **MI** and has limitations that come from infrequent words. This problem did not arise when we were dealing with grammatical categories, since these by definition cannot be *hapax legomenon*.

In order to show the general correlation among all these measures (plus raw frequency), the Kendall rank correlation was calculated for each query and then averaged across all queries. The result is presented in Figure 7.2, where the darker color represents a stronger correlation between two measures. As can be seen, the measures discussed in this paper can be divided into two groups: **MI** and **wFR**, and **log-likelihood**, **t-score**, and **Dice**. The strong correlation between **MI** and **wFR** (94 percent) was fully expected, since **MI** is equivalent to the simple frequency ratio when there is a constant value in a query (i.e., *f(p)* is a constant).



	FREQ	wFR	MI	DICE	TS	LL
LL	0.53	0.70	0.65	0.79	0.80	1.00
TS	0.71	0.56	0.51	0.75	1.00	0.80
DICE	0.49	0.75	0.69	1.00	0.75	0.79
MI	0.15	0.95	1.00	0.69	0.51	0.65
wFR	0.21	1.00	0.95	0.75	0.56	0.70
FREQ	1.00	0.21	0.15	0.49	0.71	0.53

Figure 7.2 The Kendall rank correlation between measures.

These measures have almost no correlation with frequency, however, which was expected, since they prefer collocates that rarely occur within the pattern. **Log-likelihood**, **t-score**, and **Dice** all correlate with each other (the correlation value is higher than 80 percent) and form the second group. Of all the measures, the **t-score** has, as expected, the strongest correlation with frequency. **Dice** is closer to **log-likelihood** and the **t-score**, but also demonstrates a relatively high correlation with **MI** (71 percent) and **wFR** (77 percent).[4]

These correlation results illustrate that different measures discover different groups of collocations, and, as was pointed out in Pecina (2009, pp. 147–150), a possible way to improve performance in collocation extraction is to combine the measures into more complex models. Since some of the measures produce very similar results, this task requires finding a delicate balance between how many and exactly which ones are to be combined, an undertaking beyond the scope of this article.

Of course, there are other features that distinguish collocations. One is transparency, i.e., the disposition of collocates to stay edge-to-edge or to be wedged between other words. Another is a tendency to reduce the full morphological paradigm of the collocates to a limited number of tokens. The last (but not least) factor is semantic non-compositionality, which shifts collocations toward more idiomatic items. In dealing with language, we must analyze a continuum of similar, near-similar, and dissimilar items. Collocation extraction thus presents a Herculean task, even if we develop methods that can effectively extract parts of this wide spectrum. All of the measures discussed above enable the extraction of collocations that are relevant in different ways and are used either separately or in combinations. In any case, the lists of obtained *empirical collocations* should be investigated further, because they represent not the last step, but rather a first step toward *lexical collocations*, which are the subject of semantic studies and theoretical linguistics.

5. Conclusion

Scholars today have a variety of methods at their disposal for automatic empirical collocation extraction. Yet, these methods are often used without a full understanding of the underlying concepts or of which method is best suited to collect lexical collocations with different properties. In this article, we have examined five frequently used methods using Russian data and have endeavored to explain the concepts and the optimum use of each. In order to do this, we used two different approaches, namely, a dictionary evaluation and a survey, which provided us with similar results. In both cases, a **t-score**, **log-likelihood**, and **Dice** gave the best performance and showed similar results. The **MI** and **wFR** performed less well in both evaluations. In the dictionary experiment, for which we used *A Russian-English Collocational Dictionary of the Human Body*, the difference between **log-likelihood** and the **t-score** was more reasonable, with the **Dice** score producing results very similar to **log-likelihood.** Also in the dictionary evaluation, the raw **frequency** showed the best result; however, it has a serious disadvantage, namely, it does not filter out highly frequent words, e.g., light verbs. In general, although each of these methods is far from ideal and all have their limitations, they produced similar results that are worth using in further linguistic analysis.

Acknowledgements

This work was partly supported by a grant from the Research Foundation of the University of Helsinki and by the University's BAULT (Building and Use of Language Technology) community. We would also like to thank the developers of the Russian National Corpus, especially Ekaterina Rakhilina and Olga Lyashevskaya, for sharing the data, and Anastasia Bonch-Osmolovskaya, of the Higher School of Economics, Moscow, for her assistance with the survey. We also deeply thank the anonymous reviewers for their inspirational comments and suggestions.

Notes

[1] Many standard software programs also allow the extraction of collocations, e.g., WordSmith (see Anagnostou and Weir, 2006, for further reference).

[2] **Log-likelihood** was primarily introduced as an alternative to the commonly used **χ^2-test** in Dunning (1993), where the advantage of **log-likelihood** to χ^2 was also shown

experimentally. The paper argued that χ^2 is not appropriate for the collocation extraction task, since it assumes that data are normally distributed, while text data are highly skewed; thus χ^2 overestimates low-frequency collocations. For further discussion on likelihood, see Evert (2005).

[3] In this dictionary, the 'body-part' meaning is understood in a very broad sense, e.g., *sleza* 'tear' and *golos* 'voice' are both included. This decision might be seen as open to debate; however, we have accepted it at face value, since this point is not relevant to our work.

[4] These are averaged numbers with all queries, where one word is given and taken into account. When the query is non-restricted, the correlations may look different. In our experiments, **Dice** demonstrated the greatest variety; for some queries, it is close to **MI** and **wFR**, while for others, it is closer to **log-likelihood** and the **t-score**. However, the general correlation pattern remains the same.

References

Anagnostou, Nikolaos, and George R. S. Weir. "Review of software applications for deriving collocations." *ICT in the Analysis, Teaching and Learning of Languages, Preprints of the ICTATLL Workshop* (2006): 91–100.

Antoine, Jean-Yves, Jeanne Villaneau, and Anaïs Lefeuvre. "Weighted Krippendorff's alpha is a more reliable metrics for multi-coders ordinal annotations: Experimental studies on emotion, opinion and coreference annotation." *EACL* (2014): 550–559.

Artstein, Ron, and Massimo Poesio. "Inter-coder agreement for computational linguistics." *Computational Linguistics* 34:4 (2008): 555–596.

Braslavskij, Pavel, and Evgeny Sokolov. "Sravnenie četyreh metodov avtomatičeskogo izvlečenija dvuhslovnyh terminov iz teksta" [Comparing four methods for automatic extraction of two-word terms from a text]. *Computational Linguistics and Intellectual Technologies: Papers from the Annual conference "Dialogue"* (2006): 88–94 [In Russian].

Bratus, B. V., I. B. Bratus, and E. A. Dancig. *Russian Word-Collocations: Learner's Dictionary*. Moscow: Russkij iazyk, 1979.

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): 1934–1940.

Church, Kenneth W., William Gale, Patrick Hanks, and Donald Hindle. "Using statistics in lexical analysis." *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon* (1991): 115.

Church, Kenneth W., and Patrick Hanks. "Word association norms, mutual information, and lexicography." *Computational Linguistics* 16:1 (1990): 22–29.

Cicchetti, Domenic V., and Alvan R. Feinstein "High agreement but low kappa: II. Resolving the paradoxes." *Journal of Clinical Epidemiology* 43:6 (1990): 551–558.

Daudaravicius, Vidas. "Automatic identification of lexical units." *Informatica* 34:1 (2010): 85–92.

Denisov, P. N., V. V. Morkovkin, and N. K. Zelenova. *Učebnyj slovar' sočetaemosti slov russkogo iazyka* [A Learner's Combinatorial Dictionary of the Russian Words]. Moscow: Russkij iazyk, 1978.

Di Eugenio, Barbara, and Michael Glass. "The kappa statistic: A second look." *Computational Linguistics* 30:1 (2004): 95–101.

Dunning, Ted. "Accurate methods for the statistics of surprise and coincidence." *Computational Linguistics* 19:1 (1993): 61–74.

Evert, Stefan. *The statistics of Word Co-Occurrences: Word Pairs and Collocations*. PhD dissertation, Institut für maschinelle Sprachverarbeitung Universität Stuttgart, 2005. Available at http://elib.uni-stuttgart.de/opus/volltexte/2005/2371/pdf/Evert2005phd.pdf (accessed February 1, 2016).

Evert, Stefan. "Corpora and collocations." In *Corpus Linguistics. An International Handbook*. Vol. 2, edited by A. Lüdeling and M. Kytö, 1212–1248. Berlin: Mouton de Gruyter, 2008.

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): 188–195.

Feinstein, Alvan R., and Domenic V. Cicchetti. "High agreement but low kappa: I. The problems of two paradoxes." *Journal of Clinical Epidemiology* 43:6 (1990): 543–549.

Fleiss, Joseph L. "Measuring nominal scale agreement among many raters." *Psychological Bulletin* 76:5 (1971): 378–382.

Gries, Stefan Th. "50-something years of work on collocations: What is or should be next. . . ." *International Journal of Corpus Linguistics* 18:1 (2013): 137–166.

Iordanskaia, Lidiia Nikolaevna, Slava Paperno, Lesli LaRocco, Jean MacKenzie, and Richard L. Leed. *A Russian-English Collocational Dictionary of the Human Body*. Slavica Pub, 1996.

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

Kopotev, Mikhail, Lidia Pivovarova, Natalia Kochetkova, Roman Yangarber. 'Automatic detection of stable grammatical features in n-grams.' *Papers from the 9th Workshop on Multiword Expressions* (2013): 73–81.

Kopotev, Mikhail, Daria Kormacheva, and Lidia Pivovarova. "Constructional generalization over Russian collocations." In *Collocations Cross-Linguistically: Corpora, Dictionaries and Language Teaching* [=Mémoires de la Société Néophilologique de Helsinki], edited by Begoña Sanromán Vilas, 121–140. Helsinki: Unigrafia, 2016.

Krippendorff, Klaus. "Reliability in content analysis." *Human Communication Research* 30:3 (2004): 411–433.

Manning, Christopher, and Schütze, Hinrich. *Foundations of Statistical Natural Language Processing*. Vol. 999. Cambridge: MIT Press, 1999.

Mel'čuk, Igor. "Phrasemes in language and phraseology in linguistics." In *Idioms: Structural and Psychological Perspectives*, edited by Martin Everaert, Erik-Jan van der Linden, André Schenk, and Rob Schreuder, 167–232. Hillsdale, NJ: Lawrence Erlbaum, 1995.

Mel'čuk, Igor. *Semantics: From Meaning to Text*. Vol. 1–3. John Benjamins Publishing Company, 2012–2015.

Mitrofanova, Olga A., V. V. Belik, and V. V. Kadina "Korpusnoe issledovanie sočetaemostnyh predpočtenij častotnyh leksem russkogo jazyka" [Corpus analysis of selectional preferences of frequent words in Russian]. *Computational Linguistics and Intellectual Technologies: papers from the Annual conference "Dialogue"* (2008): 361–367. [In Russian].

Pecina, Pavel. "Lexical association measures: Collocation extraction." In *Studies in Computational and Theoretical Linguistics*. Vol. 4. UFAL: Praha, Czech Republic, 2009.

Pecina, Pavel, and Pavel Schlesinger. "Combining association measures for collocation extraction." *Proceedings of the COLING/ACL on Main Conference Poster Sessions* (2006): 651–658.

Stubbs, Michael. "Collocations and semantic profiles: On the cause of the trouble with quantitative studies." *Functions of Language* 2:1 (1995): 23–55.

Teufel, Simone, and Marc Moens. "Summarizing scientific articles: Experiments with relevance and rhetorical status." *Computational Linguistics* 28:4 (2002): 409–445.

Toldova, S. Y., Y. S. Akinina, and I. O. Kuznetsov. "The impact of syntactic structure on verb-noun collocation extraction." *Computational Linguistics and Intellectual Technologies: papers from the Annual conference "Dialogue"* (2013): 2–16.

Wiechmann, Daniel. "On the computation of collostruction strength: Testing measures of association as expressions of lexical bias." *Corpus Linguistics and Linguistic Theory* 4:2 (2008): 253–290.

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* 44:3 (2010): 164–175.